FACULTE DES SCIENCES ET TECHNOLOGIES – UNIVERSITE DE LILLE

FACULTY OF MEDICINE AND HEALTH SCIENCES – GHENT UNIVERSITY

VIB - UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

ÉCOLE DOCTORALE DE BIOLOGIE-SANTÉ

DOCTORAL THESES

In order to obtain the degree of:

Doctor of Science (Molecular and Cellular Aspects of Biology) from the University of Lille

and

Doctor of Health Sciences from Ghent University

Presented by

DIEGO FERNANDO GARCIA DEL RIO

Studying Protein Complexes for Assessing the Function of Ghost Proteins (Ghost in the Cell)

Lille at 20th of December 2023

The present Jury is composed by:

Reporter	Andreas Tholey	Professor (Kiel University)
President/Reporter	Marianne Fillet	Professor (Université de Liège)
Examiner	Kathryn Lilley	Professor (University of Cambridge)
Examiner	Virginie Redeker	Chargée de recherche (INSERM)
Thesis director	Michel Salzet	Professor (Université de Lille)
Thesis co-director	Kris Gevaert	Professor (Ghent University)
Supervisor	Amélie Bonnefond	Directeur de recherche (Université de Lille)
Supervisor	Sven Eyckerman	Professor (Ghent University)

FACULTE DES SCIENCES ET TECHNOLOGIES – UNIVERSITE DE LILLE

FACULTY OF MEDICINE AND HEALTH SCIENCES – GHENT UNIVERSITY

VIB - UGENT CENTER FOR MEDICAL BIOTECHNOLOGY

ÉCOLE DOCTORALE DE BIOLOGIE-SANTÉ

THÈSE DE DOCTORAT

En vue de l'obtention du grade de :

Docteur en Sciences (aspects moléculaires et cellulaires de la biologie) de l'Université de Lille

et

Doctor of Health Sciences from Ghent University

Presenté par

DIEGO FERNANDO GARCIA DEL RIO

Etudier des complexes protéiques pour évaluer la fonction des protéines fantômes (Ghost in the cell)

À Lille le 20 décembre 2023

Présenté devant le jury composé de :

Rapporteur	Andreas Tholey	Professeur (Kiel University)
Président/Rapporteur	Marianne Fillet	Professeur (Université de Liège)
Examinateur	Kathryn Lilley	Professeur (University of Cambridge)
Examinateur	Virginie Redeker	Chargée de recherche (INSERM)
Directeur	Michel Salzet	Professeur (Université de Lille)
Co-Directeur	Kris Gevaert	Professeur (Ghent University)
Co-Encadrante	Amélie Bonnefond	Directeur de recherche (Université de Lille)
Co-Encadrante	Sven Eyckerman	Professeur (Ghent University)

"Science is not about being right or wrong, it's about being willing to ask the right questions and follow the evidence wherever it leads."

Neil deGrasse Tyson

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor and laboratory director, Prof. Michel Salzet, for his guidance and support throughout this research journey. Additionally, I would like to thank the laboratory co-director, Prof. Isabelle Fournier, for opening the lab's facilities and giving me the first opportunity to enter the world of mass spectrometry five years ago.

I would also like to extend my thanks to my thesis supervisors from the University of Ghent, Prof. Kris Gevaert and Prof. Sven Eyckerman, for their invaluable feedback and honest remarks which allowed me to improve my work. To my supervisor, Prof. Amelie Bonnefond for her mentorship and help during performing the RNA-seq experiments.

I am grateful to the reporters of my thesis, Prof. Andreas Tholey and Prof. Marianne Fillet, for their insightful comments and constructive feedback. I am thankful for their time and effort in reviewing my thesis and providing valuable suggestions for improvement. Additionally, I would like to express my sincere gratitude to the examiners of my thesis, Prof. Kathryn Lilley and Prof. Virginie Redeker, for their evaluation of my work. I am thankful for their time and effort in attending this defense and providing valuable comments on my work.

I would like to especially acknowledge Dr. Tristan Cardon for his invaluable day-by-day supervision of my thesis. His guidance, expertise, and daily support have been crucial and instrumental in shaping the direction of my research and ensuring its success. I am grateful for his dedication, commitment, and confidence throughout this journey. Without all our weekly or daily discussions, I'm sure this journey would have been a totally different story. To Dr. Antonella Milagros Raffo-Romero, I am incredibly grateful for the friendship and unwavering support she has provided me from the very first day we met. Since the beginning, she offered me her unconditional support inside and outside the laboratory (Latino style). I am truly fortunate to have both them as friends and mentors, and I am forever grateful to you for your kindness and generosity. Thank you for opening your home and family to me.

To my fellow PhD companion and FRIEND, Kamel Bachiri. For your friendship and always welcoming attitude. For all the silly and “some” good jokes that help to cope with the most stressful and down moments. Thanks for all the songs, series, books,

and molecular biology recommendations. Thanks for always trying to include me and making me feel welcomed. Thanks also to Romane for sharing some drinks and sharing her passion for reading.

To my other FRIEND and not only PhD colleague Lucas Roussel. From the beginning, he showed genuine concern about my well-being in a new country, city, and language. I will be forever grateful for those gestures. Thank you for all the cell culture talks, drinks, laughs, serious moments, and especially for sharing with me your motivations and passions. I hope this journey helped you find “happiness”. Also, I want to thank Dr. Melanie Rose for sharing her insights during the multiple times we shared a drink and for enduring my jokes about Happy.

To Dr. Diala Kantar, for her patience in teaching me molecular biology and WB. For her friendship displayed as reels shared. For encouraging me during low times and always sharing a smile and laughs during the day. To Dr. Alice Capuz (I wrote it correctly) for all the effort she put into talking to me and helping to learn new things. For always displaying her dedication and will to excel. To Dr. Nina Orgrinc, for being my first supervisor five years ago when I arrived in France. Thanks for all the knowledge and insights you have shared with me. For always having comforting words and advice to pursue my career. For trusting in me and always giving a good referral about my work.

I would like to thank Lea, Laurine, and Lydia for always being concerned about how things are going. For the scientific and non-scientific talks that we have shared in the hallway or in the lab. And for sharing their struggles, doubts, and goals in life.

To Soulaimane, Sumeyye, Lucie, Etienne, Estelle, Yanis, Jean-Pascal, Kodhor, Tala, Louise, Marie, Maheul, Christelle, Atef, Christophe, Frank, Julien, and Maxence. For their help and advice every time that I needed it and for always being patient and friendly with me.

To my former supervisor, Dr. Maria Fedorova, for teaching me how to be confident in the things that I do and for showing me how to be rigorous with my work. Also, to Dr. Isabel Ruiz, for always pushing me to be better and for always believing in my potential. To my first research supervisor and mentor, Dr. Marco Loza-Mejia, for showing me the true love for science and teaching, for always believing in me, and considering me his friend. He was my first source of admiration and example, and he taught me that the goal of teaching and research is to watch the students grow and excel.

To my ASC friends: Gary Cooney, Daniel Sinausia, Milena Barp and Ettore Paltanin. For their support and encouragement during these three years, even though it was from a distance. For their true friendship that started in this same city and campus five years ago. To my friends Nabil Georges, Juan Rojas, Yesid Roman, Daniel Mireles, Emilie Chantraine and Elena Giaretta. For their friendship and support through this journey.

A mi amigo Jorge Covarrubias por abrirme las puertas de su casa desde el primer día. Por darme el ejemplo de adaptación y como levantarse de cualquier tropiezo. A Adán José, por ser un ejemplo de vida, perseverancia, tenacidad y carrera. A los dos por su amistad que se convirtió en mi familia Lillois. Porque sin ustedes, esta estancia en Lille hubiera sido muy diferente. Por todas las trasnochadas de José José, Vicente Fernández y Juan Gabriel. Por cada nueva aventura en un nuevo bar, en Boulogne o en París. Por cada salida que hacía que recordara que no todo era el trabajo y que lo más importante son mis seres queridos. A Raquel Escalante, por su amistad, alegría y, sobre todo, por enseñarme a afrontar las adversidades con una sonrisa. Y más aún, por recordarme que esta carrera la inicié para aportar un granito de arena en esta lucha. A Rocío, Nikitas, y Claudia, por todas las tardes de baile, comida, bebida y fiesta que compartimos durante estos años en Lille.

A mis amigos que están en México: Alam, Luis, Oswaldo, Miguel, Chito, Gabo, Mariana, Gina, Ilse, y Adri; por su apoyo incondicional desde la distancia, amistad incondicional y siempre creer en mí. A mi terapeuta la Dra. Adriana Patiño, porque sin el camino que comenzamos a la par de este viaje, el resultado de este doctorado hubiera sido completamente diferente. Por ayudarme a creer en mi y darme las herramientas para equilibrar mi vida.

A mis tías, tíos, primos y primas; que siempre han estado ahí apoyando a mis papás y a mi hermana desde siempre, pero más desde que llegué a Europa. A todos ellos por su apoyo, estar al pendiente de mí y sus oraciones.

A mi futura esposa Ivonne Hernández, por ser la persona que siempre ha creído en mi sobre todo cuando yo no lo hacía. Por estos 13 años de amistad, amor y crecimiento. Por alentarme a perseguir mis sueños, aunque eso pondría distancia entre nosotros. Por alentarme a crecer como persona y académico. Por ser el complemento, mi lugar seguro y aceptar compartir tu vida conmigo. Por atreverte a empezar una nueva etapa de nuestras vidas lejos de nuestros seres queridos y fuera de nuestras zonas de confort. Te amo mucho y gracias por estar en mi vida.

A mi abuelita Carmen, por haberme cuidado desde pequeño. Por haber entregado la última etapa de su vida a mi hermana y a mí. Porque donde esté sé que está orgullosa de mí.

A mi hermana Mariana García, por su amor incondicional, incluso en los momentos donde la he defraudado. Por creer en mí, ser mi confidente y amiga, aunque sea desde la distancia. Por apoyar a mis papás y tener más responsabilidades. Espero que este trabajo te inspire y te demuestre que tú puedes hacer lo que tú quieras. Te amo y vuela muy alto.

Finalmente, quiero agradecer a mis padres, Fernando García y Carmen del Río, por todo el sacrificio que han hecho para darnos a Mariana y a mí las herramientas necesarias para desarrollarnos personal y profesionalmente. Por haber trabajado tanto para que pudiéramos vivir una vida cómoda, tener la mejor educación posible y sin preocupaciones. Por ser un ejemplo de trabajo duro, amabilidad, respeto y solidaridad. Por todas las veces que fueron estrictos conmigo para que pudiera sacar lo mejor de mí y no perder mis objetivos. Por estar siempre ahí cuando los necesito, aunque sea a la distancia. Por siempre apoyarme en cumplir este sueño y objetivo. Por enseñarme a valorar y que lo más importante son nuestros seres queridos. Por inculcarme que la ética profesional no está peleada con dignificar a las personas. Por inculcarme valores que me han llevado a estar en el lugar donde estoy parado. Por su amor desinteresado y, sobre todo, por nunca frenarme y siempre respetar mis decisiones. Los amo y gracias por todo lo que me han dado en esta vida.

Scientific productions

Publications

- Garcia-del Rio, D. F.; Fournier, I.; Cardon, T.; Salzet, M. Protocol to Identify Human Subcellular Alternative Protein Interactions Using Cross-Linking Mass Spectrometry. STAR Protocols 2023, 4 (3), 102380. https://doi.org/10.1016/j.xpro.2023.102380.
- Garcia-del Rio, D. F.; Cardon, T.; Eyckerman, S.; Fournier, I.; Bonnefond, A.; Gevaert, K.; Salzet, M. Employing Non-Targeted Interactomics Approach and Subcellular Fractionation to Increase Our Understanding of the Ghost Proteome. iScience 2023, 26 (2). https://doi.org/10.1016/j.isci.2023.105943.
- Criscuolo, A.; Nepachalovich, P.; Garcia-del Rio, D. F.; Lange, M.; Ni, Z.; Baroni, M.; Cruciani, G.; Goracci, L.; Blüher, M.; Fedorova, M. Analytical and Computational Workflow for In-Depth Analysis of Oxidized Complex Lipids in Blood Plasma. Nat Commun 2022, 13 (1), 6547. https://doi.org/10.1038/s41467-022-33225-9.
- Espinosa-Valdés, M.; Borbolla-Alvarez, S.; Delgado-Espinosa, A.; Sánchez-Tejeda, J.; Cerón-Nava, A.; Quintana-Romero, O.; Ariza-Castolo, A.; García-Del Río, D.; Loza-Mejía, M. Synthesis, In Silico, and In Vitro Evaluation of Long Chain Alkyl Amides from 2-Amino-4-Quinolone Derivatives as Biofilm Inhibitors. Molecules 2019, 24 (2), 327. https://doi.org/10.3390/molecules24020327.

Publications submitted

- Garcia-del Rio, D. F.; Derhourhi, M.; Bonnefond, A.; Leblanc, S.; Guilloy, N.; Roucou, X.; Eyckerman, S.; Gevaert, K.; Salzet, M.; Cardon, T. Deciphering the ghost proteome in ovarian cancer cells by deep proteogenomic characterization. Submitted to Nucleic Acid Research.

Oral presentations

- Journées Françaises de Spectrométrie de Masse, 2023, Marseille, France.
- Advanced Spectroscopy in Chemistry Master welcome, 2021, 2022 and 2023, Lille, France.
- Professional internship workshop UVM, 2022, online.

Poster presentations

- Human Proteome Organization Congress, 2022, Cancun, Mexico (travel grant).
- 3i Forum, 2022, Lille, France.
- Next-Generation Protein Analysis and Detection (4th edition), 2022, Ghent, Belgium.
- European Proteome Association Congress, 2022, Leipzig, Germany.
- Journée Stratégique de la SFR Technologie de la Santé et Médicaments, 2022, Lille, France.
- Human Proteome Organization Congress, 2021, online.

Flash presentations

- Journée André Verbert, 2022, Lille, France.
- Journées Françaises de Spectrométrie de Masse, 2021, online.
- Meeting for Young Scientists in Proteomics, 2021, online.
- Journée Stratégique de la SFR Technologie de la Santé et Médicaments, 2020, online.

Associations

- Société Française de Spectrométrie de Masse (SFSM) since 2021
- Initiative on Model Organism Proteomics (iMOP) since 2021
- French Proteomics Society (2021)

Supervision

- Sarah Saiche: M1 Omics.
- Antonina Gonet: M1 Omics.

Industrial exprerience

- Laboratorios Silanes, Mexico (2016), Regulatory affairs chemist.
- Laboratorios Servier, Mexico (2015-2016), Regulatory affairs analyist.
- Laboratorios Servier, Mexico (2015), Clinical research intern.

Summary

Ovarian cancer (OvCa) has the highest mortality rate among female reproductive cancers worldwide. OvCa is often referred to as a stealth killer because it is commonly diagnosed late or misdiagnosed. Once diagnosed, OvCa treatment options include surgery or chemotherapy. However, chemotherapy resistance is a significant obstacle. Therefore, there is an urgent need to identify new targets and develop novel therapeutic strategies to overcome therapy resistance.

In this context the ghost proteome is a potentially rich source of biomarkers. The ghost proteome, also known as the alternative proteome, consists of proteins translated from alternative open reading frames (AltORFs). These AltORFs originate from different start codons within mRNA molecules, such as the coding DNA sequence (CDS) in frameshifts (+1, +2), the 5'-UTR, 3'-UTR, and possible translation products from non-coding RNAs (ncRNA).

Studies on alternative proteins (AltProts) are often limited due to their case-by-case occurrence and complexity. Obtaining functional protein information for AltProts requires complex and costly biomolecular studies. However, their functions can be inferred by profiling their interaction partners, known as "guilty by association" approaches. Indeed, assessing AltProts' protein-protein interactions (PPIs) with reference proteins (RefProts) can help identify their function and set them as research targets. Since there is a lack of antibodies against AltProts, crosslinking mass spectrometry (XL-MS) is an appropriate tool for this task. Additionally, bioinformatic tools that link protein functional information through networks and gene ontology (GO) analysis are also powerful. These tools enable the visualization of signaling pathways and the grouping of RefProts based on their biological process, molecular function, or cellular localization, thus enhancing our understanding of cellular mechanisms.

In this work, we developed a methodology that combines XL-MS and subcellular fractionation. The key step of subcellular fractionation allowed us to reduce the complexity of the samples analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS). To assess the validity of crosslinked interactions, we performed molecular modeling of the 3D structures of the AltProts, followed by docking studies and measurement of the corresponding crosslink distances. Network analysis indicated potential roles for AltProts in biological functions and processes. The

advantages of this workflow include non-targeted AltProt identification and subcellular identification.

Additionally, a proteogenomic analysis was performed to investigate the proteomes of two ovarian cancer cell lines (PEO-4 and SKOV-3 cells) in comparison to a normal ovarian epithelial cell line (T1074 cell). Using RNA-seq data, customized protein databases for each cell line were generated. Differential expression of several proteins, including AltProts, was identified between the cancer and normal cell lines. The expression of some RefProts and their transcripts were associated with cancer-related pathways. Moreover, the XL-MS methodology described above was used to identify PPIs in the cancerous cell lines.

This work highlights the significant potential of proteogenomics in uncovering new aspects of ovarian cancer biology. It enables us to identify previously unknown proteins and variants that may have functional significance. The use of customized protein databases and the crosslinking approach have shed light on the "ghost proteome," an area that has remained unexplored until now.

Resumé

Le cancer de l'ovaire (OvCa) est le cancer le plus mortel parmi les cancers féminins. Il est souvent diagnostiqué tardivement ou mal diagnostiqué, ce qui le rend difficile à traiter. Les options de traitement incluent la chirurgie ou la chimiothérapie, toutefois la résistance à la chimiothérapie est un problème majeur. Il est donc urgent de trouver de nouvelles cibles et de développer de nouvelles stratégies pour surmonter cette résistance.

Dans ce contexte le protéome fantôme est une source potentiellement riche de biomarqueurs. Le protéome fantôme, ou protéome alternatif, est composé de protéines traduites à partir de cadres de lecture ouverts alternatifs (AltORFs). Ces AltORFs proviennent de différents codons START issus de différente région de l'ARNm, tels qu'un décalage de cadre de lecture (+1, +2) dans la séquence codante de l'ADN (CDS), dans le 5'-UTR, 3'-UTR et éventuellement de la traduction d'ARN non codants (ncRNA).

Les études sur les protéines alternatives (AltProts) sont souvent complexes et nécessite des études biomoléculaires coûteuses. Cependant, leurs fonctions peuvent être déduites en identifiant leurs partenaires d'interaction, la détection des interactions protéine-protéine (PPI) entre AltProts et protéines de référence (RefProts) peut aider à identifier leur fonction. La stratégie de pontage chimique (*crosslink*) combiné à la spectrométrie de masse (XL-MS) est un outil approprié à cet objectif. De plus, les outils bioinformatiques qui relient les informations fonctionnelles des RefProt et les analyses d'ontologie génique (GO) permettent la visualisation des voies de signalisation et le regroupement des RefProts en fonction de leur processus biologique, de leur fonction moléculaire ou de leur localisation cellulaire, et ainsi y placer certaine AltProt.

Dans ce travail, nous avons développé une méthodologie combinant XL-MS et le fractionnement subcellulaire. L'étape de fractionnement subcellulaire nous a permis de réduire la complexité des échantillons analysés par chromatographie liquide et spectrométrie de masse (LC-HRMS/MS). Pour évaluer la validité des interactions, nous avons réalisé une modélisation moléculaire des structures 3D des AltProts, suivie d'une prédiction informatique de l'interaction et de mesure des distances de pontages identifiés expérimentalement. L'analyse a révélé des rôles d'AltProts dans

les fonctions et les processus biologiques tel que la réparation de l’ADN ou encore la présentation d’antigène.

La protéogénomique a été utilisée pour générer des bases de données protéiques personnalisées à partir des données de séquençage ARN afin d’étudier les protéomes de deux lignées cellulaires de cancer de l'ovaire (PEO-4 et SKOV-3) en comparaison avec une lignée cellulaire ovarienne normale (T1074). L'expression différentielle de plusieurs protéines a ainsi été identifiée entre les lignées cellulaires cancéreuses et normales, avec une association aux voies de signalisation connues pour le cancer. Des PPI ont également été identifiées dans les lignées cellulaires cancéreuses en utilisant la méthodologie XL-MS.

Ce travail met en évidence le potentiel de l’approche protéogénomique pour découvrir de nouveaux aspects de la biologie du cancer de l'ovaire. Il nous permet d'identifier des protéines et des variants auparavant inconnus qui peuvent avoir une signification fonctionnelle. L'utilisation de bases de données protéiques personnalisées et de l'approche de réticulation a mis en lumière le "protéome fantôme", une vision du protéome restée inexplorée jusqu'à présent.

Contents

List of figures

List of tables

List of abbreviations

3'-UTR	3’-untranslated region
5'-UTR	5’-untranslated region
A	Adenine
ACN	Acetonitrile
AltGNL1	Alternative G protein nucleolar 1
AltORF	Alternative open reading frame
AltProts	Alternative proteins
APEX-MS	Ascorbic acid peroxidase proximity-labeling MS
AP-MS	Affinity purification coupled to MS
Apols	Amphipols
AT1R	Angiotensin type 1a receptor
B2M	B2 microglobulin
BAM	Binary alignment map
BioID	Proximity dependent biotin identification
BirA*	Biotin protein ligase
BLAST	Basic local alignment search tool
BS3	Bis(sulfosuccinimidyl)suberate
C	Cytosine
CBDPS	Cyanurbiotindipropionylsuccinimide
CBP	Calmodulin binding protein
CDS	Coding sequence
ChIP	Chromatin immunoprecipitation
COFRADIC	Combined fractional diagonal Chromatography
coIP-MS	Co-immunoprecipitation couple to MS
CT	Computed tomography
DAVID	Database for annotation, visualization and integrated discovery
DBD	DNA-binding domain
DDA	Data-dependent acquisition
DGE	Differential gene expression
DIA	Data-independent acquisition
DNA	Deoxyribonucleic acid
DS	Differential solubility
DSBSO	Disuccinimidyl bissulfoxide
DSBU	Disuccinimidyl dibutyric urea
dsDNA	Double-stranded DNA
DSS	Disuccinimidylsuberate
DSSO	Disuccinimidyl sulfoxide
DTT	Dithiothreitol
eIFs	Eukaryotic initiation factors
EMBL-EBI	European bioinformatics institute
Erk1/2	Extracellular signal-regulated kinases 1/2
ERLIC	Electrostatic repulsion-hydrophilic interaction chromatography
ETD	Electron-transfer dissociation
EThCD	Electron-transfer/higher-energy collision dissociation
FASP	Filter aided sample preparation
FDR	False discovery rate

FIGO	International Federation of Gynecology and Obstetrics
FPKM	Fragments per kilo base of transcript per million mapped fragments
FRET	Fluorescent molecule energy transfer
FT	Fourier transformation
FTICR	Fourier-transform ion cyclotron resonance
G	Guanine
GAP	Gtpase-activating protein
GDF-1	Growth/differentiation factor 1
GELFrEE	Gel eluted liquid fraction entrapment electrophoresis
GFP	Green fluorescent protein
GO	Gene ontology
GSEA	Gene set enrichment analysis
GST	Glutathione-S transferase
H_2O_2	Hydrogen peroxide
H3K4me3	Trimethylation of H3K4
HAltORF	Human alternative Open Reading Frame
HATs	Histone acetyltransferases
HCD	Higher-energy collisional dissociation
HDACs	Histone deacetylases
HES	Hematoxylin eosin and saffron
HILIC	Hydrophilic interaction chromatography
HPLC	High-performance liquid chromatography
HRMS	High-resolution mass spectrometer
IAA	Iodoacetamide
IEX	Ion-exchange chromatography
IF	Immunofluorescence
IHC	Immunohistochemical
IMAC	Immobilized metal ion-affinity chromatography
iMet	Initiator methionine
indels	Insertions/deletions
iST	In-stagetip
I-TASSER	Iterative Threading assembly Refinement
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	Liquid chromatograph mass spectrometry
LC-MS/MS	Liquid chromatography tandem mass spectrometry
LFQ	Label-free quantification
LINC-PINT	Long intergenic non-protein-coding RNA p53-induced transcript
lncRNA	Long non-coding RNA
LOPIT	Localization of organelle proteins by isotope tagging
MALDI	Matrix-assisted laser desorption/ionization
MBP	Maltose-binding protein
MHC	Major histocompatibility complex
MINION	Microprotein inducer of fusion
MLN	Myoregulin
MOXI	Micropeptide regulator of β-oxidation
mRNA	Messenger RNA
MS	Mass spectrometry
MSI	Mass spectrometry imaging

MWCO	Molecular weight cut-off
nanoESI	Nanoelectrospray
ncRNAs	Non-coding RNA
NER	Nucleotide excision repair
NGS	Next-generation sequencing
NHS	N-hydroxysuccinimide
Nt	N-terminal
ONT	Oxford nanopore
ORF	Open reading frame
OvCa	Ovarian cancer
PacBio	Pacific biosciences
PAF1c	Polymerase associated factor complex
PAGE	Polyacrylamide gel electrophoresis
PANTHER	Protein analysis through evolutionary relationship
PET	Positron emission tomography
Phox	Disuccinimidyl phenyl phosphonic acid
PIR	Protein information resource
PLA	Proximity ligation assay
PPIs	Protein-protein interactions
PRM	Parallel reaction monitoring
PSM	Peptide-to-spectrum match
PTMs	Post-translational modifications
QC	Quality control
q-TOF	Quadrupole time-of-flight
RCA	Rolling circle amplification
RefORF	Reference open reading frame
RefProts	Reference proteins
RFP	Red fluorescent protein
Ribo-seq	Ribosome profiling
RIN	RNA integrity number
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
ROS	Reactive oxygen species
RPC	Reversed phase chromatography
RPFs	Ribosome-protected fragments
RPKM	Reads per kilobase of transcript per million reads mapped
RPL10	Ribosomal protein 10
rRNA	Ribosomal RNA
SAX	Strong anion exchange
SCX	Strong cation exchange
SDS	Sodium dodecyl sulphate
SEC	Size exclusion chromatography
SEPs	Short open reading frame-encoded peptides
SFINX	Straightforward filtering index
SNV	Single nucleotide variants
SP3	Solid-phase-enhanced sample preparation
SP4	Solvent precipitation SP3
SPAR	Small regulatory polypeptide of amino acid response

SPE	Solid phase extraction
SPEED	Sample Preparation by Easy Extraction and Digestion
SQLE	Squalene epoxidase
SRSP	Splicing regulatory small protein
ssDNA	Single-stranded DNA
STAR	Spliced Transcripts Alignment to a Reference
S-Trap	Suspension trapping
SVMs	Support vector machines
T	Thymine
TAD	Transcription-activating domain
TAFs	TATA-binding protein associated factors
tbu-Phox	Tert-butyl disuccinimidyl phenyl phosphonate
TCGA	The Cancer Genome Atlas
TCR	T cell receptors
TFA	Trifluoroacetic acid
TIS	Translation initiation sites
TMT	Tandem mass tag
TPM	Transcripts per million
tRNA	Transfer RNA
$\text{tRNA}^{\text{iMet}}$	Initiator methionyl (Met)-transfer RNA
U	Uracil
UHMR	Ultra-high mass range
UICC	Union for International Cancer Control
UniProtKB	Uniprot Knowledgebase
upstream ORF	Upstream open reading frame
UVPD	Ultraviolet photodissociation
VCF	Variant calling files
VLPs	Virus-like particles
WHO	World health organization
WT	Wild-type
Xcorr	Correlation score
XIC	Extracted ion chromatograms
XL-MS	Cross-linking mass spectrometry
Y2H	Yeast two-hybrid
β-ME	B-mercaptoethanol

PART I STATE OF THE ART

Ovarian cancer

In 2020, on the occasion of the World Cancer Day, The Union for International Cancer Control (UICC) conducted a global survey on public perception of cancer. Specifically, the survey aimed to gather data on the effects of cancer on people's lives and their concerns about the future. The results were quite significant, with approximately 60% of individuals reporting that they have been directly or indirectly affected by cancer. Moreover, the same percentage of people expressed their worry about developing cancer in the future[1].

Cancer is a major health issue that affects millions of people worldwide. According to the World Health Organization (WHO), in 2020 alone, more than 19 million new cases of cancer were reported, while 10 million people have succumbed to the disease. These statistics highlight the importance of raising public awareness about cancer and its prevention. Additionally, it encourages researchers to investigate novel pathological mechanisms that can result in new and more efficient therapeutic options. By educating people about the risk factors and encouraging them to adopt a healthy lifestyle, we can reduce the incidence of cancer and improve the quality of life for those affected by this disease[2]. Cancer is a group of diseases characterized by the rapid and uncontrolled growth of abnormal cells that surpass the normal boundaries of healthy cells and invade nearby tissues and organs. This can lead to the formation of malignant tumors or neoplasms. These abnormal cells can affect any part of the body. Cancer is a complex and multifaceted disease that can manifest in various forms and have different causes and risk factors. Some of the common causes of cancer include genetic mutations, exposure to environmental factors, infections, unhealthy diet and lack of exercise. Cancer can have a significant impact on an individual's physical, emotional and social well-being, and can require a comprehensive and multidisciplinary approach to treatment and management.

Among these diseases, ovarian cancer (OvCa) is the tenth leading cause of death in women worldwide and has the highest mortality rate among female reproductive cancers. OvCa is considered a stealth killer due to its late diagnosis and misdiagnosis. Shockingly, in 2020, 207,253 women succumbed to this cancer[3]. Neoplasms have different origins: 3% from ovarian germ cells, 2% from germ cell stroma, and the vast majority from ovarian epithelium. The histopathological classification of epithelial tumors is divided into the following types: serous, mucinous, endometrioid, clear cell, Brenner and undifferentiated carcinomas. Among them, the most common subtype of this cancer is high-grade serous carcinoma, accounting for 70-80% of cases. Less common subtypes include low-grade serous carcinoma (<5%), endometrioid (10%), clear cell (10%) and mucinous (3%)[4].

Epithelial OvCa is caused by the accumulation of genetic mutations. There are five common gene mutations that have been identified in epithelial OvCa: *TP53*, *BRCA1*, *BRCA2*, *PIK3CA* and *KRAS*. Each of these gene mutations plays a different role in the development and progression of OvCa. The *TP53* gene mutation is commonly found in high-grade serous carcinoma and associated with poor prognosis. On the other hand, hereditary and recurrent high-grade serous carcinoma OvCa are characterized by mutations in the *BRCA1* and *BRCA2* genes, which predispose women to an increased risk of developing OvCa. Endometrioid and clear cell carcinoma, on the other hand, are known to present a high frequency of mutations in the *PIK3CA* gene that are associated with the activation of the PI3K/Akt/mTOR pathway, which promotes cell growth and proliferation. The *KRAS* mutation plays a key role in the development of low-grade serous and mucinous carcinoma. This mutation results in the activation of the RAS/MAPK pathway, which promotes cell survival and proliferation[5].

In addition to these main gene mutations, other gene mutations have also been found to contribute to the development of ovarian cancer. These include *CHEK2*, *ATM*, *BRIP1*, *BARD1*, *PALB2*, *RAD50*, *RAD51C*, *RAD51D*, *MRE11A*, *MSH6* and *NBN*. Somatic mutations can interact with each other and with germline mutations to further increase the risk of developing ovarian cancer[5].

OvCa has been classified into different stages based on operative findings. The International Federation of Gynecology and Obstetrics (FIGO) established a four-stage classification (**Table 1**), which determines the precise histologic diagnosis and prognosis of the patient based on these findings. The general characteristics of each stage are shown in **Figure 1**.

Table 1. International Federation of Gynecology and Obstetrics (FIGO) OvCa classification. *Adapted from Berek et al.[6].*

STAGE	FINDING
I	**Tumor confined to ovaries or fallopian tube(s)**
IA	Tumor limited to one ovary (capsule intact) or fallopian tube; no tumor on ovarian or fallopian tube surface; no malignant cells in the ascites or peritoneal washings
IB	Tumor limited to both ovaries (capsules intact) or fallopian tubes; no tumor on ovarian or fallopian tube surface; no malignant cells in the ascites or peritoneal washings
IC	Tumor limited to one or both ovaries or fallopian tubes, with any of the following
IC1	Surgical spill
IC2	Capsule ruptured before surgery or tumor on ovarian or fallopian tube surface
IC3	Malignant cells in the ascites or peritoneal washings
II	**Tumor involves one or both ovaries or fallopian tubes with pelvic extension (below pelvic brim) or peritoneal cancer**
IIA	Extension and/or implants on uterus and/or fallopian tubes and/or ovaries
IIB	Extension to other pelvic intraperitoneal tissues
III	**Tumor involves one or both ovaries or fallopian tubes, or peritoneal cancer, with cytologically or histologically confirmed spread to the peritoneum outside the pelvis and/or metastasis to the retroperitoneal lymph nodes**
IIIA1	Positive retroperitoneal lymph nodes only (cytologically or histologically proven)
IIIA1(i)	Metastasis up to 10 mm in greatest dimension
IIIA1(ii)	Metastasis more than 10 mm in greatest dimension
IIIA2	Microscopic extrapelvic (above the pelvic brim) peritoneal involvement with or without positive retroperitoneal lymph nodes
IIIB	Macroscopic peritoneal metastasis beyond the pelvis up to 2 cm in greatest dimension, with or without metastasis to the retroperitoneal lymph nodes
IIIC	Macroscopic peritoneal metastasis beyond the pelvis more than 2 cm in greatest dimension, with or without metastasis to the retroperitoneal lymph nodes (includes extension of tumor to capsule of liver and spleen without parenchymal involvement of either organ)
IV	**Distant metastasis excluding peritoneal metastases**
IVA	Pleural effusion with positive cytology
IVB	Parenchymal metastases and metastases to extra-abdominal organs (including inguinal lymph nodes and lymph nodes outside of the abdominal cavity)

A key part of diagnosing ovarian cancer is establishing risk factors. The most strongly correlated risk factors are age, menopause, obesity, late or no pregnancy, hormone therapy after menopause, family history of ovarian, breast, or colorectal cancer, fertility treatment, smoking and mutations in cancer related genes e.g. BRCA. Other factors that are less clear but may contribute include androgenic therapy, talcum powder usage, and a diet high in red and processed meats, sugary drinks, and highly processed foods[7].

Unfortunately, 4 out of 5 patients are diagnosed at late stages of the disease, when metastasis has already occurred in the abdominal cavity and other organs[8]. This late diagnosis is due to the non-specific nature of symptoms, which persist for longer periods of time. In the early stages of ovarian cancer, the most common symptoms include unusual bloating, fullness, pressure in the abdomen, unusual abdominal pain, lower back pain and lack of energy[9].

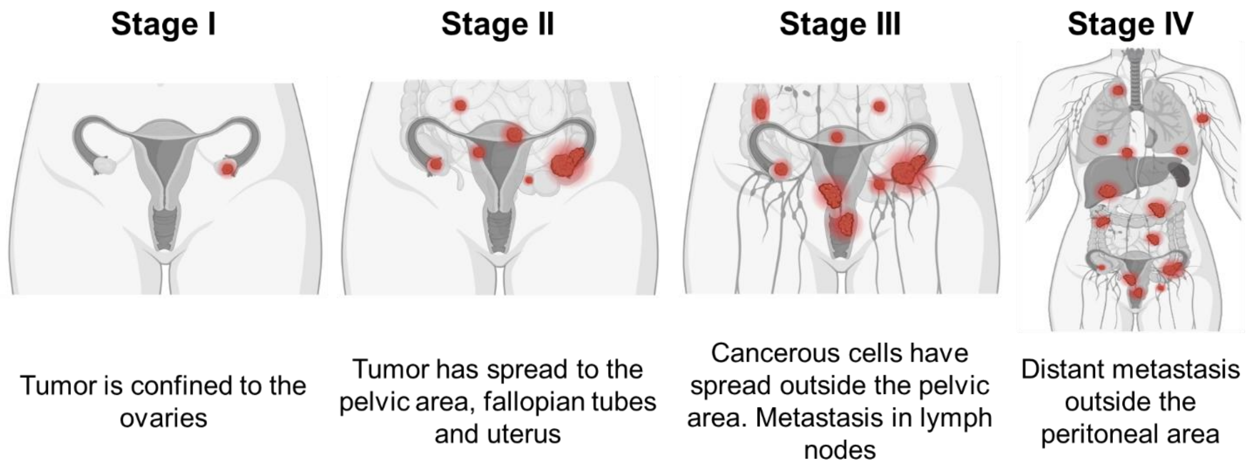

Stage I
Stage II
Stage III
Stage IV
Tumor is confined to the ovaries
Tumor has spread to the pelvic area, fallopian tubes and uterus
Cancerous cells have spread outside the pelvic area. Metastasis in lymph nodes
Distant metastasis outside the peritoneal area

Figure 1. Schematic representation of OvCa FIGO classification. *Each illustration displays the general characteristics of each FIGO stage. Tumors are displayed in red.*

The screening for ovarian cancer involves measuring the serum levels of tumor epithelial antigen CA125, pelvis ultrasound and pelvis examination. However, this screening is not recommended for asymptomatic patients due to the high rate of false positive cases. Instead, it is endorsed for patients with symptoms[6].

The first stage of OvCa diagnosis is performed by a positron emission tomography (PET) scan or a computed tomography (CT) scan. These imaging methods detect tumor markers by analyzing the body slice by slice and determining if a tumor mass is present and if a histological biopsy is necessary.

Laparoscopic surgery is recommended to evaluate the characteristics of the lesion, stage the cancer and explore the presence of metastases. Depending on the findings, the course of treatment is decided, especially whether the patient is operable or not. Laparoscopic surgery also enables the removal of the tumor appendix in early-stage OvCa. If the patient is not operable, chemotherapy is required to reduce the tumor before surgery.

Histological analysis can be performed by sampling the tumor through a minimally invasive incision or the removed tumor. The extracted specimen is given to a pathologist to determine the nature of the tissue and type of cancer. Hematoxylin eosin and saffron (HES) staining is performed, followed by microscopic examination (**Figure 2**) and immunohistochemical (IHC) staining for tumor markers. While this method can stage and differentiate most cases, more complicated cases can lead to errors in the diagnosis. The major weakness of this diagnostic methodology is the availability of the pathologist. Therefore, new technologies and biomarkers are needed to improve and facilitate the diagnosis.

Treatment for OvCa involves surgery and chemotherapy. Before starting treatment, multiple factors must be considered, including the FIGO stage, tumor burden and general condition of the patient. For patients at stage IA and IB, surgical removal of the tumor is the recommended treatment. When the stage is elevated to IC, platinum-based chemotherapy is recommended, followed by surgery. All patients with stage II disease should receive adjuvant chemotherapy. The recommended treatment for these two stages is carboplatin/paclitaxel in six to eight cycles. For higher stages (III and IV), chemotherapy is used to reduce the size of the tumor before surgery.

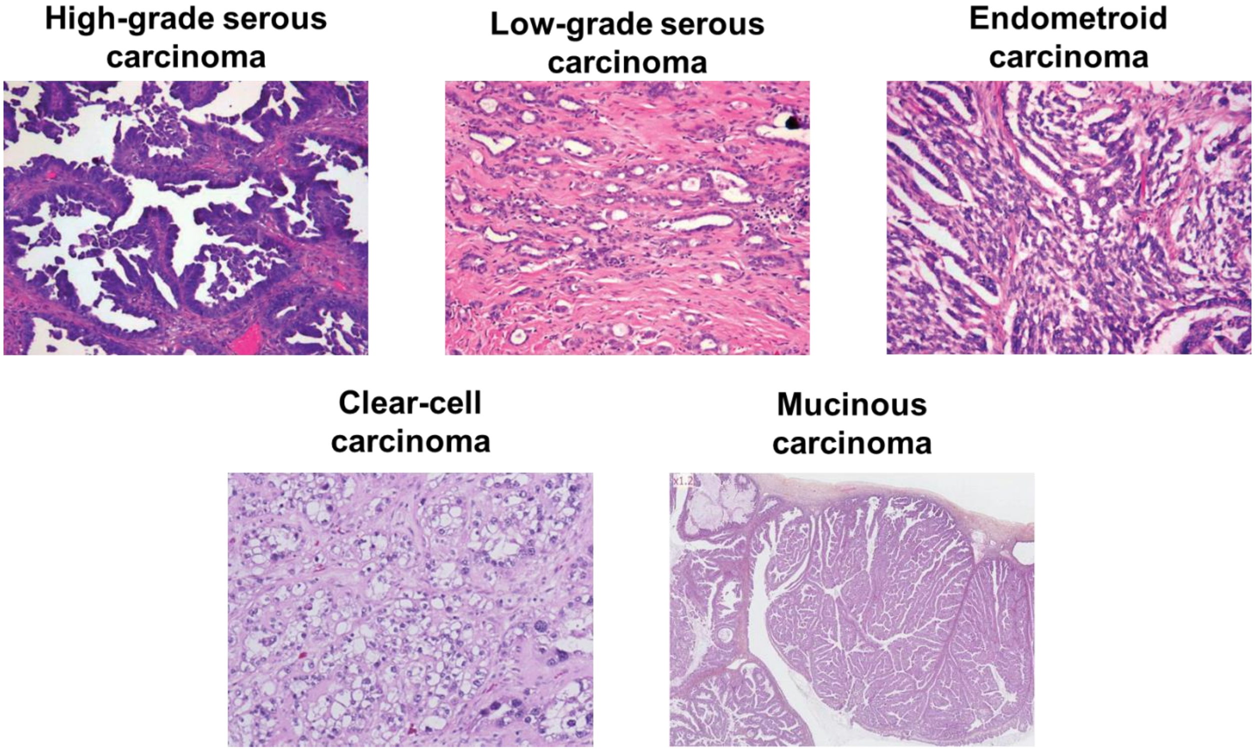


Figure 2. HES staining of each OvCa subtypes. *Cell nuclei are stained blue, basophils are purple, eosinophils are red, and collagen is pink. The micrographies were obtained from Soslow[11], Lewin et al.[12], and Genestie et al.[13].*

Paclitaxel is a powerful drug that affects the normal function of microtubule growth. Unlike other drugs that simply depolymerize microtubules, paclitaxel hyper-stabilizes their structure, rendering the cell's cytoskeleton inflexible. This is achieved by binding to the β subunit of tubulin, the building block of microtubules, and locking these building blocks in place. As a result, the microtubule/paclitaxel complex is unable to disassemble[10].

Carboplatin works by attaching alkyl groups to the nucleotides, leading to the formation of monoadducts and DNA fragmentation when repair enzymes attempt to correct the error. Additionally, 2% of its activity is attributed to DNA cross-linking from a base on one strand to a base on another, which prevents DNA strands from separating for synthesis or transcription[14].

To avoid the recurrence of ovarian cancer, it is recommended to implement radical strategies such as oophorectomy and salpingectomy[15]. While chemotherapy has shown efficacy, resistance to chemotherapy remains a significant barrier to successful treatment. Thus, it is imperative to identify new targets and develop novel therapeutic strategies to overcome chemotherapy resistance.

Several new and upcoming treatments for ovarian cancer are currently being studied in clinical trials. One example of a targeted therapy is PARP inhibitors, which block a protein called PARP that is involved in DNA repair. They are now being used as first-line maintenance therapy for patients with platinum-resistant disease[16]. Immunotherapies for ovarian cancer include checkpoint inhibitors, which block certain proteins that enable cancer cells to evade the immune system[17]. Another approach to immunotherapy is CAR T-cell therapy, where a patient's own T cells are genetically modified to recognize and eliminate cancer cells. CAR T-cell therapy has shown promise in treating leukemia and lymphoma, and investigations are underway for its potential in ovarian cancer treatment[18]. Overall, research of ovarian cancer is rapidly advancing, with new and upcoming treatments showing promise. However, it is important to acknowledge that ovarian cancer remains a complex and challenging disease and new pathological pathways need to be studied.

The dogma of molecular biology

Biomolecules are essential components of all living organisms. Two major types of biomolecules are polymers of nucleic acids and proteins. Nucleic acids are responsible for storing and transmitting genetic information, and they play a vital role in the organization and functioning of cells. They are closely connected, by the dogma of molecular biology, to their functional partners, proteins. Proteins, on the other hand, perform an incredible variety of functions in a cell, from providing structural support to catalyzing biochemical reactions and responding to internal and external stimuli[19].

Within a cell, there are two types of nucleic acid-containing biomolecules: DNA and RNA. DNA, or deoxyribonucleic acid, is a macromolecule that carries genetic information that forms the blueprint for RNA and proteins. It has a double-stranded helix structure made of deoxyribonucleotides. A deoxynucleotide consists of a deoxyribose sugar, a phosphate, and one of the purine bases adenine (A) or guanine (G), or one of the pyrimidine bases cytosine (C) or thymine (T).These four nucleobases are paired in a specific, complementary way through hydrogen bonds, with adenine pairing with thymine, and guanine with cytosine[20]. RNA, or ribonucleic acid, on the other hand, is a single-stranded polymer of nucleic acid that ribonucleotides. Each ribonucleotide contains a ribose, a phosphate and a nucleobase adenine, guanine, cytosine, and uracil (U). Compared to DNA, RNA can have varying lengths and structures. Three main types of RNA are directly involved in protein synthesis: messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). mRNA is a transcript of DNA that is translated into protein by ribosomes. tRNA delivers amino acids to the ribosome during protein synthesis. Finally, rRNA forms the structure of the ribosome itself and, amongst others, helps catalyze the formation of peptide bonds between amino acids during protein synthesis[21].

Proteins are involved in a multitude of biological processes. They have an intricate structure and are made up of 20 different amino acids. These amino acids are thus the building blocks of proteins, and their unique combinations determine a protein's structure, function and properties[22].

As stated above, these three biomolecules are connected by the dogma of molecular biology. Sixty-six years ago, Francis Crick described what would become the central dogma of molecular biology at the Society for Experimental Biology Symposium on the Biological Replication of Macromolecules, held at University College London. In his presentation entitled "ON PROTEIN SYNTHESIS", the 1962 Nobel Prize recipient explained the transfer of information in a cell (**Figure 3**)[23]. To him, the information is transferred from DNA to DNA by DNA replication, from DNA to mRNA by transcription, and, finally, during translation, the information in the nucleotide code is transferred from mature mRNA to proteins[24].

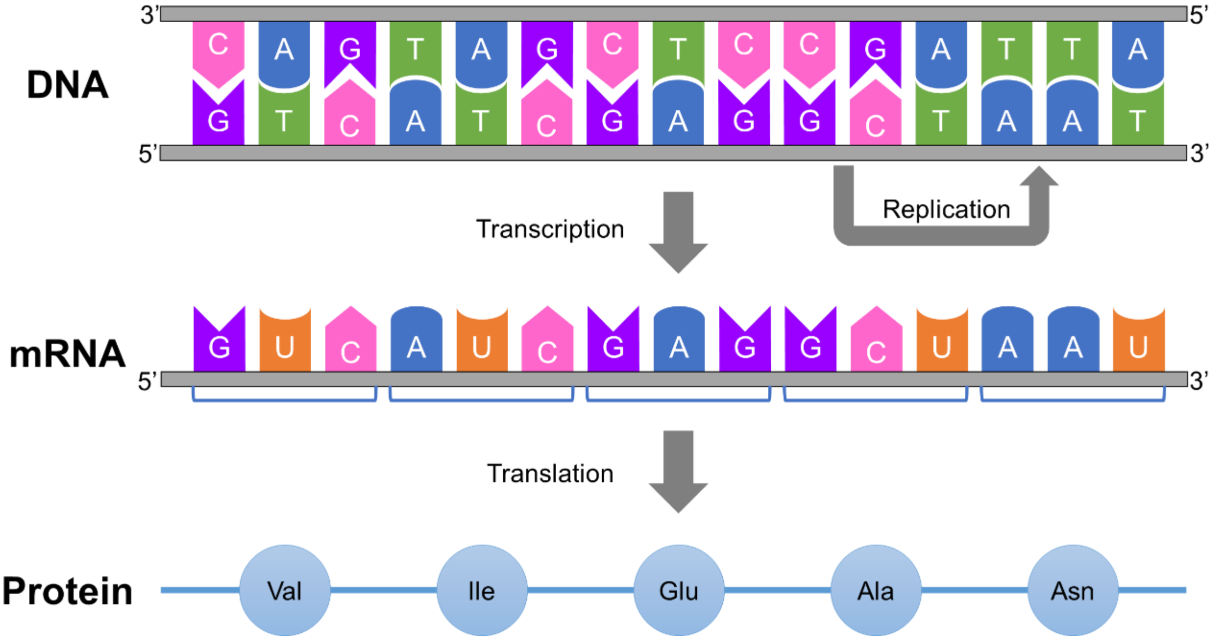

DNA
3'
5'
C A G T A G C T C C G A T T A
G T C A T C G A G G C T A A T
5'
3'
Transcription
Replication
mRNA
G U C A U C G A G G C U A A U
5'
3'
Translation
Protein
Val
Ile
Glu
Ala
Asn

Figure 3. Central dogma of molecular biology described by Francis Crick. *Information is transferred from DNA to DNA by replication. From DNA to RNA by transcription and from RNA to proteins by translation.*

Protein synthesis

Translation is a complex biological process which takes place in ribosomes and has four mayor stages: initiation, elongation, termination and ribosome recycling. In eukaryotes during the initiation step, the small (40S) ribosomal subunit attaches to the specific initiator methionyl (Met)-transfer RNA $(tRNA)^{iMet}$ and the mRNA. Once the small subunit has attached to the mRNA, it begins to scan the sequence in search of the start codon that will initiate translation in a 5' to 3' direction. Once the start codon is recognized, the large (60S) ribosomal subunit joins to form a functional ribosome[25]. The elongation phase of protein synthesis is a complex process that involves the codon-dependent addition of amino acids to the growing polypeptide chain. Each amino acid is added to the chain in a specific order as dictated by the sequence of codons in the mRNA molecule[26]. The termination steps of protein synthesis involve the release of the completed polypeptide chain from the ribosome, which is accomplished by the recognition of a stop codon that signals the end of the protein coding sequence[27]. Finally, the recycling phase refers to the dissociation of the ribosome and tRNA from the mRNA, which allows the ribosome to be reused in subsequent rounds of protein synthesis[28].

1.1. Protein translation initiation process

The translation initiation phase (**Figure 4**) is one of the four stages in protein synthesis. It regulates the initiation of protein synthesis and ensures that it occurs accurately and efficiently. During this stage, the post-termination ribosomal complexes dissociate during the recycling phase, and the free 40S ribosomal subunit recruits the eukaryotic initiation factors (eIFs) 1 and 1A[29].

eIF1 and eIF1A play an essential role in protein synthesis as they open the ribosomal channel through conformational changes[31,32]. Furthermore, the recruitment of eIF3, which promotes the attachment of the eIF2–GTP–Met-tRNA-^{iMet} anticodon loop to the P-site of the 40S subunit, occurs. These components then bind to the previous eIFs, forming the 43S complex[30].

Once the 43S complex has been formed, it requires the assistance of eIF4F and either eIF4B or eIF4H. These proteins work together to unwind the region of the mRNA that is close to the 5' cap, so that it can be ready for ribosomal attachment. The eIF4F protein complex comprises of a scaffold protein (eIF4G), a cap-binding protein (eIF4E), a DEAD-box RNA helicase (eIF4A), the poly(A)-binding protein (PABP) and eIF3[33]. eIF4F interacts with both the cap (through eIF4E) and the ribosome-associated eIF3 (through eIF4G), bridging the mRNA and the ribosome[34]. Thus, recruitment of the 43S complex is ultimately achieved by the cap$_{mRNA}$–eIF4E–eIF4G–eIF3–40S chain of interactions[30].

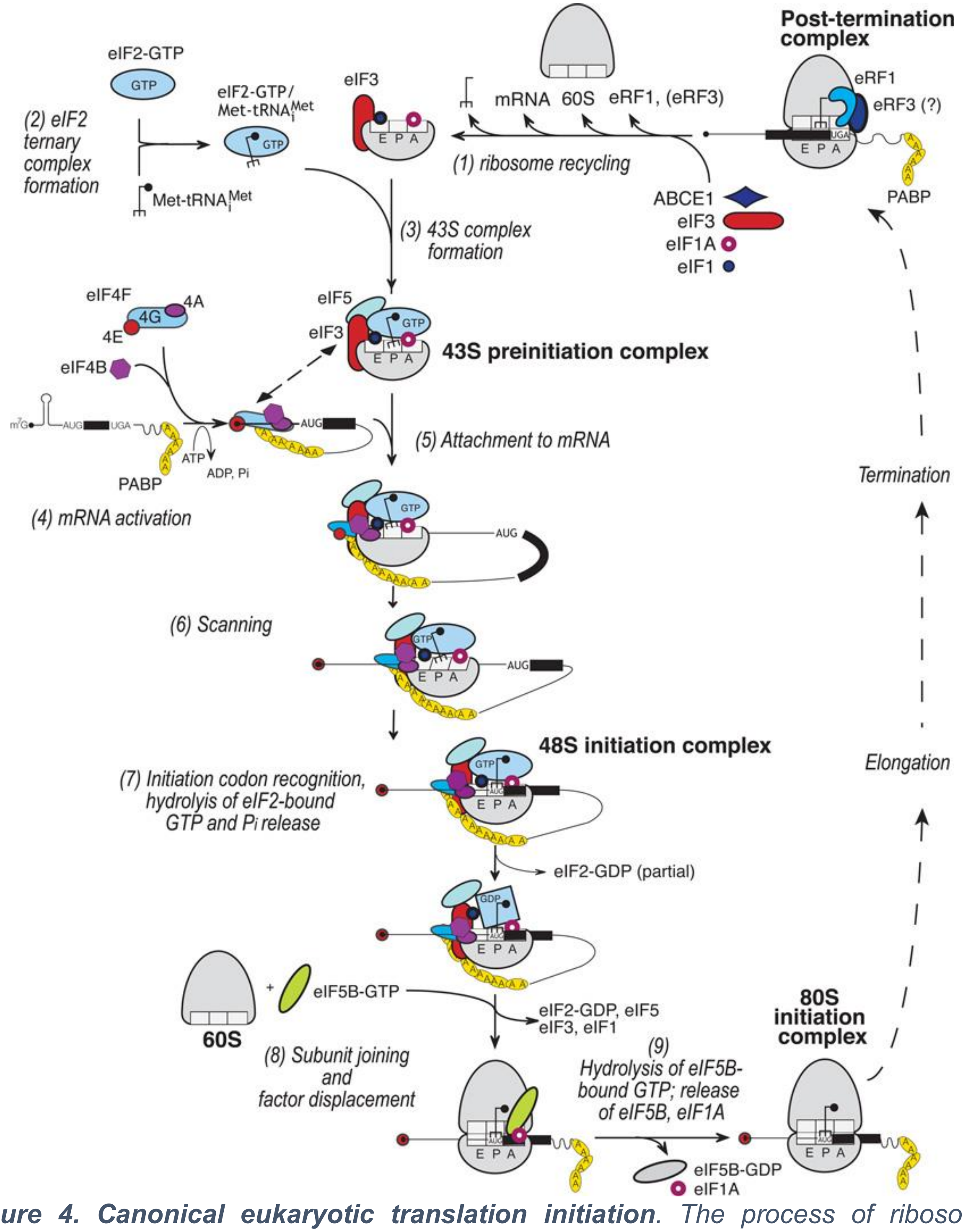


Figure 4. Canonical eukaryotic translation initiation*. The process of ribosomal complex formation involves initiation, elongation and recycling. Initiation occurs in nine steps, which include the recycling of components following elongation and termination. The eIF2 complex is formed in steps 2 and 3, followed by the formation of the 43S complex. In step 4, the mRNA is activated by eIF4, and the 43S complex binds to the mRNA in steps 5 and 6, scanning from 5' to 3'. Step 7 involves the recognition of the start codon, and in steps 8 and 9, the 60S ribosome subunit binds to the 40S, forming the 80S complex after the release of translation factors. Obtained from Jackson et al.[30].*

After the formation of the 43S complex, the scanning stage of the initiation phase begins. During this stage, the ribosomal subunit scans the 5'-untranslated region (5'-UTR) for the start codon. This scanning process is facilitated by the unwinding of the secondary structure of the 5'-UTR, allowing the ribosomal subunit to move along it[31]. In addition, eIF3 plays a crucial role in this stage by extending the mRNA binding channel and interacting with the mRNA next to the E-site. This interaction helps to position the mRNA in the right orientation for the ribosomal subunit to bind to the start codon[35]. The scanning process itself is an energy-intensive process that requires ATP hydrolysis[36]. The need for ATP is proportional to the complexity of the secondary structure of the mRNA being scanned[37]. During scanning, the involvement of eIFs that activate mRNA, like eIF4A, eIF4G, and eIF4B, remains a point of discussion. One proposed mechanism suggests that eIF4G is positioned near the E-site[38], which aligns with the postulation that a helicase-mediated "ratcheting" of mRNA unfolds the mRNA secondary structure by 40S subunits at their leading edge[39]. On the other hand, another suggestion involves eIF4A, eIF4G and eIF4B in the loosening of mRNA before it enters the ribosomal channel[40]. Scanning occurs at approximately 8 bases per second in the 5' to 3' direction[41].

To ensure that the Met-tRNA^{iMet} anticodon and the translation initiation codon are accurately paired, Kozak proposed a recognition site known as the "Kozak sequence" in 1987. This sequence is comprised of 10 nucleotides, GCC(A/G)CCAUGG[42], and recognized by eIF1. eIF1 plays a crucial role in the recognition of initiation codons by allowing the 43S subunit to distinguish against non-Kozak AUG codons[43,44]. Furthermore, eIF2 interacts with the purine bases that surround the AUG start codon, promoting a more stable conformation of the initiation complex[45]. Two nucleotides located at positions -3 and +4 (with the A of the AUG codon designated as +1) of the Kozak motif are highly conserved and their optimal composition greatly enhances the translation efficiency. On the other hand, the remaining consensus sequence surrounding the AUG codon has a relatively minor contribution to the overall efficiency[46].

Once the ribosome recognizes the start codon the locked ribosome is ready to begin the process of protein synthesis. This commitment step is facilitated by eIF5, which is an eIF2-specific GTPase-activating protein (GAP)[47]. eIF5 hydrolyzes GTP from the eIF2–

GTP–Met-tRNA-^{iMet} complex, causing eIF2 to lose its affinity for Met-tRNA-^{iMet}. This leads to a partial dissociation of the eIF2–GDP from the 40S subunit[48].

eIF5B facilitates the coupling of the 60S subunit and dissociation of eIF1, eIF1A, eIF3 and residual eIF2-GDP[49]. eIF5B displaces eIF2-GDP from the 40S subunit[50] and promotes 60S subunit joining by burying large solvent-accessible surfaces on both subunits[51]. After the 80S ribosomal complex is assembled, the elongation, termination, and recycling phases of the translation process occur.

Non-canonical protein synthesis

In eukaryotes, protein translation is considered to be monocistronic[52]. This means that each mRNA molecule is translated into a single protein product associated with only one open reading frame (ORF) or reference ORF (RefORF). The RefORF is also known as the coding sequence (CDS), which is delimited by a start (Kozak sequence) and a termination codon[53].

This assumption was fundamentally challenged by Lee in 1991 by identifying two distinct non-homologous proteins as being coded from the same transcript of Growth/differentiation factor 1 (GDF-1)[54]. Lee's work demonstrated that our understanding of mRNA translation was incomplete and, specifically, discovered that an upstream ORF coded for a 350-amino-acid protein of unknown function that was conserved in humans and mice. In exploring the mechanisms that might explain these alternative ORFs (AltORF)-derived proteins, two key possibilities, reinitiation and leaky scanning, were suggested[55].

1.2. Reinitiation and leaky scanning mechanisms

Ribosomal reinitiation is a process in which the ribosome, after the translation of an ORF terminates, moves downstream to reinitiate the translation of another ORF[56–58]. This mechanism takes place when the 80S ribosome is formed upstream the Kozak AUG site. The ribosome begins the translation process and produces a non-homologous protein at this upstream ORF (upORF). The ribosome then resumes scanning until it finds a downstream AUG. This mechanism occurs when the eIF4F complex participates in the primary initiation event at the uORF initiation codon[59].

While the function of the uORF region is not fully understood, researchers have found that it regulates the transcription of the downstream protein and some examples have demonstrated this functionality[60]. The 5'-UTR region of *Saccharomyces cerevisia*e mRNAs is known to have multiple uORFs, which play a crucial role in the expression of the reference ORF as they act as a transcription factor for other genes[61].

Another mechanism is leaky scanning which involves ribosomes bypassing the first AUG codon of an mRNA and starting translation at a downstream AUG codon[62]. This mechanism has been observed in eukaryotes, which have long 5′-UTRs with frequent uORFs[63]. To facilitate this process, eIF4G2, a homologue of the canonical translation initiation factor eIF4G1, is involved in leaky scanning for a subset of mRNAs. eIF4G2 thus takes the place of eIF4G1 during scanning of the 5′-UTR, and the need for eIF4G2 arises only when eIF4G1 dissociates from the scanning complex. This can occur when leaky scanning complexes interfere with initiating or elongating 80S ribosomes within a translated uORF. Moreover, recent studies have shown that leaky scanning is involved in the regulation of gene expression, particularly in stress response pathways[64]. Recent findings have shown that PRRC2 proteins play a role in the facilitation of leaky scanning. Through their interaction with eukaryotic translation initiation factors and preinitiation complexes, PRRC2 proteins actively participate in the translation of mRNAs containing uORFs. These uORFs are known to negatively impact translation efficiency. However, PRRC2 proteins counteract this inhibition by promoting leaky scanning, thereby enhancing the translation process of uORF-containing mRNAs. The abundance of PRRC2 proteins on ribosomes engaged in uORF translation further highlights their significance in regulating translation initiation and overall protein synthesis[65].

1.3. Translation from near-cognate AUG codons

Near-cognate AUG codons refers to codons that are similar to the AUG start codon but differ by one or two nucleotides. These codons can be used as alternative translation initiation sites (TISs) when the main AUG start codon is not available or is inefficient. Generally, during the elongation phase codon-anticodon pairing is done at the strict-monitored A-site of the ribosome. While the recognition of the start codon is mediated by the interaction of the codon AUG and Met-tRNA^{iMet} at the P-site. As the P-site is less restrictive it opens the possibility of a non-AUG recognition of the Met-tRNA^{iMet}[66]. The selection process involves an initial checkpoint at the open conformation of the 48S PIC, where almost all noncognate codons are rejected. For the four near-cognate codons with lower energy penalties (ACG, CUG, GUG, and UUG), the 48S PIC may proceed to the closed conformation and execute a second accuracy check to reject these near-cognate triplets and achieve stringent AUG selection[67]. Among the near-cognate AUG codons the most efficient and most abundant non-AUG codon is CUG[68]. Near-cognate codons have lower initiation efficiency than annotated AUGs and may be able to compete with annotated AUGs when located in 5'-UTR but not in CDS. Additionally, a good Kozak sequence has been shown to enhance the efficiency of translation initiation from near-cognate AUG codons[46]. It was observed that near-cognate AUG codons possessing a good Kozak sequence exhibited higher translation efficiency compared to near-cognate AUG codons with a poor Kozak sequence[69].

The usage of near-cognate AUG codons can have significant implications in the regulation of gene expression. This flexibility allows for a more precise control over gene expression and adaptation to changing environments[66].

1.4. Long non-coding RNA

A finding that shows a spread wrong annotation and goes against the dogma of molecular biology is the increasing evidence of translation from long non-coding RNAs (lncRNAs)[70–72]. These are transcripts that exceed 200 nucleotides and lack a RefORF[73,74]. In total, more than 100,000 lncRNAs have been annotated in humans[75], with around 15,000 derived from pseudogenes[76]. However, they generally show lower expression levels than mRNAs, with ~10-fold lower abundance[77].

Despite the fact that this subgroup of RNA molecules have been described as “transcripts of unknown function”[78], lncRNAs play important roles in regulating gene expression[79] and various physiological processes. They are particularly involved in regulating cell differentiation[80,81] and cell development[82]. Moreover, epigenetic modifications and control of chromatin architecture (protein-RNA condensates) are attributed to them[83–85]. They have also been described as enhancers for transcription factors, which indicates an intimate link between lncRNA expression and the spatial control of gene expression during development[86]. In addition to their roles in the nucleus, lncRNAs have also been found to have important functions in the cytoplasm and beyond. For instance, they are involved in the regulation of translation, metabolism and signaling[74].

The coding potential of lncRNAs has been identified following advances in deep transcriptomic sequencing. In particular, ribosome profiling (Ribo-seq), which involves deep sequencing of ribosome-protected fragments, has provided evidence of interactions between ribosomes and lncRNAs[87,88], and in certain studies, the protection patterns exhibited by the ribosomes suggest that translation into small proteins is feasible in the absence of ideal Kozak sequences[72,89,90]. Additionally, many lncRNAs present structures similar to mature mRNAs (transcribed by RNA polymerase II, 5'-capped, 3'-polyadenylated, splicing and found in the cytoplasm), which points to their protein-coding potential and, importantly, also to gene misannotations.

1.5. Prediction of coding sequences

With the advancements of genomic annotation, it became possible to identify potential coding genes in large eukaryotic genomes[91,92]. Genome annotation methods usually involve analyzing statistical information on codon usage, splicing sites, sequence similarity to other known proteins, and experimental evidence of transcript-derived sequences[93–95]. CDS predictions identify one open reading frame per transcript that has a statistically significant signature of a protein-coding region. This has paved the way for further research into the mechanisms of gene expression and has provided scientists with a better understanding of the genetic makeup of organisms. In cases where no significant protein-coding region is found, the longest ORF (>100 codons) is considered the most probable CDS. Currently, the most used protein sequence databases are still based on the mRNA monocistronic idea, use a 100-codon cut-off and do not contain the predicted small, non-RefORFs derived proteins.

Alternative open reading frames mRNA paradigm

In proteomic experiments, around 75% of the spectra remains unidentified[97]. From this it is frequently observed that a non-negligible fraction (about 10%[98]) of good-quality MS/MS spectra (high number of high intensity fragment ions) does not match predicted MS/MS spectra. Sometimes this can be attributed to proteoforms such as post-translationally modified proteins, genetic variants and alternative splicing-derived forms[99]. However, many unmatched MS/MS spectra cannot be attributed to such protein variants. Such observations led to the interest into searching for novel, non-expected proteins which are not yet included in traditional databases[100].

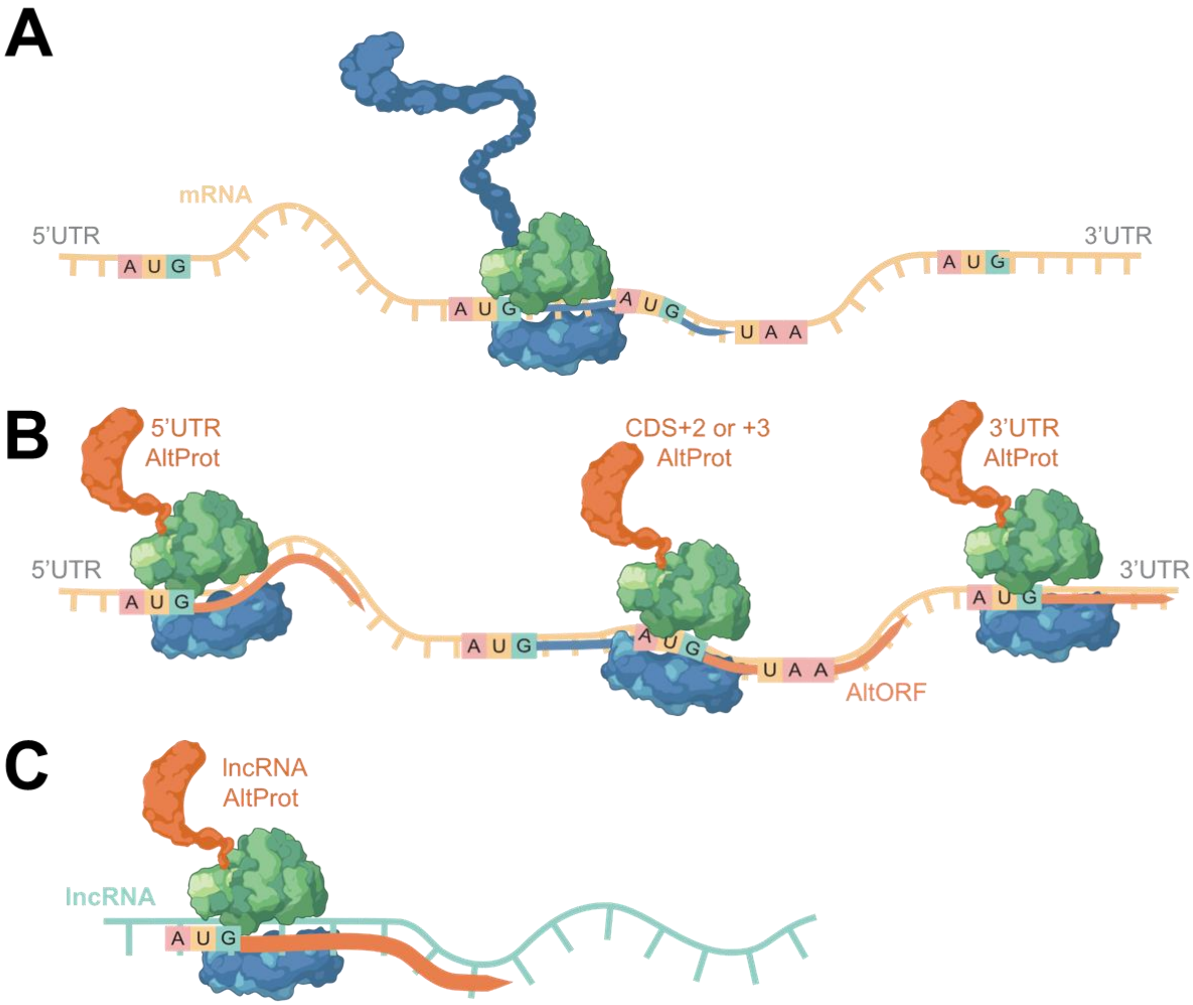


Figure 5. Schematic representation of the translation of RefProts and AltProts. *(A) Translation of RefProts from a RefORF (CDS) region. (B) Translation of AltProts from 5'- & 3'-UTRs and CDS +2, +3 frames. (C) Translation of AltProts from lncRNAs. Obtained from Garcia-del Rio et al*[96]

Bioinformatics studies allowed for significant advances in our understanding of the complexity of mRNA. Amongst others, the presence of alternative open reading frames AltORFs[101,102] was predicted, which can lead to the translation of alternative proteins (AltProts)[53]. As a result, we now have to consider that mature mRNA contains multiple open reading frames (ORFs). The RefORF is associated with the coding DNA sequence (CDS) and leads to the translation of a single, unique protein known as the reference protein (RefProt; **Figure 5A**). In addition to the RefORFs, AltORFs can originate from different start codons within mRNA molecules and can be found in various mRNA regions, including the 5'-UTR upstream of the RefORF, the 3'-UTR downstream of the RefORF, or frameshift (+1, +2) in the RefORF[103] (**Figure 5B**). Furthermore, it was shown that long non-coding RNAs (lncRNAs) can be translated into proteins, making them a type of AltORF and highlighting their wrong annotation (**Figure 5C**). These AltProts, also known as short open reading frame-encoded peptides (SEPs)[104], small proteins[105] or ghost proteins[106], have significant implications for our understanding of protein synthesis and the regulation of gene expression.

Roles of AltProts physiological processes.

Additionally to humans, AltProts have been identified across different species such as green algae[107], rice[108], *Arabidopsis thaliana*[109], *Saccharomyces cerevisiae*[110], mice[111], *Drosophila melanogaster*[112] and zebrafish[113]. This hints to the fact that AltProts can be involved in physiological processes. According to Samandi *et al*., several AltProts have been conserved throughout evolution[114]. Thus, human AltProts have homologs in other species.

Evolutionary origins of AltProts can be explained by a polymorphism of initiation and stop codons. A premature stop codon at the beginning of a coding sequence could lead to the emergence of a new, independent ORF in the 3'-UTR of the original gene if another translation initiation site can be used downstream. This way of AltORF formation, which is similar to gene fission, would possibly create a new altORF with the same protein domains as the annotated CDS. Another process that can explain AltORF existance is the '*de novo* ORF origin mechanism'[115]. In this concept, new ORFs can be transcribed and translated, leading to AltProts with novel functions. Alternatively, such new ORFs may expect the evolution of new functions through mutations[114].

The following examples (**Table 2**) showcase the involvement of AltProts in physiological processes. While the examples discussed here are just a small fraction of the known AltProts, they highlight important roles of AltProts in various biological processes.

Table 2. AltProt-mediated physiological processes.

AltProt	Amino acids	RNA type	Function	Same gene AltProt-RefProt functional pairs
ALEX	356	3'-UTR	Negative regulation of the activity of the G-protein XLalphas subunit, enhancing receptor-mediated cAMP formation.	Yes
Alt-ATXN1	185	CDS +2 ORF	Interacts with ATXN1, which is a Notch signaling repressor and is involved in brain development.	Yes
$A_{2A}R$ uORF5	134	5'-UTR	Unknown. Its expression is regulated by A2AR-mediated cAMP signaling.	Yes
MKKS 5'-UTR	43	5'-UTR	Localization inside the mitochondrial membrane.	Yes
MINAS-60	60	CDS +1 ORF	Down-regulation of the assembly of the pre-60S ribosomal unit.	No
PEP7	7	5'-UTR	Inhibition of the non-G protein-coupled signaling pathway of angiotensin II.	No
DWORF	34	lncRNA	Interaction with the sarco-/endoplasmic reticulum Ca^{2+}-ATPase. Enhances muscle contractility function.	No
MINION	84	lncRNA	Induces rapid cytoskeletal arrangement, involved in cellular fusion and muscle development.	No
Mitoregulin	56	lncRNA	Binding to cardiolipin increases calcium retention. Reduction of ROS.	No
Myoregulin	46	lncRNA	Inhibition of SERCA and impeding calcium uptake.	No
28aa-pncr003:2L	29	lncRNA	Uptake of cardiac Ca^{2+}.	No
29aa-pncr003:2L	29	lncRNA	Uptake of cardiac Ca^{2+}.	No
MOXI	56	lncRNA	Fatty acid metabolism and carbohydrate oxidation.	No
MRI-2	69	lncRNA	Ligation of DNA double-strand breaks.	No
NoBody	27	lncRNA	Interaction with mRNA decapping proteins. Protein translation.	No
AltATAD2	139	CDS +1 ORF	Interact with RPL10 as a potential 5S ribosomal RNA regulator.	No
Toddler	58	lncRNA	Embryonic development.	No

The existence of AltProts in humans has been established in around 15% of the protein identifications of different cell lines, tissues, and biological fluids such as cerebrospinal, urine, plasma and serum, highlighting their widespread presence in the human body[100]. It is interesting to note however, that despite the clear establishment of AltProts in different organisms, the function of these proteins remains largely unknown. This presents an opportunity for further exploration of the role of AltProts. On the other hand, several examples of AltProt-RefProt pairs originating from the same gene have been described and such pairs were investigated in molecular approaches[103]. Functional interaction between such AltProt-RefProt pairs has been shown: for instance, the pair XLalphas/ALEX, which are localized in the plasma membrane, and it was observed that ALEX negatively regulates the activity of the G-protein XLalphas subunit, enhancing receptor-mediated cAMP formation[116,117]. Another example of such a functional interaction is Alt-ATXN1. The AltProt is localized at the nucleus and directly interact with its genomic neighbor ATXN1[118]. A2AR and uORF5 were also found to functionally interact: when A2AR is stimulated by adenosine, the levels of uORF5 are upregulated[119]. Finally, a 5’UTR-coded protein was found at the mitochondrial membrane of HeLa cells along with the RefProt MKKS, and it was observed that upon the knockout of the AltProt, translation of MKKS increased, pointing to a regulatory role of the 5’UTR and AltProt expression[120].

Other AltORFs-coded proteins have been identified that do not involve RefProt-AltProt pairs. For instance, MINAS-60, a CDS +1-frameshift encoded protein that down-regulates the assembly of the pre-60S ribosomal unit. Depletion of this AltProt was shown to increase protein synthesis and cell proliferation, which suggests that it may have a role in regulating cell growth and division. On the other hand, the overexpression of the AltProt decreases the cytoplasmic 60S ribosomal subunit. Therefore, it slows the functional assembly of the large ribosomal subunit[121]. A peptide encoded by the 5'-UTR (PEP7) of the angiotensin type 1a receptor (AT_1R) gene was found to inhibit the non-G protein-coupled, β-arrestin signaling pathway of angiotensin II, while leaving the G protein-coupled pathway unaffected. This leads to decreased angiotensin II-stimulated phosphorylation of extracellular signal-regulated kinases 1/2 (Erk1/2). As a result, the consumption of salt is increased in rat models[122]. For the authors, this finding opens the question if PEP7 could potentially serve as a therapeutic agent to decrease salt craving and consumption in people with hypertension that is exacerbated by high salt intake. This discovery demonstrates the potential for AltProts to pave the way for new therapeutic implications and alternative treatments.

The last few years have seen an increase in the number of lncRNA-encoded proteins discovered that were found to function in a range of physiological processes. Among these proteins, one particular group was found to be involved in cardiac and skeletal muscle. One of the first identified members of this group is a 34-amino acid AltProt named DWORF, which was found at the sarcoplasmic reticulum in muscular cells. DWORF was shown to interact with the sarco/endoplasmic reticulum Ca^{2+}-ATPase and therefore, enhance muscle contractility function[123]. Other examples of lncRNA-encoded AltProts include MINION (microprotein inducer of fusion), an 84-amino acid protein found in skeletal muscle which plays an essential role in inducing rapid cytoskeletal arrangement, and which is necessary for cellular fusion and muscle development[124]. Similarly, mitoregulin (MTLN), an AltProt found in the inner mitochondrial membrane and which binds to cardiolipin, influences protein complex assembly(respiratory super complexes and fatty acid β-oxidation), increase respiration rate and calcium retention while decreasing reactive oxygen species (ROS)[125]. Myoregulin (MLN) is another example of a lncRNA-encoded protein found in skeletal muscle, where it inhibits SERCA and impedes calcium uptake. Upon its inhibition in mice, enhanced exercise performance was demonstrated[70]. Additionally, two AltProts which are encoded from a lncRNA of *Drosophila sarcolamban*, were found involved in cardiac Ca^{2+} uptake and exhibit ancient conservation between species[126]. These findings demonstrate that AltProts function in vital organs such as the heart and skeletal muscle.

Another example is the micropeptide regulator of β-oxidation (MOXI), which is encoded by a noncoding RNA and is a 56-amino acid peptide that is localized at the inner mitochondrial membrane. Interestingly, hearts of MOXI knockout mice showed a decrease in the ability to metabolize fatty acids and increased carbohydrates oxidation[127]. Another AltProt is MRI-2, which is a 69-amino acid encoded protein involved in DNA double-strand break ligation. It interacts with Ku heterodimer and plays a crucial role in this process[128]. In addition, NoBody, a protein encoded by a lncRNA, interacts with mRNA decapping proteins involved in protein translation[129]. Furthermore, AltATAD2 originates from the mRNA coding for RE/poly(U)-binding/degradation factor 1 (AUF1) and was found to interact with ribosomal protein 10 (RPL10) as a potential 5S ribosomal RNA regulator in HeLa cells[130]. This interaction sheds new light on the role of AltProts in ribosomal RNA regulation and highlights the complex interplay between AltProts and mRNA regulation. Toddler is yet another interesting example of an AltProt. This lncRNA-derived 58-amino acid secreted peptide was described in zebra fish embryogenesis where it promotes the movement of mesodermal cells during the formation of zebra fish gastrulation. Upon Toddler's knockout, 0 of 25 embryos survived, highlighting the critical role of AltProts in embryonic development[131].

Roles of AltProts pathology processes.

As mentioned in the previous section, AltProts appear to play significant roles in various physiological processes. In recent years, there has been a growing body of evidence pointing to their involvement in different pathological mechanisms. One AltProt that received particular attention is SPAR (small regulatory polypeptide of amino acid response), which is encoded by the lncRNA LINC00961. Inactivation of this peptide is involved in muscle regeneration by modulating mTORC1[132] and, while the therapeutic potential of SPAR is under investigation, its pathological involvement is also being explored[133].

AltProts have also been identified to play a role in cancer development. For example, Cardon *et al*. demonstrated that glioma cells express three different AltProts, AltMAP2, AltTRNAU1AP and AltEPHA5, which were shown to interact with each other in a way that might be associated with cellular mobility and tRNA regulation[134]. In colorectal cancer, the AltProt Splicing Regulatory Small Protein (SRSP) was identified. SRSP is a lncRNA-derived protein that interacts with splicing regulators. When SRSP is upregulated, it is associated with tumorigenesis and poor prognosis of colorectal cancer patients[135], which suggests that SRSP might play a key role in cancer progression. Another example is CASIMO1. This 10 kDa protein interacts with squalene epoxidase (SQLE), which is a known oncogene in breast cancer. Upon overexpression, SQLE accumulates, leading to lipid droplet clustering[136]. Given this finding, the authors concluded that CASIMO1 is involved in carcinogenesis and cell lipid homeostasis. Additionally, a study employing mass spectrometry (MS) imaging and top-down microproteomics identified four AltProts in tumor regions of serous ovarian cancer biopsies[137]. Such findings suggest that AltProts have a profound impact on cancer development and progression, yet further research on their functions and interactions is warranted.

On the other hand, several AltProts exhibit tumor suppression characteristics. For instance, Huang *et al*. discovered that a peptide encoded by a lncRNA known as HOXB-AS3 plays a critical role in suppressing tumor growth. This peptide is particularly important for colorectal cancer patients, as low levels of it are associated with poor prognosis and disease progression[138]. Another example of an AltProt with anti-cancer properties is ASRPS. This is a 60-amino acid, lncRNA (LINC00908) encoded AltProt that was shown to reduce angiogenesis, the process by which new blood vessels are formed, thus inhibiting the growth of cancer cells. Interestingly, ASRPS expression was found to be downregulated in triple-negative breast cancer and is associated with poor overall survival[139]. Other AltProts that were described as tumor suppressors include a 146-amino acid protein encoded by the non-coding SNF2 histone linker PHD RING helicase, which was found to be downregulated in glioblastoma. This AltProt protects against ubiquitin-proteasome degradation of full-length SHPRH. In addition, its downregulation is associated with increased tumorigenicity and cell proliferation in glioblastoma, which suggests that it plays a critical role in cancer development and progression[140]. Recent studies identified an AltProt, FBXW7-185aa, involved in glioblastoma tumorigenesis. Liquid chromatograph mass spectrometry (LC-MS) analysis showed that FBXW7-185aa is downregulated in glioblastoma cells and its suppression was found to enhance malignant phenotypes *in vitro* and *in vivo*, while its upregulation prevented proliferation and cell cycle acceleration[141]. This suggests that FBXW7-185aa may be a promising target for the development of novel cancer therapies.

Finally, researchers have discovered an 87-amino-acid AltProt from the long intergenic non-protein-coding RNA p53-induced transcript (LINC-PINT) that is downregulated in glioblastoma. This AltProt interacts with the polymerase associated factor complex (PAF1c) and prevents several oncogenes from extending their transcription, which suppresses glioblastoma cell proliferation[142]. Therefore, LINC-PINT may play a crucial role in regulating gene expression in glioblastoma and thus have potential as a therapeutic target for the treatment of this disease.

All these findings (**Table 3**) suggest that some AltProts also play a crucial role in the regulation of cell proliferation and the development of tumors and could be targeted for the development of new treatments for pathologies and cancer.

Table 3. Summary table of the AltProts referenced in pathological processes.

AltProt	Amino acids	RNA type	Pathology	Function
SPAR	90	LncRNA	Muscular injury	Upon inactivation, muscle regenerates via mTORC1.
AltMAP2	103	CDS +3 ORF	Glioblastoma	Interaction with TPM3. Associated with cellular mobility and transfer RNA regulation.
AltTRNAU1AP	47	3'-UTR	Glioblastoma	Interaction with TPM3. Associated with cellular mobility and transfer RNA regulation.
AltEPHA5	56	CDS +3 ORF	Glioblastoma	Interaction with TPM3. Associated with cellular mobility and transfer RNA regulation.
SRSP	130	LncRNA	Colorectal cancer	Associated with the tumorigenesis and poor prognosis.
CASIMO1	83	LncRNA	Breast cancer	Interaction with SQLE. Enrollment in carcinogenesis and cell lipid homeostasis.
HOXB-AS3	53	LncRNA	Colorectal cancer	Tumor suppressor and regulator for PKM splicing.
ASRPS	60	LncRNA	Breast cancer	Reduction of angiogenesis.
SHPRH-146aa	146	ncRNA	Glioblastoma	Protects SHPRH from degradation. Its down regulation is associated to increased cell proliferation.
FBXW7-185aa	185	ncRNA	Glioblastoma	Prevents proliferation and cell cycle acceleration
LINC-PINT	87	LncRNA	Glioblastoma	Interacts with PAF1c. This prevents several oncogenes from extending their transcription.
AltCMBL	42	3'-UTR	Ovarian cancer	Unknown. Potential serous ovarian cancer marker.
AltGNL1	64	CDS	Ovarian cancer	Unknown. Potential serous ovarian cancer marker.
AltRP11-576E20.1	31	LncRNA	Ovarian cancer	Unknown. Potential serous ovarian cancer marker.
AltCSNK1A1L	44	3'-UTR	Ovarian cancer	Unknown. Potential serous ovarian cancer marker.

Research into alternative proteins

1.6. Transcriptomics approaches

Transcriptomics is the study of the transcriptome, which is the complete set of RNA molecules produced by the genome in a single cell or a tissue. It has revolutionized our understanding of gene expression, alternative splicing, single-cell and spatial dynamics.

1.6.1. RNA sequencing

RNA sequencing (RNA-seq) is a high-throughput sequencing method that enables the analysis of the expression of multiple transcripts in different physiological or pathological conditions[143]. RNA-seq is particularly useful for studying differential gene expression (DGE), which allows to compare transcript levels between different conditions and identify transcript that are regulated in response to stimuli, pathology or environmental factors.

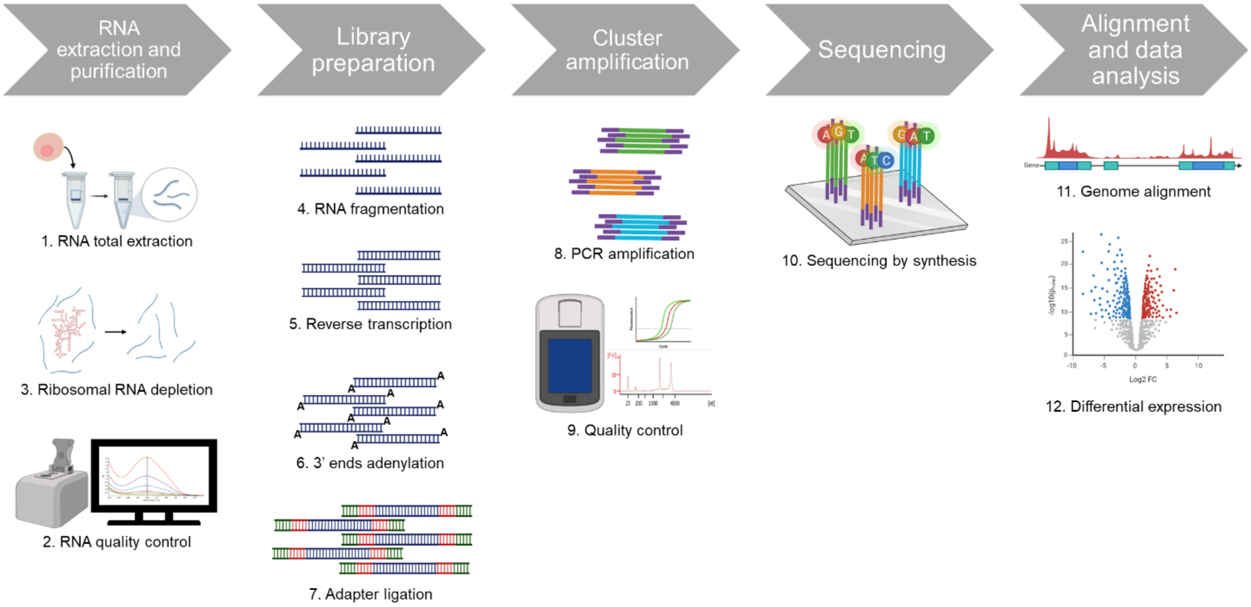


***Figure 6. RNA-seq workflow overview**. Steps 1-3 correspond to RNA extraction and purification. Steps 4-7 correspond to the library preparation phase. Steps 8 and 9 are part of the cluster amplification process. Step 10 corresponds to sequencing by synthesis. The final phase is sequence analysis, alignment, transcript annotation and analysis.*

RNA-seq can generate high-resolution transcriptome maps that provide detailed information on the transcript structure and level of expression[144]. This is made possible by next-generation sequencing (NGS) platforms such as Illumina, MGI, Pacific Biosciences (PacBio) and Oxford Nanopore (ONT). The choice of the NGS platform depends on the organism to study, features, benefits, experimental design and research questions. For example, Illumina's sequencing technology is widely used for its accuracy, scalability and speed, while the ONT and PacBio platforms are known for their long reads and ability to sequence DNA in real-time. MGI's sequencing technology is known for its affordability and high-throughput, making it a popular choice among researchers who require large amounts of data.

1.6.1.1. RNA extraction and purification

The general RNA-seq workflow comprises of several stages, each of which is critical to obtaining high quality data. The first stage involves RNA extraction and purification (**Figure 6**, steps 1-3). To extract RNA, different extraction methods can be used, such as TRIzol extraction[145], silica spin columns-based[146], magnetic beads-based[147], modified guanidinium thiocyanate-phenol-chloroform extraction[148] and ultrafiltration[149].

Once RNA has been extracted, its quality must be assessed, which is determined by several factors, including the integrity and purity of the RNA molecules. The most commonly used methods to evaluate RNA quality are gel electrophoresis and spectrophotometric analysis. With agarose gel electrophoresis the ratio of the quantities of the ribosomal RNA molecules are observed as a degradation status control (18S and 28S subunits). However, this method has limitations as it only provides a rough estimate of the RNA quality. A more objective approach was described by Schroeder *et al*.[150], who calculated an RNA integrity number (RIN) from the electropherograms generated by Agilent Bioanalyzers. The RIN value considers the area occupied by 18S and 28S rRNA, the height of the 28S peak, the presence or absence of RNA degradation products, the fast area ratio and marker height. Generally, a RIN value of one corresponds to fully degraded RNA and a value of ten indicates intact RNA[150].

Besides electrophoresis, spectrophotometric methods are used to quantify RNA and evaluate its quality. RNA absorbance is measured at λ= 260 and 280 nm, and the ratio between both values (A260/A280) is an indicator of chemical contamination; if below 1.8, it generally indicates that the RNA is not pure[151]. Therefore, it is important to consider both the RIN and spectrophotometric methods when assessing RNA quality.

The next important step in RNA-seq is either mRNA isolation or rRNA depletion. This step is crucial as it ensures that only the desired mRNA or total (ncRNA and mRNA) molecules will be sequenced and analyzed. For mRNA isolation, the most common method is using beads coated with oligo(dT) primers. These primers capture the 3’ poly-A tail of mRNA, which distinguishes it from other RNA molecules[152]. However, this technique has a limitation in that it cannot enrich non-polyadenylated RNA molecules (many lncRNAs) or degraded samples. To overcome this limitation, rRNA depletion kits are employed that remove rRNA, which is the major RNA type (90%) in samples[153]. Indirectly, this leads to an enrichment of mRNA and ncRNA, allowing more in-depth RNA-seq. Two main approaches are used in commercially available rRNA depletion kits. In the first approach, beads coated with rRNA complementary oligonucleotides capture rRNA, which is then precipitated[153]. The second approach involves hybridizing rRNA to single-stranded DNA oligonucleotides, followed by the degradation of the DNA-RNA molecules using ribonuclease H (RNase H) and DNAse I enzymes[154]. When selecting an appropriate enrichment method, it is important to consider the project aims, platform and the characteristics of the RNA mixture. By carefully selecting the right method, accurate and reliable results can be obtained.

1.6.1.2. Sequencing library preparation

After RNA has been extracted and depleted, the next step is to convert it into a library of cDNA fragments that can be sequenced. This is done through a library preparation process that involves several steps (**Figure 6**, steps 4-9).

The first step of this process is the fragmentation of RNA. This is a crucial step, especially for Illumina short-read sequencing. RNA molecules are typically very long and sequencing them directly would not be feasible. By fragmenting the RNA, the molecules become more manageable, allowing for better sequencing results. Then, fragmented RNA is reverse transcribed to cDNA. Reverse transcriptases are used here, which read the RNA molecule and create a complementary strand of DNA. This new molecule, cDNA, is more stable and better suited for sequencing.

Once cDNA has been generated, the library preparation phase starts. This consists of the addition of sequencing adapters to both ends of the cDNA through a process called adapter ligation, which involves the addition of short, platform-specific nucleotide adapter sequences to the ends of cDNA fragments. These adapter sequences allow the cDNA to bind to the sequencing platform and are necessary for the amplification and sequencing steps of NGS[155]. An alternative to adapter ligation is tagmentation, which combines the fragmentation and adapter ligation steps by using an enzyme called transposase. Transposase cuts the DNA fragment and ligates the adapters in one step, speeding up the process and reducing the risk of errors[156]. Adapter-ligated fragments are then PCR-amplified and gel-purified.

The next step is cluster generation which involves the amplification and clustering of the DNA fragments. First, the library is denatured and loaded into the Illumina flow cell. There, single-stranded DNA (ssDNA) is immobilized across a flow cell which contains two different immobilized oligonucleotides that are complementary to one of the two adapter sequences on the ssDNA. Following complementary binding, any unbound DNA is washed away to ensure that only the ssDNA fragments that have bound to the oligonucleotides remain. These ssDNA fragments are elongated by a DNA polymerase from the immobilized oligonucleotide, which results in an immobilized double-stranded DNA (dsDNA) fragment. These fragments are then bent, which causes the other adapter sequence, not attached to the flow cell, to bind to a nearby complementary oligonucleotide on the flow cell. The bent ssDNA is then elongated, resulting in a dsDNA bridge, which is then denatured, leading to the formation of two ssDNA strands bound to the flow cell. This process results in several clusters of hundreds of millions of ssDNA strands clonally amplified. When cluster generation is complete, the templates are ready for sequencing.

1.6.1.3. Sequencing

For the Illumina platform, the sequencing technology developed is called sequencing by synthesis (**Figure 6**, step 10). This technology uses four fluorescently labeled nucleotides to detect single bases as they are incorporated into growing DNA strands.

Sequencing begins with the extension of the first sequencing primer to produce the first read. With each cycle, fluorescently tagged nucleotides compete for addition to the growing chain. Only one is incorporated based on the template sequence. After each nucleotide addition, the clusters are excited by a light source and emit a characteristic fluorescent signal. The length of the read is determined by the number of cycles, while base calling is determined by the emission wavelength and signal intensity. All identical strands in each cluster are read simultaneously, and hundreds of millions of clusters are sequenced in a massively parallel process.

After completion of the first read, the read product is washed away. In this step, the first index read primer is introduced and hybridized to the template. After completion of the index read, the read product is washed off and the 3'-ends of the template are deprotected. The template now folds over and binds the second adapter on the flow cell. The second index read is performed and polymerases extend the second flow cell oligo, forming a double-stranded bridge. This double-stranded DNA is then linearized and the 3'-ends are blocked. The original forward strand is cleaved off and washed away, leaving only the reverse strand.

The paired read begins with the introduction of a second sequencing primer. Like the first read, the sequencing steps are repeated until the desired read length is achieved. The paired read product is then washed away. This entire process generates millions of reads representing all fragment sequences. Sequences are separated based on the unique indexes introduced during sample preparation. For each sample, reads with similar stretches of base calls are locally clustered. Forward and reverse reads are paired, creating contiguous sequences. These contiguous sequences are aligned back to the reference genome for variant identification. An advantage of pair reads is that the resulting information is used to resolve ambiguous alignments.

1.6.1.4. RNA-seq data analysis

Once the RNA-seq data are acquired, the analysis phase starts (**Figure 6**, steps 11-12). The quality of the data affects the analysis, hence the need for a thorough quality control (QC), which involves analyzing different parameters such as read quality, presence of adaptors, GC content, over-representation of *k*-mers and duplication levels, length and N-bases content[157]. This helps in assessing the quality of the data and removing low-quality reads. There are many QC tools available for NGS data, with the most commonly used ones being FASTQC[157], NGS QC[158] and Trimmomatic[159].

After the quality is assessed, the reads are aligned to the annotated reference genome. Such alignment is necessary to discover the reads origins with respect to the intron and exon annotated reference sequence. Various software have been developed to perform this task and can be divided in two categories: unspliced aligners and spliced aligners[160]. Unspliced aligners map the reads against the annotated reference transcriptome, while spliced aligners map the reads to a reference genome. The latter allows to match intronic sequences during the alignment. RNA-seq alignment tools typically consist of two main steps. The first step, indexing, involves structuring the reference genome to enable fast matching of reads to specific regions. This is accomplished by creating a data structure that allows the aligner to efficiently locate genome regions that align with a given read. The second step is the alignment itself, where the reads are compared to the indexed genome to find the best match. RNA-seq alignment tools employ various algorithms for this purpose. Some common algorithms include hashing, which involves creating a hash table of the reference genome to speed up read matching[161] and suffix trees that enable quick identification of all occurrences of a specific pattern in a string[162]. Additionally, RNA-seq alignment tools offer adjustable parameters to control the alignment process. These parameters include the allowance for mismatches and gaps in the alignment. Algorithms used include TopHat[163], Spliced Transcripts Alignment to a Reference (STAR)[162], Bowtie[164] and HISAT[165]. After the alignment, QC is recommended to evaluate the successful reference-based alignment. The most used algorithm to assess the quality of this alignment is MultiQC[166], which supports different alignments and processing tools.

The main application of RNA-seq is transcript quantification, which involves measuring the transcript and its levels by computing the number of reads of a sequence. In this process, two types of algorithms can be used. Union-exon base algorithms merge all the overlapping exons of the same gene. Examples of such algorithms are FeatureCounts[167], easyRNASeq[168] and HTSeq[169]. The second category of algorithms is transcript-based and comprises the majority and most popular tools such as Kallisto[170], RSEM[171] and Salmon[172]. These algorithms were developed to improve the accuracy and efficiency of transcript quantification.

RNA-seq normalization is a crucial step in RNA-seq data analysis that adjusts raw transcriptomic data to account for various technical factors that may mask actual biological effects and lead to incorrect conclusions[173]. There are several factors that affect transcript quantification in RNA-seq data, such as sequencing depth, GC-content, transcript length, sequencing error rate, and sample-to-sample and batch-to-batch variability[174]. Several normalization methods exist to minimize these variables and ensure reliable transcriptomic data.

Reads per kilobase of transcript per million reads mapped (RPKM) is a normalized gene expression unit, normalized to correct the gene (transcript) lengths and library sizes (sequencing depth). It is calculated by dividing the number of reads that align to a particular gene by the length of the gene in kilobases, and then dividing that number by the total number of reads in the sample (in millions)[175]. Fragments per kilo base of transcript per million mapped fragments (FPKM) is analogous to RPKM but considers the length of the sequenced fragments rather than the length of the gene itself. FPKM is calculated by dividing the number of fragments that align to a particular gene by the length of the gene in kilobases, and then dividing that by the total number of fragments in the sample (in millions)[176]. Transcripts Per Million (TPM) is the most wide-spread unit of measurement used to quantify gene expression levels. It is similar to RPKM and FPKM but takes into account the number of transcripts rather than the number of reads or fragments. TPM is calculated by dividing the number of reads that align to a particular gene by the length of the gene in kilobases, dividing this number by the total number of reads in the sample (in millions), and then dividing it by the effective length of the gene in kilobases[176].

Once transcript quantification and normalization are done, the next step is DGE assessment, which is the actual comparison of gene expression levels between different conditions. It identifies genes that are differentially expressed and play a significant role in the biological process under investigation. Several software tools are available for DGE analysis, each with its own strengths and limitations. Some of the widely used algorithms include edgeR[177], DESeq[178] and limma-voom[179,180]. EdgeR is a statistical package designed for analyzing RNA-seq data. It uses a negative binomial distribution to model the count data and provides a robust and accurate method for identifying differentially expressed genes[177]. DESeq is another popular tool that uses a similar approach to EdgeR; it uses a negative binomial distribution to model the count data and allows for normalization of the data[178]. Limma-voom uses a linear model to identify differentially expressed genes and it is known for its speed and accuracy and widely used in gene expression analysis[179,180].

RNA-seq is thus a comprehensive approach to study gene expression. Moreover, it can demonstrate the existence of predicted transcripts and their abundance in different organisms (bacteria, yeast and virus), tissue, cell lines, primary tumors and single-cells. Moreover, sequenced transcriptomes can be employed to generate sample-specific, *in silico* translated protein sequence databases. Due to the alignment against a reference genome, mutations can be detected and introduced in such databases, allowing to identify the presence of mutated proteins.

1.6.2. Variant identification

The process of analyzing identifying genomic variants from RNA-seq reads typically involves several steps. First, the reads are mapped to a reference genome to determine their origin. This mapping helps establish the relationship between the reads and the reference genome, providing insights into the genetic information contained within the reads. Then, a variant calling algorithm is used to identify single nucleotide variants (SNVs) and small insertions/deletions (indels) that differ from the reference genome. A variant caller algorithm will differentiate a sequencing error from a nucleotide variant based on the sequencing depth. This sequencing depth refers to having multiple reads covering a specific position. Even if one of the reads contains a "wrong base" due to alignment errors, the consensus base of the remaining reads can identify it as a sequencing error. Variant caller software, such as GATK[181], Samtools[182], FreeBayes[183] and Platypus[184] and play a crucial role in distinguishing between errors and genuine genetic variants. These tools enable accurate identification and characterization of variations in the genome, aiding in the analysis and interpretation of genomic data. These identifications of genomic variants is an interesting field because they can impact in transcripts that encode AltProts. OpenVar is a tool designed to annotate genomic variants in AltORFs and predict they functional effect[185].

1.6.3. Identification of novel splicing events

Identification of novel splicing events in RNA-Seq data allows us to detect splice junctions that have not been previously annotated. By aligning the RNA-Seq reads to a reference genome using specialized spliced aligners[162,186], like we can effectively identify reads that span splice junctions. This step is important as it helps us generate assembled transcripts using transcript assemblers. These assembled transcripts can then be compared to annotated transcripts to identify novel isoforms, thus adding to our understanding of splicing diversity. However, careful filtering is needed during this analysis to distinguish true events from artifacts or contamination. Factors such as the number of sequencing dept, splicing motifs, and exon overhang length can be considered during the filtering process. By incorporating these novel junctions into gene annotation databases, we enhance the completeness of these databases and provide a more comprehensive representation of splicing diversity.

1.6.4. Ribosome profiling

Ribo-seq is a powerful technique to investigate protein synthesis and translation efficiency on a systems-wide level. This technique was developed in 2009 by Ingolia *et al*.[71] and involves the sequencing of ribosome-protected fragments (RPFs) that are generated from the mRNA transcripts during translation. These RPFs are usually 30 nucleotides of length. Ribo-seq involves treating the cells with cycloheximide, pateamine A, lactimidomycin, harrintonine, or puromycin; translation inhibitors that freeze elongating ribosomes along the mRNA molecules. Also, TIS detection is enabled if specific antibiotic treatments are used (harringtonine or lactimidomycin). After treatment, cells are lysed and lysates are treated with high-salt buffers or nonionic detergents, which stabilize the ribosome-mRNA interactions and disrupt non-specific interactions[187]. Then, RNases are used to digest the unprotected single-stranded RNA molecules, separating these from the RPFs[188]. Once the RPFs are isolated, RPF mRNA is separated by polyacrylamide gel electrophoresis (PAGE), size exclusion chromatography (SEC) or magnetic beads[189]. To isolated and purified RPF-derived mRNA, NGS adapters are ligated to start RNA-seq as described in the previous section. The resulting Ribo-seq data consists of short reads that correspond to the ribosome-protected mRNA fragments. These reads can be aligned to a reference genome or transcriptome to determine the positions of ribosomes along the mRNA molecules and analyze various aspects of translation.

Algorithms have been developed for the sole purpose of performing QC, visualizing and statistical analysis of Ribo-seq data. RiboVIEW offers visualization tools to explore and interpret ribosome positions and translation dynamics. It allows visualization of ribosome occupancy profiles along mRNA transcripts, detect translated regions, identify translation initiation sites, ribosome speed and density, and significance of observed translation events[190]. Another tool is Trips-Viz, which allows to study translation dynamics and ribosome movement[191]. Additionally, It also includes statistical methods for identifying differentially translated genes and for performing enrichment analysis[192]. This technique has been key for the research of AltProts and has been used to discover novel ORFs[193] and functional micropeptides[194].

1.7. Mass spectrometry-based proteomics

Proteomics is a highly specialized discipline that comprehensively studies the proteins present in a given system of interest, such as an organelle, cell, tissue, organ, fluid or species. It encompasses the analysis of protein abundance and physiological/pathological activity.

In the post-genomics era, significant progress has been made in the development of techniques and computational tools for proteomic research. These developments aim to address complex biomedical challenges that were previously difficult to solve. One such development is high-throughput proteomics, which allows for large-scale protein characterization, achieved by reducing the analysis time of protein samples, whilst increasing the accuracy and depth of proteome coverage[195]. Mass spectrometry-based proteomics is a powerful method for identifying proteins in different (disease) situations. Over the past 40 years, it has become indispensable in cellular and molecular life sciences[196] as it is used to identify and quantify proteins in purified and complex mixtures. Additionally, it can provide information on protein structure, post-translational modifications and protein-protein interactions.

To enhance MS-based proteomics, peptide separation techniques and enhanced MS instruments have been developed. For instance, separation techniques based on high-performance liquid chromatography (HPLC) are commonly used in the field. HPLC enables the continuous separation of many peptides from highly complex mixtures and can be paired with MS as LC-MS for their identification.

There are two main categories of MS-based proteomic experiments: (1) top-down proteomics, (2) bottom-up, shotgun proteomics[195]. In top-down proteomics, an entire full size protein is fragmented inside the mass spectrometer and the resulting fragments are analyzed. In bottom-up, shotgun proteomics (a bottom-up technique), involves the extraction of proteins from a well-characterized system, such as cell lines. The extracted proteins are then digested by a protease, such as trypsin, and the resulting peptides are fractionated via HPLC, which is online coupled to a high-resolution mass spectrometer (HRMS). After the acquisition of the LC-MS/MS data, it is matched to identify the target proteins and their associated modifications in a protein sequence database by discovery engines (database-driven approach). Finally, the identifications are statistically scored.

1.7.1. Top-down alternative proteins research

As already mentioned, top-down proteomics avoids using enzymatic protein digestion, which can create concerns with the identification of proteoforms due to the homology of some peptides. These concerns are known as the protein inference problem. Instead, top-down proteomics analyses intact proteins[197] and, therefore, more accurate proteoform identification should be possible by distinguishing between variants and isoforms, localizing post-translational modifications (PTMs) and quantifying isoform expression levels.

Analyzing intact proteins by MS is a complex process and requires the use of sophisticated technology and analysis methods. LC-MS top-down analysis, in particular, is a major challenge due to the wide variation in protein physicochemical properties (size, charge and hydrophobicity). Further, it is necessary to have access to mass spectrometers that are capable of resolving very large protein species at sufficiently high resolution. Fourier-transform ion cyclotron resonance (FTICR)[198] and quadrupole time-of-flight (q-TOF)[199] mass spectrometers have traditionally been used to analyze large proteins. However, in recent years, efforts by several groups have enabled Orbitrap mass spectrometers to increase their range and analyze larger proteins[200]. These advances were achieved by tuning different acquisition parameters in the most commonly used mass spectrometers for proteomic research. Additionally, ultra-high mass range (UHMR) Orbitraps have been developed, which are capable of analyzing large protein species (up to 70,000 m/z) due to hardware improvements in ion transmission and detection[201]. Top-down proteomics also requires special ion activation methods for efficient fragmentation of intact proteins such as higher-energy collisional dissociation (HCD), electron-transfer dissociation (ETD), electron-transfer/higher-energy collision dissociation (EThCD) and ultraviolet photodissociation (UVPD)[202].

Studying AltProts by top-down proteomics is interesting as it could help to annotate larger protein sequences and not only partial sequences (peptide sequences) which may fully overlap with RefProt sequences. Identifying protein variants could thus lead to identifying AltProts, which can be mutated in a pathological context. This is particularly relevant in cancer, where the identification of potential markers that have never been considered could lead to new diagnostic and therapeutic approaches. The potential of discovering novel proteins and AltProts by top-down analysis was shown in 2009. Five sORF-encoded proteins were detected in *M. acetivorans* using a Thermo Scientific 12 T LTQ-FT Ultra[203]*.* In a second study, 12 AltProts were identified by solid phase extraction (SPE) followed by 2D-LC-MS top-down analysis. Moreover, using both HCD and EThcD ion activation, 36 proteoforms were mapped to these 12 AltProts.

In human tissues, two examples of this approach are the studies conducted by Prof. Isabelle Fournier in 2017[204] and 2018[137]. In both, her team described a strategy that employed non-targeted molecular classification by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI). This technique enables the localization of regions of interest based on their molecular protein signature on the surface of a thin section of tissue. Within the tissues, regions of interest defined by MALDI-MSI were used for microproteomics by micro-extraction of the tryptic peptides after on-tissue enzymatic digestion. For glioma, more than 2,500 proteins including 22 AltProts were identified by shotgun microproteomics[204]. A similar approach was used in their study of ovarian cancer. First, the regions of interest were delimited by MALDI-MSI and then liquid micro-junction and parafilm-assisted manual microdissection were used as methods for microextraction. With this approach, 15 AltProts were identified, including alternative G protein nucleolar 1 (AltGNL1) found in the tumor, and translated from an AltORF nested within the GNL1 canonical coding sequence[137]. The authors described that the study of AltProts by spatially resolved top-down proteomics is a means to evaluate protein changes in the case of serous ovarian cancer, allowing the detection of potential markers that had not been considered.

1.7.2. Bottom-up, shotgun proteomics to identify alternative proteins.

Shotgun proteomics is widely used to identify and quantify proteins in complex biological samples. **Figure 7** displays the general workflow of shotgun proteomics which involves cell lysis, sample preparation and protein digestion. The resulting peptides are then separated and analyzed using LC-MS/MS[205].

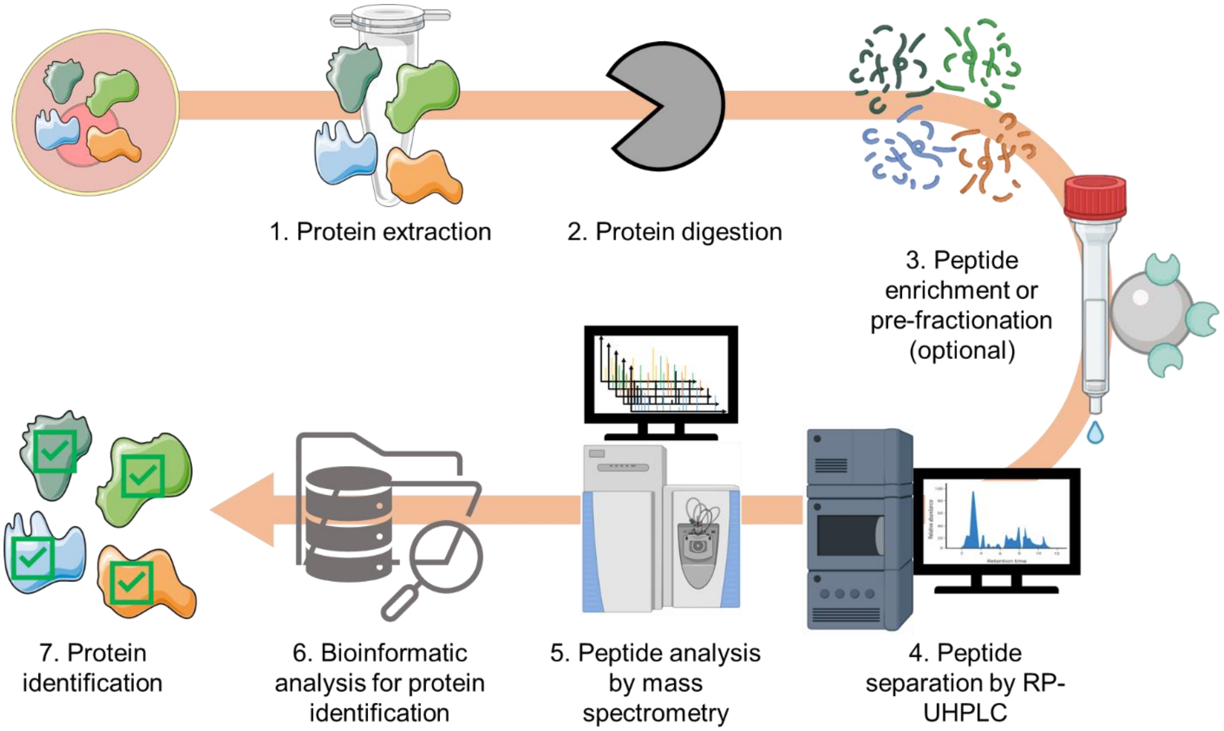

1. Protein extraction
2. Protein digestion
3. Peptide enrichment or pre-fractionation (optional)
4. Peptide separation by RP-UHPLC
5. Peptide analysis by mass spectrometry
6. Bioinformatic analysis for protein identification
7. Protein identification

Figure 7. General workflow of a bottom-up shotgun proteomics experiment. *The workflow involves several steps, including cell lysis, protein extraction and purification, and protein digestion. Depending on the experimental design, peptide enrichment or pre-fractionation may also be performed. After peptide separation, the peptides are analyzed by MS before being identified using a bioinformatic platform.*

1.7.2.1. Sample preparation

Cell lysis is a crucial first step for shotgun proteomics. The goal of this step is to disrupt cellular membranes, which is the barrier separating the inner contents of the cells from the exterior. To obtain good quality and unbiased results, it is important to assess the experimental conditions and objectives when choosing a sample preparation method.

Two main categories of lysis methods are used. Physical disruption methods involve breaking the cells using external forces, such as shearing, and include sonication, freeze-thaw and manual grinding. The second category are non-physical methods that employ detergent and salt-based buffers to disrupt cellular membrane by breaking the lipid membrane and solubilizing membrane proteins. Choosing the right detergent is crucial for efficient lysis. Commonly used detergents are sodium dodecyl sulphate (SDS), which is a strong lysis agent that also denatures proteins, and non-ionic detergents, such as Triton X-100, NP-40, and Tweens 20 and 80. The latter allow for milder lysis and are non-denaturing.

In general, sample preparation is critical as it determines the proportion of the proteome that will be available for analysis. An ideal sample preparation method should be reproducible, efficient and robust, and should isolate and clean-up all proteins[206]. When chaotropic agents are used during sample preparation, protein denaturation occurs during which proteins unfold by using chaotropic agents. Also, cysteine disulfide bonds are reduced, e.g., using dithiothreitol (DTT)[207], and cysteine thiols are alkylated to avoid the (re)formation of (new) disulfide bonds. Alkylation is commonly performed using iodoacetamide (IAA), which might however produce several side-products[208]. This step is key in order to study extracellular proteins.

Enzymatic digestion is the process of breaking proteins into peptides using proteases for subsequent LC-MS/MS analysis. Trypsin is the gold standard for this purpose in shotgun proteomics[205] and cleaves C-terminal to arginine and lysine, resulting in peptides with an average size of 700 to 1,500 Da (~6 to 14 amino acids). Adding lysyl endopeptidase (Lys-C) to trypsin (Trypsin/Lys-C) was shown to decrease the number of missed cleavages observed with conventional trypsin digestion. Note that other proteases are sometimes used to identify PTM sites, proteoforms and proteotypic peptides[209]. Among these proteases, chymotrypsin performs C-terminal cleavage at Tyr, Phe, Leu and Trp, Lys-N performs N-terminal cleavage at Lys, Asp-N performs N-terminal cleavage at Asp, Glu-C performs C-terminal cleavage at Glu and Asp, and Arg-C performs C-terminal cleavage at Arg.

Varnavides *et al.*[206] described two main categories of sample preparation and protein digestion methods: in-solution and device-based methods. In-solution methods include classical in-solution digestion, which utilize different buffers, detergents and chaotropic agents[210], as well as protein precipitations in organic solvents, and Sample Preparation by Easy Extraction and Digestion (SPEED)[211], which solubilizes proteins in trifluoroacetic acid (TFA). The device-based methods use specific beads or "reactors" to clean up, trap or concentrate proteins. Examples of these approaches include Filter Aided Sample Preparation (FASP), which uses a "reactor" containing a membrane with a specific molecular weight cut-off (MWCO)[212], and Suspension Trapping (S-Trap), which traps the proteins in a 3D-porous quartz filter[213]. Among the on-bead purification and digestion approaches are single-pot, solid-phase-enhanced sample preparation (SP3)[214] and solvent precipitation SP3 (SP4)[215]. In addition to these methods, in-StageTip (iST) utilizes C18-coated disks inside a pipette tip or column to capture proteins and subsequently clean them up, digest and desalt the resulting peptides[216].

Commercially available kits based on these methodologies and covering all sample preparation and digestion steps are available and include EasyPep (Thermo Scientific), S-Trap (ProtiFi), and iST (PreOmics).

1.7.2.2. Peptide separation

Once the proteins are digested into peptides, HPLC is used for their separation, and this technique has been extensively employed for the last 45 years[217].

HPLC separates mixtures of peptides based on their physicochemical properties. The side chains of amino acids are categorized according to their polarity, distinguishing between nonpolar or hydrophobic, and polar or hydrophilic. Acidic and basic peptides contain ionizable side chains. In an aqueous solution, the net charge and polarity of a peptide will vary with the pH due to these ionizable side chains. Therefore, both hydrophilicity/hydrophobicity and the presence of charged groups are crucial factors in peptide separation.

Four main chromatographic methods have been used for peptide separation: size exclusion chromatography (SEC), ion-exchange chromatography (IEX), hydrophilic interaction chromatography (HILIC) and reversed phase chromatography (RPC)[217]. Usually, SEC, IEX and HILIC are used as pre-fractionation or enrichment techniques prior to RPC.

SEC separates peptides based on their size[218] and SEC resins consist of porous beads or a gel with a defined pore size. Molecules larger than the largest pore will pass through and peptides with partial access to the pores are eluted from the column in decreasing size order.

IEX separates peptides according to differences in their net charge. Both strong cation exchange (SCX)[219] and strong anion exchange (SAX)[220] can be used. As peptides may consist of basic and/or acidic amino acids and functional groups, their net charge will vary with changes in the pH of the solution in which they reside. In IEX, peptides are typically eluted by increasing the ionic strength (salt concentration). Indeed, when the ionic strength increases, salt ions compete with the charged peptides for complementary charges on the chromatographic resin. As a result, less charged peptides start to elute from the column while peptides with higher numbers of charge will be more strongly retained and only elute at higher salt concentrations.

HILIC separates peptides based on the differences in their surface hydrophilicity by utilizing a reversible interaction between peptides and hydrophilic resins that have a strong affinity for hydrophilic molecules[221]. The HILIC mobile phase contains a high percentage of organic solvent and more hydrophilic peptides interact more strongly with the hydrophilic stationary phase, resulting in increased retention. Thus, in HILIC, peptides with higher hydrophobicity elute sooner. Additionally, HILIC is well-suited for separating polar and hydrophilic peptides. HILIC retains polar peptides more effectively than RPC, making it useful for the analysis of PTMs.

RPLC is the most used chromatographic method for LC-MS-based proteomics. In this technique, the separation is based on the hydrophobicity of peptides. The hydrophobic stationary phase is typically an n-alkyl, aromatic hydrocarbon or a hydrophobic polymer matrix. The sample is loaded using an aqueous solution containing an ion-pairing agent, such as TFA to enhance hydrophobic interactions, and a low concentration of organic solvent (e.g., acetonitrile). Peptide elution starts by increasing the concentration of organic solvent, with more hydrophilic peptides eluting first. By using an organic solvent gradient, peptides are gradually eluted[222].

1.7.2.3. Peptide analysis by mass spectrometry

One of the main methods that allows peptides to be charged and transferred into the gas phase is nanoelectrospray (nanoESI)[223]. This ionization method involves the application of high voltage to analytes that exist as ions in solution. Then, this charged liquid is sprayed through a 1-2 μm diameter spraying orifice, creating a fine mist of charged droplets (plume). These droplets are then evaporated, leaving behind (multiple) charged ions that can be analyzed by MS[224].

The main advantages of this method are that operates at atmospheric pressure, there is no limitation on the analyte mass (due to the multiple charging), quasi-molecular ions are formed ($[M+H]^+$ and $[M-H]^-$), and it allows the ionization of analytes dissolved in aqueous or organic solvents and can therefore be coupled online to HPLC systems. Its major drawback is the ion suppression, which is the competition and interference with analyte ionization by other endogenous matrix species (salt and polymers), which results in decreased ionization of the actual sample. Thus, sample preparation, clean-up and desalting are crucial to avoid this ion suppression.

MS peptide analysis has driven the development of new mass spectrometers in the past few years. Various hybrid instruments have emerged, featuring distinct mass analyzers, ion optics and fragmentation sources. Among this new generation of mass spectrometers, the QExactive platform from Thermo Fisher Scientific has been extensively used for proteomics. This mass spectrometer combines a quadrupole mass filter with an Orbitrap mass analyzer[225]. In the quadrupole, the entire ion package is either transmitted to the Orbitrap (MS1 mode) or only certain mass windows around a precursor ion are filtered and transmitted (MS/MS mode). The Orbitrap mass analyzer was developed by Alexander Makarov[226] and consists of a compact electrostatic apparatus where ion packets are introduced with significant energy, causing them to revolve around a spindle-shaped electrode at its core. The detector captures the current generated by the axial movement of the ions, which is then subjected to Fourier transformation (FT) producing high-resolution mass spectra[227]. Additionally, an intermediate device called a C-trap is used to inject ions into the Orbitrap.

In this configuration, a higher-energy C-trap dissociation or higher-energy collision-induced dissociation (HCD) cell is used for ion fragmentation[228]. This occurs by subjecting the ions to collisions with an inert gas, typically helium or nitrogen, at high energies[229]. This type of dissociation induces the fragmentation of the peptide bond, resulting in the formation of b and y fragment ions, leaving a positive charge on either the N-terminal or C-terminal peptide fragment, respectively. To label b ions, the index points are determined from the N-terminal side. The b ion including the first peptide residue will be termed as b1. The b2 ion will represent the second and first amino acids. The remaining b ions are numbered in progression towards the C-terminus. The same numbering occurs for the y ions, now towards the N-terminus.

Hybrid mass spectrometers can perform data-dependent acquisition (DDA) analysis. In this method, the mass spectrometer scans parent ions and selects the top-N most abundant ions for fragmentation. However, low abundant peptides are often omitted, resulting in a bias towards a low dynamic range of detection[230]. An alternative approach is data-independent acquisition (DIA), in which all precursor ions in an MS1 window are fragmented[231]. DIA thus collects fragment ions from all precursor ions within such predefined isolation windows, avoiding the exclusion of low abundance peptides based on their precursor ion intensity.

1.7.2.4. Bioinformatic analysis for protein identification

The manual interpretation of MS/MS spectra has become impossible due to the enormous amount of data generated nowadays. Therefore, automatic proteomic identification pipelines and search engines have been developed since 1994[232]. Among the most used algorithms one finds SEQUEST[233], Mascot[234], X!Tandem[235] , Andromeda[236], MS Amanda[237], MS-GF+[238] and IONBOT[239]. The core concept of a search engine is to match an experimental acquired MS/MS spectrum to an *in silico* predicted one based on peptide sequences derived from protein sequences stored in databases.

The algorithm used in this work is SEQUEST[240]. First, SEQUEST simplifies the experimental spectrum by summing multiple scans of fragment ions to one, rounding each peak to the nearest integer and removing noise peaks. Then, the algorithm makes a first selection of a candidate peptide sequences based on the precursor ion mass. Each sequence in the set of plausible candidates is compared to the observed spectrum. The peptide sequence is converted into a list of m/z values corresponding to predicted fragment ions. Then, SEQUEST searches the experimental spectrum for ions corresponding to these predicted m/z values, summing the intensities of matched peaks. The algorithm assesses the continuity of each sequence and calculates the percentage of expected fragment ions found in the spectrum. These three factors are combined to generate a score that provides a rapid pre-evaluation of each sequence against the experimental spectrum. From this pre-evaluation, SEQUEST constructs a predicted MS/MS spectrum from each of the 500 best scored sequences. This is done by calculating the m/z values of each possible ion predicted from the fragmentation of each peptide bond. The theoretical ion abundances are adjusted for each fragment ion.

Afterwards, the normalized virtual and experimental MS/MS spectra are compared to produce a correlation score (XCorr). The virtual spectrum with the highest score points to the best match to the experimental spectrum, which will be termed a peptide-to-spectrum match (PSM).

Open search algorithms, such as IONBOT, have been developed to improve peptide identification. Open search algorithms allow the consideration of hundreds of PTMs, chemical artifacts, amino acid substitutions, N-terminal and mis-cleavages. This algorithm uses deeper exploits data acquired in an LC-MS/MS experiment. By matching information from peptide retention time prediction, precursor m/z and fragment ion intensity prediction, confidence in peptide identification can be improved. Additionally, a machine learning algorithm used for PSM rescoring inside Ionbot produces a “candidate match” that is rescored to produce both reproducible and tailored to the experimental data, leading to better performance than other search engines[239]. Another rescoring algorithm is INFERYS Rescoring for Sequest HT. It predicts fragment ion intensities to calculate additional scores. Due to its deep learning algorithm it can help in the construction of deep learning-based spectral libraries and rescue or discard incorrectly annotated spectra[241].

Subsequent filtering methods and software tools have been created to statistically assess the confidence of PSMs. One of these tools is Percolator[242], which recalibrates PSMs based on the learned decision boundary between targets and decoys using support vector machines (SVMs). Percolator trains a machine learning model using a subset of high-quality PSMs as positive examples and a larger set of decoys or negative PSMs as negative examples. This database of decoys is created by reversing the sequences from the normal database. It uses these examples to learn the discriminant features that separate correct from incorrect PSMs. After this training step, the model will re-score all PSMs in a dataset. A q-value will be assigned to each PSM as an estimation of the false discovery rate (FDR) at a given score threshold. This q-value can be used as a criterion to control FDR and select a desired level of confidence[243]. Then, an algorithm groups the redundant PSMs into peptide groups. Statistical methods are applied to provide greater significance analysis of peptide matches, based on the FDR targets selected. The peptides are filtered and matched to protein sequences in a database. Generally, two PSMs are required for identification of a protein even though evidence has shown that a single unique peptide identification can be a useful to identify the presence of a protein in the sample[244,245]. Finally, a last algorithm is used to calculate protein scores, validate the FDR scores, and group the identifications.

Ongoing research and advancements in deep learning continue to push the boundaries of peptide and protein identification by mass spectrometry and pave the way for more accurate, efficient and confident analyses in the field of proteomics[246].

1.7.3. Alternative proteins database development

There are different types of databases used in protein research, including sequence databases, structure databases and genomic databases. Sequence databases, such as GenBank[247], RefSeq[248] and Ensembl[249], store nucleotide and protein sequences. Protein databases such as UniProt[250], Protein Data Bank[251] and Protein Atlas[252] store sequences, 3D structures, localization and functionality of proteins. These databases have revolutionized the way data are analyzed. For instance, they allow to share data among researchers, which promotes collaboration and reduces duplication of efforts. Databases have also made it possible to identify relationships between different sets of data, leading to new discoveries and insights.

In MS based proteomics the gold-standard database is the UniProt Knowledgebase (UniProtKB)[250], which is a joint effort from the European Bioinformatics Institute (EMBL-EBI), SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR) in the United States. Swiss-Prot, created in 1986, is a high-quality, manually annotated protein sequence database[253]. TrEMBL is a computer-annotated protein sequence database that contains all translations of EMBL nucleotide sequence entries not yet integrated into Swiss-Prot[254]. PIR contains protein sequences and functional information[255]. The main characteristic of this database is that it follows the single reading frame of mRNA and the Kozak dogma. Therefore, only RefProts and their isoforms are found in this database.

Recent efforts were made to develop databases that allow researchers to study AltProts (**Figure 8**). The first database that considered AltProts was HAltORF[53] (2012), which only contained frame-shift AltProts. The need for a repository containing short ORFs led to the publication of sORFs.org in 2016[193], which contains only short ORFs (identified using Ribo-seq). In 2018, smProt was published[256], and this database exclusively contains small AltProts of less than 100 amino acids. In the same year, OpenProt was introduced[257]. OpenProt contains predicted sequences on all transcripts reported by both Ensembl and NCBI RefSeq. Over 400,000 new protein sequences were predicted, outnumbering their canonical counterparts (UP000005640; 82,427 sequences), and thus making up the hidden part of the proteome iceberg.

As new AltProts continue to be discovered and the amount of available data grows, it has become necessary to update and expand the AltProt databases. Since their initial release, these databases have undergone updates. sORFs.org, smProt, and OpenProt have been updated in 2018[258], and 2021[105,259] respectively. In the following sections, I will describe the AltProt-containing databases in detail.

1.7.3.1. HAltORF

In 2012, Prof. Xavier Roucou led the first efforts to construct an AltProt database. His team's work resulted in the creation of the Human alternative Open Reading Frame (HAltORF)[53], which was the first web-based searchable AltProt database. HAltORF was developed based on the idea of multiple frame translation at the CDS region of a mRNA. To generate the database, the full mRNA annotations of GenBank were used and associated to their RefProts. After this association, *in silico* translation was performed and the resulting protein sequences were mapped to the matching RefProts to identify the initiation and stop codons. This step established the frameshift based on the RefORF of the mRNA. To annotate AltProt sequences, a 24-amino acid cut-off was used. The arbitrary threshold of 24 amino acids was selected to reduce the database size. Finally, the database was filtered by keeping the sequences which corresponding AltORF possess a strong Kozak context surrounding their AUG codon and their stop codon should be before the RefORF stop codon. This rigorous approach resulted in the generation of around 17,000 AltProts, with 83% of them originating from the +2 ORF. It is important to note that this database does not contain any AltProt from 5’-UTRs, 3’-UTRs and lncRNAs.

The development of this database was a significant step towards advancing AltProt research. Indeed, researchers had now access to a comprehensive database that held AltProt sequences.

1.7.3.2. sORF.org

A second databases that contains a comprehensive annotation of AltProts is sORF.org[258]. In this repository, the smORFs identified by Ribo-seq experiments can be retrieved. In the first version human, mouse and fruit fly were integrated, resulting in 263,354 sequences. This sequences comprises 5′-UTR, exonic, intronic, 3′-UTR, ncRNA, or intergenic AltProts, The Ribo-seq reads were treated following the PROTEOFORMER pipeline[98]. First the reads were aligned to the iGenomes repository. Then the TIS were determined. The sORFs were assembled from the identified TIS to the following stop codon. Additionally, the general characteristics of each AltProt were calculated alongside their coding potential, sequence variation functionality and homology. In 2018 an update of this database included three more species (zebrafish, rat, and *Caenorhabditis elegans*) and supplementary Ribo-seq reads. More stringent noise filters, inner BLAST and PRIDE-ReSpin MS data reprocessing pipeline were added to the tool.

1.7.3.3. SmProt

Another database that contains information on AltProt sequences is SmProt[256]. In this database, only AltProts with less than 100 amino acids are stored and called small proteins. The creators of this database curated different datasets to construct it, including low- and high-throughput literature mining, database queries (such as UniProt and CCDS[260]), MS data and Ribo-seq data. In fact, they followed a comprehensive workflow. First, for literature-derived data, a manual review was conducted, filtering and classification to identify research articles containing strong experimental evidence of AltProts. From a total of 5,475 articles, the sequences, start codons, characteristics, functions and probable pathological associations of the AltProts were retrieved. Publicly available Ribo-seq and their paired RNA-seq datasets were trimmed and aligned to the genome, following which the *in silico* predicted small protein sequences were obtained using RiboTaper[261]. To retrieve ncRNA-derived small proteins from MS data sets, the authors first matched raw MS data to their reference genome localization using Peppy[262]. They then filtered the ncRNA-derived proteins by matching the MS data genomic localization to the non-coding sequences in the NONCODE database[263].

The first version of SmProt[256] contained 255,010 AltProts indexed for eight different species, each with their own protein ID, sequence, genomic location, type of AltORF, gene symbol, transcripts, predicted function and evidence. For humans, a total of 167,785 sequences were annotated. In 2021, an update was released, incorporating new Ribo-seq and MS raw data, which also allowed for the inclusion of genetic variants and disease specific AltProts, as well as AltProts obtained from translation starting at non-AUG codons. The updated database now contains a total of 638,958 annotated AltProts, representing a 2.5-fold increase from the original release[7] (for humans, 327,995 small AltProts were annotated). These new additions greatly enhanced the functionality of the database, making it an even more valuable resource for researchers in the AltProt field.

1.7.3.4. OpenProt

After the release of HAltORF, the Roucou team expanded their database for AltProt research by ensuring that it was as extensive and robust as possible. This led to the creation of OpenProt in 2018[257]. OpenProt follows the full polycistronic dogma of mature mRNA, which means it contains AltProts encoded from 5'- and 3'-UTRs, frame shifts and lncRNAs without the 100-codon cut-off. OpenProt was also designed to include novel isoforms produced from an unannotated ORF with significant homology to a RefProt from the same gene.

To build this database, the authors retrieved and merged the transcriptome from NCBI RefSeq and Ensembl, performed translation in all reading frames, starting with an AUG codon and a minimum of 30 codons. These translated sequences were then matched against NCBI RefSeq, Ensembl and Uniprot and matched entries were termed RefProts. To annotate novel isoforms, the corresponding sequences had to meet two conditions. The first was to have over 80% of protein sequence identity over 50% of the length as revealed by Basic Local Alignment Search Tool (BLAST)[264]. The second condition was to have the same genomic coordinates of the start or end codon with a protein sequence identity over 20% of the length. Finally, all the ORFs that were not annotated as RefProts or novel isoforms were classified as AltProts.

In addition to predicting AltProts, translational evidence of these predicted AltProts was included. A total of 114 publicly available MS-based proteomic and Ribo-seq datasets were re-interrogated using OpenProt. For each AltProt, RefProt and novel isoform, the identification results were added. Also, data on the conservation of AltProts between the ten different species and protein functional domains were added to the database. As a result of this extensive workflow, a total of 2,019,609 sequences were predicted for ten different species. For humans, 461,462 AltProts were predicted in the initial release of OpenProt and, in general, in this database, multiple pieces of information can be queried for each AltProt, including predicted protein characteristics, transcript and gene sources, localization within the transcript, genomic and transcript coordinates, initiation motifs, protein and DNA sequences, and protein evidence.

As genomic annotation evolves, the latest release of OpenProt was published in 2021[259]. In this release, the database was updated according to the 2019 annotation of NCBI RefSeq and Ensembl, and 627 AltProt sequences were added. In addition, the translation evidence was updated by the re-analysis of 125 Ribo-seq and 171 MS-based proteomic datasets with this new version of OpenProt. Overall, OpenProt has become an extensive and valuable tool for AltProt researchers as it contains an enormous amount of data which is useful in the discovery and characterization of AltProts.

1.7.3.5. OpenCustomDB

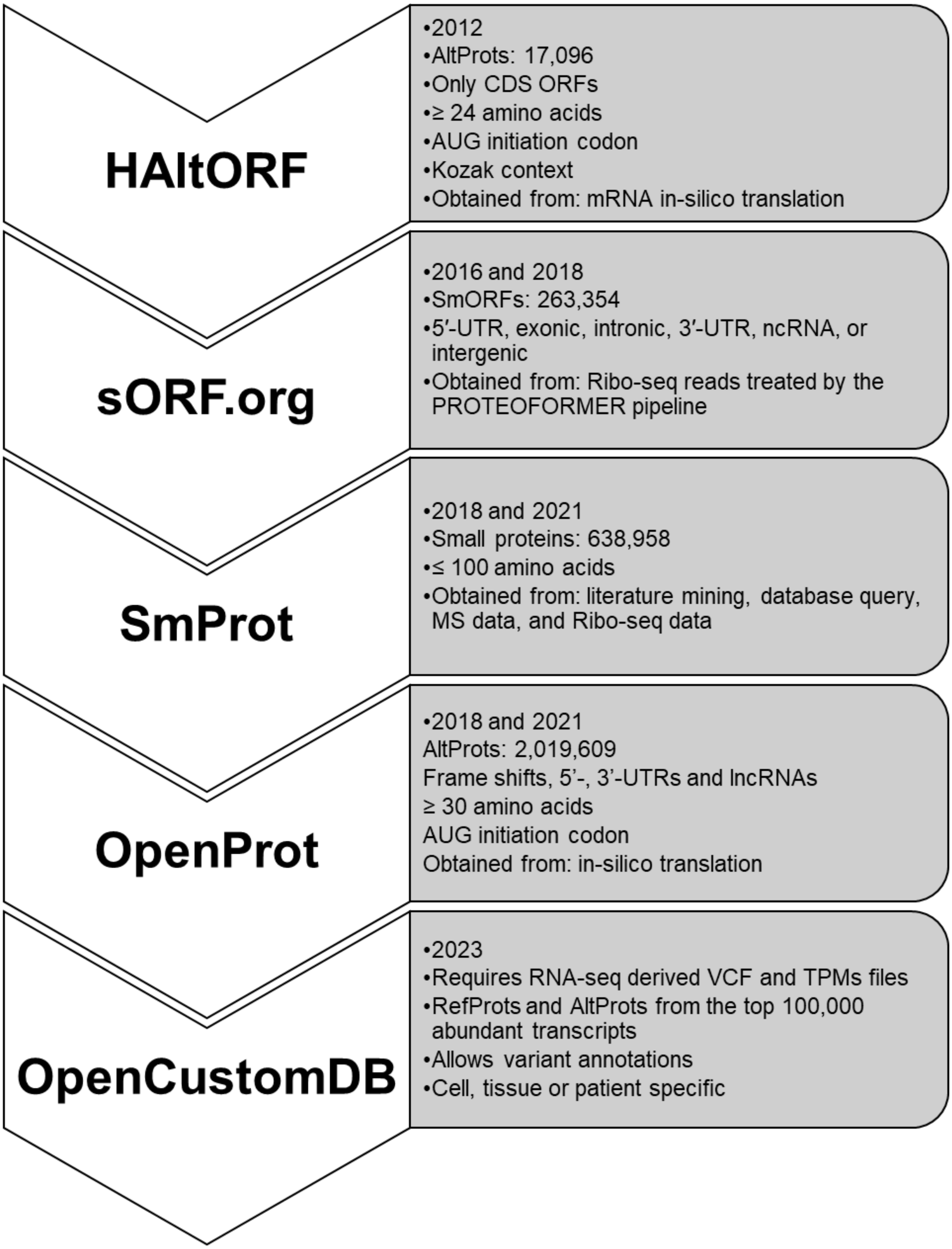
An elegant and growing approach to study AltProts is by using proteogenomic workflows. Proteogenomics allows the construction of sample-specific genomic or transcriptomics-derived databases. These databases can be expanded to include single or multiple nucleotide variations, frameshifts, novel isoforms, gene fusions, and novel proteins and their variants. This enables a more thorough analysis of the genomic makeup of a sample.

One tool that can be used to generate RNA-seq based custom databases using OpenProt annotations is OpenCustomDB[265],which is specifically designed to create databases that contain the RefProt, novel isoforms, AltProt annotations and their variants coded by the transcripts of the sample of interest. By utilizing OpenCustomDB, the identification of AltProts and their variants can be performed with greater precision and accuracy.

The process of generating RNA-seq derived databases consists of several steps. First, the reads resulting from RNA-seq experiments are aligned to the reference genome. Once this is complete, transcript expression is quantified and normalized in TPM. After these steps, variant calling files (VCF) are generated from the binary alignment map (BAM) files.

These VCF and transcript expression files are then imported into OpenCustomDB. Here, the VCF files are used by OpenVar[185], which is designed to annotate the variants of RefProts, novel isoforms and AltProts. Then, the transcripts are ranked from highest to lowest expression and the top 100,000 transcripts are added to the database. If a variant is found, the wild-type (WT) protein sequences are also added to the database, ensuring access to both the original and the variant sequences.

OpenCustomDB thus enables the study of AltProts and their variants as part of precision medicine studies on a routine basis, allowing to gain a deeper understanding of the underlying mechanisms of various diseases and develop targeted therapies that are tailored to individual patients.



•2012
•AltProts: 17,096
•Only CDS ORFs
•≥ 24 amino acids
•AUG initiation codon
•Kozak context
•Obtained from: mRNA in-silico translation

•2016 and 2018
•SmORFs: 263,354
•5′-UTR, exonic, intronic, 3′-UTR, ncRNA, or intergenic
•Obtained from: Ribo-seq reads treated by the PROTEOFORMER pipeline

•2018 and 2021
•Small proteins: 638,958
•≤ 100 amino acids
•Obtained from: literature mining, database query, MS data, and Ribo-seq data

•2018 and 2021
AltProts: 2,019,609
Frame shifts, 5’-, 3’-UTRs and lncRNAs
≥ 30 amino acids
AUG initiation codon
Obtained from: in-silico translation

•2023
•Requires RNA-seq derived VCF and TPMs files
•RefProts and AltProts from the top 100,000 abundant transcripts
•Allows variant annotations
•Cell, tissue or patient specific

Figure 8. Evolution of AltProt databases. *There are four main databases that contain AltProts: HAltORF, SmProt, OpenProt, and OpenCustomDB. The characteristics of the latest version of each database are shown.*

1.7.4. High-throughput identification of AltProts by shotgun proteomics

The use of shotgun proteomics for identifying AltProts has been growing over the years, which can be attributed to the development of more robust and powerful algorithms that allow the use of larger sequence databases. In addition, different techniques have been used to enrich and separate small proteins like several AltProts. These techniques include the use of physicochemical fractionations, polyacrylamide gel electrophoresis and two dimensional HPLC. Combining these techniques has revolutionized the field of AltProt proteomics. Several examples of this have been described. For instance, ten years ago, Slavoff *et al*. identified 86 AltProts in human K562 cells using a 10 KDa MWCO filter to enrich for small proteins[104]. In addition, they used 2D LC-MS/MS with a pre-fractionation step using electrostatic repulsion-hydrophilic interaction chromatography (ERLIC). Each resulting fraction was then separated by RPLC coupled to a LTQ Orbitrap Velos and a DDA top 20 method was employed. In another study, 24 AltProts were identified among ovarian, fallopian tube and endometrial formalin-fixed paraffin-embedded tissues, by using in-tissue digestion and peptide extraction followed by RPLC coupled to a Thermo Scientific Orbitrap Elite mass spectrometer[100]. A similar approach combined cell and tissue protein extraction, Tricine PAGE, 30 KDa MWCO, ERLIC prefractionation and RPLC separation, and 237 AltProts in K562, MCF10A and MDAMB231 cell lines were identified, as well as in human breast tumor samples[266].

A related workflow was applied to identify AltProts in hepatic cancer cells (HEP3B). Two different extraction methods were used: HCl extraction and acetonitrile (ACN) precipitation. Then, Tricine in-gel digestion was performed, followed by ERLIC fractionation and RPLC-MS. In this comparison of methods, HCl extraction outperformed ACN precipitation. The overlap of identifications by the two methods was low, demonstrating that they were complementary, and a total of 365 AltProts were identified[267]. Another example of how combining different techniques boosts identification of AltProts was described by Cassidy *et al*. Two approaches were used, with the first consisting of a peptide separation by high/low pH RPLC-MS. The second approach involved gel eluted liquid fraction entrapment electrophoresis (GELFrEE) of protein extracts, in-solution digestion and RPLC-MS. With this methodology, 28 AltProts were identified[268]. Yet another approach combining Tricine PAGE, in-gel digestion and RPLC-

MS (LTQ Orbitrap QExactive) quantified 28 AltProts in two human leukemia cell lines (K562 and MOLT4). Among them, 12 were found differentially expressed between these cell lines[269].

To enrich for AltProts, depletion of high molecular weight proteins was used by Cassidy *et al*.[270]. The authors employed a differential solubility (DS) method. DS is based on precipitation under denaturing conditions followed by re-solubilization. First the sample was diluted in denaturing buffer. Then, the sample was slowly dropped into ice-cold acetone, stirred and centrifugated. The precipitate was resuspended in ACN/HCl and re-centrifugated. With this technique, the low molecular weight proteins have a higher tendency to remain in the supernatant. By this approach 70% of high molecular weight proteins were depleted and 11 AltProts were identified.

In 2017, Ma *et al*. compared four different extraction methods: (1) 50 mM HCl, 0.1% β-mercaptoethanol (β-ME); 0.05% Triton X-100 at room temperature (lysis buffer); (2) 1 N acetic acid/0.1 N HCl at room temperature; (3) boiling water; and (4) boiling lysis buffer. Additionally, they compared three different enrichment methods for AltProts: (1) acid precipitation, (2) 30 kDa MWCO filter, and (3) reverse-phase (C8) cartridge enrichment[271]. The enriched samples were digested in-solution, separated by C18 RPLC and analyzed by a QExactive Plus and an Orbitrap Fusion Tribrid mass spectrometer. Their results indicate that acid precipitation and C8 cartridge enrichment outperformed the 30 kDa MWCO filter. In addition, the researchers concluded that the best combination was the extraction with lysis buffer and C8 column enrichment. By this approach (lysis buffer and C8 column enrichment), 169 AltProts were identified.

Another study comparing different extraction and enrichment methods was published in 2020 by Cardon *et al*. In this study, the authors compared four protein extraction methods in human NCH82 stage IV glioma cells: (1) 4% SDS buffer, (2) RIPA lysis buffer, (3) methanol acid buffer, and (4) boiling water. Then, three enrichment methods were compared: (1) gel fractionation, (2) acetic acid precipitation and (3) trichloroacetic acid precipitation. For the gel enrichment, in-gel digestion was performed, and liquid digestion for the two other methods. The tryptic peptides were separated by RPLC and analyzed by a Thermo Scientific QExactive. In this comparative study, the best method was found to be the combination of boiling water or RIPA buffer followed by acetic acid precipitation. By the RIPA and acetic acid approach, 21 AltProts were identified[245].

Overall, shotgun proteomics is becoming increasingly popular for identifying AltProts. As demonstrated by the examples provided, the combination of various techniques has resulted in the successful identification of a wide range of AltProts in different biological contexts. These findings have the potential to shed light on the discovery phase of uncharacterized proteins and lead to new insights in various fields of biology and medicine.

1.7.5. N-terminomics

N-terminal proteomics is an approach used to identify the uttermost N-terminal peptide of a protein, mainly following depletion of non-N-terminal peptides. Moreover, most proteins can be identified only by their N-terminal peptide as such peptides are very often proteotypic[272]. Additionally, N-terminomics allows to identify the TIS used and the origin of (N-terminal) proteoforms. The presence of an initiator methionine (iMet) and the acetylation status of the N-terminus are used to confirm if a proteoform or AltProts originates from translation or protein processing pathways.

An approach presented by Bogaert *et al.*[273] employed an Ribo-seq derived protein database to analyze MS/MS spectra and thus to identify potential translation products from noncoding regions. Here, N-terminal peptides were enriched by COmbined FRActional DIagonal Chromatography (COFRADIC)[274]. After LC-MS/MS analysis at an Orbitrap Fusion Lumos mass spectrometer, stringent filtering of the identified peptides was applied to find evidence for novel translation events. Evidence of only 19 peptides from noncoding regions was recovered. Finally, the functional analysis of a novel AltProt was demonstrated through Virotrap-based interactome analysis of two of its N-terminal proteoforms.

1.8. Interactomics

As reviewed in the previous chapter, MS-based proteomics is used for the discovery of AltProts. However, their cellular roles and their involvement in pathological conditions remain poorly understood and unexplored. It is important to highlight that for these uncharacterized proteins, specific antibodies are infrequently used due to their development costs and possible lack of specificity for small-sized proteins.

An interesting approach to functionally characterize AltProts is to connect and map them in signaling pathways and, for instance, to assign GO terms to AltProts. The use of such terms to predict protein function is particularly popular in statistical approaches based on the observation of PPIs[275]. By identifying the network to which a protein-of-interest belongs, one gains valuable information about the signaling pathways it participates in, the so-called “guilt-by-association” principle. This, in turn, enables us to put forward hypotheses about the possible functions of a protein in the identified pathways. It is important to note that proteins in a cell interact with each other, and the binding of a ligand to a receptor can initiate a chain reaction. This reaction often involves the phosphorylation of a receptor partner and the subsequent modification of the players in the signaling pathway. Within a signaling pathway, there are a variety of components, including activators, inhibitors, contact partners, enzymes, and more. The sequence of protein modifications ultimately leads to changes in the cell's phenotype, and it is the modification of these signaling pathways that is often the root cause of many pathologies. By identifying the partners of a signaling pathway, one not only gains insight into the function of these partners, but also additional clues about the origin of a pathology. This makes the identification of interaction partners essential for studying the role of proteins in cellular processes.

Numerous techniques have been developed to detect PPIs, each with its own advantages and limitations. Some of the most commonly used techniques include yeast two-hybrid (Y2H)[276] and fluorescent molecule energy transfer (FRET)[277], which require the construction of fused proteins from the target proteins of interest. Such techniques are useful for detecting binary interactions between two proteins, but ill-suited for detecting protein complexes involving more than two proteins. Other methods, such as proximity ligation assay (PLA)[278] and affinity purification coupled to MS (AP-MS)[279], are based on the use of antibodies and can detect a range of protein interactions, including those involving protein complexes. PLA is particularly useful for detecting interactions between proteins within a specific subcellular compartment, while AP-MS is the most widely used method for detecting PPIs in complex biological systems. Recently, advances in MS technology, transfection and expression systems have led to the development of new proteomics-based methods for PPI detection, such as ascorbic acid peroxidase proximity-labeling MS (APEX-MS)[280], proximity-dependent biotin identification (BioID)[263], Virotrap[282], and crosslinking-MS (XL-MS)[283]. These methods offer advantages over traditional techniques, such as the ability to detect in higher throughput PPIs.

Finally, the meta-data analysis in such experiments is also an important issue, and powerful network analysis software is now available (e.g., Cytoscape[284], ClueGO[285], STRING[286,287], IntAct[288], and BioGRID[289]). These can be used to locate signaling pathways, GO-terms or even the types of links referenced for the proteins observed.

1.8.1. Yeast two-hybrid assays (Y2H)

Y2H is a classic and one of the earliest PPI research techniques. It was introduced in 1989 by Fields and Song for the study of paired protein interactions[290]. It uses a DNA-binding domain (DBD) to bind a specific DNA sequence but cannot activate gene transcription. To activate transcription, a transcription-activating domain (TAD) is needed. In Y2H, a bait (protein-of-interest) is fused with a DBD and a possible prey (interaction partner) is fused with a TAD. Yeast cells are transformed with these fusion constructs, resulting in the production of bait-DBD and prey-TAD fusion proteins. If the bait and prey proteins interact within the yeast nucleus, the DBD and TAD come close together and the TAD interacts with the transcription machinery to initiate transcription of a reporter gene. The reporter gene produces a signal, indicating bait-prey interaction[291]. Y2H can be used for genome-wide screens by constructing a library of preys using a cDNA library.

A common issue with Y2H is the high number of false positive interactions caused by non-specific PPIs. To address this, Y2H variants were developed that use proteins with two separate structural domains. These proteins can reassemble to form a functional reporter system when brought together through bait-prey interaction.

Y2H can be employed as a targeted approach to identify AltProt PPIs as proposed by Inchingolo *et al*. in 2021[292] who described the characterization of SEP^{53BP1}. This AltProt is coded from an overlapping ORF in the CDS region of the gene. The authors used ULTImate Y2H^{TM} (performed by Hybrigenics Services) to determine the interactome of

SEP^{53BP1} (bait). Five high confidence interactions (PSMA7, UBQLN4, TRIP12, MAPRE1 and BCOR) were identified from a 51 million peptide library generated from a human B cell Lymphoma_RP1 (Hybrigenics Services). By STRING analysis and its co-IP validation, they proposed the involvement with the α4 subunit of the 20S proteasome barrel and it plays a key role in its assembly[292].

1.8.2. Proximity Ligation Assays (PLA)

Another technique to identify binary PPIs is the proximity ligation assay (PLA). This antibody-based technique visualizes interacting proteins and their localization in cells or tissue. The antibodies used in PLA contain an oligonucleotide tag that allows for ligation, replication, and reporting signaling[278,293].

The general workflow to identify the interaction of two proteins of interest consists of five main stages. First, protein-specific primary antibodies are selected, raised in different hosts. Then, the samples (cells or tissue) are incubated with the primary antibodies, allowing them to bind to their respective target proteins. In the second stage, two PLA probes (secondary antibodies conjugated with oligonucleotides) are added and incubated. After incubation, if the two proteins of interest are in close proximity, the DNA strands attached to the secondary antibodies will form a bridge. A ligase is applied and the two probes hybridize into a circular DNA molecule. This DNA molecule serves as a template for rolling circle amplification (RCA). One of the PLA probes acts as a primer for the polymerase, generating many copies of the DNA sequence, which is still joined to the secondary antibody. Fluorescently labeled complementary oligonucleotides are used to yield a signal that can be detected by fluorescence or confocal laser scanning microscopy. Visualization only occurs if the two proteins are within a 40 nm range.

Sandmann *et al*. used PLA to validate the interaction of the AltProt PVT1-MP with the RefProt SRSF2[294]. The PLA reaction was performed using the Duolink In Situ Proximity Ligation Assay Starter Kit (Red, Mouse/Rabbit), between V5-tagged PVT1-MP and FLAG-tagged SRSF2 in HeLa cells[294].

1.8.3. Co-immunoprecipitation couple to MS (coIP-MS)

One of the most used techniques for identifying PPIs is coIP-MS. This technique involves capturing a protein of interest and its interactors employing antibodies specific for the chosen baits. These antibodies are then immobilized to functionalized magnetic or agarose beads.

First, a key step is to perform a mild cell lysis and protein solubilization need to be performed to maintain the native PPIs. Note that non-denaturing, low ionic strength and non-ionic detergents (such as NP-40 and Triton X-100) are less likely to disrupt PPIs[295].After incubation with the antibody and immobilization of the antibody-protein complexes, extensive washing is performed to remove nonspecifically bound proteins and contaminants. This step reduces the background noise and increases the specificity of the co-IP. Finally, the complexes are digested, and LC-MS/MS is applied to identify the proteins involved in the bait-containing protein complexes.

Antibodies have high affinity for their bait protein, allowing for the purification of endogenous proteins under native conditions. Purification of endogenous proteins in this way may come with some drawbacks. Antibodies can disrupt PPIs by disturbing the PPI-interface and a proper control condition is necessary to avoid too many false positives from cross-reactivity or contaminants. Additionally, the currently available antibodies cover only a limited number of proteins from the entire genome, and there are no antibodies targeting predicted AltProts. Moreover, due to the size of an antibody or bulky epitope tags compared to an AltProt, PPIs can be blocked. The cost of producing and maintaining these antibodies is also significant. Currently efforts are made by The Human Protein Atlas to map different tissues and cells using antibodies[296]. Therefore, the characterization of AltProts needs to be boosted so these “unknown” or misannotated proteins can be recognized and analyzed in a wide-spread manner.

1.8.4. Affinity purification coupled to MS (AP-MS)

This technique involves capturing a protein of interest using a ligand (such as oligonucleotides, chemicals, lipids, peptides or proteins) coupled to an immobilized solid matrix (such as agarose or magnetic beads)[279]. Epitope tagging involves fusing the DNA sequence of a protein of interest to an ORF to express a peptide or protein tag that can be efficiently purified on a support material.

For these reasons, epitope tagging provides an advantage in that usually the tag is not endogenously found, yet serves as a molecular handle for specific detection and purification. Common tags include c-Myc (EQKLISEEDL), FLAG (DYKDDDDK), HA (YPYDVPDYA), His-Tag (HHHHHH), Strep-tag II (WSHPQFEK), green fluorescent protein (GFP), Red Fluorescent Protein (RFP) and mCherry. Even though some endogenous tags can be used including biotin, Glutathione-S transferase (GST), Calmodulin Binding Protein (CBP), and Maltose-binding protein (MBP). To generate an epitope-tagged protein, the DNA encoding the epitope tag is added to the DNA sequence encoding the protein-of-interest using molecular cloning techniques. Tags are usually inserted at either the N-terminus or C-terminus of the protein and the plasmid containing the tagged sequence is then transfected into appropriate host cells for protein expression. Once expression of the tagged protein is confirmed, co-IP is performed to isolate intact PPIs (protein complexes).

As stated above, no specific antibodies have been produced against novel AltProts. Therefore, epitope tagging of AltProts is a convenient approach to perform co-IP or immunofluorescence (IF) studies. This technique was employed by Ichingolo *et al*. to identify the interactors of SEP^{53BP1}. For this experiment, HEK293T cells were transfected with pcDNA expressing SEP^{53BP1}-3xHA.

1.8.5. Ascorbic acid peroxidase proximity-labeling MS (APEX-MS)

Enzyme-catalyzed proximity-labelling techniques are based on genetically encoded enzymes, which, upon treatment and biotin supply, allow for covalent labeling of proteins in a radius of approximately 20 nm.

For APEX, a protein-of-interest is fused to a peroxidase. After transfecting the genetical construct in a cell and stimulation with hydrogen peroxide (H_2O_2), APEX catalyzes the conversion of exogenously supplied biotin-phenol to highly reactive biotin-phenoxyl radical. These radicals biotinylate nearby proteins on electron-rich amino acids (e.g., tyrosine) in less than 1 ms, thus taking a snapshot of the bait environment. Biotinylated proteins are subsequently enriched on streptavidin beads and digested for analysis by

LC-MS/MS[297]. By comparing protein levels between the APEX-fusion samples and proper control samples, one can identify proteins that directly interact or associate with the bait protein. The second generation of APEX enzymes (APEX2) allowed to identify highly specific interactions in shorter distances (1-10 nm) as well as the characterization of subcellular compartments[298].

APEX-MS has uncovered novel AltProt-RefProt interactions. It was used by Chu *et al*. to characterize the interaction partners of the 123-amino acid microprotein encoded by the C11orf98 smORF[299]. In this study, N-terminal or C-terminal APEX fused C11orf98 was expressed in HEK293T cells. From the C-terminal fusion, 112 interactors were identified. Additionally, 137 interactors were identified by the N-terminal construct.

Of these, 99 interactors overlapped, which provided additional confidence in data reliability. The authors found that the C11orf98 microprotein interacts with NCL and NPM1, and several other nucleolar proteins, and C11orf98 was suggested to participate in the synthesis and maturation of ribosomes.

1.8.6. Proximity dependent biotin identification (BioID)

A second proximity labeling technique is BioID. In this technique, a mutant prokaryotic biotin protein ligase (BirA*) is used to covalent biotinylate the proximal interactors of a bait protein. This enzyme adenylates biotin to generate a reactive intermediate, biotin–5'-AMP, which diffuses from its active site and reacts with lysine side chains on proximal protein within a 10 nm radius of the bait-BirA* fusion protein[300]. Key for this workflow is the generation of the BirA*-fusion protein vector for transient of stable transfection in the desired cells. By generating stable cell lines, a construct can be expressed without repeated transfections. Stable transfection involves the integration of transfected DNA into the host cell genome, allowing transfected cells to pass this exogenous DNA during passages. Some expression systems are designed to induce protein expression upon cell stimulation[301].

Once the bait-BirA* fusion protein is expressed, biotin is added to the cell medium. As in the APEX-studies, covalent biotinylation enables harsh cell lysis conditions to be used to solubilize (hydrophobic) proteins. Two techniques have been developed to increase the temporary resolution of BioID. In TurboID, a biotin ligase was engineered such that the labeling time is decreased to less than one hour[302]. A second technique is split-TurboID, which consists of two inactive TurboID fragments, with one fragment fused to bait A and the second to bait B. If these two baits interact, TurboID reassembled and biotinylates surrounding proteins[302].

A high-throughput TurboID approach was developed to map AltProts and microproteins to their subcellular locations (MicroID). By performing TurboID on six baits (fibrillarin, histone H2B, laminin B1 and a nuclear localization signal) more than 150 AltProts were found associated with subnuclear organelles[303].

Some of the advantages of these techniques and APEX are that it can identify transient PPIs, overcome solubility issues, identify neighboring proteins and stringent washes eliminate nonspecific bindings. On the other hand, the main disadvantage is that fusing the bait protein to a large epitope tag as BirA* or APEX can induce conformational changes and may disturb its cellular neighborhood.

1.8.7. Virotrap

Virotrap is an interactomics technique developed by the Eyckerman lab in 2015. It relies on the formation of virus-like particles (VLPs) containing protein complexes-of-interest that bud from mammalian cells[282].

In Virotrap , a bait protein is fused to the HIV-1 GAG protein, which is responsible for the production of VLPs due to its high mobility and accumulation in cholesterol-rich regions of the membrane. Once sufficient GAG proteins accumulate, GAG multimerization starts, triggering budding and release of VLPs. A key advantage of Virotrap is the trapping and protection of PPIs inside VLPs, which can then be purified from the medium, thus avoiding cell lysis. The purification step is enabled by co-expression of the vesicular stomatitis virus G (VSV-G) protein and a FLAG-tagged version of it. These proteins relocate to the plasma membrane, with trimers of these (FLAG-tagged) proteins found on the surface of the VLPs. The FLAG tag assists VLP enrichment using anti-FLAG antibodies coupled to paramagnetic beads[304]. A critical step of this technique is the lysis, clean-up and digestion of the PPIs engulfed inside the VLPs. Although the combination of SDS-based lysis and HiPPR detergent-removal spin columns performs well, the use of amphipols (APols) and acid-based precipitation was proven to be an elegant and robust approach for this step[304,305]. Further, to analyze the data generated by Virotrap, the straightforward filtering index (SFINX) was proposed to separate true interactors (preys) from the background[306].

Bogaert *et al*. demonstrated the use of this technology in the quest for identifying the interaction partners of AltProts. In their paper, the interactome of two AltProts encoded from a non-coding region was retrieved by Virotrap[307] . The ACTB pseudogene 8 and its N-terminal (Nt) proteoform were fused to the HIV-1 GAG protein and expressed in HEK293T cells. After MS analysis of VLPs, 11 and 10 potential interactors were identified for the full length and Nt-proteoform, respectively. These interaction partners had their main function in vesicle/protein transport and are localized in the membrane. Thus, this technique was used to identify the possible function of these proteins.

1.8.8. Cross-linking mass spectrometry (XL-MS)

Chemical cross-linking coupled to mass spectrometry (XL-MS) is a versatile technique that has been used to elucidate protein conformation and PPIs. XL-MS evolved from elucidating the structure of a purified protein to large-scale PPI studies in cell lysates and tissues[283,308].

The general principle of this technique is simple. By adding a reactive molecule that bridges two amino acids or functional groups, one may obtain structural and interaction information from two proteins. These reagents, called crosslinkers, have a defined length and covalently bind to two amino acid side chains/functional groups. After protein digestion, the crosslinked peptide can be identified by MS. By identifying the position of the crosslinked amino acids and considering the length of the crosslinkers, one can propose structural constraints in the protein's 3D conformation and assess the distance between interactors.

Crosslinking between lysines will only occur if both lysines are at the correct bridging distance. **Figure 9** shows the reaction products that can be obtained during protein crosslinking: (1) intrapeptide crosslinks or Type 1, (2) interpeptide crosslinks or Type 2, and (3) dead-end crosslinks (mono links) or Type 0. Intrapeptide crosslinks are generated when both lysines are in proximity inside the same peptide. Interpeptide crosslinks occur when two lysines are found in two different peptide sequences or located in different proteins. Dead-ends happen when the crosslinker reacts with one lysine but a second one is not present in close proximity hence, the remaining NHS ester will be hydrolyzed. This modification remains attached to the reacted lysine and it will not give any interaction information, but it can give information of the solvent-accessible surface area of the protein.

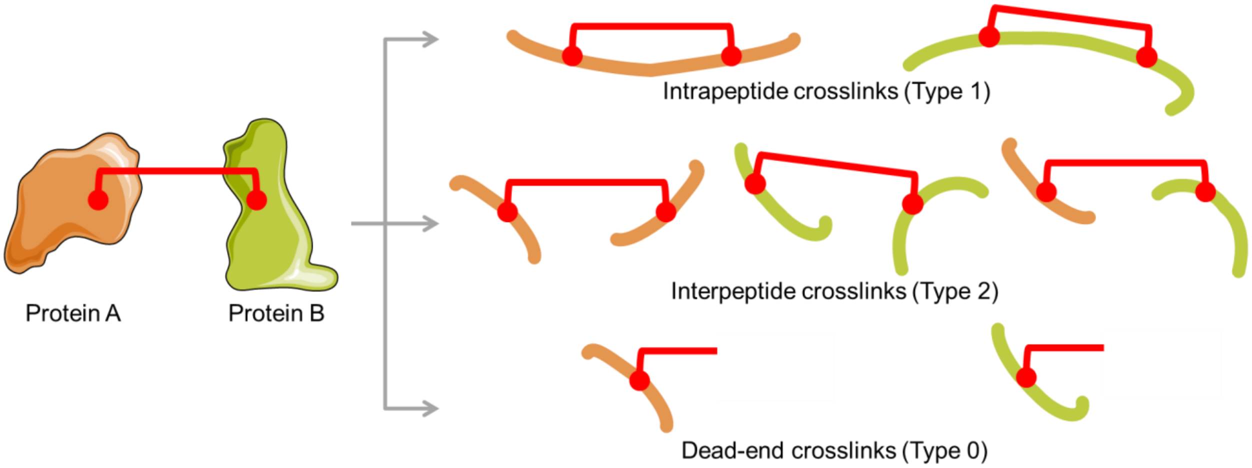

Intrapeptide crosslinks (Type 1)
Protein A
Protein B
Interpeptide crosslinks (Type 2)
Dead-end crosslinks (Type 0)

Figure 9. Schematic representation of crosslinking reaction products. *The top section displays intrapeptide crosslinks between proteins A and B. The middle section shows interpeptide crosslinks between proteins A-A, B-B, or A-B. The bottom section displays mono-links or dead-end products of proteins A and B.*

The general workflow of interactomics using XL-MS consists of several stages. First, crosslinking is performed by adding the crosslinker to the selected system, such as protein complexes, organelles, (lysates of) cells or tissue, followed by incubation and crosslinker quenching. Then, protein extraction and digestion are performed. Usually, crosslinked peptide enrichment is performed by size-exclusion chromatography (SEC) or by utilizing the molecular handlers of some crosslinkers, which is crucial due to the low abundance of crosslinked peptides compared to non-crosslinked ones. Afterwards, the crosslinked peptides are analyzed by LC-MS/MS). Finally, data analysis is performed using (different) algorithms designed for identifying crosslinked peptides and, therefore, crosslinked proteins[309]. One of the main limiting factors of the technique is the low intensity of the crosslinked spectra which increases the challenge in correcting identifying crosslinked peptides. Additionally, the increment of the search space (database expansion) and increased FDR add a higher level of complexity to this technique. Depending on the biological question, computational molecular modeling can be used to refine protein 3D-structures from the crosslink distance constraints. Moreover, combining the crosslinker distance constraint with protein-protein docking can be used to validate protein-protein interactions in a large(r)-scale approach.

The first aspect to consider when designing a XL-MS experiment is which crosslinker to use. A variety of crosslinkers are available that possess different reactivities, lengths, enrichable handlers and cleavability.

Amine-reactive crosslinkers are the most commonly used crosslinkers. They rely on the reactivity of N-hydroxysuccinimide (NHS) esters with nucleophiles such as free amines or hydroxyl groups (**Figure 10A**). The main advantage of targeting lysines is that they are hydrophilic and accessible at the protein surface. However, lysine side chains are also highly flexible hence, the distance constraints are not rigid. Additionally, NHS-esters react with hydroxyl groups such as in serine, threonine, and tyrosine. Moreover, some NHS-esters are not highly soluble in aqueous buffers. This can be addressed by solubilizing them in organic buffers and then diluting the stock solution in the reaction buffer. Another way to address their solubility is to modify the NHS-group with sulfonic acid. Finally, NHS ester are rapidly hydrolyzed in water, and this have a short half-life.

Among the NHS-based crosslinkers, two main non-cleavable homobifunctional crosslinkers are used: disuccinimidylsuberate (DSS)[310] and its sulfonate twin, bis(sulfosuccinimidyl)suberate (BS^3)[311]. Both possess a 11.4 Å spacer arm. DSS has a lipophilic character, allowing it to pass through membranes, which is useful for intracellular crosslinking. In contrast, BS^3 has a charged group and is not membrane-permeable, making it suitable for crosslinking cell-surface proteins. Well-established protocols utilizing these crosslinkers are widely employed in structural biology to study proteins and relatively small protein assemblies[312–316].

In recent years, crosslink experiments have shifted towards the study of PPIs. This shift was made possible by the introduction of MS-cleavable crosslinkers which contain labile bonds as MS-cleavage sites in their spacer chains. These bonds produce characteristic fragment ions upon CID fragmentation. The advantage of such crosslinkers is that, after cleavage of the crosslinker, the linear peptides can be accurately identified, reducing the quadratic search space (n^2) to a linear search space ($2n$)[317]. A homobifunctional, MS-cleavable crosslinker was developed by the Sinz lab; disuccinimidyl dibutyric urea

(DSBU)[318]. This crosslinker has a 12.5 Å spacer length. The particularity of this molecule is that the fragmentation energy of the urea group is similar to the amide bond of the peptide backbone. This allows the detection of the crosslinking indicative ions and the b- and y- ions on the MS/MS level. On the other hand, disuccinimidyl sulfoxide (DSSO)[319] has a 10.1 Å spacer arm and possesses a carbon-sulfur bond adjacent to the sulfoxide which is cleaved by lower energy than amide bonds. For this reason, the use of MS^3 or the combination of CID and ETD is recommended. In this case, following MS/MS, crosslink-specific fragments are detected, while in MS^3, b and y ions are generated. **Figure 10B** illustrates the fragmentation patterns observed for DSSO-crosslinked peptides. When a peptide (α-β) undergoes fragmentation, the carbon-sulfur bond adjacent to the sulfoxide is cleaved, resulting in two peptide fragments, α_A and β_S. The α peptide fragment remains with an alkene (A) moiety (+54 Da), while the β peptide fragment is modified with a sulfenic acid (S) moiety (+104 Da). If the α and β peptides have different sequences, two possible pairs of fragments (α_A/β_S and α_S/β_A) will be observed, resulting in four individual peaks in the MS/MS spectrum. These fragments are then further analyzed using MS^3 to identify the peptide sequences. DSSO-modified peptides that are dead-end modified have a defined mass modification (+176 Da) due to the half-hydrolyzed DSSO (**Figure 10B**). Dead-end modified peptides (α_{DN}) will yield two possible fragment ions: α_A and α_S. The difference in mass between these fragments, correlates to the difference in remnants of DSSO attached to the fragments. For intralinked peptides (α_{intra}), a defined mass modification (+158 Da) is present due to DSSO crosslinking of two distinct lysines in the same peptide sequence (as shown in **Figure 10B**). Cleavage of the carbon-sulfur bond will result in only one fragment peak with the same mass as the parent ion observed in MS. In both cases, performing MS^3 analysis will lead to the detection of b and y ions.

Enrichable (trifunctional) crosslinkers were developed to contain an affinity handle that allows enrichment of crosslinked peptides. Thereby a more sensitive analysis can be achieved due to the removal of high abundance signals of non-crosslinked peptides. One enrichable crosslinker is cyanurbiotindipropionylsuccinimide (CBDPS)[320]. This crosslinker possesses a spacer arm of 14 Å, biotin, allows for isotopic coding and is CID-cleavable. The biotin tag allows the enrichment of crosslinked peptides on streptavidin beads, the drawback being that biotinylated endogenous peptides can be co-enriched with the crosslinking peptides. Additionally, the release of the peptides from the streptavidin beads can be difficult due to the very high affinity between biotin-streptavidin (Kd ~10^{-14} mol/L).

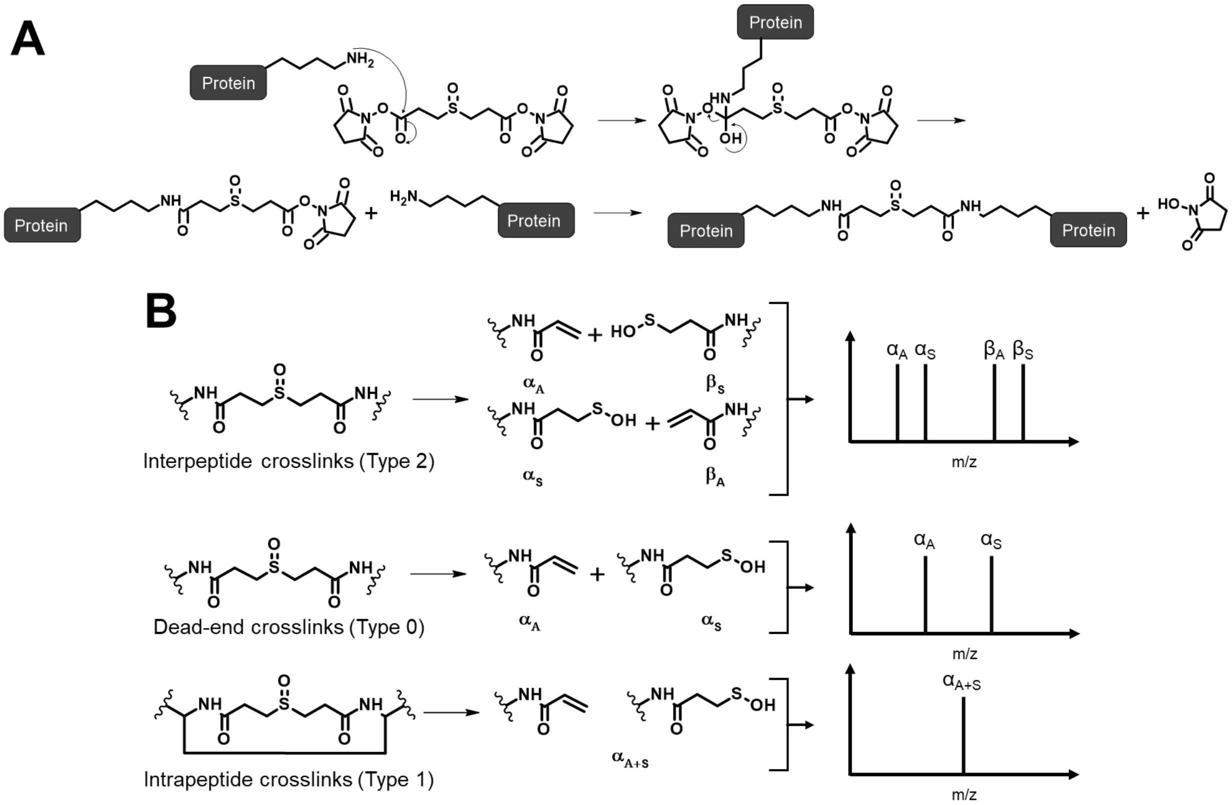
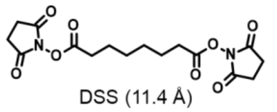
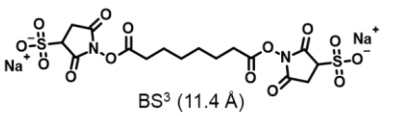
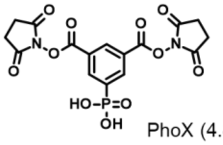
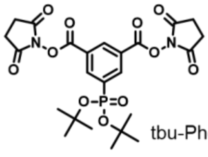
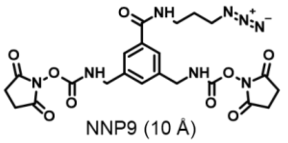
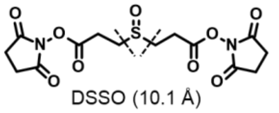
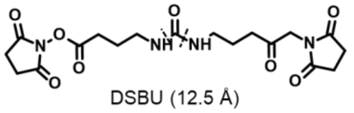
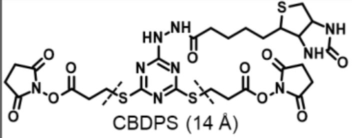
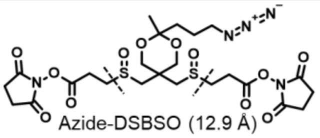
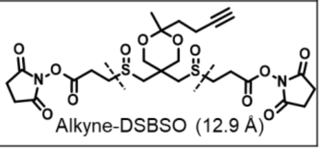

A
Protein
NH_2
Protein
B
Interpeptide crosslinks (Type 2)
α_A
β_S
α_S
β_A
α_A α_S β_A β_S
m/z
Dead-end crosslinks (Type 0)
α_A
α_S
m/z
Intrapeptide crosslinks (Type 1)
α_{A+S}
m/z

Figure 10. Crosslinking reaction mechanism and CID fragmentation of DSSO. (A) The reaction mechanism of the DSSO NHS ester with lysine involves the ester reacting with nucleophiles to release NHS. (B) Proposed fragmentation patterns of DSSO crosslinked peptides are as follows: interpeptide crosslinks result in four signals, dead-end modifications result in two fragment signals, and intrapeptide crosslinks result in one signal. Adapted from Kao et al.[319].

Azido and alkyne handlers can also be used to enrich crosslinked peptides by click-chemistry reactions. The latter are one-step copper-catalyzed cycloadditions that produce a stable triazole scaffold. Enrichment is then based on beads coated with the crosslinker counterpart group of a click-chemistry reaction. Alkyne/azide tagged disuccinimidyl bissulfoxide (DSBSO)[321] are crosslinkers designed to be enriched by click-chemistry. These crosslinkers have a 12.9 Å spacer arm and were designed to be able to elute from resin by acid-based cleavage. Additionally, these crosslinkers can be CID-cleaved. Another crosslinker containing an azido handle is NNP9[322]. This crosslinker has a spacer arm length of ~10 Å and possesses two carbamate moieties, a phenyl rigid core, but is not CID-cleavable. For this specific crosslinker, it was recommended to immobilize crosslinked peptides on an ultraviolet (UV)-cleavable support. After washing off free peptides, the beads were irradiated with UV light by which crosslinked peptides were eluted[323].

Table 4. Structure and spacer arm length of NHS-ester-based crosslinkers. *The crosslinkers are divided by their CID behavior and if they possess an enrichable handler. Dashed lines indicate CID-cleavable sites.*

Non-CID cleavable	Non-enrichable	 DSS (11.4 Å)  BS³ (11.4 Å)
	Enrichable	 PhoX (4.8 Å)  tbu-PhoX (4.8 Å)  NNP9 (10 Å)
CID cleavable	Non-enrichable	 DSSO (10.1 Å)  DSBU (12.5 Å)
	Enrichable	 CBDPS (14 Å)  Azide-DSBSO (12.9 Å)  Alkyne-DSBSO (12.9 Å)

Finally, two new-generation crosslinkers, disuccinimidyl phenyl phosphonic acid (Phox)[324] and tert-butyl disuccinimidyl phenyl phosphonate (tbu-Phox)[325], were developed to be enriched by immobilized metal ion-affinity chromatography (IMAC). Both crosslinkers are non-CID-cleavable, possess a 4.8 Å spacer and a phosphonic acid group as a molecular handle which is enriched in the same way as phosphopeptides. An advantage is that phosphonate is not cleaved by phosphatase hence, this enzyme can be used to elute endogenous phosphopeptides prior IMAC enrichment. Phox is more water-soluble and non-membrane-permeable, which is more suited for crosslinking membrane proteins. On the contrary, tbu-Phox is membrane-permeable and suited for in-cellulo crosslinking. In addition, tbu-Phox contains two tert-butyl protective groups on the phosphonic acid handle. This allows the peptide mixture to be pre-cleared of endogenous phosphopeptides with a first IMAC. Then, the flow-through is acidified to deprotect the phosphonic acid for a second IMAC enrichment. This approach avoids treating samples with phosphatase and capturing other IMAC-enrichable species.

XL-MS data analysis has made significant advances in recent years, with a focus on improving identification and reducing the false positive rate. The development of software tools has revolutionized comprehensive analysis of XL-MS datasets, enabling continuous improvement despite the increase in data complexity due to the field moving from individual protein complexes to complex biological systems. Thanks to the rapid progress in computing power, almost every research group in the XL-MS field relies on their own software adapted to meet their specific needs. Among these software tools, one finds MaxLynx[326], MeroX (StavroX)[317,327], OpenPepXL[328], XlinkX[283] and XiSearch/xiFDR[329], which reflect the increased interest in developing better algorithms to made this technology more available and improve the confidence of researchers in this cutting-edge field.

The use of XL-MS to identify the function of AltProts was shown by Cardon *et al.*[130] who isolated HeLa cell nuclei, used DSSO for crosslinking followed by protein digestion, SCX peptide fractionation and LC-MS/MS. Data analysis was performed in Proteome Discoverer with the XlinkX node and the database used was HaltORF. With this approach, 1,679 crosslink interactions were found, of which 292 involved AltProts. The AltProt AltATAD2 was found crosslinked to the RE/poly(U)-binding/degradation factor 1 (AUF1) and the ribosomal protein 10 (RPL10). To validate such interactions, protein docking was performed. The authors also described a mechanism in which AUF1 attaches on the external part of RPL10 and the interaction of AltATAD2 on the RPL10 region interacting with 5S ribosomal RNA as a mechanism of regulation of the ribosome. This particular example played a pivotal role in paving the way for the utilization of this technique in identifying the involvement of AltProts in various physio/pathological processes. Recent advancements in mass spectrometry gave rise to the development of new hybrid and tribrid mass spectrometers, which made XL-MS more accessible for large-scale interatomic studies across different systems. These technological advances have also enabled us to identify a wider range of AltProts than before, thus broadening our understanding of their role in biological processes.

PART II GENERAL OBJECTIVES

After 45 years since the introduction of mass spectrometry-based protein sequencing, shotgun proteomic analysis can now identify and quantify over 8,000 human proteins in just 24 minutes[330]. This high-throughput approach is widely used in proteomics. However, it can only detect proteins that are already stored in databases, so it cannot identify "new proteins". For example, the alternative prion protein (UniProtKB: F7VJQ) was added to the UniProtKB reviewed (Swiss-Prot) database 20 months after its discovery was published by Vanderperre *et al.*[331].

In 2010, the PRISM laboratory and Prof. Roucou from the University of Sherbrooke collaborated to investigate proteins not found in traditional databases. Prof. Roucou's groundbreaking work on the PRNP gene[331] led to the creation of a new database called HaltORF[53], which later became OpenProt[257]. This work introduced the term AltProts and pioneered the proteogenomic approach for characterizing them.

AltProts are conserved in evolution, especially among mammals[114]. RefProts, on the other hand, show lower conservation. Evidence of AltProt expression was found in green algae[107], rice[108], *Arabidopsis thaliana*[109], *Saccharomyces cerevisiae*[110], mice[111], *Drosophila melanogaster*[112] and zebrafish[113]. Proteome-wide studies in humans have revealed that AltProts make up approximately 15% of protein identifications in various cell lines, tissues, and biological fluids (such as cerebrospinal fluid, urine, plasma, and serum)[100].

First objective: Interaction mapping of AltProts

Despite the increasing evidence of AltProts' physiological and pathological functions, there is a lack of commercially available antibodies to study their expression. To understand their effects on intracellular pathways, inhibiting AltProts is one approach. Techniques like CRISPR-Cas9 or shRNA can be used to target the AltProt gene or transcript, respectively, to investigate the impact of AltProt inhibition on cellular phenotype, signaling pathways, or the regulation of other proteins. While these strategies are limited to a specific target, our goal is to understand the function of AltProts on a larger scale. Therefore, we have focused on large-scale, non-targeted strategies to identify PPIs. Through these studies, we aim to identify AltProts that interact with RefProts, allowing us to place an AltProt in a pathway or link it to a GO term, and thus assign a possible AltProt function through "guilt by association".

Various approaches can be used to identify PPIs. Although targeted AP-MS methods such as coIP-MS remain widely used, other strategies, such as APEX, BioID and Vitrotrap based on fusing the target protein (or bait protein) and afterwards enriching its partners, are becoming increasingly popular. However, all these methods are targeted and protein or organelle specific. In contrast, protein crosslinking strategies coupled to mass spectrometry (protein crosslinking analyzed by mass spectrometry or XL-MS) enable non-targeted searches for PPIs on a large scale in complex mixtures. XL-MS is based on chemical bridging between amino acid side chains of two proteins in close spatial proximity. This bridging freezes the interacting proteins and, after MS-based analysis, enables the identification of interacting partners and the site of interaction. However, XL-MS processes two main challenges. First, enriching the crosslinked peptides from the non-crosslinked. Finally, the need for specific computational algorithms that allow the identification of specific crosslinking peaks within a complex spectrum.

A methodology devoted to analyzing the cellular proteomes will be developed by XL-MS to add information on the function of the identified AltProts and their subcellular compartment. This approach is based on the necessity to reduce the complexity of the cellular proteomes after crosslinking to increase identification of crosslinked peptides. To gain information on AltProts beside their protein interactions, this approach can propose a localization for the identified AltProts.

Second objective: Cellular proteogenomic characterization

Besides MS-based proteomics, RNA-seq enables the assessment of expression levels of all transcripts. As a result, high resolution transcriptome maps are generated that provide transcript structures and levels of expression. Additionally, RNA-seq allows the annotation of novel transcripts[332]. Moreover, the innovative field of proteogenomics employs the transcriptomes to generate sample-specific, *in silico* predicted protein sequence databases[266]. By alignment to a reference transcriptome or genome, mutations can be detected and introduced in such protein sequence databases, allowing to identify the presence of mutated proteins.

The proteomic and transcriptomic landscapes of SKOV-3 and PEO-4 OvCa cell lines will be compared to those of immortalized ovary cells (T1074 cells). Based on the RNAseq data of each cell line, a dedicated database containing mutated RefProts and AltProts will be generated thanks to OpenCustomDB. RNAseq analysis will thus provide a source of data to compare with the proteomic results and to identify the different variations of AltProts, novel isoforms and RefProts. Additionally, we will be able to map the variations in different pathways assessing the differences between the three cell lines. The main goal of this objective is to realize a multi-omic characterization of OvCa cells, to compare with clinical data as the one present in COSMIC and TCGA or bibliography.

PART III DEVELOPMENT OF A PROTOCOL TO IDENTIFY HUMAN SUBCELLULAR ALTERNATIVE PROTEIN INTERACTIONS

System-wide crosslinking mass spectrometry

Originally designed for structural biology, XL-MS has now become a widespread technique for system-wide, high-throughput experiments. New developments, such as MS-cleavable, trifunctional crosslinkers and improved computational algorithms, have enabled XL-MS to move beyond its structural biology origin and expand to more complex systems like cell lysates, fruit fly embryos, organelles, tissue, plants, organs, living bacteria and human cells. This improvement is evident from the comparison of the first 2,427 crosslinks reported by Schweppe *et al*.[333] in 2017 and the 9,319 crosslinks reported by Yugandhar *et al*.[334] in 2020.

A key advantage of MS-cleavable crosslinkers is that they produce specific fragment ions, reduce the computational search space and enable more confident identification of crosslinked peptides in larger systems. This boosts the detection of crosslinks in complex mixtures. Using crosslinkers, interactions, including weak and transient protein-protein interactions in their native states, can be "frozen" and studied *in situ* in the system of interest. Overall, XL-MS snapshots provide insights into the protein interaction networks operating in diverse biological systems.

A key aspect of designing an XL-MS experiment is choosing the appropriate crosslinker, with longer crosslinkers being more suitable for identifying PPIs. Larger spacer arms have a larger radius of reactivity. This allows the crosslinker to bridge longer distances between interacting proteins they are more flexible and have less conformational contrains. Additionally, it is important to consider the membrane permeability of a crosslinker. If one aims to identify protein interactions of cellular membrane-bound proteins, a non-permeable crosslinker is advised. On the other hand, a permeable crosslinker like DSSO is recommended if the aim of the experiment is to identify *in-cellulo* PPIs. Finally, it is essential to consider that the addition of a chemical crosslinker will affect the diffusion and the interface of the proteins.

Reducing the complexity of the sample injected into the LC-MS system is another key factor in a crosslinking experiment. Crosslinked peptides are present in much lower quantities than their non-crosslinked counterparts. Thus, enriching the crosslinked proteins or peptides at is necessary as this will increase confidence and decrease the computational processing time to identify crosslinked peptides. One option to enrich crosslinked proteins or peptides is to use tri-functional crosslinkers as these allow for a pull-down enrichment due to the presence of a molecular handle in the crosslinkers. Such a pull-down can be performed with CBDPS, alkyne/azide-DSBSO, NNP9 and tbu-Phox crosslinkers. Although this approach eliminates (most of) the non-crosslinked peptides, method development and optimization need to be carefully considered, which can be a long and challenging phase. Another drawback is that some trifunctional crosslinkers contain bulky molecular handles that may interfere with the protein interface during the crosslinking reaction.

A second approach to enrich for crosslinked peptides is based on pre-fractionation of the peptide mixture by SCX or SEC. SCX pre-fractionation relies on the higher number of positive charges in crosslinked peptides compared to non-crosslinked peptides which thus elute before the crosslinked peptides. On the contrary, using SEC, higher molecular weight crosslinked peptides elute before linear non-crosslinked peptides.

Sequential digestion is an approach where proteins are digested by two different proteases to generate more effective protein cleavage. This is commonly performed by trypsin followed by another protease like chymotrypsin, or Glu-C. Using another protease that cleaves at different sites generates more diverse peptides, which improves protein sequence coverage and provides more effective digestion of proteins that are resistant to trypsin. However, additional proteases add more complexity to the peptide mixture, and a search engine that supports different proteases is required. The capability of applying this technique to XL-MS has been demonstrated by Mendes *et al.*[335]. They demonstrated that sequential digestion increased the number of identified crosslinked peptides when using trypsin followed by Glu-C, chymotrypsin or Asp-N.

Another parameter that can be optimized in XL-MS experiments is HCD. Stieger *et al.* demonstrated that applying stepped-collision energies allows the identification of a larger number of DSSO crosslinked peptides, avoiding the need for an MS^n instrumentation[336]. This is possible due to the different collision energy required to fractionate C-S and peptide bonds.

Subcellular fractionation

Subcellular fractionation involves separating cellular components into different fractions. The methodology involves a mild cell lysis step that keeps organelles intact. The cell homogenate is then centrifuged at different speeds (differential centrifugation) to pellet organelles based on their size and density. Larger organelles like nuclei pellet at lower speeds (1000 x g), while smaller ones like ribosomes require higher speeds (100,000 x g). Another technique for subcellular fractionation is equilibrium density centrifugation, in which a gradient of sucrose or glycerol is used to separate organelles with closely related densities. During centrifugation, zones of different densities are generated in which organelles or subcellular fractions of equal density can be found. Finally, major biotechnology companies have developed differential detergent fractionation kits. Additionally, compartment-specific or organelle-specific kits are available. Such kits allow the fractionation of cell pellets or tissues without ultracentrifugation or gradients in a ~3-hour timeframe and typically fractionate the cell into three to five compartments.

Subcellular fractionation is an interesting approach to decrease the complexity of a sample prior to analysis. As mentioned above, such fractionation or enrichment step can be exploited in XL-MS workflows. Moreover, cellular fractionation result in a greater coverage of the proteome compared to analyzing whole lysates of the cell[337]. This is a key characteristic that enhances detection of low-abundance proteins (AltProts and crosslinked proteins). Further, due to the lack of information on AltProts, such an approach also helps to define the subcellular localization of AltProts and possibly monitor their translocalization under physiological and stimulated conditions[338].

One main limitation however is the potential cross-contamination (overlap) that can occur between fractions, which can lead to inaccuracies in downstream analysis and data interpretation. Another limitation is the loss of weak PPIs, which can impact the sensitivity and specificity of the technique. Therefore, it is important to carefully optimize the conditions for each sample and to perform adequate quality control to ensure reliable results. Additionally, it is important to note that different buffers and detergents used in differential detergent fractionation might be incompatible with downstream analysis, which can limit the scope of the study. As such, it is essential to consider the compatibility of different techniques and to select the most appropriate approach for a given experimental question.

Objective

Cardon *et al*. introduced a methodology capable of identifying AltProt-RefProt PPIs, which also allowed for the identification of biochemical pathways and GO terms in which an AltProt is possibly involved[130]. However, current strategies using extensive fractionation or enrichment require large amounts of starting material, with minimums around 60 million cells[339] or 2 mg of protein[314], to identify significant numbers of crosslinked peptides.

In an era where MS-based proteomics aims to study proteomes at the single-cell level or in clinical samples of limited quantity, strategies to increase the identification of crosslinked peptides from small amounts of material are needed. Therefore, I proposed using subcellular fractionation to increase the identification of crosslinked peptides and simultaneously provide information on the cellular localization of identified AltProts in a human ovarian epithelial cell line (T1074). By fractionating a reasonable number of cells (3 million) into membrane, cytoplasmic, nuclear, chromatin and cytoskeletal proteomes, I reduced sample complexity and enriched crosslinked peptides for improved detection. Fractionating a cell's proteome increases the overall sensitivity and enables identification of crosslinked peptides from subcellular proteomes derived from limited starting material. In addition, based on the improvements in the identification of crosslinked peptides by sequential digestion and optimization of the stepped NCE for DSSO crosslinked peptides, I decided to optimize this parameter to boost the identification of crosslinking peptides.

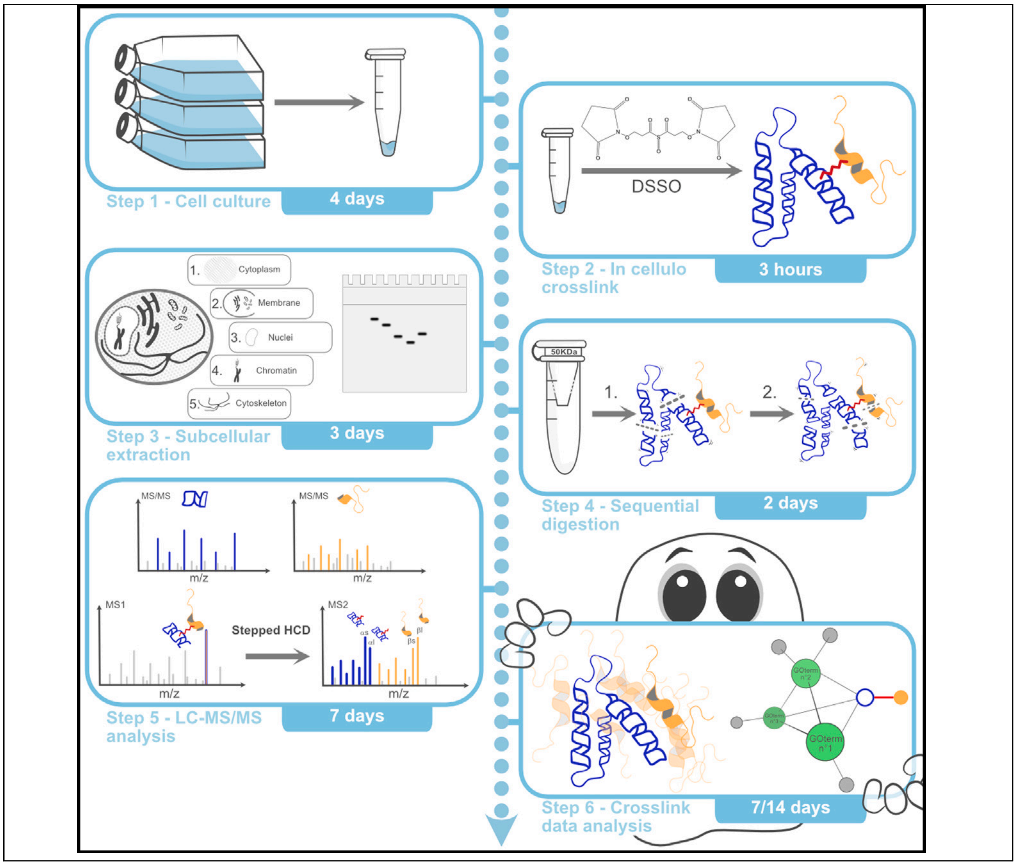
Overall, this protocol aims to enable high-throughput AltProt characterization, interaction mapping and functional assignment in cells. Moreover, it intends to overcome the limitations posed by the lack of reagents (antibodies) and references for these hidden proteins.

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol to identify human subcellular alternative protein interactions using cross-linking mass spectrometry



Since the start of mass-spectrometry-based proteomics, proteins from non-referenced open reading frames or alternative proteins (AltProts) have been overlooked. Here, we present a protocol to identify human subcellular AltProt and deciphering some interactions using cross-linking mass spectrometry. We describe steps for cell culture, *in cellulo* cross-link, subcellular extraction, and sequential digestion. We then detail both liquid chromatography-tandem mass spectrometry and cross-link data analyses. The implementation of a single workflow allows the non-targeted identification of signaling pathways involving AltProts.

Publisher's note: Undertaking any experimental protocol requires adherence to local institutional guidelines for laboratory safety and ethics.

Diego Fernando Garcia-del Rio, Isabelle Fournier, Tristan Cardon, Michel Salzet

isabelle.fournier@univ-lille.fr (I.F.)
tristan.cardon@univ-lille.fr (T.C.)

Highlights

Identify details related to possible function of unknown proteins

Steps described for using the cross-link to identify protein interactions

Determine subcellular localization information for unreferenced proteins

Highlight alternative proteins in cells by cross-link and subcellular fractionation

Garcia-del Rio et al., STAR Protocols 4, 102380
September 15, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.xpro.2023.102380

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol to identify human subcellular alternative protein interactions using cross-linking mass spectrometry

Diego Fernando Garcia-del Rio,[1,2,3,5] Isabelle Fournier,[1,*] Tristan Cardon,[1,4,5,*] and Michel Salzet[1,4,6]

[1]Université de Lille, Univ. Lille, CHU Lille, Inserm U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France

[2]VIB Center for Medical Biotechnology, VIB, Ghent 9052, Belgium

[3]Department of Biomolecular Medicine, Ghent University, Ghent 9052, Belgium

[4]These authors contributed equally

[5]Technical contact: diego.garciadelrio@univ-lille.fr; tristan.cardon@univ-lille.fr

[6]Lead contact

*Correspondence: isabelle.fournier@univ-lille.fr (I.F.), tristan.cardon@univ-lille.fr (T.C.)
https://doi.org/10.1016/j.xpro.2023.102380

SUMMARY

Since the start of mass-spectrometry-based proteomics, proteins from non-referenced open reading frames or alternative proteins (AltProts) have been overlooked. Here, we present a protocol to identify human subcellular AltProt and decipher some interactions using cross-linking mass spectrometry. We describe steps for cell culture, *in cellulo* cross-link, subcellular extraction, and sequential digestion. We then detail both liquid chromatography-tandem mass spectrometry and cross-link data analyses. The implementation of a single workflow allows the non-targeted identification of signaling pathways involving AltProts.

For complete details on the use and execution of this protocol, please refer to Garcia-del Rio et al.[1]

BEFORE YOU BEGIN

The protocol described below outlines the detailed steps and resources required for a high throughput interactomic study of alternative proteins (AltProts). The study of this kind of protein has been disregarded because mRNA was considered monocistronic. AltProts, also known as ghost proteins[2] or short open reading frames (sORF)-encoded proteins (SEPs),[3] are translated from alternative open reading frames (AltORFs), such as 3′ and 5′ UTRs, reading frame shifts or long non-coding RNAs (LncRNAs). Despite their physiological presence in the cell, studying these proteins has been difficult due to the absence antibodies and databases that cover this type of proteins. The number of potential Human AltProts has been estimated to be around 450,000 sequences,[4,5] five times larger than the actual reference proteome available in Uniprot. However, these proteins represent a vast source of potential physiopathological biomarkers.[2,6–10] To overcome this challenge, we propose a methodology based on cross-linking mass spectrometry (XL-MS), subcellular fractionation, and bioinformatic tools, which enables the retrieval of functional information through network and gene ontology (GO) analysis.

In this protocol we propose the exploration of AltProt in non-pathological cells, however it can be adapted to any cell, adaptations in terms of quantity of cells will be expected, especially due to their size and their cellular content.

Immortalized human ovarian epithelial cells (SV40) culture

Timing: 4 days

1. Seed immortalized human ovarian epithelial cells (SV40) into 25 cm^2 flasks.
   a. Thaw a cryogenic vial with immortalized human ovarian epithelial cells (SV40) at 37°C.
   b. Transfer the contents of the cryogenic vial to 9 mL of complete medium (Prigrow I medium with 10% of Hi-FBS and 100 U/mL of Penicillin-Streptomycin).
   c. Centrifuge at 100 × *g* for 5 min at 20°C.
   d. Remove supernatant by pipetting. Wash the cells gently pipetting up and down 5 mL of Dulbecco's phosphate-buffered saline (DPBS).
   e. Pellet the cells by centrifuging at 100 × *g* for 5 min at 20°C and remove the supernatant.
   f. Suspend cells with 5 mL of complete medium and thoroughly seed them in a 25 cm^2 flask.
   g. Incubate cells at 37°C with 5% CO_2.
   h. Observe the cells in a microscope every day until the cells reach 80%–90% confluency.

Note: A cryogenic vial of immortalized human ovarian cells typically contains 1 million viable cells in 1 mL FBS /DMSO 10% and is made from cells that are approximately 80%–90% confluent. Pre-warm all the reagents in a water bath at 37°C, for 20 min.

2. Passage of immortalized human ovarian cells into 75 cm^2 flasks.
   a. Once the cells reach ~80%–90% of confluency, remove the medium and wash the cell with 2.5 mL of DPBS.
   b. Detach the cell from the flask using 0.5 mL of 0.05% Trypsin-EDTA (1×), phenol red.
   c. Incubate for 5 min at 37°C, 5% CO_2.
   d. Add 1.5 mL of complete medium to inactivate Trypsin.
   e. Transfer the cells into a conical centrifuge tube and spin at 100 × *g* for 5 min at 20°C.
   f. Remove supernatant and wash the cells with 5 mL DPBS.
   g. Pellet the cells by centrifuging at 100 × *g* for 5 min at 20°C and remove the supernatant.
   h. Suspend cells with 10 mL of complete medium and thoroughly seed them in a 75 cm^2 flask.

Note: Repeat this step until the desired number of cells are reached with a confluency of ~80%–90%.

Pre-warm all the reagents in a water bath at 37°C, for 20 min.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Goat Anti-Chicken IgG IgY (IgG) (H + L) (HRP) (1/5000)	Jackson Immuno Research	Cat# 103-035-155; RRID: AB_2337381
Monoclonal Mouse anti-Cytokeratin 18 (1/1000)	Dako	Cat# M7010; RRID: AB_2133299
Monoclonal Mouse anti-Histone H3 (1/1000)	Santa Cruz Biotechnology	Cat# sc-517576; RRID: AB_2848194
Monoclonal Mouse anti-Hsp70 (1/1000)	Abcam	Cat# ab2787; RRID: AB_303300
Monoclonal Mouse anti-SP1 (1/200)	Santa Cruz Biotechnology	Cat# sc-420; RRID: AB_628271
Peroxidase AffiniPure Goat Anti-Mouse IgG (H + L) (1/5000)	Jackson Immuno Research	Cat# 115-035-146; RRID: AB_2307392
Polyclonal Chicken anti-Calreticulin (1/200)	Abcam	Cat# ab2908; RRID: AB_303403
Chemicals, peptides, and recombinant proteins		
Acetonitrile	Carlo Erba Reagents	Cat# 412341
Acrylamide: Bis acrylamide 29:1 (40% solution / electrophoresis)	Euromedex	Cat# EU0063-B

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Amersham Protran Western blotting membranes, nitrocellulose	Merck	Cat# GE10600002
Ammonium persulfate (APS), BioUltra, for molecular biology, ≥98.0%	Sigma-Aldrich	Cat# 09913
Ammonium bicarbonate, BioUltra ≥99.5%	Sigma-Aldrich	Cat# 09830
Bovine serum albumin	Merck	Cat# A3059
Bromophenol blue sodium salt	Sigma-Aldrich	Cat# B5525
Chymotrypsin, sequencing grade	Promega	Cat# V1062
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich	Cat# D5879
Disuccinimidyl sulfoxide (DSSO)	Thermo Fisher Scientific	Cat# A33545
DL-Dithiothreitol (DTT)	VWR Life Science	Cat# 97063-760
DPBS, no calcium, no magnesium	Thermo Fisher Scientific	Cat# 14190-094
Fetal bovine serum, qualified, heat inactivated, E.U.-approved, South America origin	Gibco	Cat# 10500064
Formic acid (for LC-MS)	TCI America	Cat#F0654
Glycerol, 99+%, extra pure	Thermo Fisher Scientific	Cat# 10562524
Glycine, 99+%, for analysis	Acros Organics	Cat# 10358210
Hydrochloric acid 37%,	VWR Chemicals	Cat# 20252.420
Iodoacetamide (IAA)	Sigma-Aldrich	Cat# I1149
2-Mercaptoethanol	Sigma-Aldrich	Cat# M6250
Methanol ≥98.5%, technical	VWR Chemicals	Cat# 20903.368
PageBlue™ Protein Staining Solution	Thermo Fisher Scientific	Cat# 24620
Penicillin-Streptomycin (10,000 U/mL)	Gibco	Cat# 15140122
Prigrow I Medium	Applied Biological Materials	Cat# TM001
Sodium chloride	Fisher Chemical	Cat# S/3161/60
Sodium dodecyl sulfate (SDS), UltraPure™	Invitrogen	Cat# 15525017
SuperSignal™ West Dura Extended Duration Substrate	Thermo Fisher Scientific	Cat#34075
Tetramethylethylenediamine (TEMED)	Bio-Rad	Cat# 1610801
TG-SDS 10×	Euromedex	Cat# EU0510
Trifluoroacetic acid	Sigma-Aldrich	Cat# 302031
TRIS biotech grade	Interchim	Cat# UP031657
Trypsin-EDTA (0.05%), phenol red	Gibco	Cat# 5300054
Trypsin/Lys-C Mix, Mass Spec Grade	Promega	Cat# V5073
Tween 20	Sigma-Aldrich	Cat# P2287
Urea Ultra-Pure	Euromedex	Cat# EU0014B
Water, UHPLC-MS	Thermo Fisher Scientific	Cat# 15339865
Critical commercial assays		
Detergent Removal Spin Columns HiPPR™	Thermo Fisher Scientific	Cat# 88306
Subcellular protein fractionation for cultured cells	Thermo Fisher Scientific	Cat# 78840
Experimental models: Cell lines		
Human Immortalized Ovarian Epithelial Cell line (SV40)	Applied Biological Materials	Cat# T1074
Software and algorithms		
Biological General Repository for Interaction Datasets (BioGRID)	Oughtred et al.[34]	RRID:SCR_007393; http://www.thebiogrid.org/
ClueGO	Bindea et al.[32]	RRID:SCR_005748; https://apps.cytoscape.org/apps/cluego
CluePedia	Bindea[33]	RRID:SCR_015784; https://apps.cytoscape.org/apps/cluepedia
ClusPro 2.0	Kozakov et al.[24]	RRID:SCR_018248; https://cluspro.bu.edu/login.php
Cytoscape 3.9.1	Shannon et al.[30]	RRID:SCR_003032; https://cytoscape.org
IntAct	Orchard et al.[35]	RRID:SCR_006944; http://www.ebi.ac.uk/intact

(*Continued on next page*)

Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
I-TASSER (Iterative Threading ASSEmbly Refinement)	Yang et al.[21]	RRID:SCR_014627; https://zhanggroup.org/I-TASSER/
NetMHC - 4.0	Andreatta and Nielsen 2016	RRID:SCR_021651; https://services.healthtech.dtu.dk/service.php?NetMHC-4.0
OpenProt Protein Database 1.6	Brunet et al.[5]	https://www.openprot.org/p/ng/Home
OriginPro, Version 2022b	OriginLab Corporation	RRID:SCR_014212; https://www.originlab.com/
Proteome Discoverer 2.5	Thermo Fisher Scientific	RRID:SCR_014477; https://www.thermofisher.com/order/catalog/product/OPTON-31040
STRING app	Doncheva et al.[31]	http://apps.cytoscape.org/apps/stringapp
UniProtKB	The UniProt Consortium	RRID:SCR_004426 https://www.uniprot.org/uniprotkb?facets=model_organism%3A9606&query=%2A
XlinkX 2.5 nodes for Proteome Discoverer 2.5	Thermo Fisher Scientific	https://www.thermofisher.com/order/catalog/product/OPTON-31047
YASARA view	YASARA Biosciences	RRID:SCR_017591; http://www.yasara.org/
yFiles Layout Algorithms	yWorks	https://apps.cytoscape.org/apps/yfileslayoutalgorithms
Other		
Amicon Ultra-0.5 Centrifugal Filter Unit 50 kDa	Merck	Cat# UFC505024
BB15 CO_2 Incubator	Thermo Fisher Scientific	Cat# 51023121
ECLIPSE Ts2 inverted microscope	Nikon	Cat# Ts2-FL
iBright CL750 Imaging System	Thermo Fisher Scientific	Cat# A44116
LD79 Digital Test-Tube Rotator	Labinco	Cat# 79000
Mini orbital shaker	ClearLine	Cat# 060956CL
Mini-PROTEAN® Tetra Handcast System	Bio-Rad	Cat# 1658003FC
Mini tube rotator	Thermo Fisher Scientific	Cat# 15534080
Oven 100-800	Memmert	Cat# 200718
Refrigerated centrifuge 5804 R	Eppendorf	Cat# 5805000010
PowerPac 1000	Bio-Rad	Cat# 4006038
Trans-Blot Cell	Bio-Rad	Cat# 20179
TW8 Water Bath	Julabo	Cat# 9550108
Vacufuge Concentrator System 5301	Eppendorf	Cat# 000210
ZipTip with 0.6 µL C18 resin	Merck	Cat# ZTC18S096

MATERIALS AND EQUIPMENT

Complete cell culture media		
Reagent	Final concentration	Amount
Hi-FBS (Heat-Inactivated Fetal Bovine Serum)	10%	5 mL
Penicillin/streptomycin	100 U/mL	500 µL
Prigrow I medium	N/A	Up to 50 mL
Total	**N/A**	**50 mL**
Store at 4°C up to 1 month.		

Tris/HCl 1.5 M		
Reagent	Final concentration	Amount
Tris base	1.5 M	908 mg
HCl	Up to pH 8.1	N/A
ddH_2O	N/A	Up to 5 mL
Total	**N/A**	**5 mL**
Store at 20°C up to 6 months.		

2× Laemmli buffer

Reagent	Final concentration	Amount
Tris base	125 mM	747 mg
SDS	4%	2 g
Glycerol	20%	10 mL
2-mercapto-ethanol	10%	5 mL
Bromophenol blue	N/A	100 mg
HCl	N/A	Up to pH ||6.8
ddH$_2$O	N/A	Up to 50 mL
Total	**N/A**	**50 mL**

Store 500 μL aliquots at −20°C up to 1 year.

2-mercapto-ethanol is seriously irritating and toxic if swallowed or inhaled, so it is advised to handle it in an active fume hood.

4% concentration SDS-PAGE gel

Reagent	Final concentration	Amount
Acrylamide: Bis Acrylamide 29:1 (40% Solution / Electrophoresis)	4%	1 mL
Tris-HCl (0.5 M) pH 6.8	125 mM	2.5 mL
SDS (10%)	0.1%	100 μL
APS (10%)	0.05%	50 μL
TEMED	0.1%	10 μL
ddH$_2$O	N/A	Up to 10 mL
Total	**N/A**	**10 mL**

Once the gel is solid, it can be stored for one week at 4°C keeping it humid.

Invitrogen™ Novex™ 4%–12% Tris-Glycine Plus, 1.0 mm, Midi Protein Gels (Cat# WXP41220BOX) can be also use.

12% migration SDS-PAGE gel

Reagent	Final concentration	Amount
Acrylamide: Bis Acrylamide 29:1 (40% Solution / Electrophoresis)	12%	3 mL
Tris-HCl (1.5 M) pH 8.8	375 mM	2.5 mL
SDS (10%)	0.1%	100 μL
APS (10%)	0.062%	62.2 μL
TEMED	0.062%	6.2 μL
ddH$_2$O	N/A	Up to 10 mL
Total	**N/A**	**10 mL**

Once the gel is solid, it can be stored for one week at 4°C keeping it humid.

Towbin transfer buffer

Reagent	Final concentration	Amount
Tris base	25 mM	3.03 g
Glycine	192 mM	14.4 g
Methanol	20%	200 mL
ddH$_2$O	N/A	Up to 1 L
Total	**N/A**	**1 L**

Store at 4°C up to 3 months.

10× TBS-T buffer

Reagent	Final concentration	Amount
Tris base	200 mM	24.23 g
NaCl	1.5 M	87.66 g
Tween 20	1%	10 mL
ddH$_2$O	N/A	Up to 1 L
Total	**N/A**	**1 L**

Store at 20°C up to 6 months.

Denaturing buffer		
Reagent	Final concentration	Amount
Urea	8 M	24 g
Tris base	0.1 M	600 mg
HCl	N/A	Up to pH 8.5
ddH_2O	N/A	Up to 50 mL
Total	**N/A**	**50 mL**

Store at 20°C up to 6 months.

- 50 mM DSSO: 1 mg of DSSO (pre-weighted tube) in 51.5 µL of DMSO.
- 10 µM BSA for cross-linking control: 6.6 mg of BSA in 10 mL of DPBS. Store at 4°C for up to 6 months.
- 1× Tris-Glycine-SDS running buffer: Dilute 100 mL of 10× TG-SDS in 900 mL of ddH_2O.
- 1× TBS-T buffer: Dilute 100 mL of 10× TBS-T buffer in 900 mL of ddH_2O.
- 5% milk blocking buffer: Dissolve 5 g of powder milk in 100 mL of 1× TBS-T.
- 100 mM reduction buffer: Dissolve 15.4 mg of dithiothreitol (DTT) in 1 mL of denaturing buffer.
- 50 mM alkylation buffer: Dissolve 9.3 mg of iodoacetamide (IAA) in 1 mL of denaturing buffer.
- 50 mM ammonium bicarbonate buffer: Dissolve 197.6 mg of ammonium bicarbonate in 50 mL of ddH_2O.
- 1% trifluoroacetic acid (TFA): Dilute 100 µL of TFA in 10 mL of UHPLC grade water.
- 0.1% TFA: Dilute 1 mL of 1% TFA in 9 mL of UHPLC-MS grade water.
- Mobile phase A: 0.1% formic acid in UHPLC-MS grade water.
- Mobile phase B: 0.1% formic acid in HPLC grade acetonitrile.

STEP-BY-STEP METHOD DETAILS

In cellulo cross-linking

Timing: 3 h

This step consists of detachment, collection of the cells and *in-cellulo* cross-linking.

1. Cell harvesting
   a. Once the cells reach ~80%–90% of confluency, remove the medium and wash the cells with 5 mL of DPBS.
   b. Detach the cell from the flask using 1 mL of 0.05% Trypsin-EDTA (1×), phenol red.
   c. Incubate for 5 min at 37°C, 5% CO_2.
   d. Add 2 mL of complete cell culture media to inactivate Trypsin-EDTA (0.05%), phenol red.
   e. Transfer the cells into a centrifuge tube and centrifugate at 100 × *g* for 5 min at 20°C.
   f. Remove supernatant and wash the cells with 5 mL of DPBS.
   g. Pellet the cells by centrifuging at 100 × *g* for 5 min at 20°C and remove the supernatant by aspiration.
   h. Repeat the DPBS wash two more times.
   i. Count the cells and aliquot to 3 million cells.
   j. Pellet the cells by centrifuging at 100 × *g* for 5 min at 20°C and aspirate the supernatant.
   k. Keep the dry pellet on ice.

   Note: In our experience the optimal number of cells is 3 million, this number has to be adapted and tested for other kinds of cells.

2. Cross-linking reaction.
   a. Resuspend 3 million cells in 196 µL of DPBS.
   b. Prepare two 10 µM BSA solution as positive and negative control.

c. Prepare a 50 mM stock solution of disuccinimidyl sulfoxide (DSSO) in DMSO.
d. Add 4 μL of 50 mM DSSO to the suspended cells and positive BSA control (Final DSSO concentration of 2 mM).
e. Incubate at 37°C with continuous shaking (10 RPM) for 1 h.
f. To quench the reaction, add 10 μL of Tris-HCl 1.5 M.
g. Incubate for 30 min at 20°C with continuous shaking (10 RPM).
h. Centrifuge at 2,000 × *g* for 10 min at 4°C. Remove the supernatant.
i. Wash the cells with 200 μL of ice-cold DPBS.
j. Pellet the cells by centrifuging at 100 × *g* for 5 min at 4°C and remove as much supernatant as possible.

Note: It is recommended to use a BSA cross-linking positive control (96 μL of 10 μM BSA). Negative control is performed using 4 μL of DMSO instead of DSSO. The optimal working final concentration of DSSO is between 1–5 mM.

⚠ CRITICAL: Cross-linkers are moisture sensitive. Prepare these cross-linkers immediately before use. Use amine-free buffers (PBS, 20 mM HEPES, 100 mM carbonate/biocarbonate, or 50 mM borate). Cross-linking reactions (acylation) are favored near neutral pH (pH 6– 9) and with concentrated protein solutions.

Subcellular protein fractionation of cross-linked and non-cross-linked cells.

Timing: 3 days

The following methodology describes the steps of subcellular protein fractionation after the cross-linking reaction. This methodology is based on the instructions from the Subcellular Protein Fractionation Kit for Cultured Cells (Thermo Scientific, Cat# 78840). For more details and troubleshooting, please refer to the manual on Thermo website.

DMSO control of cells are treated according to the same protocol. These controls are key to determining the experimental subcellular location of the AltProts.

Other subcellular fractionation kits in the market are Abcam's Cell Fractionation Kit - Standard (Cat# ab109719) and Cell Signaling's Cell Fractionation Kit (Cat# 9038). These kits are designed to fractionate the cells in three subcellular fractions which will not decrease the complexity of the samples as much.

Note: Thaw all buffers using a 20°C water bath. Keep CEB, MEB, and NEB buffers on ice until use. Use a rotary shaker to avoid clumping of insoluble material during incubations.

⚠ CRITICAL: Immediately before use, add Thermo Scientific Halt Protease Inhibitor Cocktail at a 1:100 dilution into each volume of buffer required. Keep all protein extracts on ice.

3. Subcellular Protein Fractionation of 3 million cells (Figure 1).
   a. Lyse the cells by adding 300 μL of Cytoplasmic Extraction Buffer (CEB). Incubate at 4°C for 10 min with gentle shaking (10 RPM).
   b. Centrifuge at 2,000 × g for 5 min. Aspirate by pipetting and immediately transfer the supernatant (cytoplasmic extract) to a clean, pre-chilled (4°C in ice) 1.5 mL microcentrifuge tube.
   c. Resuspend the pellet in 300 μL of ice-cold Membrane Extraction Buffer (MEB). Vortex for 5 s and incubate at 4°C for 10 min with gentle shaking (10 RPM).
   d. Centrifuge at 5,000 × g for 5 min. Aspirate and immediately transfer the supernatant (membrane extract) to a clean, pre-chilled 1.5 mL microcentrifuge tube.

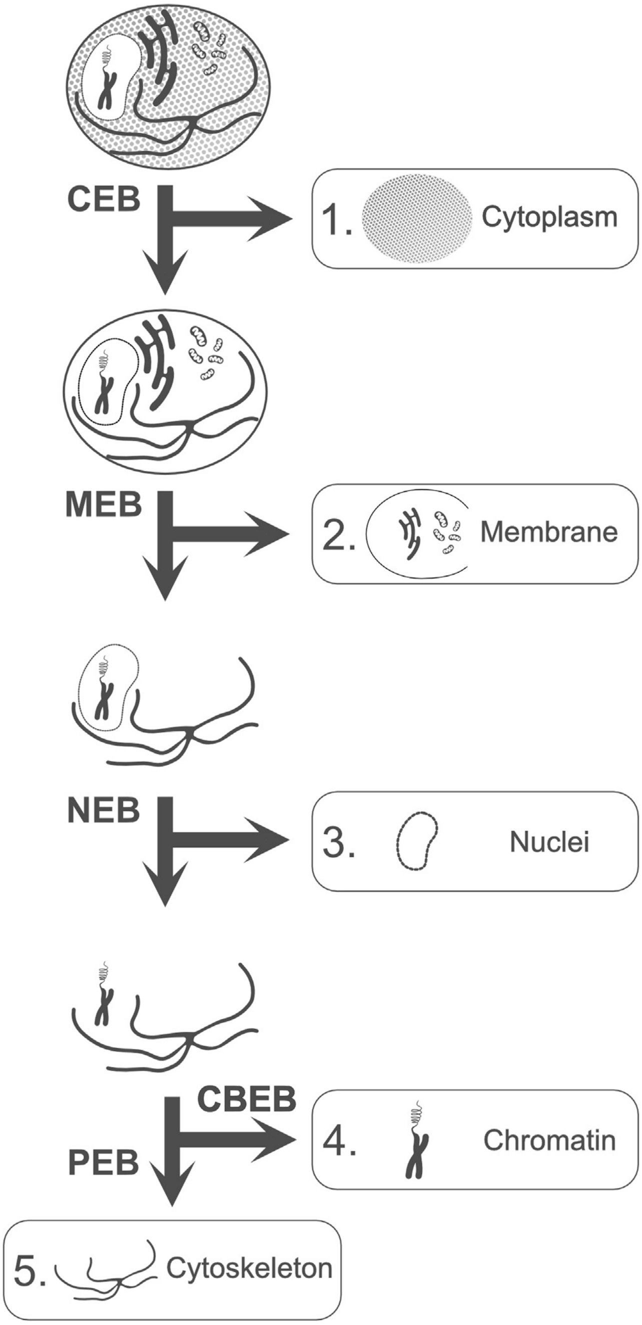


Figure 1. Subcellular protein fractionation workflow

Pelleted cells, both cross-linked and non-cross-linked, are resuspended in CEB buffer. After incubation and centrifugation, the resulting cytoplasmic extract is removed and stored. The remaining pellet is then retaken in MEB buffer, incubated, and centrifuged to obtain the membrane extract, which is also removed and stored. Next, the pellet is resuspended in NEB buffer, incubated, and centrifuged to obtain the nuclear extract, which is similarly removed and stored. The pellet from the previous step is then retaken in CBEB buffer, incubated, and centrifuged to obtain the chromatin-bound extract, which is also removed and stored. Finally, the remaining pellet is taken back in PEB buffer, incubated, and centrifuged to obtain the cytoskeletal extract, which is also removed and stored.

e. Add 150 μL of ice-cold Nuclear Extraction Buffer (NEB). Roughly vortex (highest vortex setting) for 15 s and incubate at 4°C for 30 min with gentle shaking (10 RPM).

Note: During the 30-min incubation time, prepare the Chromatin-Bound Extraction Buffer (CBEB) by adding 15 μL of 100 mM $CaCl_2$ and 9 μL of Micrococcal Nuclease (300 units) in 150 μL of 20°C NEB.

f. Centrifuge at 7,000 × g for 5 min. Aspirate and immediately transfer the supernatant (nuclear extract) to a clean, pre-chilled 1.5 mL microcentrifuge tube.
g. Resuspend the pellet in 150 μL of 20°C CBEB. Roughly vortex for 15 s and incubate at 20°C for 15 min with gentle shaking (10 RPM).
h. After incubation, roughly vortex 15 s and centrifuge at 16,000 × g for 5 min. Aspirate and immediately transfer the supernatant (chromatin-bound extract) to a clean, pre-chilled 1.5 mL microcentrifuge tube.
i. Add 150 μL of Pellet Extraction Buffer (PEB) to the remaining pellet. Roughly vortex for 15 s and incubate at 20°C for 10 min with gentle shaking (10 RPM).
j. After incubation, roughly vortex for 15 s and centrifuge at 16,000 × g for 5 min. Aspirate and immediately transfer the supernatant (cytoskeletal extract) to a clean pre-chilled 1.5 mL microcentrifuge tube.
k. Aliquot 10 μL of each extract for SDS-PAGE and western blotting.

Note: Performing protein quantification employing Thermo Scientific Pierce BCA Protein Assay (Cat# 23225) is recommended to calculate the correct protein: protease ratio, for the sequential enzymatic digestion.

Keep the extracts in ice or for long-term storage keep them at −80°C.

4. Cross-linking and subcellular fractionation confirmation.
   a. Mix 10 μL of 2× Laemmli buffer with the 10 μL protein aliquot.
   b. Load each sample of subcellular protein fraction and BSA on to a 4%–12% SDS-PAGE gel.
   c. Migrate the gels for 15 min at 70 V and for 90 min at 120 V in Tris-glycine-SDS buffer.
   d. After migration, stain the gels with PageBlue™ Protein Staining Solution (Coomassie blue) for 1 h.
   e. Destain the gels by discarding the excess staining solution. Rince the gels two times with water.
   f. Wash the gel for 16 h in an orbital shaker at 60 RPM. Placing a folded Kimwipes Tissue in the container to absorb excess dye will accelerate the destaining process.
   g. Visualize the destained gels using your preferred system (Figure 2A).

Note: Only non-cross-linked samples will continue to the western blot analysis.

⚠ CRITICAL: Use one membrane with the five subcellular protein fractions for one compartment specific primary antibody.

   h. Transfer the gels onto a 0.45 μm nitrocellulose membrane. Employ a tank transfer system for 2 h at 290 mA in Towbin buffer.
   i. Wash the membranes for 5 min in an orbital shaker three times with 20 mL of 1× TBS-T buffer.
   j. Block the membranes for 1 h in an orbital shaker with 20 mL of milk blocking solution.
   k. Meanwhile, prepare the primary antibody dilution in milk blocking solution. The concentration of antibody has been adjusted as following: Cytokeratin 18 (1/1000), SP1 (1/200), Histone H3 (1/1000), Hsp70 (1/1000) and Calreticulin (1/200).

⚠ CRITICAL: Depending on the antibody's supplier, the dilution of it must be adjusted.

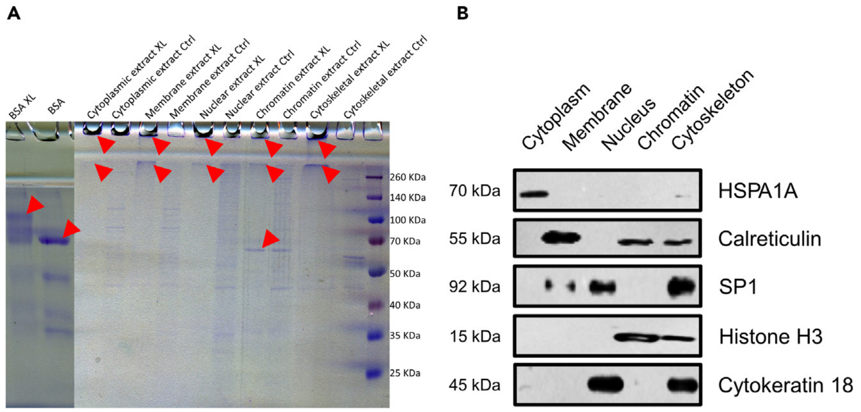


Figure 2. Cross-linking reaction and subcellular fractionation confirmation

(A) The Coomassie blue stained SDS-PAGE displays each cross-linked subcellular fraction, compared to a non-cross-linked fraction. BSA cross-linked and not cross-linked are used as controls. Red arrows display cross-linked signals demonstrating that the reaction takes place.

(B) The subcellular fractionation was confirmed by Western blot with compartment specific markers. The cytoplasm fraction showed the presence of HSPA1A signal. Calreticulin signal was detected at chromatin, cytoskeleton, and exhibited a stronger signal at the membrane-bounded fraction. SP1 was observed in the nucleus and cytoskeleton, while Histone H3 was found in chromatin and cytoskeleton. Similarly, Cytokeratin 18 was detected in the nucleus and cytoskeleton. These findings are consistent with the results reported in UniProtKB, COMPARTMENTS, and the literature.

l. Incubate each membrane for each antibody for 16 h at 4°C in an orbital shaker.
m. After incubation, wash the membrane for 5 min in an orbital shaker three times with 20 mL of 1× TBS-T buffer.
n. Incubate for 1 h with the matched HRP anti-Chicken (1/5000) or anti-Mouse (1/5000) secondary antibody.
o. After incubation, wash the membrane for 5 min in an orbital shaker three times with 1× TBS-T buffer.
p. Perform the horseradish-peroxidase reaction, by preparing the SuperSignal™ West Dura Extended Duration Substrate (1 mL of Luminol/Enhancer Solution and 1 mL of Stable Peroxide Solution).
q. Incubate the membrane with the substrate working solution for 5 min in the dark.
r. Remove the membrane from the substrate working solution and place it in a plastic sheet protector.
s. Remove the excess liquid with an absorbent tissue pressing out bubbles.
t. Scan the membranes using the Invitrogen iBright Imaging System or other compatible imaging system (Figure 2B).
    i. Mode: Chemi Blots.
    ii. Exposure mode: Normal.
    iii. After the autoexposure, the exposure time adjusted for each membrane: Cytokeratin 18 (5085 ms), SP1 (30 s), Histone H3 (4106 ms), Hsp70 (8213 ms) and Calreticulin (10 s).
    iv. Resolution: 4 × 4.
    v. Optical zoom: 1×.
    vi. Digital zoom: 2×.
    vii. Focus Level: 220.
    viii. Sensitivity: Frame 1:100.

Sequential enzymatic digestion

Timing: 2 days

The subsequent steps described the filter aided sample preparation (FASP)[11] sequential enzymatic digestion using LysC/Trypsin followed by Chymotrypsin for cross-linked samples. A 50 kDa cut-off Amicon filter is suggested to eliminate as many as possible non-cross-linked proteins.

5. FASP and sequential digestion
   a. Concentrate the five subcellular fractions in the 50 kDa Amicon filter by centrifugation at 4°C for 15 min at 14,000 × *g*. As a result, a 20 μL protein concentrate will remain as dead volume in the filter.
   b. Add 80 μL of denaturing buffer to the filter and pipette up and down gently inside the filter.
   c. Add 100 μL of reduction buffer.
   d. Incubate at 56°C for 40 min.

   Note: Don't incubate at 95°C. At above 60°C urea can produce protein carbamylation. Additionally, the filter could melt.

   e. Centrifuge 15 min at 14,000 × *g*.
   f. Add 200 μL of denaturing buffer and centrifugate 15 min at 14,000 × *g*.
   g. Repeat step 5e at least two times.
   h. Add 100 μL of alkylation buffer.
   i. Incubate for 20 min at 20°C in the dark.
   j. Centrifugate 15 min at 14,000 × *g*.
   k. Add 200 μL of ammonium bicarbonate buffer and centrifugate 15 min at 14,000 × *g*.
   l. Repeat the previous step at least two times.
   m. Add Trypsin/Lys-C Mix Mass Spec Grade to the vendor recommended 25:1 protein: protease ratio (w/w). Incubate for 16 h at 37°C.
   n. After incubation, add Chymotrypsin, Sequencing Grade at a 100:1 protein: protease ratio (w/w). Incubate for 4 h at 20°C.
   o. Place the Amicon filter into a new clean tube.
   p. Add 50 μL of ammonium bicarbonate buffer and centrifugate 15 min at 14,000 × *g*.
   q. Repeat the previous step.
   r. Discard the Amicon filter.
   s. Acidify the filtered peptides with TFA 1% until pH < 7.
   t. Vacuum dry the samples in a SpeedVac concentrator and store them at −20°C if needed.
   u. The membrane fraction contains a large amount of polymer. It requires the use of a HiPPR™ Detergent Removal Resin column (Thermo Scientific, Cat# 88305) following the vendor's protocol to be compatible for MS analysis.

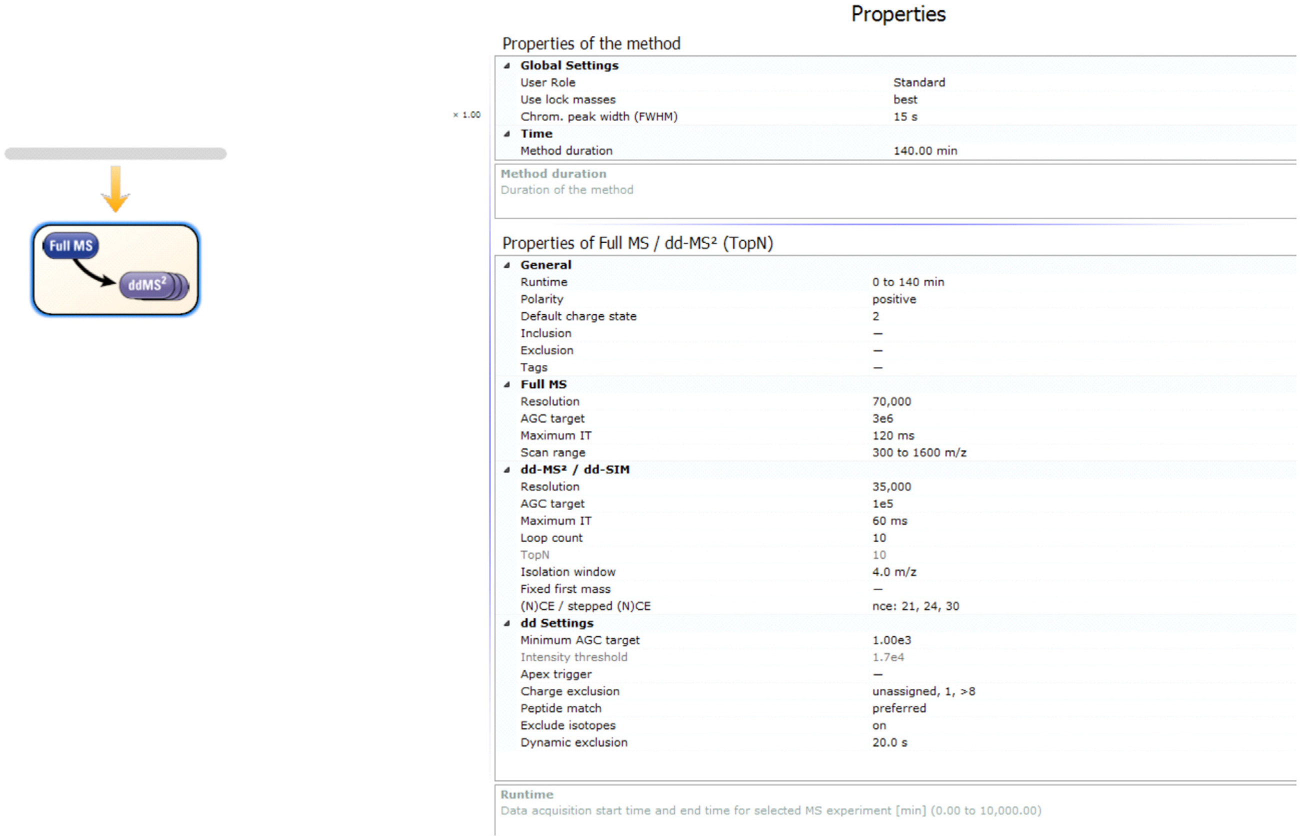
   Note: −80°C is recommended for long term sample storage.

NanoLC-MS/MS analysis

Timing: 1 week

The following section describes the parameters used in the nanoLC-MS/MS sample analysis and shotgun protein interrogation of non-cross-linked samples. We recommend the use of Sequest HT[12] search algorithm at Thermo Fisher's Proteome Discoverer.

6. NanoLC-MS/MS
   a. Resuspend the dried samples in 0.1% TFA.
   b. Desalt the peptides using C18 resin ZipTips.



Properties

Properties of the method

Global Settings	
User Role	Standard
Use lock masses	best
Chrom. peak width (FWHM)	15 s
Time	
Method duration	140.00 min

Method duration
Duration of the method

Properties of Full MS / dd-MS² (TopN)

General	
Runtime	0 to 140 min
Polarity	positive
Default charge state	2
Inclusion	–
Exclusion	–
Tags	–
Full MS	
Resolution	70,000
AGC target	3e6
Maximum IT	120 ms
Scan range	300 to 1600 m/z
dd-MS² / dd-SIM	
Resolution	35,000
AGC target	1e5
Maximum IT	60 ms
Loop count	10
TopN	10
Isolation window	4.0 m/z
Fixed first mass	–
(N)CE / stepped (N)CE	nce: 21, 24, 30
dd Settings	
Minimum AGC target	1.00e3
Intensity threshold	1.7e4
Apex trigger	–
Charge exclusion	unassigned, 1, >8
Peptide match	preferred
Exclude isotopes	on
Dynamic exclusion	20.0 s

Runtime
Data acquisition start time and end time for selected MS experiment [min] (0.00 to 10,000.00)

Figure 3. Settings for MS and MS^2 acquisition method
Parameters used for data dependent acquisition method.

Note: we recommend following the protocol described in the product insert (Merck Cat# ZTC18S096). Another alternative is to use Affinisep AttractSPE®Tips - C18, 200µL (Cat #Tips-C18.T1.200.96); Thermo Fisher Pierce™ C18 Tips (Cat # 87782), or home-made stage tips based on C18 membrane.

c. Vacuum dry the desalted peptides.
d. Resuspend in 20 µL of ACN/0.1% FA (2:98, v/v) and then transfer to a clean autosampler vial.
e. Inject 5 µL of sample onto the nanoLC-MS/MS system.
f. Analyze samples on a nanoAcquity (Waters) coupled to a Q Exactive mass spectrometer (Thermo Fisher Scientific).
   i. The injected sample is trapped on a ACQUITY UPLC M-Class Symmetry C18 Trap Column (100 Å, 5 µm, 180 µm × 20 mm, 2G, V/M, Waters Part No: 186007496).
   ii. The peptides are separated on a ACQUITY UPLC M-Class Peptide BEH C18 Column (75 µm × 250 mm, 1.7 µm, 130 Å, Waters Part No: 186007484) with a flow of 300 nL/min.
   iii. Mobile phase A is ultrapure water, 0.1% formic acid, while mobile phase B is acetonitrile, 0.1% formic acid.
   iv. The eluent gradient is set to go from 5% to 20% of mobile phase A in 100 min, then from 20% to 30% in 20 min, and finally to 90% in 20 min.
   v. Settings for MS and MS/MS acquisitions (Figure 3): range is m/z 300–1,600, resolution 70,000 at FWHM (m/z 400), positive mode, AGC target of 3 × 10^6 and stepped NCE of 21, 24 and 30.

MS/MS spectra are acquired using a Top-10 DDA (data-dependent acquisition) method, with a resolution of 35,000 FWHM. Dynamic exclusion is enabled, and only MS/MS spectra from peptide ions with charge states between +2 and +8 are selected.

7. Shotgun data analysis of non-cross-linked samples

Note: as previously described, the databased OpenProt[5] can be oversized for some identification nodes, we recommend the use of the limited AltProt database with at least 1 identification in other MS data or Riboseq analysis.[5] According to size limitation we recommend the use of SequestHT (Thermo ProteomeDiscoverer V2.5) which is not size limited.

Two different databases and consensus steps are employed to analyze AltProts (in a FASTA file combined AltProt, new isoforms and RefProt) and RefProts alone. The common consensus and processing parameters are enumerated at 7a. Specific parameters for RefProts (7b) and AltProts (7c) are displayed below.

a. Analyze the RAW LC-MS/MS data using Proteome Discoverer V2.5 (Thermo Fisher Scientific) with the Sequest HT search engine.
   i. Select LysC-trypsin and chymotrypsin as cleaving enzymes and 2 possible missed cleavages.
   ii. Variable modifications: methionine oxidation and protein N-terminus acetylation.
   iii. Static modifications: carbamidomethylation of cysteines.
   iv. Minimum peptide length: six amino acids.
   v. Minimum precursor tolerance: 10 ppm.
   vi. PSM and peptide validator: between 0.01 and 0.05 FDR.
   vii. Fragment mass tolerance: 0.02 Da.
   viii. Validation is done with Percolator using strict FDR = 0.01 and relaxed FDR = 0.05
b. For RefProts identification.
   i. Protein database: UniProtKB v.2022_02 reviewed and unreviewed. (77,895 sequences, downloaded from Uniprot website, 25 feb 2022)
   ii. At least two peptides per sequence.
c. For AltProts identification.
   i. Protein database: Homo sapiens OpenProt v1.6 (184,706 sequences), containing RefProts and predicted AltProts detected in mass spectrometry experiments with at least one unique peptide.
   ii. At least one peptide per sequence.

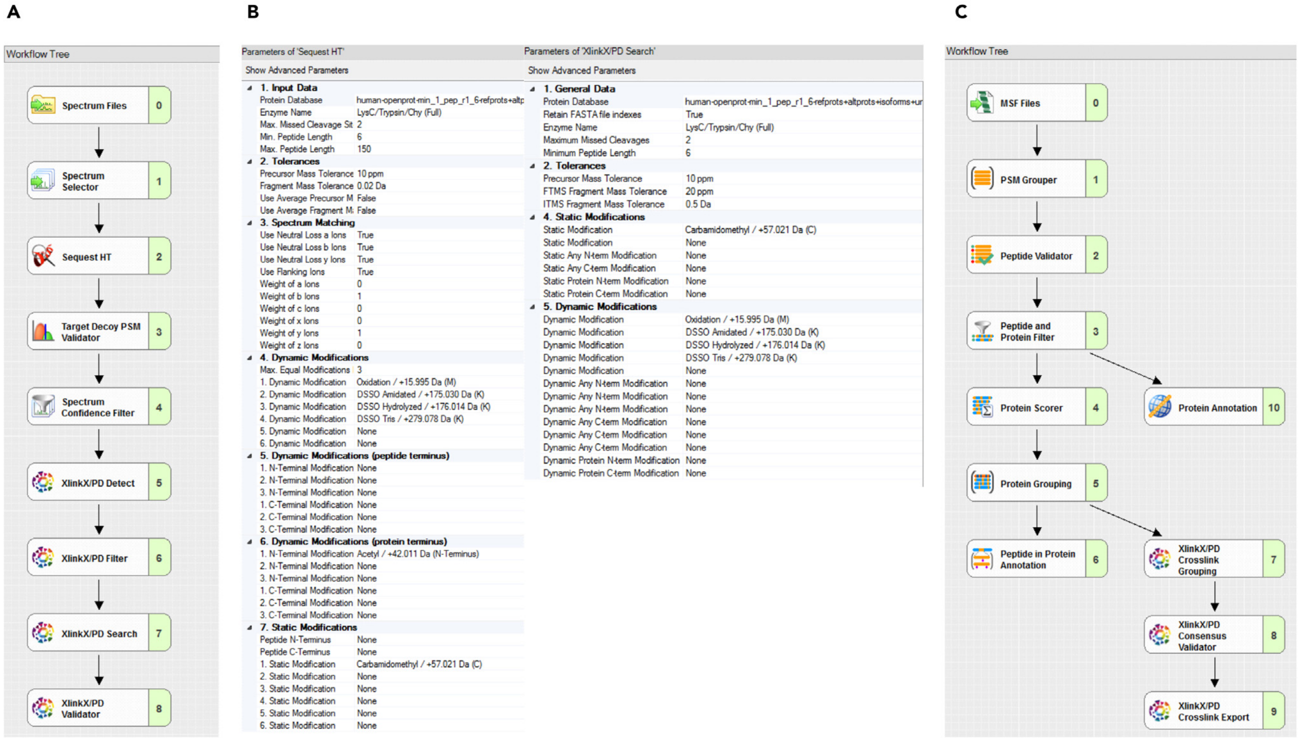
⚠ CRITICAL: To eliminate false positives, use protein BLASTP[13]

The basic parameters of BlastP are preserved (or the automated adaptation for small protein) and the database used for the comparison is "non-redundant protein sequences". An AltProt is considered "false positive" if it presents a sequence homology (identity+ coverage) > 80%, ideally no identification should be detected, if an alignment <80% homology is identified, it should be checked that the peptide/PSM identified in MS is specific to the AltProt.

Additionally, check the identified PSMs of the AltProt with the NextProt Peptide uniqueness checker[14] tool. Parameters from the NextProt uniqueness tool have been kept unchanged and the expected result is no sequence homology for the specific peptide of the AltProt previously identified and tested.

Cross-link data analysis and interaction modeling

Timing: 1 week



Parameters of 'Sequest HT'

Show Advanced Parameters

1. Input Data	
Protein Database	human-openprot-min_1_pep_r1_6-refprots+altp
Enzyme Name	LysC/Trypsin/Chy (Full)
Max. Missed Cleavage Sit	2
Min. Peptide Length	6
Max. Peptide Length	150
2. Tolerances	
Precursor Mass Tolerance	10 ppm
Fragment Mass Tolerance	0.02 Da
Use Average Precursor M	False
Use Average Fragment M:	False
3. Spectrum Matching	
Use Neutral Loss a Ions	True
Use Neutral Loss b Ions	True
Use Neutral Loss y Ions	True
Use Flanking Ions	True
Weight of a Ions	0
Weight of b Ions	1
Weight of c Ions	0
Weight of x Ions	0
Weight of y Ions	1
Weight of z Ions	0
4. Dynamic Modifications	
Max. Equal Modifications	3
1. Dynamic Modification	Oxidation / +15.995 Da (M)
2. Dynamic Modification	DSSO Amidated / +175.030 Da (K)
3. Dynamic Modification	DSSO Hydrolyzed / +176.014 Da (K)
4. Dynamic Modification	DSSO Tris / +279.078 Da (K)
5. Dynamic Modification	None
6. Dynamic Modification	None
5. Dynamic Modifications (peptide terminus)	
1. N-Terminal Modification	None
2. N-Terminal Modification	None
3. N-Terminal Modification	None
1. C-Terminal Modification	None
2. C-Terminal Modification	None
3. C-Terminal Modification	None
6. Dynamic Modifications (protein terminus)	
1. N-Terminal Modification	Acetyl / +42.011 Da (N-Terminus)
2. N-Terminal Modification	None
3. N-Terminal Modification	None
1. C-Terminal Modification	None
2. C-Terminal Modification	None
3. C-Terminal Modification	None
7. Static Modifications	
Peptide N-Terminus	None
Peptide C-Terminus	None
1. Static Modification	Carbamidomethyl / +57.021 Da (C)
2. Static Modification	None
3. Static Modification	None
4. Static Modification	None
5. Static Modification	None
6. Static Modification	None

Parameters of 'XlinkX/PD Search'

Show Advanced Parameters

1. General Data	
Protein Database	human-openprot-min_1_pep_r1_6-refprots+altprots+isoforms+ur
Retain FASTA file indexes	True
Enzyme Name	LysC/Trypsin/Chy (Full)
Maximum Missed Cleavages	2
Minimum Peptide Length	6
2. Tolerances	
Precursor Mass Tolerance	10 ppm
FTMS Fragment Mass Tolerance	20 ppm
ITMS Fragment Mass Tolerance	0.5 Da
4. Static Modifications	
Static Modification	Carbamidomethyl / +57.021 Da (C)
Static Modification	None
Static Any N-term Modification	None
Static Any C-term Modification	None
Static Protein N-term Modification	None
Static Protein C-term Modification	None
5. Dynamic Modifications	
Dynamic Modification	Oxidation / +15.995 Da (M)
Dynamic Modification	DSSO Amidated / +175.030 Da (K)
Dynamic Modification	DSSO Hydrolyzed / +176.014 Da (K)
Dynamic Modification	DSSO Tris / +279.078 Da (K)
Dynamic Modification	None
Dynamic Any N-term Modification	None
Dynamic Any N-term Modification	None
Dynamic Any N-term Modification	None
Dynamic Any C-term Modification	None
Dynamic Any C-term Modification	None
Dynamic Any C-term Modification	None
Dynamic Protein N-term Modification	None
Dynamic Protein C-term Modification	None

Figure 4. Proteome Discoverer cross-link identification workflow

(A and B) (A) Displays the processing step used to identify cross-links. The parameters used in the Sequest HT are displayed in the (B) panel. (B) The parameters used at Sequest HT and XlinkX/PD search are displayed.

(C) Shows the consensus step for cross-link identification.

To identify the cross-links, we employed the XlinkX[15–17] node in Proteome Discoverer. Additionally, the validation of the cross-links can be performed by docking the protein-protein interactions (PPIs) and measuring the distances between the residues involved in each cross-link. Here, we describe the modeling of the 3D structure of AltProts and docking them to the RefProts to which they were cross-linked.

8. Cross-linking identification
   a. Analyze the RAW LC-MS/MS data using Proteome Discoverer V2.5 (Thermo Fisher Scientific) with the Sequest HT search engine at the processing step (Figure 4A).
      i. Protein database: Homo sapiens OpenProt v1.6, which contains RefProts and predicted AltProts detected in mass spectrometry experiments with at least one unique peptide.
      ii. Select LysC-trypsin and chymotrypsin as cleaving enzymes and allow for 2 possible missed cleavages.
      iii. Set minimum peptide length to six amino acids and at least one peptide per sequence.
      iv. Set minimum precursor tolerance to 10 ppm.
      v. Set fragment mass tolerance to 0.02 Da.
      vi. Maximum equal modifications per peptide: 3
      vii. Maximum dynamic modifications per peptide: 4
      viii. Variable modifications: methionine oxidation and N-terminus acetylation, DSSO amidated, hydrolyzed, and Tris form.
      ix. Static modification: carbamidomethylation of cysteines.
   b. Set the Target Decoy PSM Validator:
      i. Target/decoy selection: concatenated
      ii. FDR set between 0.01 and 0.05.

c. Set a spectrum confidence filter: worse than high.
d. Detect the cross-links using the XlinkX/PD Detect node in Proteome Discoverer V2.5 with the following parameters (Figure 4B):
   i. Set acquisition strategy as MS2.
   ii. Set DSSO (158.0037 Da) as cross-linker.
e. Set the following parameters at the XlinkX/PD Search node:
   i. Set same parameters as the Sequest HT node.
   ii. Precursor mass tolerance of 10 ppm.
   iii. FTMS fragment of 20 ppm.
   iv. ITMS fragment of 0.5 Da.
f. Set XlinkX/PD Validator to FDR: 0.05.
g. At the consensus step set a peptide validator node with a target FDR for PSMs and peptides between 0.01 to 0.05 (Figure 4C).
h. Add a Peptide and Protein Filter with the next parameters:
   i. Peptide Confidence At Least: High
   ii. Minimum Number of Peptide Sequences: 1
i. At the XlinkX/PD Consensus Validator set the cross-link spectrum match (CSM) and cross-link FDR threshold as 0.05.
j. Perform manual curation of the identified cross-links:
   i. Verify the quality of the CSMs. The cross-linking, b and y ions should be visible and describe the amino acid sequence of the two peptides identified in the CSM. A clear example can be observed at Garcia-del Rio et al.[1]

Note: As cross-linking technology has been evolving in the last 20 years, a community-wide effort has been done to development of methodological standards which are available for the reader[18–20]

   ii. Eliminate the cross-link spectrum matches that involved N-terminal residues.

⚠ CRITICAL: Verify that the peptides identified at the CSMs correspond to the attributed proteins using the NextProt peptide uniqueness checker tool.

9. Modeling and prediction of interactions between AltProts and RefProts (Figure 5A).
   a. Retrieve the AltProts sequences from OpenProt database.
   b. Generate the 3D models at I-TASSER (Iterative Threading ASSEmbly Refinement).[21]

Note: I-TASSER generates five models with the lowest free energy and highest confidence, and the first model usually has the highest score and better quality. However, lower-ranked models might have better quality. For more information, visit the I-TASSER server website.

   c. Download the Alphafold[22] or PDB[23] 3D structures of the RefProts involved in the cross-links.
   d. For RefProt-AltProt docking ClusPro[24] tool is used, submit the RefProt as a receptor and the AltProt as a ligand at the ClusPro protein-protein docking server. Do not use any restraints in the docking.
   e. After the docking is finished, display all the balanced models, and download them. Additionally, download the coefficients for these models.

Note: If you do not have any prior knowledge of what forces dominate in your complex, use the balanced coefficients models.

   f. Open the complex in YASARA[25] view and identify the number of the atoms involved in the cross-linked complex.
   g. To verify the cross-linking distance, use the command:

```
>DistanceAtomA,AtomB,bound=No
```

h. If the distance corresponds to constrains of DSSO, use the following command to label and join both cross-linked atoms (Figure 5B):

```
>LabelDisAtomA,AtomB,Format=DIS,Height=0.7,Color=Black,X=0.0,Y=0.0,Z=0.0,bound=Yes
```

Note: For DSSO, the distances described in the literature are from 5.3 Å[26] to 30 Å.[27] Other molecular viewers such as Pymol[28] or ChimeraX[29] can also be used.

Cross-linking network analysis

⏲ Timing: 1 week

For the visualization and gene ontology (GO) enrichment of the network obtained by cross-link identification, we recommend the use of Cytoscape.[30] Additional apps have to be downloaded at Cytoscape App Store: STRING,[31] ClueGO,[32] CluePedia[33] and yFiles Layout Algorithms.

⚠ CRITICAL: Before starting your network analysis, we recommend performing the Cytoscape and ClueGo tutorials found on their websites. This will provide you with a general

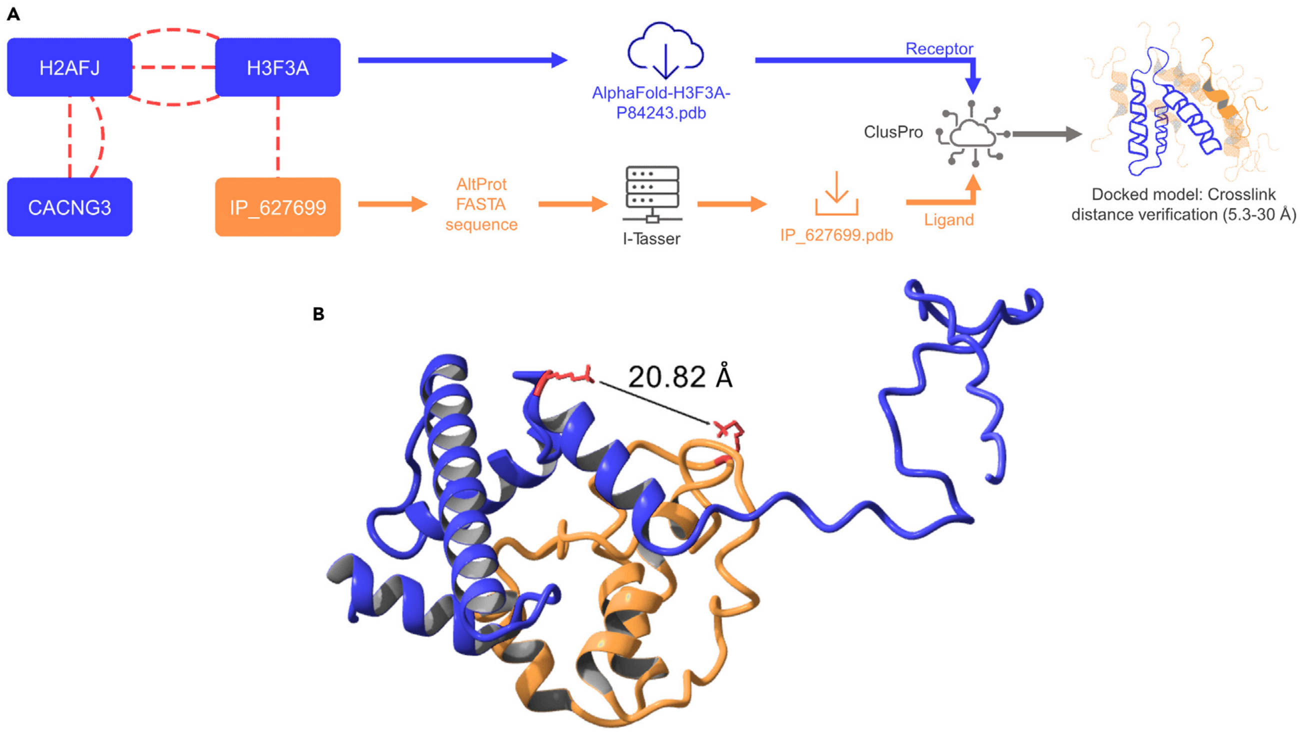
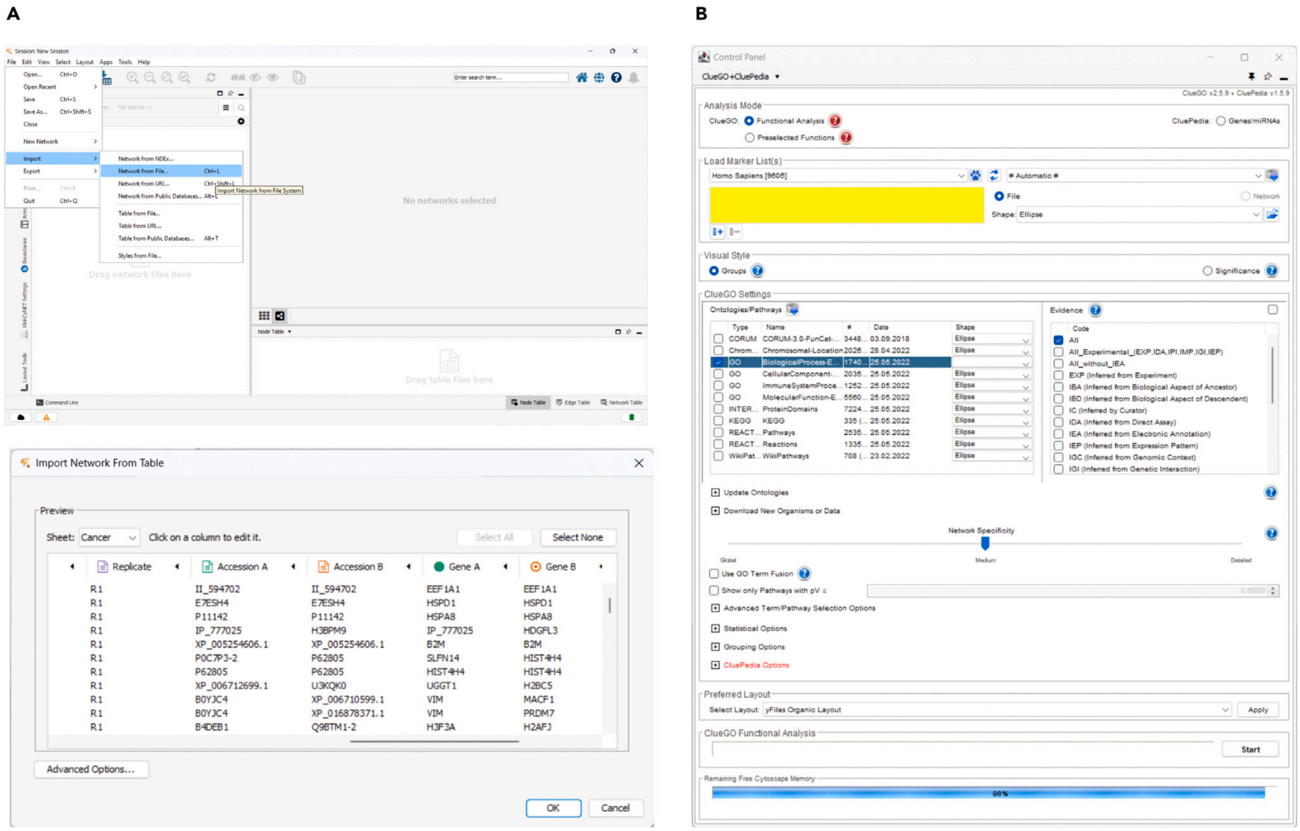


Figure 5. Cross-link interaction modeling workflow and result of the interaction found between H3F3A and the AltProt IP_6276699

(A) Workflow used to generate an interaction model. H3F3A is identified cross-linked to IP_627699 in MS analysis. H3F3A 3D model is obtained from Alphafold databased. IP_627699 FASTA sequence is obtained from OpenProt and is used to generate a 3D model at I-Tasser. Then the two models are docked in ClusPro. Finally, the best interaction model processed, the distances between the cross-linked residues are measured.

(B) 3D ribbon model showing the interaction between H3F3A (blue) and IP_627699 (orange). The cross-linked residues are displayed in red and the distance of the interaction [20.82 Å], confirmed the possible cross-link identification as its fits in the restricted distance of DSSO [5 to 30 Å].



Import Network From Table

Preview

Sheet: Cancer Click on a column to edit it. Select All Select None

Replicate	Accession A	Accession B	Gene A	Gene B
R1	II_594702	II_594702	EEF1A1	EEF1A1
R1	E7ESH4	E7ESH4	HSPD1	HSPD1
R1	P11142	P11142	HSPA8	HSPA8
R1	IP_777025	H3BPM9	IP_777025	HDGFL3
R1	XP_005254606.1	XP_005254606.1	B2M	B2M
R1	P0C7P3-2	P62805	SLFN14	HIST4H4
R1	P62805	P62805	HIST4H4	HIST4H4
R1	XP_006712699.1	U3KQK0	UGGT1	H2BC5
R1	B0YJC4	XP_006710599.1	VIM	MACF1
R1	B0YJC4	XP_016878371.1	VIM	PRDM7
R1	B4DEB1	Q9BTM1-2	H3F3A	H2AFJ

Advanced Options... OK Cancel

Analysis Mode
ClueGO: ● Functional Analysis ○ Preselected Functions CluePedia: ○ Genes/miRNAs

Load Marker List(s)
Homo Sapiens [9606] # Automatic #
● File ○ Network
Shape: Ellipse

Visual Style
● Groups ○ Significance

ClueGO Settings

Ontologies/Pathways

	Type	Name	#	Date	Shape
□	CORUM	CORUM-3.0-FunCat-...	3448...	03.09.2018	Ellipse
□	Chrom...	Chromosomal-Location	2026...	28.04.2022	Ellipse
☑	GO	BiologicalProcess-E...	1740...	25.05.2022	
□	GO	CellularComponent-...	2035...	25.05.2022	Ellipse
□	GO	ImmuneSystemProce...	1252...	25.05.2022	Ellipse
□	GO	MolecularFunction-E...	5560...	25.05.2022	Ellipse
□	INTER...	ProteinDomains	7224...	25.05.2022	Ellipse
□	KEGG	KEGG	335 (...	25.05.2022	Ellipse
□	REACT...	Pathways	2535...	25.05.2022	Ellipse
□	REACT...	Reactions	1335...	25.05.2022	Ellipse
□	WikiPat...	WikiPathways	708 (...	23.02.2022	Ellipse

Evidence □

	Code
☑	All
□	All_Experimental_(EXP,IDA,IPI,IMP,IGI,IEP)
□	All_without_IEA
□	EXP (Inferred from Experiment)
□	IBA (Inferred from Biological Aspect of Ancestor)
□	IBD (Inferred from Biological Aspect of Descendent)
□	IC (Inferred by Curator)
□	IDA (Inferred from Direct Assay)
□	IEA (Inferred from Electronic Annotation)
□	IEP (Inferred from Expression Pattern)
□	IGC (Inferred from Genomic Context)
□	IGI (Inferred from Genetic Interaction)

⊞ Update Ontologies
⊞ Download New Organisms or Data

Network Specificity: Global — Medium — Detailed

□ Use GO Term Fusion
□ Show only Pathways with pV ≤ 0.05000
⊞ Advanced Term/Pathway Selection Options
⊞ Statistical Options
⊞ Grouping Options
⊞ CluePedia Options

Preferred Layout
Select Layout: yFiles Organic Layout Apply

ClueGO Functional Analysis
Start

Remaining Free Cytoscape Memory 98%

Figure 6. Cytoscape and ClueGo interface windows

(A) Displays how to import a cross-linking network from a file of identification by Proteome Discoverer 2.5. Additionally, it shows how to label the columns as target and source nodes during this process.

(B) Presents the control panel of the ClueGo app at Cytoscape. The loading marker square, where to load the query protein list, is highlighted in yellow.

panorama of the commands, formatting options, and different analyses that can be performed in the software.

10. PPIs network treatment.
    a. Export the cross-links identified from Proteome Discoverer 2.5 as an Excel file (Figure 6A).
    b. Open the file and split the column description to obtain the gene symbols of the proteins identified in cross-link.
    c. Write in two new columns (Gene A and Gene B) the gene annotations.
    d. Import this network from the file to Cytoscape.
        i. Assign the source node to the Gene A column and the target node to Gene B column.
    e. Select the network (Figure 7A) and STRINGify it (STRING App).

Note: Verify that all RefProt nodes are now STRING nodes. If not (plain gray nodes at Figure 7B), verify that there are no spaces in the accession numbers or genes. For Cytoscape a space after the gene/accession creates a duplicate non-referenced node.

    f. Once all the RefProts have a STRING node, copy the column of accession numbers, and input them in the STRING protein query. This step will identify the already described interactions between the RefProts (Figure 7C).

Note: If there are RefProt nodes that do not have any interactors, select them one by one and add known interactors in the STRING menu (Figure 7C, bright gray STRING nodes).

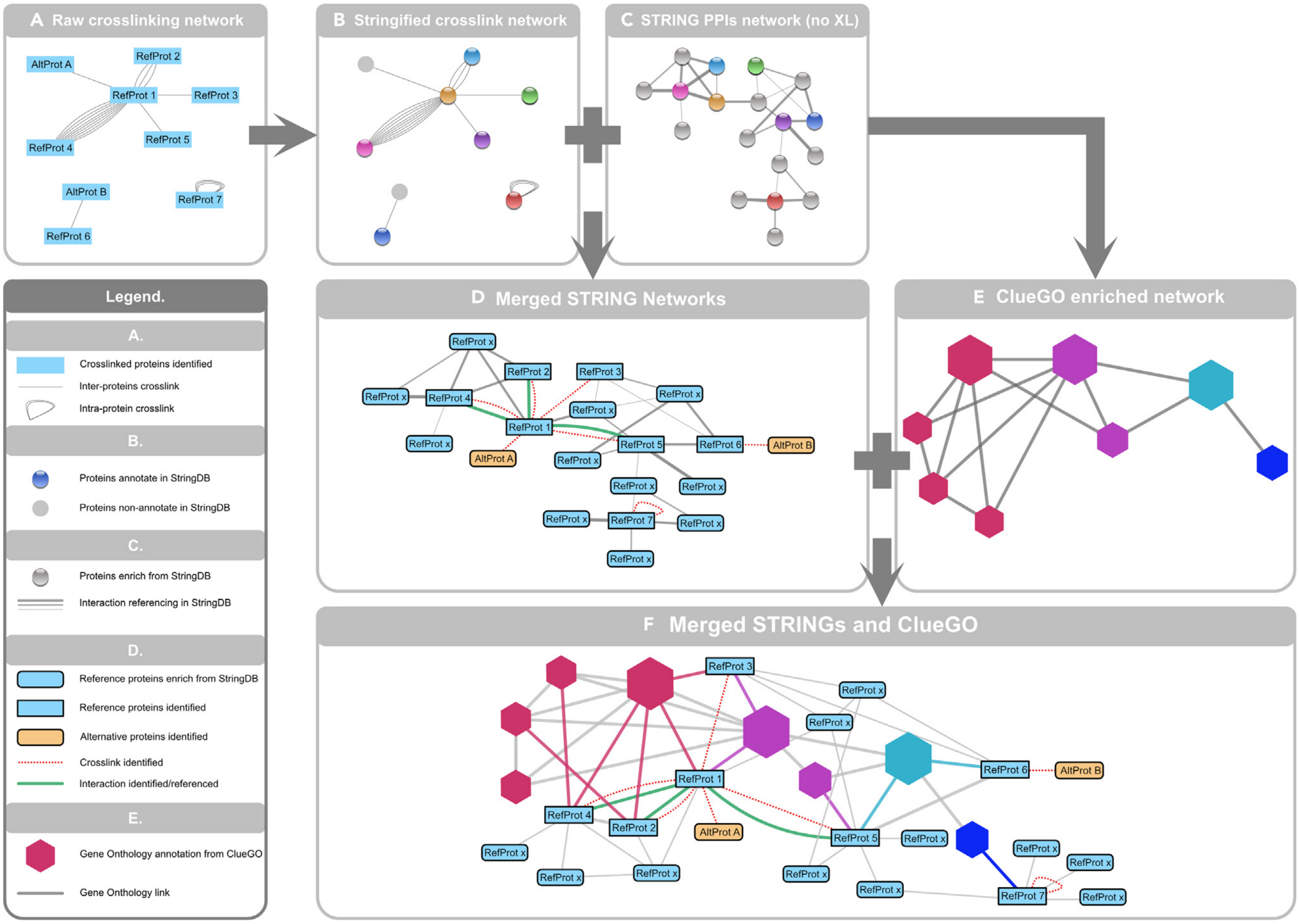

A Raw crosslinking network
AltProt A
RefProt 2
RefProt 1
RefProt 3
RefProt 4
RefProt 5
AltProt B
RefProt 7
RefProt 6
B Stringified crosslink network
C STRING PPIs network (no XL)
Legend.
A.
Crosslinked proteins identified
Inter-proteins crosslink
Intra-protein crosslink
B.
Proteins annotate in StringDB
Proteins non-annotate in StringDB
C.
Proteins enrich from StringDB
Interaction referencing in StringDB
D.
Reference proteins enrich from StringDB
Reference proteins identified
Alternative proteins identified
Crosslink identified
Interaction identified/referenced
E.
Gene Onthology annotation from ClueGO
Gene Onthology link
D Merged STRING Networks
RefProt x
RefProt 2
RefProt 3
RefProt 4
RefProt 1
RefProt 5
RefProt 6
RefProt 7
AltProt A
AltProt B
E ClueGO enriched network
F Merged STRINGs and ClueGO
RefProt 3
RefProt 1
RefProt 4
RefProt 2
AltProt A
RefProt 5
RefProt 6
AltProt B
RefProt 7
RefProt x

Figure 7. Cross-linking network evolution during the processing steps

(A) Shows a raw cross-linking network after import.

(B) Stringified cross-linking network. Note that the AltProts are not in the STRING format, meaning they are not indexed at string database.

(C) Displays only the interaction between the RefProts without cross-links. The enriched nodes are shown in gray STRING nodes.

(D) After merging B and C, all the interactions are displayed. Formatting is done for the nodes and edges.

(E) Presents a ClueGo enriched network. Only GO term nodes are displayed, the protein nodes are kept hidden.

(F) Resulting network after merging D and E. Combined the information of enriched proteins, query RefProts and AltProt, connected to the GOterm. As well as the cross-link identified interactions, StringDB and other databased enriched interaction existing between the proteins of the network.

g. Merge the networks.

Note: The resulting merged network (Figure 7D) will simplify the redundant cross-links. We recommend formatting this network to visualize the interactions of interest.

h. Cross-validate the RefProt interactions in other databases. We recommend BioGrid[34] and IntAct.[35]

i. GO term enrichment (Figure 6B).
   i. Open ClueGo App.
   ii. Select the functional analysis as analysis mode.
   iii. Input the whole list of RefProt genes at the load marker list box.
   iv. Select the ontologies/pathways to use for the enrichment.
   v. Select the specificity of the network.
   vi. Enable the GO term fusion.
   vii. Run the app.

j. Merge the ClueGO network (Figure 7E) and the merged STRING network (Figure 7D).
k. At the resulting network (Figure 7F) select the preferred network layout and edit the node's properties. We recommend formatting the network in a way that you can obtain the information you need.

EXPECTED OUTCOMES

This methodology enables the identification of AltProts, their subcellular localization, and protein-protein interactions (PPIs) in cell lines. An easy way to verify that a cross-linking reaction has occurred is by using SDS-PAGE. After a cross-linking reaction, larger protein complexes will be present in the sample. These larger complexes cannot enter the concentration gel and will be observable at the top of the wells and some at the interface between concentration and migration gels. Additionally, the disappearance of some protein bands in the cross-linked sample, compared to a negative cross-linking control, can indicate that a cross-linking reaction has taken place (Figure 2A). To validate the subcellular protein fractionation, we recommend using western blotting and compartment-specific antibodies. The optimal outcome should be that the signal of the antibody is present in just one fraction, but traces of specific markers may appear in other fractions (Figure 2B).

For AltProts identification, an AltProt/RefProt ratio of 5%–10% can be expected. It is always recommended to validate identifications by performing a BLAST search and using the NextProt peptide uniqueness checker. Additionally, different properties and characteristics of the AltProts identified can be retrieved from the OpenProt database. Although peptide fractionation using SEC[36] and SCX[37] columns can increase the number of cross-links identified, our fast and non-fractionated method can only be expected to identify a couple of hundred interactions at most. Even a small number of AltProt-RefProt cross-links is sufficient to infer their possible function or pathway involvement. This methodology was exemplified in a study of PPIs in immortalized human ovarian epithelial cells (SV40), and the characterization of the AltProts identified in the cell line.[1] In this study, the subcellular localization of 112 AltProts was observed, and subcellular protein fractionation decreased the complexity of the cross-linked sample, allowing us to identify a network of 220 cross-links without peptide enrichment, 16 of which were AltProt-RefProt interactions. Furthermore, the possible involvement of these AltProts in some cellular processes, such as antigen processing and presentation of peptide antigen via MHC class I, mRNA transcription by RNA polymerase II, and regulation of mitochondrial outer membrane permeabilization involved in apoptotic signaling pathway, was investigated.

LIMITATIONS

The methodology described above has some limitations. The first limitation (I) is related to the lack of information on AltProts. Still, the concept of alternative proteins has limited spread, and few tools and databases have been developed for the analysis of this ghost proteome. Therefore, the methodology used here provides us with a snapshot of some possible functions of a limited number of AltProts. Targeted studies need to be conducted to confirm or expand the information about these AltProt "hits".

The second limitation (II) is linked to the protein fractionation technique. The kit employed in this protocol is based on pelleting the non-extracted fraction and removing the supernatant. If the removal of the supernatant is not optimal, traces of proteins that don't correspond to the next fraction could remain. Additionally, there is limited information about the buffer composition and detergents employed. To avoid the use of the kit, other subcellular fractionation techniques can be used (e.g., sucrose gradient), since they are compatible with MS.

The third limitation (III) is the need to decrease the complexity of the cross-linking sample before injection into the nLC-MS/MS system, increasing the detection of cross-linked peptides. For this, fractionation techniques like SEC and SCX chromatography or the use of enrichable cross-linkers like NNP9,[38] tBu-PhoX,[39] and alkyne-A-DSBSO[40] can help identify more cross-linked peptides.

TROUBLESHOOTING

Problem 1

No cross-linking patterns are observed in the SDS-PAGE gels (related to Step 1).

Potential solution

- Verify the BSA positive control. If there is no cross-link in BSA positive control, it means DSSO was hydrolyzed.
- Avoid buffers that contain primary amines.
- Repeat the cross-linking reaction with a new vial of DSSO. Follow the vendor's storage and use recommendations.

Problem 2

No extraction is observed in the SDS-PAGE gels after subcellular fractionation (related to Step 2).

Potential solution

- Verify that the volumes used for the extraction are appropriate (read the vendor's manual).
- Remove DPBS completely before starting and keep the pellet as dry as possible (according to the kit manufacturer).
- Increase the incubation times.
- Vortex at the highest setting.
- Add the appropriate volume of Halt Protease Inhibitor Cocktail.

Problem 3

The extracted proteins are not compartmentalized (related to Step 2).

Potential solution

- Vortex longer to disperse completely the cell pellets.
- Increase the incubation times.
- Carefully remove all extracts before proceeding to the next step. Remove the remaining buffer with a smaller pipette.
- Re-centrifuge sample and remove excess extract.
- Primary antibodies are not specific.
- Verify if the protein selected as compartment-specific is reported in literature to be in other compartments.

Problem 4

Urea is not dissolved at the Denaturing solution (related to Step 3).

Potential solution

- Place the solution in an ultrasonic bath and sonicate it for 5–10 min.
- Freeze and thaw the buffer to solubilize the urea.

Problem 5

Presence of polymer traces in the membrane fractions (related to Step 4).

Potential solution

- Verify by MALDI-MS that all the solutions prepared are not contaminated by polymers.
- Repeat the HiPPR™ Detergent Removal Resin protocol and dilute the sample.

Problem 6

No cross-links identified by XlinkX at Proteome Discoverer (related to Step 5).

Potential solution

- If the confirmation of the cross-linking reaction by SDS-PAGE was skipped, there is no certainty that the cross-linking reaction happened.
- If only cross-link dead ends are found in the sample, perform peptide fractionation (SEC or SCX) of the samples after the digestion.

Problem 7

Unable to find the lysine residue involved in the cross-link after modeling the interaction in ClusPro (related to Step 5).

Potential solution

- The cross-linking description found in Proteome Discoverer is based on the sequence found in the FASTA file. However, some PDB accessions do not present all the amino acids described in the protein databases. Therefore, we recommend finding the 3D structures in which the residues involved are present. In UniProt, under the structure menu, we can observe the different 3D models and coverage for each protein.
- Please note that after docking in ClusPro, the model numbering starts with the receptor, followed by the ligand protein. To find the position of the ligand residue involved in the cross-link, simply add the residues of the receptor to the position of the cross-link.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Michel Salzet (michel.salzet@univ-lille.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The mass spectrometry proteomics data from Garcia-del Rio et al.[1] have been deposited to the ProteomeXchange Consortium via the PRIDE[41] partner repository with the dataset identifier PXD035764, study following this protocol.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This research was supported by funding from I-SITE of ‘‘Université Lille Nord-Europe’’ (ULNE), the European metropole of Lille (MEL), the Regional Council of ‘‘Hauts de France’’, the ‘‘Institut National de la Santé et de la Recherche Médicale’’ (Inserm), and Université de Lille.

AUTHOR CONTRIBUTIONS

Conceptualization, M.S., T.C.; Methodology, D.F.G.-d.R., T.C.; Software, D.F.G.-d.R., T.C.; Formal Analysis, D.F.G.-d.R.; Resources, I.F., M.S.; Data curation, D.F.G.-d.R., T.C.; Writing - Original Draft, D.F.G.-d.R., T.C.; Writing - Review & Editing, I.F., M.S.; Supervision, T.C., M.S., I.F.; Project Administration, M.S., I.F.; Funding Acquisition, M.S., I.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Garcia-del Rio, D.F., Cardon, T., Eyckerman, S., Fournier, I., Bonnefond, A., Gevaert, K., and Salzet, M. (2023). Employing non-targeted interactomics approach and subcellular fractionation to increase our understanding of the ghost proteome. iScience *26*, 105943. https://doi.org/10.1016/j.isci.2023.105943.

2. Cardon, T., Salzet, M., Franck, J., and Fournier, I. (2019). Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation. Biochim. Biophys. Acta. Gen. Subj. *1863*, 1458–1470. https://doi.org/10.1016/j.bbagen.2019.05.009.

3. Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. Nat. Chem. Biol. *16*, 458–468. https://doi.org/10.1038/s41589-019-0425-0.

4. Brunet, M.A., Brunelle, M., Lucier, J.-F., Delcourt, V., Levesque, M., Grenier, F., Samandi, S., Leblanc, S., Aguilar, J.-D., Dufour, P., et al. (2019). OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res. *47*, D403–D410. https://doi.org/10.1093/nar/gky936.

5. Brunet, M.A., Lucier, J.-F., Levesque, M., Leblanc, S., Jacques, J.-F., Al-Saedi, H.R.H., Guilloy, N., Grenier, F., Avino, M., Fournier, I., et al. (2021). OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. Nucleic Acids Res. *49*, D380–D388. https://doi.org/10.1093/nar/gkaa1036.

6. Aboulouard, S., Wisztorski, M., Duhamel, M., Saudemont, P., Cardon, T., Narducci, F., Lemaire, A.-S., Kobeissy, F., Leblanc, E., Fournier, I., and Salzet, M. (2021). In-depth proteomics analysis of sentinel lymph nodes from individuals with endometrial cancer. Cell Rep. Med. *2*, 100318. https://doi.org/10.1016/j.xcrm.2021.100318.

7. Hajjaji, N., Aboulouard, S., Cardon, T., Bertin, D., Robin, Y.-M., Fournier, I., and Salzet, M. (2021). Path to clonal theranostics in luminal breast cancers. Front. Oncol. *11*, 802177. https://doi.org/10.3389/fonc.2021.802177.

8. Le Rhun, E., Duhamel, M., Wisztorski, M., Gimeno, J.-P., Zairi, F., Escande, F., Reyns, N., Kobeissy, F., Maurage, C.-A., Salzet, M., and Fournier, I. (2017). Evaluation of non-supervised MALDI mass spectrometry imaging combined with microproteomics for glioma grade III classification. Biochim. Biophys. Acta. Proteins Proteom. *1865*, 875–890. https://doi.org/10.1016/j.bbapap.2016.11.012.

9. Cardon, T., Fournier, I., and Salzet, M. (2021). Shedding light on the ghost proteome. Trends Biochem. Sci. *46*, 239–250. https://doi.org/10.1016/j.tibs.2020.10.003.

10. Leblanc, S., and Brunet, M.A. (2020). Modelling of pathogen-host systems using deeper ORF annotations and transcriptomics to inform proteomics analyses. Comput. Struct. Biotechnol. J. *18*, 2836–2850. https://doi.org/10.1016/j.csbj.2020.10.010.

11. Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362. https://doi.org/10.1038/nmeth.1322.

12. Tabb, D.L., Eng, J.K., and Yates, J.R. (2001). Protein identification by SEQUEST. In Proteome Research: Mass Spectrometry Principles and Practice, P. James, ed. (Springer), pp. 125–142. https://doi.org/10.1007/978-3-642-56895-4_7.

13. McGinnis, S., and Madden, T.L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Res. *32*, W20–W25. https://doi.org/10.1093/nar/gkh435.

14. Schaeffer, M., Gateau, A., Teixeira, D., Michel, P.-A., Zahn-Zabal, M., and Lane, L. (2017). The neXtProt peptide uniqueness checker: a tool for the proteomics community. Bioinforma. Oxf. Engl. *33*, 3471–3472. https://doi.org/10.1093/bioinformatics/btx318.

15. Liu, F., van Breukelen, B., and Heck, A.J.R. (2014). Facilitating protein disulfide mapping by a combination of pepsin digestion, electron transfer higher energy dissociation (EThcD), and a dedicated search algorithm SlinkS. Mol. Cell. Proteomics *13*, 2776–2786. https://doi.org/10.1074/mcp.O114.039057.

16. Liu, F., Rijkers, D.T.S., Post, H., and Heck, A.J.R. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat. Methods *12*, 1179–1184. https://doi.org/10.1038/nmeth.3603.

17. Liu, F., Lössl, P., Scheltema, R., Viner, R., and Heck, A.J.R. (2017). Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. Nat. Commun. *8*, 15473. https://doi.org/10.1038/ncomms15473.

18. Leitner, A., Bonvin, A.M.J.J., Borchers, C.H., Chalkley, R.J., Chamot-Rooke, J., Combe, C.W., Cox, J., Dong, M.-Q., Fischer, L., Götze, M., et al. (2020). Toward increased reliability, transparency, and accessibility in cross-linking mass spectrometry. Structure *28*, 1259–1268. https://doi.org/10.1016/j.str.2020.09.011.

19. Iacobucci, C., Piotrowski, C., Aebersold, R., Amaral, B.C., Andrews, P., Bernfur, K., Borchers, C., Brodie, N.I., Bruce, J.E., Cao, Y., et al. (2019). First community-wide, comparative cross-linking mass spectrometry study. Anal. Chem. *91*, 6953–6961. https://doi.org/10.1021/acs.analchem.9b00658.

20. Piersimoni, L., Kastritis, P.L., Arlt, C., and Sinz, A. (2022). Cross-linking mass spectrometry for investigating protein conformations and protein–protein Interactions—a method for all seasons. Chem. Rev. *122*, 7500–7531. https://doi.org/10.1021/acs.chemrev.1c00786.

21. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat. Methods *12*, 7–8. https://doi.org/10.1038/nmeth.3213.

22. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

23. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The protein data bank. Nucleic Acids Res. *28*, 235–242. https://doi.org/10.1093/nar/28.1.235.

24. Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D., Yueh, C., Beglov, D., and Vajda, S. (2017). The ClusPro web server for protein–protein docking. Nat. Protoc. *12*, 255–278. https://doi.org/10.1038/nprot.2016.169.

25. Land, H., and Humble, M.S. (2018). YASARA: a tool to obtain structural guidance in biocatalytic investigations. Methods Mol. Biol. *1685*, 43–67. https://doi.org/10.1007/978-1-4939-7366-8_4.

26. Kao, A., Chiu, C.I., Vellucci, D., Yang, Y., Patel, V.R., Guan, S., Randall, A., Baldi, P., Rychnovsky, S.D., and Huang, L. (2011). Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. Mol. Cell. Proteomics *10*. https://doi.org/10.1074/mcp.M110.002212.

27. Hevler, J.F., Lukassen, M.V., Cabrera-Orefice, A., Arnold, S., Pronker, M.F., Franc, V., and Heck, A.J.R. (2021). Selective cross-linking of coinciding protein assemblies by in-gel cross-linking mass spectrometry. EMBO J. *40*,

e106174. https://doi.org/10.15252/embj.2020106174.

28. Schrödinger, LLC (2015). The PyMOL Molecular Graphics System, Version 1.8.

29. Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., and Ferrin, T.E. (2021). UCSF ChimeraX: structure visualization for researchers, educators, and developers. Protein Sci. *30*, 70–82. https://doi.org/10.1002/pro.3943.

30. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

31. Doncheva, N.T., Morris, J.H., Gorodkin, J., and Jensen, L.J. (2019). Cytoscape StringApp: network analysis and visualization of proteomics data. J. Proteome Res. *18*, 623–632. https://doi.org/10.1021/acs.jproteome.8b00702.

32. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093. https://doi.org/10.1093/bioinformatics/btp101.

33. Bindea, G., Galon, J., and Mlecnik, B. (2013). CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. Bioinforma. Oxf. Engl. *29*, 661–663. https://doi.org/10.1093/bioinformatics/btt019.

34. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. *30*, 187–200. https://doi.org/10.1002/pro.3978.

35. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358–D363. https://doi.org/10.1093/nar/gkt1115.

36. Mendes, M.L., Fischer, L., Chen, Z.A., Barbon, M., O'Reilly, F.J., Giese, S.H., Bohlke-Schneider, M., Belsom, A., Dau, T., Combe, C.W., et al. (2019). An integrated workflow for crosslinking mass spectrometry. Mol. Syst. Biol. *15*, e8994. https://doi.org/10.15252/msb.20198994.

37. Fritzsche, R., Ihling, C.H., Götze, M., and Sinz, A. (2012). Optimizing the enrichment of cross-linked products for mass spectrometric protein analysis. Rapid Commun. Mass Spectrom. *26*, 653–658. https://doi.org/10.1002/rcm.6150.

38. Nury, C., Redeker, V., Dautrey, S., Romieu, A., van der Rest, G., Renard, P.-Y., Melki, R., and Chamot-Rooke, J. (2015). A novel bio-orthogonal cross-linker for improved protein/protein interaction analysis. Anal. Chem. *87*, 1853–1860. https://doi.org/10.1021/ac503892c.

39. Jiang, P.-L., Wang, C., Diehl, A., Viner, R., Etienne, C., Nandhikonda, P., Foster, L., Bomgarden, R.D., and Liu, F. (2022). A membrane-permeable and immobilized metal affinity chromatography (IMAC) enrichable cross-linking reagent to advance in vivo cross-linking mass spectrometry. Angew. Chem. Int. Ed. *61*, e202113937. https://doi.org/10.1002/anie.202113937.

40. Burke, A.M., Kandur, W., Novitsky, E.J., Kaake, R.M., Yu, C., Kao, A., Vellucci, D., Huang, L., and Rychnovsky, S.D. (2015). Synthesis of two new enrichable and MS-cleavable cross-linkers to define protein–protein interactions by mass spectrometry. Org. Biomol. Chem. *13*, 5030–5037. https://doi.org/10.1039/C5OB00488H.

41. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D.J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., et al. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. *50*, D543–D552. https://doi.org/10.1093/nar/gkab1038.

Conclusion

AltProts, derived from non-canonical open reading frames, are an important and often overlooked dimension of the proteome. AltProts can play a vital role in key cellular processes and serve as valuable biomarkers for diseases. Despite their potential, investigating AltProts has been incredibly challenging. This is due to a lack of antibody reagents, limited representation in protein sequence databases and a lack of functional information. However, the here introduced protocol describes a new workflow that combines XL-MS with subcellular fractionation. To improve coverage of the alternative proteome, the sample complexity was reduced by fractionating cells into cytoplasmic, membrane, nuclear, chromatin-bound and cytoskeletal proteomes. It allows researchers to illuminate AltProt networks and interactions in human cells in a high-throughput and unbiased manner. With this approach, investigating AltProts will no longer be a challenge but rather an opportunity to gain new insights into the proteome.

Performing *in-cellulo* crosslinking with DSSO accurately captures interactions between AltProts and endogenous proteins under native conditions. This protocol provides a major advantage by allowing for the identification of AltProt subcellular localization and protein interactors without the need for AltProt-specific antibodies. Moreover, confident identification of AltProts and mapping to resident protein binding partners is facilitated by AltProt-specific databases such as OpenProt and advanced algorithms to identify crosslinked peptides.

The workflow was applied to immortalized human ovarian epithelial cells (T1074). Besides identifying AltProt-RefProt interactions, we also performed a GO enrichment analysis on the crosslinked network and mapped AltProts with their interactors to learn more about the biological processes and pathways AltProts are part of. By inputting the list of crosslinked proteins into ClueGO[285], an app within Cytoscape[284], it is possible to visualize networks enriched in proteins linked to specific GO terms. This refines the complex interaction data down to key biological themes.

An important component of our workflow is the use of structural modeling and molecular docking to validate the identified AltProt-RefProt interactions. Since AltProts lack characterization, predicting their structure and docking to their binding partners provides crucial information on the feasibility of the crosslinked complex. We employed the Iterative Threading ASSEmbly Refinement (I-TASSER)[340] to generate structural models of AltProts by threading their sequences onto homologous proteins of known structure. For partner proteins, existing structures were obtained from PDB or predicted by AlphaFold[341] . The AltProt and protein structures were then docked using ClusPro[342] which outputs the most energetically favorable complexes. By measuring the distance between crosslinked residues in the modeled AltProt-protein complex, we validated crosslinks that fall within the expected distance constraints for DSSO. This provides independent support for the interaction and builds confidence in the identification. While modeling and docking have limitations in accurately predicting protein structures and interactions, they offer useful complementary data to strengthen cross-linking mass spectrometry studies on novel proteins like AltProts.

The simplicity and robustness of our XL-MS protocol overcomes long-standing challenges in AltProt research. While extensive follow-up studies remain needed to confirm the functions of identified AltProts, this method yields crucial foundational insights into both localization and binding partners to guide downstream investigations. The localization patterns and protein interactions can inform antibody generation, targeted validations, pharmacological modulation, and elucidation of signaling mechanisms.

In summary, our XL-MS workflow provides an unbiased and much-needed strategy to illuminate the “dark matter” of the proteome. The interactomics view enables initial integration of AltProts into known biological systems, advancing our understanding of these unknown and yet influential molecules. This protocol exemplifies the power of emerging proteomics technologies to unravel uncharted fractions of biology and push the boundaries of proteome coverage.

PART IV NON-TARGETED INTERACTOMICS APPROACH AND SUBCELLULAR FRACTIONATION TO INCREASE OUR UNDERSTANDING OF THE GHOST PROTEOME

As described in Chapter III, we developed a protocol for identifying, characterizing and locating AltProts. To identify physiologically relevant AltProt binding partners, we employed XL-MS. We conducted this protocol in an immortalized human ovarian epithelial cell line, T1074. This cell line is a cuboidal, adherent cell line that was immortalized by serial passage and transduction with recombinant lentiviruses carrying SV40 Large T antigen. It has been used as a cellular model in several studies on OvCa[343–345]. Therefore, we chose to apply our developed methodology to this cell line to assess the function of AltProts under physiological conditions.

Prediction of protein function

One challenge of biology is determining the functions of newly identified proteins. Answering several questions can help here. For example, where is the protein located? What are its targets? In which pathways is the protein involved? And in which cells or tissues can it be found?

One approach to gain insight into the function of a protein is to examine its sequence similarity. This involves comparing the sequence of the novel protein to those of proteins previously characterized and stored in databases. In principle, the amino acid sequence determines the protein structure, and this structure steers its biochemical function. Therefore, proteins with similar amino acid sequences tend to possess similar biochemical functions, even if the protein is from another organism (functional orthologs).

Searching for homologous proteins can be performed using sequence alignment software such as BLAST[264]. This tool scans the database to identify similar sequences and performs statistical calculations to determine the degree of similarity. Another approach is to search for specific motifs or protein domains within protein sequences. Often, domains are associated with a specific function and can be used to infer the protein function. InterProScan[346] is a tool that integrates multiple domain and protein family resources to aid the functional analysis of novel proteins based on conserved signatures and motifs. A deep learning algorithm, DeepGOPlus, has been developed to combine similarity-based searches and motif-based function prediction[347]. As a result, using such algorithms, a protein function can be predicted from its sequence alone.

Expression analysis is a valuable method for characterizing protein function. It involves studying the patterns and levels of protein expression by calculating correlation coefficients between the expression pattern of the uncharacterized protein and genes of known function across multiple conditions. A high positive correlation suggests that the novel protein may be co-regulated and participate in similar pathways or processes as the correlated genes[348,349].

Functional assays are other approaches to identify the function of a novel protein. For instance, enzyme assays directly test if a purified recombinant protein has biochemical activities like kinase activity, DNA or histone binding, etc.[350,351]. Another type of functional assay is rescue experiments, which involve expressing the novel protein in a knockout model organism (or cell line) where it has been deleted. If the protein can rescue and restore the normal phenotype, pathway, or metabolic process, it provides strong evidence for its endogenous function[352].

As described in PART I, interactomics identify PPIs. Once an interaction is identified and validated, a functional enrichment can be performed to place this interaction in a metabolic or biological process pathway. One tool that can be used for this goal is GO enrichment analysis, which utilizes the Gene Ontology system of classification to aid the interpretation of high-throughput gene sets. This system aims to homogenize vocabulary, assigning GO terms to molecular function, biological process and cellular component. The annotation system has a hierarchical relationship, allowing annotations to be made at different levels of specificity. This analysis aims to identify whether certain categories or terms are overrepresented in a given set of genes compared to what would be expected by chance. It involves calculating enrichment scores or p-values to determine the significance of the observed enrichment. Various statistical tests, such as Fisher's exact test or a hypergeometric test, can be used for this purpose. Some popular packages for performing GO enrichment analysis include database for annotation, visualization and integrated discovery (DAVID)[353], protein analysis through evolutionary relationship (PANTHER)[354], gene set enrichment analysis (GSEA)[355] and ClueGO[285].

Major histocompatibility complex (MHC) class I

In the past years, more specific tools have been developed to predict the function or certain characteristics of peptides and proteins, depending on the research focus and needs. One such tool is NetMHC, developed by DTU Health Tech[356]. This tool uses gapped sequence alignment to predict major histocompatibility complex (MHC) class I peptide binding affinity.

MHC class I molecules are responsible for binding peptides derived from intracellular proteins and presenting them to cytotoxic CD8$^+$ T cells. They consist of a heavy α chain and a light β2 microglobulin (β2M) chain (**Figure 11**). The α chain folds to create the peptide binding pocket, where intracellular peptides can bind. There are three major variants of MHC class I molecules: HLA-A, HLA-B and HLA-C. Each variant possesses different physicochemical properties that allow for selective binding to peptides of different lengths (8-10) and amino acid composition.

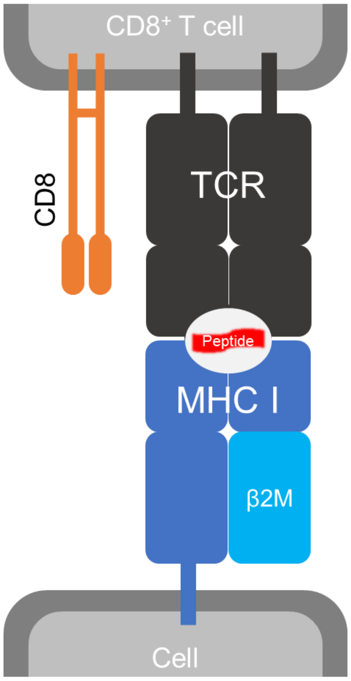


Figure 11. MHC class I complex presenting a peptide to a CD8$^+$ TCR. *Peptides presented via the MHC class I can be recognized by the TCR receptor and CD8 which is expressed in cytotoxic T cells.*

Class I molecules fold and assemble with β2M in the endoplasmic reticulum, and this dimer is then incorporated into the peptide-loading complex. In this complex, proteolysisderived peptides are loaded into the peptide binding pocket of the MHC class I molecule, a process catalyzed by tapasin and chaperone proteins. Once the peptides are bound to the complex, they travel to the cell surface, where $CD8^+$ T cells recognize them with their surface T cell receptors (TCR)[357]. This presentation event leads to T cell activation, triggering an immune response if the presented peptides have a pathogenic origin.

Histone role in gene regulation

Histones are proteins that provide structural support for chromosomes and play a role in the regulation of gene expression. DNA is packaged and wrapped around these proteins (**Figure 12**; left), and regulate gene expression, and can either promote or repress transcription. In eukaryotes most biological processes involved in the manipulation and expression of DNA rely on histone modifications[358].

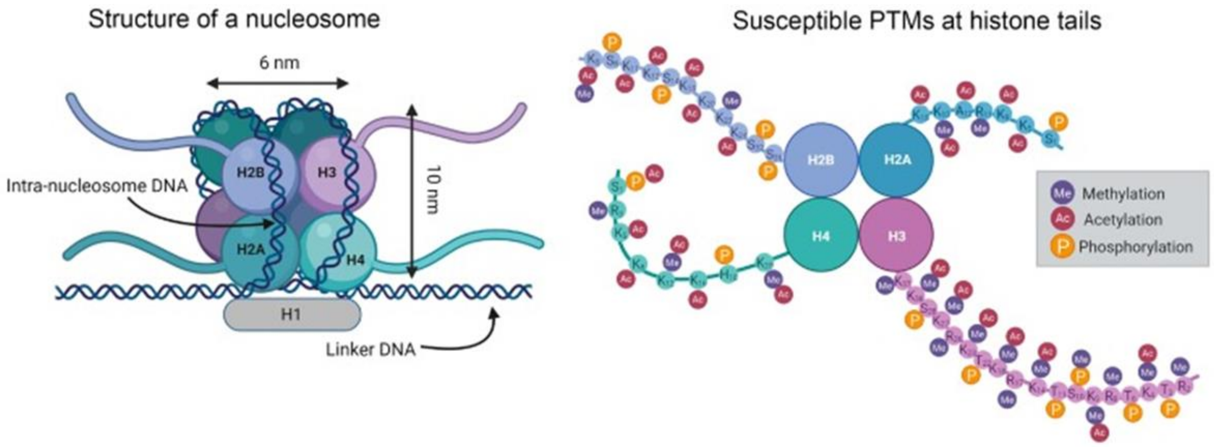


Figure 12. Histone arrangement at the nucleosome (left). *DNA is wrapped around an array of different histone proteins in a structure called nucleosome.* ***Possible PTMs at histone tails (right).*** *Histone tails undergo PTMs that play a role in DNA accessibility to transcription. Obtained and adapted from Torres-Perez et al.[359].*

Acetylation is a common histone modification that loosens the chromatin structure, allowing for transcription (**Figure 12**; right). Histone acetyltransferases (HATs) are responsible for acetylating lysine residues. They are often found in transcriptional coactivator complexes that are recruited to target genes. On the other hand, histone deacetylases (HDACs) remove acetyl groups, thereby condensing chromatin and repressing transcription. The balance of HAT and HDAC activities maintains acetylation levels.

Methylation is another important histone modification that mainly occurs on lysines and arginines (**Figure 12**; right). Lysines can be mono-, di-, or tri-methylated. For instance, trimethylation of H3K4 (H3K4me3) stimulates transcription by recruiting chromatin remodelers and HATs. In contrast, H3K9me3 accumulates in heterochromatin and silences genes by compacting chromatin[359].

TATA-binding protein associated factors (TAFs) connect histone modifications to transcription initiation by RNA polymerase II. They are subunits of the general transcription factor TFIID, which recognizes TATA box promoter elements and nucleates the pre-initiation complex[360]. Multiple TAFs bind to acetylated and methylated histones. For instance, TAF1 binds to acetylated H3K14 and H4K12, marking open chromatin regions for pre-initiation complex assembly. Additionally, TAF3 contains a zinc finger that likely recognizes H3K4me3 at active promoters. TAF3 recruitment stimulates histone acetylation via associated HATs. The histone-binding ability of TAFs helps to position the pre-initiation complex at sites where chromatin is in an active state, marked by modifications like H3K4me3.

As mentioned in Part I, some AltProts localize to the nucleus and regulate gene expression[124,128–130]. They represent a relatively unexplored class of proteins with big potential to influence many cellular processes. Further elucidating their mechanisms will be important to understanding gene regulatory networks.

Objective

The general goal of this project is to use the protocol described in chapter III as an exploratory framework for systematically characterizing the interactions, localization and functions of the alternative proteome. This approach could reveal new signaling and regulatory molecules with roles in cellular physiology and mechanisms.

The first objective is to develop an unbiased, non-targeted strategy to explore the functions and interactions of AltProts on a large scale. Subcellular fractionation will provide insights into the localization of AltProts within subcellular compartments, while also reducing sample complexity for analysis. By employing crosslinking-mass spectrometry, we aim to identify AltProt interaction partners and networks without typical enrichment steps, using only cell fractionation. Based on the interacting partners and localization, AltProt associations with processes or signaling pathways can be identified using GO enrichment algorithms.

To thoroughly evaluate the feasibility of AltProt-protein interactions, we will employ structural modeling. The resulting models will not only provide additional support for the feasibility of the identified interactions but will also give us confidence in the crosslinking results, especially in the absence of other target-directed validation approaches. Structural modeling will give us a more comprehensive understanding of the molecular interactions happening between AltProt and proteins, giving us a clearer view of the potential implications of these interactions.

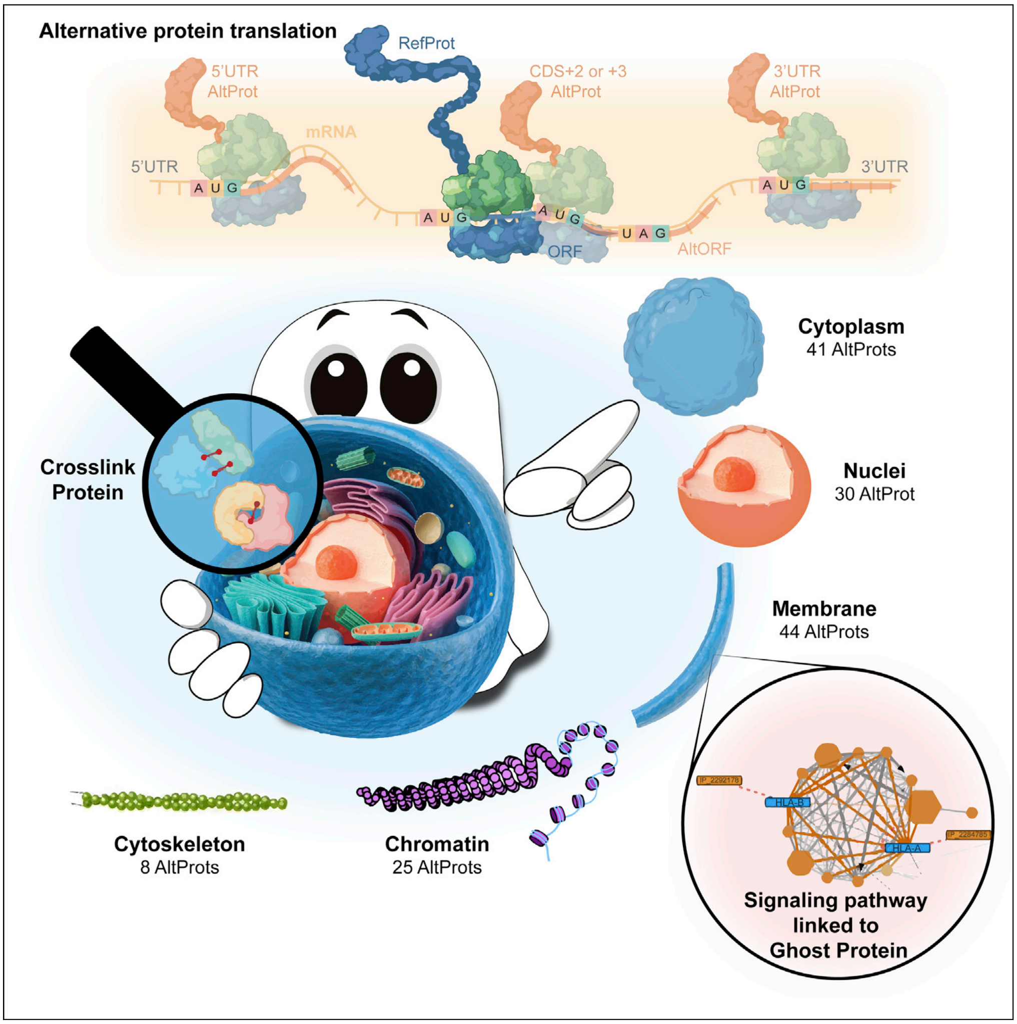
Our proposed workflow is a comprehensive approach that incorporates XL-MS, molecular modeling, docking and GO enrichment. This integrated methodology will enable us to gain a deeper understanding of the "hidden" proteome of epithelial ovarian cells, which has remained elusive until now.

iScience

CellPress
OPEN ACCESS

Article

Employing non-targeted interactomics approach and subcellular fractionation to increase our understanding of the ghost proteome



Diego Fernando Garcia-del Rio, Tristan Cardon, Sven Eyckerman, Isabelle Fournier, Amelie Bonnefond, Kris Gevaert, Michel Salzet

amelie.bonnefond@univ-lille.fr (A.B.)
kris.gevaert@ugent.be (K.G.)
michel.salzet@univ-lille.fr (M.S.)

Highlights

The ghost proteome has been neglected and its role is still unknown

Subcellular localization of 112 alternative proteins was described

220 crosslinks were identified involving 16 alternative proteins

Using this method, we identified the possible physiological role of these proteins

Garcia-del Rio et al., iScience *26*, 105943
February 17, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.105943

CellPress
OPEN ACCESS

Article

Employing non-targeted interactomics approach and subcellular fractionation to increase our understanding of the ghost proteome

Diego Fernando Garcia-del Rio,[1,3,4,5] Tristan Cardon,[1,5] Sven Eyckerman,[3,4] Isabelle Fournier,[1] Amelie Bonnefond,[2,*] Kris Gevaert,[3,4,*] and Michel Salzet[1,6,*]

SUMMARY

Eukaryotic mRNA has long been considered monocistronic, but nowadays, alternative proteins (AltProts) challenge this tenet. The alternative or ghost proteome has largely been neglected and the involvement of AltProts in biological processes. Here, we used subcellular fractionation to increase the information about AltProts and facilitate the detection of protein-protein interactions by the identification of crosslinked peptides. In total, 112 unique AltProts were identified, and we were able to identify 220 crosslinks without peptide enrichment. Among these, 16 crosslinks between AltProts and Referenced Proteins (RefProts) were identified. We further focused on specific examples such as the interaction between IP_2292176 (AltFAM227B) and HLA-B, in which this protein could be a potential new immunopeptide, and the interactions between HIST1H4F and several AltProts which can play a role in mRNA transcription. Thanks to the study of the interactome and the localization of AltProts, we can reveal more of the importance of the ghost proteome.

INTRODUCTION

Since 2011 considerable efforts have been made to shed light on unreferenced proteins also called the ghost proteome; in various biological contexts.[1–5] This ghost proteome, being a part of the total protein landscape, points to proteins not referenced in conventional databases like UniProt[6] and RefSeq.[7] Such ghost proteins, called alternative proteins (AltProts) or proteins coded by small open reading frames (smORFs),[8] were identified to be translated from regions of mRNA molecules described as non-coding, e.g. 3′ and 5′ UTR, reading frame shifts[3] or involve all kinds of non-coding RNA (ncRNA)[9] (Figure 1). AltProts have the particularity of having an average size of less than 100 amino acids,[10] likewise their sequences, despite being derived from a mRNA coding for a referenced protein (RefProt), have a completely different amino acid sequence, suggesting a different biological function. AltProts are estimated at 450,000 potential sequences,[3,11] compared to 79,038 RefProt sequences (Uniprot-01.2022), hence a five times larger proteome than currently considered. The ghost proteome is thus also a potentially rich source of biomarkers of major interest for the understanding of pathophysiology and it has already been studied on endometrial cancer[12] and breast cancer,[13] and on glioblastoma.[14,15] Indeed, ghost proteins, physiologically present in cells, can be impacted by mutations, which might impact the signaling pathways in which they are involved.[15] However, although AltProts have been identified in a wide variety of contexts and especially in cancer, their functions often remain enigmatic.[16,17] Studies on AltProts are often limited as case-by-case, complex and costly biomolecular studies to obtain functional protein information are lacking.[18–20] Few untargeted strategies have enabled the identification of the molecular function of a protein in a single analysis. Bioinformatics tools, including linking protein functional information through networks and gene ontology (GO) analysis, are powerful tools for this purpose.[21] Such tools allow to redraw the signaling pathways and group together RefProts belonging to the same biological process, molecular function, or cellular localization, increasing the information about the cellular mechanism. Such information can be obtained through databases holding information on protein-protein interactions (PPIs) such as STRING,[22] BioGrid,[23] and IntAct[24] allowing them to be applied to a large-scale protein analysis such as a bottom-up approach by chromatography coupled to mass spectrometry analysis (LC-MS/MS) of RefProts. However, similar PPI data are for AltProts are currently largely unknown and AltProts remain largely understudied as baits for identifying PPIs.

[1] Université de Lille, University Lille, CHU Lille, Inserm U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, 59000 Lille, France

[2] Inserm/CNRS UMR 1283/8199, Pasteur Institute of Lille, EGID, Lille, France University of Lille, Lille, France

[3] VIB-UGent Center for Medical Biotechnology, VIB, 9052 Ghent, Belgium

[4] Department of Biomolecular Medicine, Ghent University, 9052 Ghent, Belgium

[5] These authors contributed equally

[6] Lead contact

*Correspondence:
amelie.bonnefond@univ-lille.fr (A.B.),
kris.gevaert@ugent.be (K.G.),
michel.salzet@univ-lille.fr (M.S.)

https://doi.org/10.1016/j.isci.2023.105943

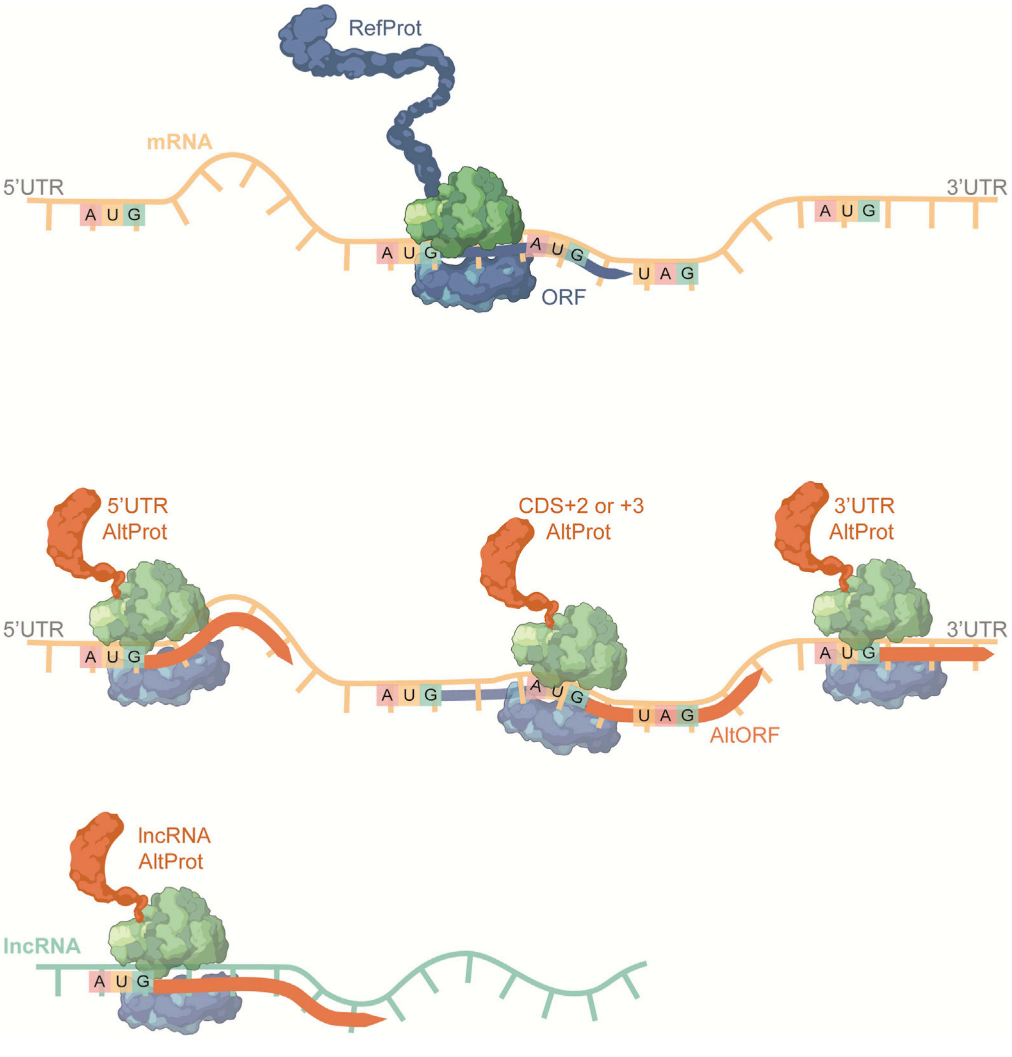


Figure 1. Schematic representation of the translation of AltProts coded from AltORFs

Top panel: translations of RefProts at the CDS region. Middle panel: AltProts translated from 5′ AND-3′ UTRs and CDS +2, +3 frames. Bottom panels: AltProts encoded from a lncRNA.

One interesting approach to obtain PPI data involving AltProts is based on the use of crosslinkers combined with analysis by mass spectrometry (XL-MS). This hypothesis-free strategy, when applied to a complex mixture such as a cell extract, fixes actual PPIs present and allows us to identify new interactions of a bait protein. When processing XL-MS data, one may search for AltProts by using a database holding AltProt sequences. XL-MS has been applied for the structural analysis of purified proteins and to identify interactions in purified protein complexes. However, XL-MS holds some limitations when applied to the large-scale exploration of cellular PPIs, the main ones being the low number of crosslinked peptides that get identified compared to non-crosslinked peptides and the identification of cross-linked peptides because of their complex spectra. To increase the identification of the former, enrichment workflows can be implemented. Such enrichment depletes non-crosslinked (or free) peptides upon sample fractionation, which is generally carried out by size exclusion chromatography (SEC) or cation exchange chromatography (SCX). However, despite a significant increase in the identification rate of crosslinked peptides, such approaches require a rather large amount of material (minimum 60 million cells[25] or 2 mg of protein.[26] Other strategies that are currently emerging are generally based on the use of customized crosslinkers (non-commercial), often tri-functional, allowing targeted enrichment of crosslinked peptides by the functionalized third arm of the molecule. However, these customized crosslinkers also require large quantities of biological material.

In an era where mass spectrometry-based proteomics aims to study proteomes at the single-cell level or is applied to clinical samples of limited quantity, we came up with a strategy to increase the identification of crosslinked peptides while using relatively small amounts of material. Our strategy is based on the decomplexation of the sample. Considering that a limiting factor of XL-MS analysis without enrichment is the (too) high signal intensity of free peptides, we have chosen to divide the cell into different fractions. Thus, from a reasonable number of cells (3 million cells), we separate the proteome into membrane, cytoplasm, nuclear, chromatin, and cytoskeleton proteomes. This strategy was chosen as it has a double advantage: it allows us to increase the number of identified crosslinked peptides without prior peptide fractionation and it provides information on the cellular localization of the identified AltProts. Very little information exists on the subcellular location of AltProt, yet, in some targeted studies, it was reported that AltProts could have a different location compared to the RefProt originating from the same mRNA. Another example shows a co-localization with the associated RefProt, for the cooperation or co-regulation of the gene via its AltProt.[27] The use of this strategy allows us to ''kill two birds with one stone'' to optimize the detection of interactions involving AltProt to assign signaling pathways in a non-targeted way and to provide information on the possible localization of AltProt in the cell.

Thus here, we propose the use of subcellular fractionation to increase the identification rate of crosslinked peptides and simultaneously provide information on the cellular localization of identified AltProts. As such, cellular functions of AltProts can be assessed in a non-targeted way, which is expected to increase our understanding of the ghost proteome.

In this study, we propose the use of subcellular fractionation in order to increase the rate of identification of crosslinked peptides all by providing information on the localization of AltProt in the cell. This is to highlight the functions of AltProt in a non-targeted way and to progress in the understanding of the ghost proteome.

RESULTS

In cellulo crosslinking, subcellular fractionation, and protein digestion

Overview of the workflow used

As the function of the vast majority of AltProts predicted from OpenProt Database[3] remains unknown, as mentioned earlier, we used crosslinking mass spectrometry to characterize AltProts in a non-targeted way. To obtain more information on AltProts on a large scale and to optimize the identification of crosslinked peptides, we set up a workflow combining *in cellulo* crosslinking with subcellular fractionation and analysis by nLC-MS/MS. Additionally, to confirm the presence of crosslinked proteins SDS-PAGE was used and Western blotting to confirm the efficiency of subcellular fractionation (Figure 2A). Finally, the generated data were integrated to identify AltProts, their partners, and the signaling pathways they are involved in.

For crosslinking, we used in cellulo DSSO treatment on replicates of 3E6 immortalized human ovarian cells (T1074-ABM). Following crosslinking and quenching, subcellular protein fractionation was used to extract five different protein fractions corresponding to cytoplasm (Cyt), membrane-bound (Memb), nuclear (Nuc), chromatin-bound (Chr), and cytoskeletal (Ske) proteins.

Characterization of the crosslink reaction

Protein crosslinking was visualized by SDS-PAGE Figure 2B. The formation of protein complexes by crosslinking prevents the migration of these complexes in the separation gel (12% acrylamide). As a result, intense protein staining is observed between the stacking and separation gels, even with protein staining in the wells at the entrance of the stacking gel pointing to the formation of protein complexes that are so large that they cannot enter the stacking gel (4% acrylamide). Note that this was only observed when analyzing crosslinked samples and for the positive control of crosslinking reaction (BSA). In crosslinking sample, a ''blur'' of migration can be observed, this could be formed by smallest structures like intra-protein crosslinks and small(er) protein complexes. Interestingly, almost complete protein crosslinking is found for most of the analyzed subcellular fractions, except for the chromatin fraction where a clear band is observed in the separation gel that is also present in the non-crosslinked control, pointing to a protein that is not affected (or only slightly) by the crosslink used.

Evaluating the efficiency of the subcellular fractionation

The efficiency of the subcellular fractionation procedure was evaluated on non-crosslinked cells. To determine if we were able to extract known-location proteins. Five protein markers were selected according to

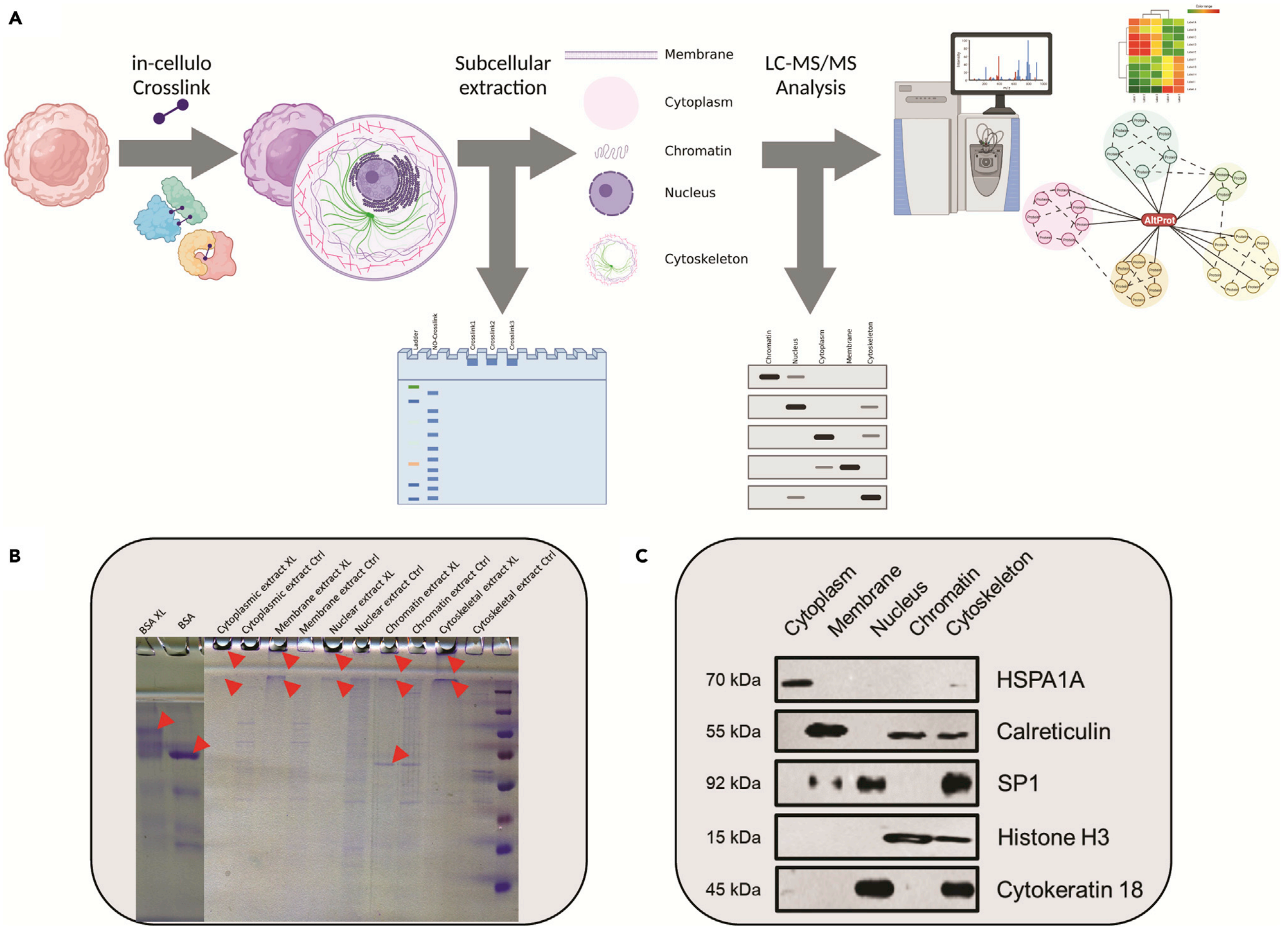


Figure 2. Description of the general workflow used

(A) The first step included harvesting *and in celullo* crosslinking, followed by subcellular fractionation and SDS-PAGE to confirm the crosslinking reaction. Additionally, Western blotting was employed to verify subcellular fractionation. Finally, nLC-MS/MS analysis and crosslinking network revision were performed.

(B) Coomassie blue stained SDS-PAGE: each crosslinked subcellular fraction was compared to a non-crosslinked fraction. BSA crosslinked or not was used as controls. Red arrows display the crosslinked signals.

(C) Western blot signals obtained from each fraction. HSPA1A signal is present in the cytoplasm fraction. For calreticulin, signals are observed at chromatin, cytoskeleton and a more intense signal at the membrane-bounded fraction. SP1 is observed at nucleus and cytoskeleton. Histone H3 is found in chromatin and cytoskeleton. Cytokeratin 18 is found at Nucleus and cytoskeleton. These results correspond to the ones found in UniProtKB, COMPARTMENTS, and the literature.

their subcellular location and already tested by the kit's vendor, HSPA1A for the cytoplasm, calreticulin for the membrane-bounded proteins, SP1 for the nucleus, Histone H3 for the chromatin, cytokeratin 18 for the cytoskeleton. To verify the subcellular location UniProtKB was used as a reference. In Figure 2C the band corresponding to HSPA1A is clearly observed in the cytoplasmic fraction, and additionally, a weak signal is found in the cytoskeletal fraction. According to UniProt: P0DMV8, HSPA1A can be found in the cytoplasm and at the cytoskeleton, which correlates with the signals observed in the blot. At UniProt, calreticulin (UniProt: P27797) is referenced in the membrane of several organelles. Furthermore, it has been described in the chromatin[28] and cytoskeleton.[29] SP1 (UniProt: P08047) was annotated to reside in the nucleus and in the cytoplasm. Here, we observed two strong signals in the nucleus and the cytoskeleton, the latter can be explained by the fact that in mitosis SP1 can be redirected toward the microtubules.[30] For histone H3, two signals can be observed: in the chromatin and cytoskeleton fractions. According to UniProt: P68431, this protein can be found in the nucleus and at chromosomes, according to the mitosis process, during the cell division chromatin is in contact with the microtubule and can explain why histone H3 is also identified in cytoskeleton fraction. Signals for cytokeratin 18 were observed in the nuclear and cytoskeleton fraction,

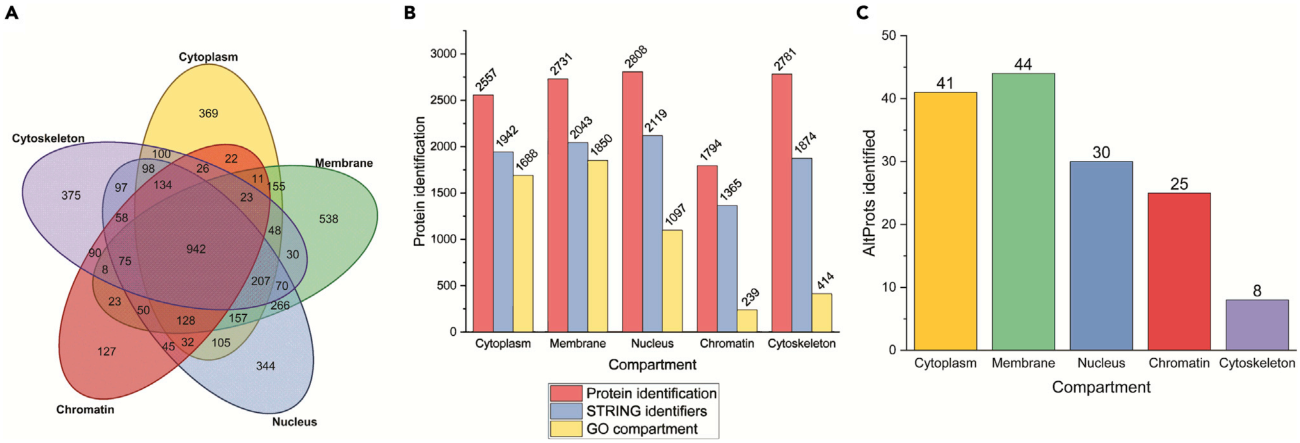


Figure 3. Subcellular fractionation analysis

(A) Venn diagram displaying the distribution of reference proteins identified in the different subcellular fractions.

(B) Bar chart showing the number of RefProts identified (red), the number of RefProts indexed in STRING (blue), and the number of RefProts that contain the GO term of the localization corresponding to the fraction where it was found.

(C) Bar chart displaying the number of AltProts identified in at least two replicates in the same subcellular compartment.

while UniProt: P05783, annotates this protein in the cytoskeleton and nucleus, and sometimes in the cytoplasm. Even if the power of compartment separation is still limited, by this method we can obtain a first view of the cellular compartments repartition of the protein. Therefore, we will be able to propose a cellular localization to the AltProts identified by this methodology.

Identification of RefProts

The MS/MS data from the crosslinked samples were analyzed by Proteome Discoverer V2.5 using Sequest HT.[31] We initially focused on the RefProts (databased Uniprot 02-2022) and could identify 4,753 unique RefProts, of which 2,557 were identified in the cytoplasmic fraction, 2,731 in the membrane, 2,808 in the nucleus, 1,794 in the chromatin fraction and 2,781 in the cytoskeleton, with a high number of proteins shared by different fractions (Figure 3A). The fraction in which more compartment-specific identifications were found was the membrane fraction (538), followed by the cytoskeleton (375), cytoplasm (369), nucleus (344), and chromatin fraction (127). A gene ontology (GO) cellular component enrichment analysis was performed using the STRING app[32] at ClueGO[33] Figure 3B. In general, and as expected, the number of indexed proteins in STRING is less than the ones identified. Moreover, the number of proteins that possess the GO term for the compartment in which it was identified is very low for the chromatin and cytoskeleton fractions.

Identification and characterization of AltProts

Following a similar approach as for the RefProts, AltProts were identified, now using the OpenProt database. A total of 112 AltProts were identified in at least two replicates in the same subcellular compartment (Figure 3C). The highest number of AltProts (44) was found in the membrane-bound fraction, followed by cytoplasmic AltProts (41), 30 in the nucleus, 25 in the chromatin fraction, and eight in the cytoskeletal fraction. Of note, 24 AltProts were identified in two or more cellular compartments. With the ability to separate subcellular proteins we can propose information about localization for the AltProts identified. Such information is important as the function of a protein depends, amongst others, on the cellular compartment or organelle where it is localized, as this provides the necessary physiological context, aiding the functional characterization of AltProts. Further analyses showed that 88.3% of the identified AltProts originate from non-coding RNAs (ncRNA), 5% from miscellaneous RNAs (misc_RNA), 3.3% from a frameshift in the mRNA CDS, and 1.7% from each of the 3′ and 5′ UTR mRNA regions (Figure 4A). Considering the distribution of the molecular weights of the identified AltProts, more than 5% of the AltProts have molecular weights below-30 kDa (Figure 4B). In Table S1, the complete description of the AltProts, protein Blast results, and the unique peptide identified in MS/MS corroborate by NextProt Peptide uniqueness checker.[34] The OpenProt database holds information on the prediction of protein domains in AltProts, made possible by comparisons with RefProt sequences and domain annotations made with algorithms like InterProScan.[35]

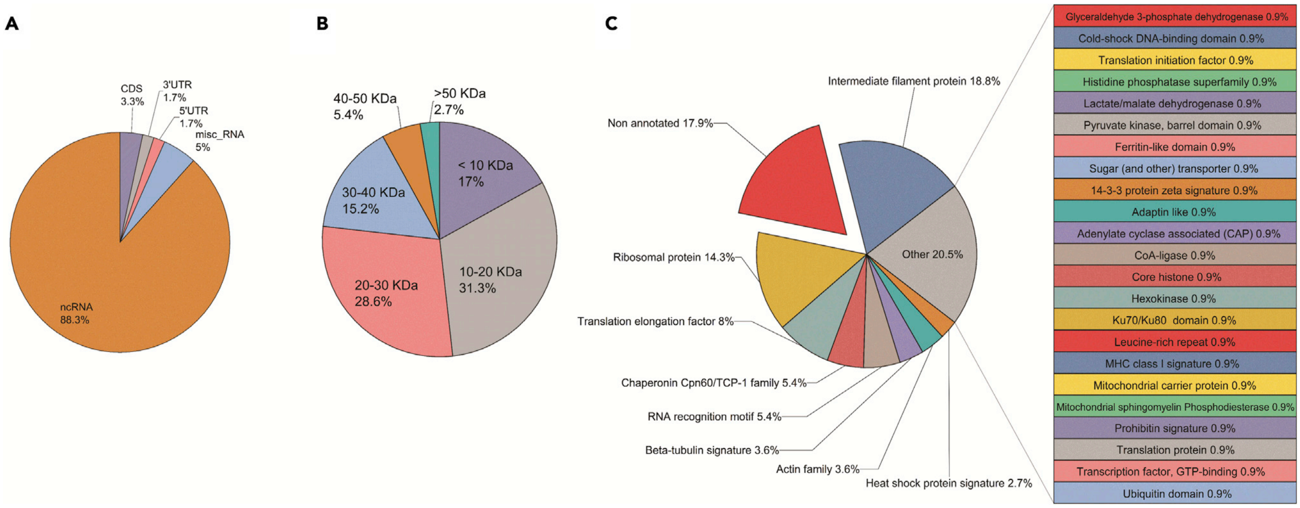


Figure 4. AltProts properties

(A) RNA type distribution found among the 112 AltProts identified.

(B) Molecular weight distribution of the AltProts identified.

(C) Predicted protein domain distribution of the AltProts, retrieved from the OpenProt database.

Domains describe a structural or functional entity that is typically evolutionary conserved among orthologs. Of the 112 AltProts identified, 17.9% do not have any annotated protein domain (Figure 4C), while the intermediate filament protein domain was the major domain found (18.8%), with beta-tubulin and actin family domains also identified. This points to the fact that some AltProts might function as structural proteins. Other retrieved protein domains relate to ribosomal proteins, translation and elongation factors and chaperonins, and an RNA recognition motif. Further, a great heterogeneity was observed, represented by the "Other" section (Figure 4C) which does not allow a proper breakdown into different domains, yet represents 20.5% of the AltProts identified.

In summary, our results show that our methodology provides robust information about AltProts. For instance, IP_596971 found in the membrane-bound fraction possesses a major histocompatibility complex (MHC) class I signature domain which is usually found at the cell membrane. Along the same line, IP_566083 identified in the same fraction has a transmembrane transport protein domain. Another example is IP_775646, identified in the cytoplasmic fraction, possessing a ribosomal protein domain (Table S1).

Crosslink network analysis

Next, the XlinkX algorithm[36] implemented to PD2.5 was used to identify the crosslinked peptides and build protein interaction maps. A total of 220 crosslinks (see Table S2) were identified without targeted crosslinked protein or peptide enrichment. Among these 220 crosslinks, 16 crosslinks were found involving an AltProt. The membrane fraction had the highest number of identified crosslinks (88, Figure 5A), which could be explained by DSSO first reacting with surface-exposed membrane proteins upon its administration to cells. A PPI network was generated in Cytoscape[37] (Figure 5B), where RefProts are identified in interaction with some AltProts. Several inter-protein crosslinks were found multiple times next to intra-protein crosslinks. In total, 16 AltProts were found to interact with RefProts (see Table S3).

To attribute functions of an AltProt from this list of PPIs, we retrieved the known interactions from STRING, BioGrid, and IntAct database and included the identified crosslinked interactions (Figure 5C). We observed (green lines) that 10 interactions were already described. These interactions found were: H3F3A-H2AFJ, ITGA5-ITGB1, YWHAZ-YWHAQ, PHB-PHB2, EMC2-EMC8, COX7B-COX4I1, ATP5A1-ATP5F1, PDIA6-PLEKHO1, HLA-B-B2M, and B2M-HLA-A.

For the RefProts that did not present referenced STRING interaction, an enrichment has been performed to expand the network (Figure S1). With this expanded network a molecular function GO term enrichment analysis was performed with the ClueGO App from Cytoscape. For the resulting network (Figure 6), the

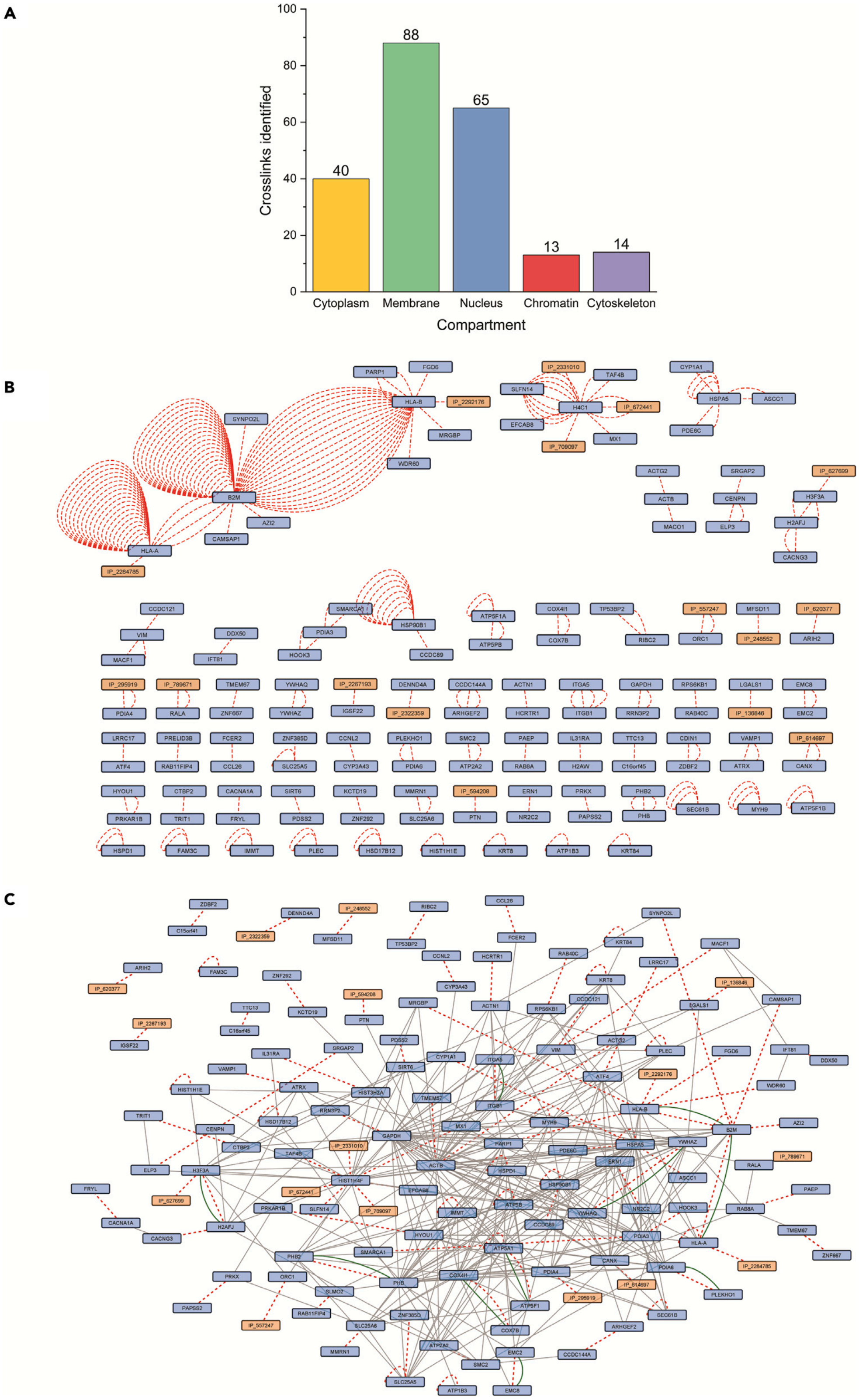

A
100
80
60
40
20
0
Crosslinks identified
40
88
65
13
14
Cytoplasm
Membrane
Nucleus
Chromatin
Cytoskeleton
Compartment
B
PARP1
FGD6
IP_2331010
TAF4B
CYP1A1
SLFN14
HLA-B
IP_2292176
H4C1
IP_672441
HSPA5
ASCC1
SYNPO2L
MRGBP
EFCAB8
PDE6C
IP_709097
MX1
WDR60
ACTG2
SRGAP2
IP_627699
B2M
ACTB
CENPN
H3F3A
AZI2
MACO1
ELP3
H2AFJ
CAMSAP1
HLA-A
CACNG3
IP_2284785
CCDC121
SMARCA1
ATP5F1A
COX4I1
TP53BP2
IP_557247
MFSD11
IP_620377
VIM
DDX50
PDIA3
HSP90B1
COX7B
RIBC2
ORC1
IP_248552
ARIH2
MACF1
IFT81
HOOK3
CCDC89
ATP5PB
IP_295919
IP_789671
TMEM67
YWHAQ
IP_2267193
DENND4A
CCDC144A
ACTN1
ITGA5
GAPDH
RPS6KB1
LGALS1
EMC8
PDIA4
RALA
ZNF667
YWHAZ
IGSF22
IP_2322359
ARHGEF2
HCRTR1
ITGB1
RRN3P2
RAB40C
IP_136846
EMC2
LRRC17
PRELID3B
FCER2
ZNF385D
CCNL2
PLEKHO1
SMC2
PAEP
IL31RA
TTC13
CDIN1
VAMP1
IP_614697
ATF4
RAB11FIP4
CCL26
SLC25A5
CYP3A43
PDIA6
ATP2A2
RAB8A
H2AW
C16orf45
ZDBF2
ATRX
CANX
HYOU1
CTBP2
CACNA1A
SIRT6
KCTD19
MMRN1
IP_594208
ERN1
PRKX
PHB2
PRKAR1B
TRIT1
FRYL
PDSS2
ZNF292
SLC25A6
PTN
NR2C2
PAPSS2
PHB
SEC61B
MYH9
ATP5F1B
HSPD1
FAM3C
IMMT
PLEC
HSD17B12
HIST1H1E
KRT8
ATP1B3
KRT84
C

Figure 5. Crosslinking network analysis

(A) Total crosslink identification distribution in each subcellular location.

(B) Raw crosslink network in which AltProts are marked in orange and RefProts are marked in blue.

(C) Crosslinked network enriched by the STRING interactions (gray lines) retrieved between these crosslinked (red dash lines) RefProts. Green lines highlight the PPIs already described in molecular interactions databases.

interactions between AltProts and RefProts were displayed along the GO terms enriched. The AltProts IP_2292176, which was found crosslinked to HLA-B, and IP_2284785, crosslinked to HLA-A, were linked to antigen processing and presentation of peptide antigens via MHC class I (GO:0002474). IP_789671, crosslinked with RALA, and IP_620377, crosslinked with ARIH2, appear to be related to the regulation of mitochondrial outer membrane permeabilization involved in the apoptotic signaling pathway (GO:1901028). IP_295919, crosslinked to PDIA4, and IP_614697, crosslinked to CANX may participate in the response to ER stress (GO:0034976). IP_136846 was identified crosslinked to LGALS1 in the membrane fraction is not annotated by a GO term, but LGALS1 is known to bind wide array carbohydrates and regulating apoptosis, cell proliferation, and differentiation.[38] As a final example, IP_627699 was found crosslinked to H3F3A, which possesses an STRING interaction with ORC1. ORC1 was also found crosslinked to IP_557247. Also, IP_2331010, IP_672441, IP_709097 and TAF4B were crosslinked to HIST1H4F, which interacts with ORC1, H3F3A, SIRT6 and CENPN. These PPIs hint that these five AltProts can be involved in mRNA transcription by RNA polymerase II (GO:0042789), protein-DNA complex subunit organization (GO:0071824), DNA dealkylation involved in DNA repair (GO:0006307), or DNA replication-independent chromatin assembly (GO:0006336).

Structural modeling of selected interactions

Since AltProts remain ill-studied, no specific antibodies are available for their monitoring in cells by immunofluorescence or for co-immunoprecipitation to confirm observed interactions with other proteins. Our objective is to set up a large-scale analysis method to identify the best signaling pathway actors to then carry out targeted characterization studies, using molecular biology to overexpress and tag the proteins of interest. Thus, in a non-targeted study context, coupled with the use of XL-MS.

We decided to confirm the probability of the interactions observed by analyzing 3D models of AltProts with unguided interaction docking between the two partners. The structures of the AltProts were predicted with I-Tasser[39] and the interactions with ClusPro.[40] The RefProt, of which the structure was predicted by AlphaFold[41] was used as a receptor of the AltProt (smaller in structure). In this way, we could confirm the interactions observed upon XL-MS by measuring the distance of the predicted interactions with a mean of 21.13 Å (Figure S2), which agrees with the distances described in the literature for DSSO, being from 5.3 Å[42] to 30 Å.[43]

DISCUSSION

AltProts remain infrequently studied and, currently, no methodology allows for the characterization of these proteins in a non-targeted way. Here, we proposed a methodology based on the identification of AltProts by mass spectrometry including XL-MS to identify their interaction partners, which allows us to place AltProts in signaling pathways, amongst others. This makes it possible to assign possible functions to yet uncharacterized proteins and it also adds such proteins to cellular pathways. Moreover, by using fractionating cells, we also proposed an intracellular localization dimension whilst allowing us to increase the number of identified crosslinks. One major advantage of our workflow is the drastic reduction of the amount of material needed. Indeed, here, we used 3E6 cells, whereas previous studies, which used or did not enrichment methods, started from at least 5E7 cells.[25] Cell fractionation reduces the complexity of the sample and therefore increases the identification of crosslinked peptides whose signals are often masked by those of free peptides. This study also reminds the fact that AltProt may be involved in the development of pathology, but like RefProt they are also present in a physiological context with involvement in signaling pathways and functions, in the same way as RefProts.

We first evaluated the efficiency of the subcellular protein fractionation kit used by Western blotting using compartment known proteins. With the RefProts identified, a gene ontology (GO) cellular component enrichment was performed. Both the signals observed in the blots and the identified GO terms seem to suggest that due to the intrinsic principle of the subcellular fractionation kit, which

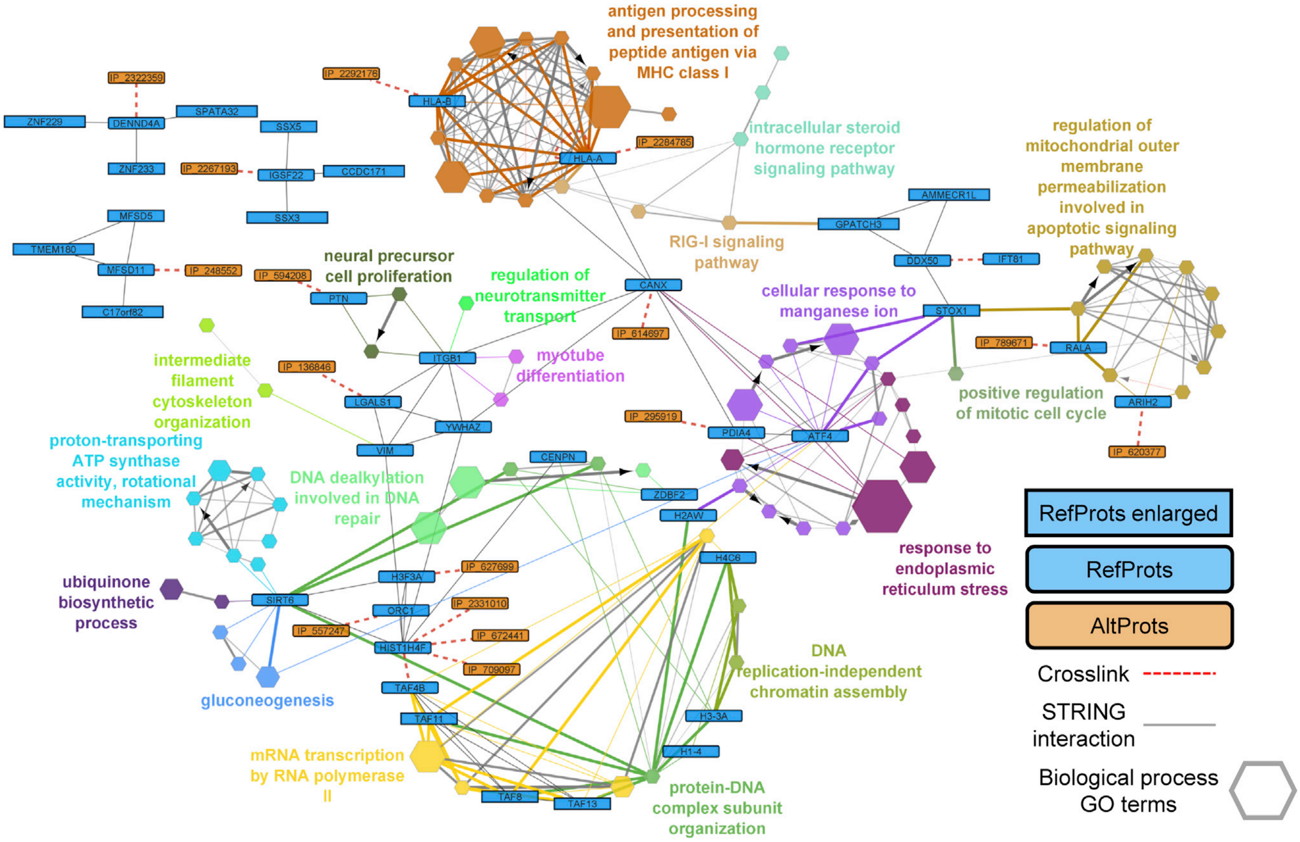

antigen processing and presentation of peptide antigen via MHC class I
IP_2322359
IP_2292176
HLA-B
ZNF229
DENND4A
SPATA32
SSX5
IP_2284785
HLA-A
ZNF233
IP_2267193
IGSF22
CCDC171
SSX3
intracellular steroid hormone receptor signaling pathway
regulation of mitochondrial outer membrane permeabilization involved in apoptotic signaling pathway
AMMECR1L
GPATCH3
MFSD5
TMEM180
MFSD11
IP_248552
C17orf82
IP_594208
PTN
neural precursor cell proliferation
regulation of neurotransmitter transport
RIG-I signaling pathway
DDX50
IFT81
CANX
cellular response to manganese ion
STOX1
IP_614697
IP_789671
RALA
intermediate filament cytoskeleton organization
IP_136846
ITGB1
myotube differentiation
LGALS1
positive regulation of mitotic cell cycle
ARIH2
IP_295919
PDIA4
ATF4
YWHAZ
proton-transporting ATP synthase activity, rotational mechanism
VIM
IP_620377
CENPN
DNA dealkylation involved in DNA repair
ZDBF2
H2AW
RefProts enlarged
response to endoplasmic reticulum stress
H4C6
RefProts
ubiquinone biosynthetic process
H3F3A
IP_627699
SIRT6
ORC1
IP_2331010
AltProts
IP_557247
IP_672441
HIST1H4F
IP_709097
DNA replication-independent chromatin assembly
Crosslink
gluconeogenesis
TAF4B
STRING interaction
TAF11
H3-3A
mRNA transcription by RNA polymerase II
H1-4
Biological process GO terms
TAF8
TAF13
protein-DNA complex subunit organization

Figure 6. GO molecular function enrichment network generated with ClueGO in Cytoscape

GO enrichment was generated from the accession numbers of Figure S1. AltProts are marked in orange and RefProts in blue. Enriched GO terms are displayed as hexagons. Crosslinks are marked in red dashed lines.

is based on centrifugation and supernatant removal, some remnant proteins from previous supernatants could be transferred to the last fraction (cytoskeleton). This increased the number of ''non-specific proteins'' identified in this fraction to 414 over the total 2781 (Figure 3B). However, one must consider that the cytoskeleton is the scaffold structure of the cell and the transport path of a large number of proteins, and one may thus identify proteins from other compartments in transit or in contact with the cytoskeleton.

Another advantage of subcellular fractionation is that one may attribute a cellular compartment to AltProts. Most AltProts were found identified in the membrane and nucleus fractions. Also, three AltProts were identified in all five cellular fractions. IP_623199 is 236 amino acids long (26.79 kDa) and coded from a lncRNA of the KRT8P25 gene. IP_774693 contains 75 residues (8.68 kDa) and coded from a lncRNA transcribed from the TUBAP2 gene. And, finally, IP_790379, 42 amino acids long (4.38 kDa) translated from a lncRNA of the AL161932.1 gene. This might point to AltProt dynamism and mobility in the cell, explaining the identification in all compartments in case that is not an artifact link to contamination between the fractions, this could be further confirmed by a targeted approach like fluorescence microscopy of these AltProts fused to Green Fluorescent Protein (GFP).

The vast majority of the identified AltProts originated from lncRNAs and a small fraction from mRNAs (Figure 4A). For a long time, lncRNAs were believed to act as transcriptional and post-transcriptional regulators without any coding potential.[44] Nowadays, and also given our data, this concept is clearly shifting.

One approach to infer functions of AltProts is based on the domains that are found in their sequence. Interestingly, one-third of the here retrieved protein domains are involved in translation. This correlates with previous observations[16] in which we have shown that the AltProt AltATAD2 can interact with the RPL10 region interacting with 5S rRNA and may thus be a mechanism of the regulation of the ribosome. It is also noteworthy that 17% of the identified AltProts have no known domain region (Figure 4C). The small size, 10 to 30 kDa for more than half of these proteins (Figure 4B), also suggests numerous functions like enzyme and protein inhibition or ligand/receptor interaction, such as the function of endogenous peptide and neuropeptide.[45,46]

Crosslinking mass spectrometry has been used since the early 2000s.[47] As of 2015, XL-MS has been used to identify PPIs in a large-scale manner.[36] As the vast majority of the AltProt functions is unknown, a PPI untargeted approach can be the first way to appoint functions to AltProts, by the guilt-by-association concept. Our methodology, which does not involve any peptide enrichment step (SCX or SEC) and consumes a small number of cells (3E6), allowed us to identify 220 and 16 crosslinks between AltProts and RefProts (Figure 5B&C). While these numbers appear not very high, according to the workflow used, they are acceptable, and already allow the future exploration of several targets.

From the previously described interactions found (10 PPIs), H3F3A (H3 histone family member 3A) and H2AFJ (H2A histone family member J) are part of the nucleosome complex in which the DNA is wrapped and arranged. Integrin alpha-5 (ITGA5) and Integrin beta-1 (ITGB1) are part of the integrins family. This family of proteins serves as cell-matrix adhesion receptors. Specifically, the Integrin alpha5beta1 binds to the fibronectin Arg-Gly-Asp motif. This interaction has been identified by high-resolution X-ray diffraction protein crystallography.[48] YWHAZ and YWHAQ are part of the 14-3-3 family of proteins that mediate signal transduction by binding to phosphoserine-containing proteins and are involved in multiple signaling pathways. The interaction between them has been identified multiple times by affinity capture-MS and co-fractionation.[49–51] Prohibitins are a family of proteins that contain a stomatin/prohibitin/flotillin/HflK/HflC domain. Moreover, PHB and PHB2 act as a frame in different cellular processes. The PPI between both has been observed in different types of experiments such as proximity label-MS,[52] co-fractionation,[53] and affinity capture-MS.[54] The ER membrane protein complex comprises nine subunits and its main function is the insertion of transmembrane domains in protein biosynthesis. The interaction between the subunits two (EMC2) and eight (EMC8) has been demonstrated by cryo-electron microscopy (EM).[55] Cytochrome *c* oxidase is a 13mer inner mitochondrial transmembrane enzyme. It is the final complex of the electron transport chain, and its main function is the reduction of molecular oxygen to water. The interaction between the subunits COX7B and COX4I1 has been proven by cryo-EM[56] and XL-MS.[57] The human mitochondrial ATP synthase complex produces ATP from ADP in the presence of a proton gradient, generated by the electron transport chain. From this complex, ATP5A1 and ATP5F1 have been found interacting by XL-MS.[57] Protein Disulfide Isomerase Family A Member 6 (PDIA6) is a member of the disulfide isomerases. These proteins catalyze the arrangement of disulfide bridges resulting in protein folding. The Pleckstrin Homology Domain Containing O1 protein (PLEKHO1) has been described to be a regulator of the cytoskeleton by its interaction with actin capping proteins. Even though the crosslink between these two proteins was already described.[57] The MHC class 1 complex is comprised of a light chain, named beta-2 microglobulin (B2M); and a heavy chain. The heavy chain belongs to the human leukocyte antigens (HLA) proteins which comprise HLA-A and HLA-B. These 2 interactions, B2M-HLA-A[58] and HLA-B-B2M,[59] have been identified by X-ray diffraction protein crystallography.

Among the PPIs found by XL-MS, IP_2292176 (AltFAM227B), which is predicted to be translated from the 5′UTR +2 ORF, giving rise to a protein of 67 amino acids (7.68 kDa), was found crosslinked to HLA-B. Upon modeling this AltProt and docking with HLA-B, we observed 20.11 Å between the two crosslinked lysines (Figure 7A), which fits with the crosslinking range described for DSSO. HLA-B is part of the MHC class 1 and oversees the presentation of antigenic peptides of 8-13 residues that are recognized by CD8$^+$ T cells driving antigen-specific immune response. Due to the importance of this system for tumor-derived antigens, informatics tools have been developed to predict the binding of peptides to this class of proteins and one of them is NetMHC-4.0,[60] this tool is based on a machine-learning algorithm that predict the capacity of binding to a protein and peptide sequence based on this size and amino acid constitution, giving the possibility to predict interaction for AltProt not referenced in other tools based on databases identification. The results obtained using the complete sequence of the AltProt divided in 8-14-mers were predicted as weak binding for the alleles HLA-B1502, HLA-B1503, HLA-B1517, HLA-B4001, HLA-B4002 and HLA-B5701. The peptide with the strongest interaction was built in I-TASSER and docking was performed in ClusPro. The distance obtained between the crosslinked residues was 16.72 Å, which validates the PPI

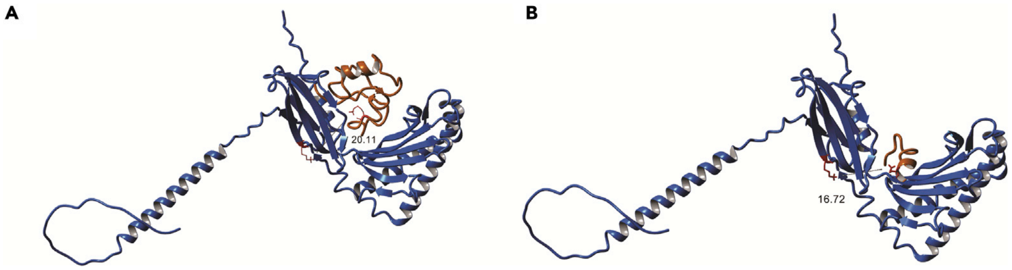


Figure 7. IP_2292176 (AltFAM227B) predicted models docked to Alpha-Fold HLA-B model

(A) displays the interaction of HLA-B and the complete IP_2292176. The distance between the two Lys residues involved at the crosslink is of 20.11 Å.

(B) Interaction between the peptide with the predicted strongest interaction (DKKESMANYPRL) and HLA-B.

found by XL-MS (Figure 7B). This allows us to make several hypotheses. This AltProt in the cell can be degraded and exposed to the surface by the MHC class I system to be presented as an antigen. This hypothesis makes AltProts potential new immunopeptides, which in the case of pathologies such as cancer can be therapeutic targets.[61],[62] A second hypothesis is that the AltProt binds the MHC-I molecule, inhibiting the presentation of other immunopeptides. The fact that this interaction was found in an XL-MS study without the enrichment of crosslinked peptides could indicate that this PPI is sufficiently represented in the studied cells. The study of AltProt in the antigenic presentation and the immune response is an axis still very poorly explored in which the identification of a new specific target has a strong potential for therapy, the AltProts are in this context a potential source of new targets not yet exploited.

The interactions found for HIST1H4F, for which we observed crosslinking to three AltProts and TAF4B, is noteworthy. The interaction of HIST1H4F with TAF4B is not referenced in STRING, but interactions with other subunits of the TATA-binding protein-associated factors (TAFs), TAF1 and TAF6L, are. As such, we may hypothesize that TAF4B indirectly interacts with HIST1H4F. TAFs are part of transcription factors that regulate RNA polymerase II transcription, which is the most flexible transcription system controlled by modified histones (acetylation), transcription factors, and chromatin structure.[63] The AltProts that were crosslinked to HIST1H4F were IP_2331010 (AltKDM4C, 3′UTR +2 ORF), IP_672441 (AltRPS15AP10, ncRNA), and IP_709097 (AltAC123769.1, ncRNA). These interactions were found in the cytoplasmic, nuclear, and membrane-bound fraction, respectively. According to the COMPARTMENTS subcellular localization database,[64] HIST1H4F is found experimentally in the nucleus and cytosol, moreover, a GO term linked to the membrane is referenced in UniProt (P62805). This could indicate that these AltProts might be involved in mRNA synthesis or in the interaction between the TAFs and the histones. Another interaction was found involving another histone; H3F3A and IP_627699 (AltSLC41A3, +3 ORF mRNA CDS). A crosslink was found between IP_557247 (AltMRRFP, ncRNA) and ORC1, which is a crucial protein in the initiation of DNA replication by the interaction with MYST histone acetyltransferase 2[65] and has annotated STRING interactions with the TAF family. ORC1 is also involved in transcription silencing.[66] These findings could indicate that these AltProts play a role in gene transcription.

In conclusion, we here described a methodology based on subcellular fractionation and crosslinking mass spectrometry to increase our knowledge of the thus far neglected alternative or ghost proteins. We were able to localize some alternative proteins and infer possible functions of some of these proteins as they were crosslinked to reference proteins. Our large-scale untargeted approach has set some bases for future research to confirm and validate the hypothesized functions of AltProts described above. Moreover, it appears interesting to employ this methodology to compare pathological to homeostatic cell states and identify disrupted pathways involving AltProts.

Limitations of the study

Our study has some limitations and the first one is related to the limited spread of the concept of alternative (ghost) proteins, resulting in a lack of information and established methodologies to unravel the function of such proteins. Secondly, by employing a detergent and microcentrifugation-based subcellular fractionation kit, cross-contamination of cellular fractions can be an issue. Hence, a more efficient technique for subcellular fractionation, like gradient-based ultracentrifugation could be employed to determine the location of AltProts and generate a finer fractionation. Additionally, given the huge database used (OpenProt), a manual check of the MS/MS spectra associated with the interaction of interest must be done. Such large databases call for more stringent analyses on (crosslinked) peptide identifications.[67] Finally, often key for the success of XL-MS is to reduce the complexity of the sample prior to LC-MS/MS analysis. Thus, employing enrichable crosslinkers like *tert*-Butyl Disuccinimidyl Phenyl Phosphonate (tBu-PhoX) and alkyne-A-DSBSO; could help to identify more crosslinked peptides. However, despite these limitations, it is clear that searching for PPIs of AltProts is opening the way to more complete systems biology.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Cell lines
- METHOD DETAILS
  - ○ Cell culture
  - ○ In cellulo chemical cross-linking
  - ○ Protein subcellular fractionation and western blotting
  - ○ Enzymatic digestion
  - ○ NanoLC-MS/MS analysis
  - ○ Shotgun data analysis
  - ○ Crosslink data analysis
  - ○ Modeling and prediction of interactions between AltProts and RefProts
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.105943.

ACKNOWLEDGMENTS

This research was supported by funding from I-SITE, Institut National de la Santé et de la Recherche Médicale (Inserm), Université de Lille and by The Research Foundation - Flanders (FWO), project number G008018N.

AUTHOR CONTRIBUTIONS

Conceptualization, M.S, T.C; methodology, DF.G, T.C; software, DF.G, T.C; validation, T.C, M.S, K.G; formal analysis, DF.G.; investigation, DF.G, T.C, M.S, K.G; resources, I.F, M.S.; data curation, DF.G, T.C.; writing - original draft, DF.G, T.C. writing - review & editing, K.G, T.C, A.B, S.E., M.S; supervision, T.C, M.S, I.F, K.G, A.B, S.E.; project administration, M.S, I.F, K.G, A.B, S.E.; funding acquisition, M.S, I.F, K.G, A.B, S.E

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: September 12, 2022
Revised: November 7, 2022
Accepted: January 4, 2023
Published: February 17, 2023

REFERENCES

1. Hanada, K., Kumagai, K., Yasuda, S., Miura, Y., Kawano, M., Fukasawa, M., and Nishijima, M. (2003). Molecular machinery for non-vesicular trafficking of ceramide. Nature *426*, 803–809. https://doi.org/10.1038/nature02188.

2. Cardon, T., Fournier, I., and Salzet, M. (2020). SARS-Cov-2 interactome with human ghost proteome: a neglected world encompassing a wealth of biological data. Microorganisms *8*, 2036. https://doi.org/10.3390/microorganisms8122036.

3. Brunet, M.A., Lucier, J.-F., Levesque, M., Leblanc, S., Jacques, J.-F., Al-Saedi, H.R.H., Guilloy, N., Grenier, F., Avino, M., Fournier, I., et al. (2021). OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. Nucleic Acids Res. *49*, D380–D388. https://doi.org/10.1093/nar/gkaa1036.

4. Wang, B., Wang, Z., Pan, N., Huang, J., and Wan, C. (2021). Improved identification of small open reading frames encoded peptides by top-down proteomic approaches and de novo sequencing. Int. J. Mol. Sci. *22*, 5476. https://doi.org/10.3390/ijms22115476.

5. Fabre, B., Choteau, S.A., Duboé, C., Pichereaux, C., Montigny, A., Korona, D., Deery, M.J., Camus, M., Brun, C., Burlet-Schiltz, O., et al. (2022). Depth exploration of the alternative proteome of Drosophila melanogaster. Front. Cell Dev. Biol. *10*, 901351. https://doi.org/10.3389/fcell.2022.901351.

6. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

7. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745. https://doi.org/10.1093/nar/gkv1189.

8. Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. Nat. Chem. Biol. *16*, 458–468. https://doi.org/10.1038/s41589-019-0425-0.

9. Cardon, T., Fournier, I., and Salzet, M. (2021). Unveiling a ghost proteome in the glioblastoma non-coding RNAs. Front. Cell Dev. Biol. *9*, 703583. https://doi.org/10.3389/fcell.2021.703583.

10. Samandi, S., Roy, A.V., Delcourt, V., Lucier, J.-F., Gagnon, J., Beaudoin, M.C., Vanderperre, B., Breton, M.-A., Motard, J., Jacques, J.-F., et al. (2017). Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. Elife *6*, e27860. https://doi.org/10.7554/eLife.27860.

11. Brunet, M.A., Brunelle, M., Lucier, J.-F., Delcourt, V., Levesque, M., Grenier, F., Samandi, S., Leblanc, S., Aguilar, J.-D., Dufour, P., et al. (2019). OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. Nucleic Acids Res. *47*, D403–D410. https://doi.org/10.1093/nar/gky936.

12. Aboulouard, S., Wisztorski, M., Duhamel, M., Saudemont, P., Cardon, T., Narducci, F., Lemaire, A.-S., Kobeissy, F., Leblanc, E., Fournier, I., and Salzet, M. (2021). In-depth proteomics analysis of sentinel lymph nodes from individuals with endometrial cancer. Cell Rep. Med. *2*, 100318. https://doi.org/10.1016/j.xcrm.2021.100318.

13. Hajjaji, N., Aboulouard, S., Cardon, T., Bertin, D., Robin, Y.-M., Fournier, I., and Salzet, M. (2021). Path to clonal theranostics in luminal breast cancers. Front. Oncol. *11*, 802177. https://doi.org/10.3389/fonc.2021.802177.

14. Le Rhun, E., Duhamel, M., Wisztorski, M., Gimeno, J.-P., Zairi, F., Escande, F., Reyns, N., Kobeissy, F., Maurage, C.-A., Salzet, M., and Fournier, I. (2017). Evaluation of non-supervised MALDI mass spectrometry imaging combined with microproteomics for glioma grade III classification. Biochim. Biophys. Acta. Proteins Proteom. *1865*, 875–890. https://doi.org/10.1016/j.bbapap.2016.11.012.

15. Cardon, T., Fournier, I., and Salzet, M. (2021). Shedding light on the ghost proteome. Trends Biochem. Sci. *46*, 239–250. https://doi.org/10.1016/j.tibs.2020.10.003.

16. Cardon, T., Salzet, M., Franck, J., and Fournier, I. (2019). Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation. Biochim. Biophys. Acta. Gen. Subj. *1863*, 1458–1470. https://doi.org/10.1016/j.bbagen.2019.05.009.

17. Leblanc, S., and Brunet, M.A. (2020). Modelling of pathogen-host systems using deeper ORF annotations and transcriptomics to inform proteomics analyses. Comput. Struct. Biotechnol. J. *18*, 2836–2850. https://doi.org/10.1016/j.csbj.2020.10.010.

18. Delcourt, V., Staskevicius, A., Salzet, M., Fournier, I., and Roucou, X. (2018). Small proteins encoded by unannotated ORFs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mRNA. Proteomics *18*, e1700058. https://doi.org/10.1002/pmic.201700058.

19. Brunet, M.A., Jacques, J.-F., Nassari, S., Tyzack, G.E., McGoldrick, P., Zinman, L., Jean, S., Robertson, J., Patani, R., and Roucou, X. (2021). The FUS gene is dual-coding with both proteins contributing to FUS-mediated toxicity. EMBO Rep. *22*, e50640. https://doi.org/10.15252/embr.202050640.

20. Dubois, M.-L., Meller, A., Samandi, S., Brunelle, M., Frion, J., Brunet, M.A., Toupin, A., Beaudoin, M.C., Jacques, J.-F., Lévesque, D., et al. (2020). UBB pseudogene 4 encodes functional ubiquitin variants. Nat. Commun. *11*, 1306. https://doi.org/10.1038/s41467-020-15090-6.

21. Omranian, S., Nikoloski, Z., and Grimm, D.G. (2022). Computational identification of protein complexes from network interactions: present state, challenges, and the way forward. Comput. Struct. Biotechnol. J. *20*, 2699–2712. https://doi.org/10.1016/j.csbj.2022.05.049.

22. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613. https://doi.org/10.1093/nar/gky1131.

23. Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. *30*, 187–200. https://doi.org/10.1002/pro.3978.

24. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358–D363. https://doi.org/10.1093/nar/gkt1115.

25. Gao, H., Zhao, L., Zhong, B., Zhang, B., Gong, Z., Zhao, B., Liu, Y., Zhao, Q., Zhang, L., and Zhang, Y. (2022). In-Depth in vivo crosslinking in minutes by a compact, membrane-permeable, and alkynyl-enrichable crosslinker. Anal. Chem. *94*, 7551–7558. https://doi.org/10.1021/acs.analchem.2c00335.

26. Ryl, P.S.J., Bohlke-Schneider, M., Lenz, S., Fischer, L., Budzinski, L., Stuiver, M., Mendes, M.M.L., Sinn, L., O'Reilly, F.J., and Rappsilber, J. (2020). In situ structural restraints from cross-linking mass spectrometry in human mitochondria. J. Proteome Res. *19*, 327–336. https://doi.org/10.1021/acs.jproteome.9b00541.

27. Nelde, A., Flötotto, L., Jürgens, L., Szymik, L., Hubert, E., Bauer, J., Schliemann, C., Kessler, T., Lenz, G., Rammensee, H.-G., et al. (2022). Upstream open reading frames regulate translation of cancer-associated transcripts and encode HLA-presented immunogenic tumor antigens. Cell. Mol. Life Sci. *79*, 171. https://doi.org/10.1007/s00018-022-04145-0.

28. Kobayashi, S., Uchiyama, S., Sone, T., Noda, M., Lin, L., Mizuno, H., Matsunaga, S., and Fukui, K. (2006). Calreticulin as a new histone binding protein in mitotic chromosomes. Cytogenet. Genome Res. *115*, 10–15. https://doi.org/10.1159/000094795.

29. Wang, X., Tao, T., Song, D., Mao, H., Liu, M., Wang, J., and Liu, X. (2019). Calreticulin stabilizes F-actin by acetylating actin and

protects microvascular endothelial cells against microwave radiation. Life Sci. *232*, 116591. https://doi.org/10.1016/j.lfs.2019.116591.

30. He, S., and Davie, J.R. (2006). Sp1 and Sp3 foci distribution throughout mitosis. J. Cell Sci. *119*, 1063–1070. https://doi.org/10.1242/jcs.02829.

31. Tabb, D.L., Eng, J.K., and Yates, J.R. (2001). Protein identification by SEQUEST. In Proteome Research: Mass Spectrometry Principles and Practice, P. James, ed. (Springer), pp. 125–142. https://doi.org/10.1007/978-3-642-56895-4_7.

32. Doncheva, N.T., Morris, J.H., Gorodkin, J., and Jensen, L.J. (2019). Cytoscape StringApp: network analysis and visualization of proteomics data. J. Proteome Res. *18*, 623–632. https://doi.org/10.1021/acs.jproteome.8b00702.

33. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093. https://doi.org/10.1093/bioinformatics/btp101.

34. Schaeffer, M., Gateau, A., Teixeira, D., Michel, P.-A., Zahn-Zabal, M., and Lane, L. (2017). The neXtProt peptide uniqueness checker: a tool for the proteomics community. Bioinformatics *33*, 3471–3472. https://doi.org/10.1093/bioinformatics/btx318.

35. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

36. Liu, F., Rijkers, D.T.S., Post, H., and Heck, A.J.R. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat. Methods *12*, 1179–1184. https://doi.org/10.1038/nmeth.3603.

37. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

38. Anginot, A., Espeli, M., Chasson, L., Mancini, S.J.C., and Schiff, C. (2013). Galectin 1 modulates plasma cell homeostasis and regulates the humoral immune response. J. Immunol. *190*, 5526–5533. https://doi.org/10.4049/jimmunol.1201885.

39. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat. Methods *12*, 7–8. https://doi.org/10.1038/nmeth.3213.

40. Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D., Yueh, C., Beglov, D., and Vajda, S. (2017). The ClusPro web server for protein–protein docking. Nat. Protoc. *12*, 255–278. https://doi.org/10.1038/nprot.2016.169.

41. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

42. Kao, A., Chiu, C.l., Vellucci, D., Yang, Y., Patel, V.R., Guan, S., Randall, A., Baldi, P., Rychnovsky, S.D., and Huang, L. (2011). Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. Mol. Cell. Proteomics. *10*. M110.002212. https://doi.org/10.1074/mcp.M110.002212.

43. Hevler, J.F., Lukassen, M.V., Cabrera-Orefice, A., Arnold, S., Pronker, M.F., Franc, V., and Heck, A.J.R. (2021). Selective cross-linking of coinciding protein assemblies by in-gel cross-linking mass spectrometry. EMBO J. *40*, e106174. https://doi.org/10.15252/embj.2020106174.

44. Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. *22*, 96–118. https://doi.org/10.1038/s41580-020-00315-9.

45. Mallah, K., Quanico, J., Raffo-Romero, A., Cardon, T., Aboulouard, S., Devos, D., Kobeissy, F., Zibara, K., Salzet, M., and Fournier, I. (2019). Mapping spatiotemporal microproteomics landscape in experimental model of traumatic brain injury unveils a link to Parkinson's disease. Mol. Cell. Proteomics. *18*, 1669–1682. https://doi.org/10.1074/mcp.RA119.001604.

46. Erady, C., Amin, K., Onilogbo, T.O.A.E., Tomasik, J., Jukes-Jones, R., Umrania, Y., Bahn, S., and Prabakaran, S. (2022). Novel open reading frames in human accelerated regions and transposable elements reveal new leads to understand schizophrenia and bipolar disorder. Mol. Psychiatry *27*, 1455–1468. https://doi.org/10.1038/s41380-021-01405-6.

47. Piersimoni, L., Kastritis, P.L., Arlt, C., and Sinz, A. (2022). Cross-linking mass spectrometry for investigating protein conformations and protein–protein Interactions A method for all seasons. Chem. Rev. *122*, 7500–7531. https://doi.org/10.1021/acs.chemrev.1c00786.

48. Xia, W., and Springer, T.A. (2014). Metal ion and ligand binding of integrin α5β1. Proc. Natl. Acad. Sci. USA *111*, 17863–17868. https://doi.org/10.1073/pnas.1420645111.

49. Kristensen, A.R., Gsponer, J., and Foster, L.J. (2012). A high-throughput approach for measuring temporal changes in the interactome. Nat. Methods *9*, 907–909. https://doi.org/10.1038/nmeth.2131.

50. Hein, M.Y., Hubner, N.C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., Gak, I.A., Weisswange, I., Mansfeld, J., Buchholz, F., et al. (2015). A human interactome in three quantitative dimensions organized by stoichiometries and abundances. Cell *163*, 712–723. https://doi.org/10.1016/j.cell.2015.09.053.

51. Huttlin, E.L., Bruckner, R.J., Navarrete-Perea, J., Cannon, J.R., Baltier, K., Gebreab, F., Gygi, M.P., Thornock, A., Zarraga, G., Tam, S., et al. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell *184*, 3022–3040.e28. https://doi.org/10.1016/j.cell.2021.04.011.

52. Go, C.D., Knight, J.D.R., Rajasekharan, A., Rathod, B., Hesketh, G.G., Abe, K.T., Youn, J.-Y., Samavarchi-Tehrani, P., Zhang, H., Zhu, L.Y., et al. (2021). A proximity-dependent biotinylation map of a human cell. Nature *595*, 120–124. https://doi.org/10.1038/s41586-021-03592-2.

53. Moutaoufik, M.T., Malty, R., Amin, S., Zhang, Q., Phanse, S., Gagarinova, A., Zilocchi, M., Hoell, L., Minic, Z., Gagarinova, M., et al. (2019). Rewiring of the human mitochondrial interactome during neuronal reprogramming reveals regulators of the respirasome and neurogenesis. iScience *19*, 1114–1132. https://doi.org/10.1016/j.isci.2019.08.057.

54. Xu, Y., Yang, W., Shi, J., and Zetter, B.R. (2016). Prohibitin 1 regulates tumor cell apoptosis via the interaction with X-linked inhibitor of apoptosis protein. J. Mol. Cell Biol. *8*, 282–285. https://doi.org/10.1093/jmcb/mjw018.

55. Pleiner, T., Tomaleri, G.P., Januszyk, K., Inglis, A.J., Hazu, M., and Voorhees, R.M. (2020). Structural basis for membrane insertion by the human ER membrane protein complex. Science *369*, 433–436. https://doi.org/10.1126/science.abb5008.

56. Zong, S., Wu, M., Gu, J., Liu, T., Guo, R., and Yang, M. (2018). Structure of the intact 14-subunit human cytochrome c oxidase. Cell Res. *28*, 1026–1034. https://doi.org/10.1038/s41422-018-0071-1.

57. Fasci, D., van Ingen, H., Scheltema, R.A., and Heck, A.J.R. (2018). Histone interaction landscapes visualized by crosslinking mass spectrometry in intact cell nuclei. Mol. Cell. Proteomics. *17*, 2018–2033. https://doi.org/10.1074/mcp.RA118.000924.

58. Zhu, S., Liu, K., Chai, Y., Wu, Y., Lu, D., Xiao, W., Cheng, H., Zhao, Y., Ding, C., Lyu, J., et al. (2019). Divergent peptide presentations of HLA-A*30 alleles revealed by structures with pathogen peptides. Front. Immunol. *10*, 1709. https://doi.org/10.3389/fimmu.2019.01709.

59. Gras, S., Wilmann, P.G., Chen, Z., Halim, H., Liu, Y.C., Kjer-Nielsen, L., Purcell, A.W., Burrows, S.R., McCluskey, J., and Rossjohn, J. (2012). A structural basis for varied αβ TCR usage against an immunodominant EBV antigen restricted to a HLA-B8 molecule. J. Immunol. *188*, 311–321. https://doi.org/10.4049/jimmunol.1102686.

60. Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics *32*, 511–517. https://doi.org/10.1093/bioinformatics/btv639.

61. Chong, C., Müller, M., Pak, H., Harnett, D., Huber, F., Grun, D., Leleu, M., Auger, A., Arnaud, M., Stevenson, B.J., et al. (2020). Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. Nat. Commun. *11*, 1293. https://doi.org/10.1038/s41467-020-14968-9.

62. Ruiz Cuevas, M.V., Hardy, M.-P., Hollý, J., Bonneil, É., Durette, C., Courcelles, M., Lanoix, J., Côté, C., Staudt, L.M., Lemieux, S., et al. (2021). Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep. *34*, 108815. https://doi.org/10.1016/j.celrep.2021.108815.

63. Bhuiyan, T., and Timmers, H.T.M. (2019). Promoter recognition: putting TFIID on the spot. Trends Cell Biol. *29*, 752–763. https://doi.org/10.1016/j.tcb.2019.06.004.

64. Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S.I., Schneider, R., and Jensen, L.J. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database. *2014*, bau012. https://doi.org/10.1093/database/bau012.

65. Iizuka, M., and Stillman, B. (1999). Histone acetyltransferase HBO1 interacts with the ORC1 subunit of the human initiator protein. J. Biol. Chem. *274*, 23027–23034. https://doi.org/10.1074/jbc.274.33.23027.

66. Bell, S.P., Mitchell, J., Leber, J., Kobayashi, R., and Stillman, B. (1995). The multidomain structure of Orc1 p reveals similarity to regulators of DNA replication and transcriptional silencing. Cell *83*, 563–568.

67. Bogaert, A., Fijalkowska, D., Staes, A., Van de Steene, T., Demol, H., and Gevaert, K. (2022). Limited evidence for protein products of noncoding transcripts in the HEK293T cellular cytosol. Mol. Cell. Proteomics. *21*, 100264. https://doi.org/10.1016/j.mcpro.2022.100264.

68. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D.J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., et al. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. *50*, D543–D552. https://doi.org/10.1093/nar/gkab1038.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Goat Anti-Rabbit IgG H&L (HRP)	Abcam	Cat# ab6721; RRID:AB_955447
Monoclonal Mouse anti-Cytokeratin 18	Dako	Cat# M7010; RRID:AB_2133299
Monoclonal Mouse anti-Histone H3	Santa Cruz Biotechnology	Cat# sc-517576; RRID:AB_2848194
Monoclonal Mouse anti-Hsp70	Abcam	Cat# ab2787; RRID:AB_303300
Monoclonal Mouse anti-SP1	Santa Cruz Biotechnology	Cat# sc-420; RRID:AB_628271
Peroxidase AffiniPure Goat Anti-Mouse IgG (H+L)	Jackson Immuno Research	Cat# 115-035-146; RRID: AB_2307392
Polyclonal Chicken anti-Calreticulin	Abcam	Cat# ab2908; RRID:AB_303403
Chemicals, peptides, and recombinant proteins		
Acrylamide / Bis-Acrylamide Sol. Ratio 29/1	Euromedex	Cat# EU0063-B
Amersham Protran Western blotting membranes, nitrocellulose	Merck	Cat# GE10600002
Chymotrypsin, Sequencing Grade	Promega	Cat# V1062
Dimethyl sulfoxide (DMSO)	Sigma-Aldrich	Cat# D5879
Disuccinimidyl sulfoxide (DSSO)	Thermo Fisher Scientific	Cat# A33545
DL-Dithiothreitol (DTT)	VWR Life Science	Cat# 97063-760
DPBS, no calcium, no magnesium	Thermo Fisher Scientific	Cat# 14190-094
Iodoacetamide (IAA)	Sigma-Aldrich	Cat# I1149
PageBlue™ Protein Staining Solution	Thermo Fisher Scientific	Cat# 24620
Pierce Detergent Removal Resin	Thermo Fisher Scientific	Cat# 87780
Prigrow I Medium	Applied Biological Materials	Cat# TM001
TG-SDS 10X	Euromedex	Cat# EU0510
TRIS Biotech grade	Interchim	Cat# UP031657
Trypsin/Lys-C Mix, Mass Spec Grade	Promega	Cat# V5073
Urea Ultra-Pure	Euromedex	Cat# EU0014B
Critical commercial assays		
Subcellular Protein Fractionation for Cultured Cells	Thermo Fisher Scientific	Cat# 78840
Deposited data		
The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository.	This paper	PRIDE: PXD035764
Experimental models: Cell lines		
Human Immortalized Ovarian Epithelial Cell line (SV40)	Applied Biological Materials	Cat# T1074
Software and algorithms		
Biological General Repository for Interaction Datasets (BioGRID)	Oughtred et al., 2021.[23]	RRID:SCR_007393 http://www.thebiogrid.org/
ClueGO	Bindea et al., 2009.[33]	RRID:SCR_005748; https://apps.cytoscape.org/apps/cluego
CluePedia	Bindea, Galon and Mlecnik, 2013.[33]	RRID:SCR_015784; https://apps.cytoscape.org/apps/cluepedia
Cluspro 2.0	Kozakov et al., 2017.[40]	RRID:SCR_018248; https://cluspro.bu.edu/login.php

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Cytoscape 3.9.1	Shannon et al., 2003.[37]	RRID:SCR_003032; https://cytoscape.org
IntAct	Orchard et al., 2014.[24]	RRID:SCR_006944; http://www.ebi.ac.uk/intact
I-TASSER (Iterative Threading ASSEmbly Refinement)	Yang et al., 2015.[39]	RRID:SCR_014627; https://zhanggroup.org/I-TASSER/
NetMHC - 4.0	Andreatta and Nielsen, 2016.[60]	RRID:SCR_021651; https://services.healthtech.dtu.dk/service.php?NetMHC-4.0
OpenProt Protein Database 1.6	(Brunet et al., 2021.).[19]	https://www.openprot.org/p/ng/Home
OriginPro, Version 2022b	OriginLab Corporation	RRID:SCR_014212; https://www.originlab.com/
Proteome Discoverer 2.5	Thermo Fisher Scientific	RRID:SCR_014477; https://www.thermofisher.com/order/catalog/product/OPTON-31040
STRING app	Doncheva et al., 2019.[32]	http://apps.cytoscape.org/apps/stringapp
XlinkX 2.5 nodes for Proteome Discoverer 2.5	Thermo Fisher Scientific	https://www.thermofisher.com/order/catalog/product/OPTON-31047
YASARA view	YASARA Biosciences	RRID:SCR_017591; http://www.yasara.org/
yFiles Layout Algorithms	yWorks	https://apps.cytoscape.org/apps/yfileslayoutalgorithms
Other		
Amicon Ultra-0.5 Centrifugal Filter Unit 50 KDa	Merck	Cat# UFC505024
ZipTip with 0.6 µL C18 resin	Merck	Cat# ZTC18S096

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Michel Salzet (michel.salzet@univ-lille.fr).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[68] partner repository with the dataset identifier PXD035764.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

This study used a human immortalized ovarian epithelial cell line (SV40) (Applied Biological Materials; female; in this study referred to ovarian cells).

METHOD DETAILS

Cell culture

SV-40 cells were cultured in Prigrow I medium with 10% fetal bovine serum and 100 U/mL penicillin-streptomycin in a humidified air incubator at 37 °C under an atmosphere of 5% CO_2. The cells were harvested by trypsinization, centrifugated at 1000 rpm for 5 min and washed three times with DPBS and aliquoted.

In cellulo chemical cross-linking

A 50 mM stock solution of disuccinimidyl sulfoxide (DSSO) was prepared by dissolving 1 mg DSSO in 51.5 μL dry DMSO. Three million ovarian cells were resuspended in 200 μL of DPBS. The crosslinking reaction was performed with a final concentration of 2 mM of DSSO, at 37 °C and under gentle end-over-end stirring. The reaction was quenched after 1 h by adding 10 μL of 500 mM Tris-HCl pH 8.5 and gentle stirring for 30 min.

Protein subcellular fractionation and western blotting

In cellulo crosslinked cells ($3E^6$) were pelleted and the supernatant was discarded, leaving the cells as dry as possible. Thermo Scientific Subcellular Protein Fractionation Kit for Cultured Cells was employed to separate five different protein cell compartments. Cytoplasmic, membrane, nuclear, chromatin-bound and cytoskeletal proteins were extracted according to the manufacturer's instructions. To confirm the crosslinking reaction, 10 μL of proteins was mixed with 2x Laemmli buffer and loaded on a 4-12% SDS-PAGE gel. Proteins were migrated for 15 min at 70 V and then for 90 min at 120 V in Tris-Glycine-SDS buffer. After migration, the gel was stained with PageBlue Protein Staining Solution (Coomassie blue) for 1 hr. The gel was decolorated by washing with water and visualized in an Invitrogen iBright system. The decolorated gel was transferred onto a 0.45 μm nitrocellulose membrane in a tank transfer system for 2 hr at 290 mA in Towbin buffer (5 mM Tris, 192 mM glycine, 20% Methanol and 0.01% SDS). The transferred membrane was blocked with 5% milk powder containing 0.1% TBS-Tween-20 and incubated at 4 °C overnight with specific primary antibodies against Cytokeratin 18 (Dako, M7010), SP1 (Santa Cruz Biotechnology, sc-420), Histone H3 (Santa Cruz Biotechnology, sc-517576), Hsp70 (Abcam, ab2787), and Calreticulin (Abcam, ab2908). The matched HRP Anti-Rabbit (Abcam, ab6721) and Anti-Mouse (Jackson Immuno Research, 115-035-146) secondary antibodies were used to visualize proteins by incubation at room temperature for 1 h. The membranes were scanned by the Invitrogen iBright Imaging Systems (Thermo Fisher Scientific).

Enzymatic digestion

Filter Aided Sample Preparation (FASP) was performed in a 50 KDa cut-off Amicon filter. The resulting fractions were transferred to the Amicon filter, concentrated by centrifugation (14,000 g x 15 min), and 100 μL of denaturing buffer (8 M Urea, 100 mM Tris-HCl, pH 8.5) was added. Reduction was performed by adding 100 μL of 100 mM Dithiothreitol (DTT) in denaturing buffer at 56 °C for 40 min. Alkylation was done by adding 100 μL of 50 mM Iodoacetamide in denaturing buffer at room temperature (RT) for 30 min in the dark. For sequential digestion, 40 μL of 40 ng/μL Trypsin/Lys-C Mix, Mass Spec Grade was added to the filter and incubated at 37 °C overnight followed by 25 μL of 40 ng/μL Chymotrypsin, Sequencing Grade at RT and for 4 h. The resulting peptides were then acidified with 0.1%TFA and vacuum dried.

NanoLC-MS/MS analysis

Dried samples were resuspended in 20 μL of 0.1% TFA and desalted on a ZipTip with C18 resin, following the manufacturer's instructions. The samples were then vacuum-dried and resuspended in 20 μL of acetonitrile (ACN)/0.1% FA (2:98, v/v). Five microliters of peptides were separated with a nanoAcquity (Waters) chromatography equipped with a C18 precolumn (180 μm × 20 mm, 5 μm DP, Waters) and BEA C18 analytical column (25 cm, 75 μm ID, 1.7 μL DP, Waters) using a gradient of ACN from 5% to 20 % in 100 min, from 20% to 30% in 20 min and then to 90% for 20 minat 300 nL/min. A Thermo Scientific Q-Exactive mass spectrometer was used for MS acquisition. The instrument was set to acquire the ten most intense precursors in data-dependent acquisition mode, with a voltage of 2.2 kV. The survey scans were set at positive mode, with a resolving power of 70,000 at FWHM (m/z 400), a scan range of 300 to 1,600 m/z, AGC target of $3x10^6$ and stepped NCE of 21, 24 and 30. For MS/MS, 1 microscan was obtained at 35,000 FWHM and dynamic exclusion was enabled. The instrument was set to perform MS/MS only from >+2 and <+8 charge states.

Shotgun data analysis

RAW data obtained by nanoLC-MS/MS analysis were analyzed using Sequest HT in Proteome Discoverer V2.5 (Thermo Scientific) with the following processing and consensus parameters: trypsin and chymotrypsin as enzymes, two missed cleavages, methionine oxidation and N-terminus acetylation as variable modifications, carbamidomethylation of cysteines as static modification, minimum peptide length of 6 amino acids, minimum precursor mass tolerance: 10 ppm and fragment mass tolerance: 0.02 Da.

For RefProts, the protein database used was *Homo sapiens* UniProtKB v.2022_02 reviewed and unreviewed. Validation of Sequest results was performed using Percolator with a strict FDR set to 1%. A consensus workflow was then applied for the filtering and results reporting. At the consensus step, the peptide validation for PSM and peptides was established between 0.01 and 0.05 FDR with a minimum peptide length of six. The minimum number of peptide sequences for a protein was selected as two. Finally, at the Protein FDR Validator validator, the target FDR was set as 0.01.

For AltProts, the protein database used was *Homo sapiens* OpenProt v1.6 database which contains RefProts and predicted AltProts detected in mass spectrometry experiments with at least one unique peptide leading to a total of 184,706 sequences. Validation of Sequest results was performed using Percolator with a strict FDR set to 1%. At the consensus step, the peptide validation for PSM and peptides was established between 0.01 and 0.05 FDR. Peptide confidence set at high with a minimum peptide length of six. The minimum number of peptide sequences for a protein was selected as one. Finally, at the Protein FDR Validator, the target FDR was set as 0.01. The Identified AltProts were Blasted against the non-redundant protein sequences. Finally, the peptides identified as unique peptides by Sequest HT were also corroborated by hand (Figure S3) and at NextProt Peptide uniqueness checker tool.

Crosslink data analysis

The obtained data were analyzed using the XlinkX algorithm of the Heck Lab (Utrecht, Netherland) at Proteome Discoverer V2.5 (Thermo Scientific). DSSO (158.0037 Da) was defined as the crosslinker. The protein database used was the *Homo sapiens* OpenProt v1.6 database which contains RefProts and predicted AltProts detected in mass spectrometry experiments with at least one unique peptide leading to a total of 184,706 sequences. First, protein identification was made by Sequest HT considering the following parameters: Trypsin/LysC and Chymotrypsin as enzymes, maximum two missed cleavages, peptide length from 6 to 150, precursor mass tolerance of 10 ppm and fragment mass tolerance as 0.02 Da. The dynamic modifications included were methionine oxidation, cysteine carbamidomethylation, N-terminus acetylation, DSSO amidated, hydrolyzed and Tris form. The validation was performed using Target decoy PSM validator with FDR set between 0.01 and 0.05. The XlinkX detections had the following parameters: precursor mass tolerance of 10 ppm, FTMS fragment of 20 ppm, ITMS fragment of 0.5 Da. The validation was performed with XlinkX/PD Validator set to 0.05.

At the consensus step, the peptide validation for PSM and peptides was established between 0.01 and 0.05 FDR. Peptide confidence set at high with a minimum peptide length of six. The minimum number of peptide sequences for a protein was selected as one. Finally, at the XlinkX consensus validator, the Crosslink Spectrum Match FDR threshold was 0.05 and the Cross-link FDR threshold of 0.05. and a minimum score of 20.

The protein-protein interactions were manually checked (Figure S4), to eliminate the crosslink spectrum matches that involved N-terminal residues (N=6). The Crosslinking network was displayed in Cytoscape 3.9.1. The protein identifiers were STRINGify using BioGrid, STRING, and IntAct app at Cytoscape, to verify existing interaction between the proteins displayed. For the identifiers that did not have any retrieved interaction, the expand network command was employed to add 3 protein interactors. The functional analysis employing biological process GO terms was performed at ClueGO app. The specificity of the network was set at medium +1 and GO term fusion was enabled. The resulting network was fused to the STRINGified network and the Organic yFiles Layout Algorithm was selected as layout.

Modeling and prediction of interactions between AltProts and RefProts

Structural models of AltProts were generated with I-TASSER (Iterative Threading ASSEmbly Refinement). Reference protein models were downloaded from the AlphaFold Protein Structure Database. AltProts models with C-score between -5 and +2 (most stable) generated by I-TASSER were considered for protein-protein interaction (PPI) prediction, which were generated by ClusPro. The RefProts were assigned as receptors and the AltProts as ligands. The docking interactions were generated without the crosslink influence. The resulting models were ranked by stability order and displayed by YASARA view. Using the data obtained from XlinkX, the distance between the lysine residues involved in the AltProt-RefProt crosslink was measured and displayed in the model.

For the interactions retrieved between AltProts and HLA family proteins, NetMHC was employed to identify if the AltProt or the identified peptide could bind to MHC proteins. The sequence of the AltProt interacting to the HLA was submitted and the length of the peptide was set between 8-14 amino acids. HLA-A or HLA-B alleles were selected respectively to each case. Strong binders were delimited by a % Rank below 0.5 and weak binders between 0.5 and 2% Rank. The results were filtered in which weak or strong binding was predicted. The modeling and docking of the peptide and the HLA protein were performed as described above.

QUANTIFICATION AND STATISTICAL ANALYSIS

To evaluate the difference between the Sequest HT Scores from RefProts and AltProts identified by at least one peptide (Figure S5). We employed the total nuclear extraction identifications (the most abundant fraction). A *t*-test with a significance P-value of 0.05 was used. We represented this difference using a boxplot, where the centerline of the boxplot indicates the median Sequest HT Score, the box edges represent the 25th and 75th percentiles, black squares represent the average, and each whisker extends to the most extreme data point that is not an outlier. Statistical analysis and boxplot were performed in OriginPro 2022b.

Supplemental Figures

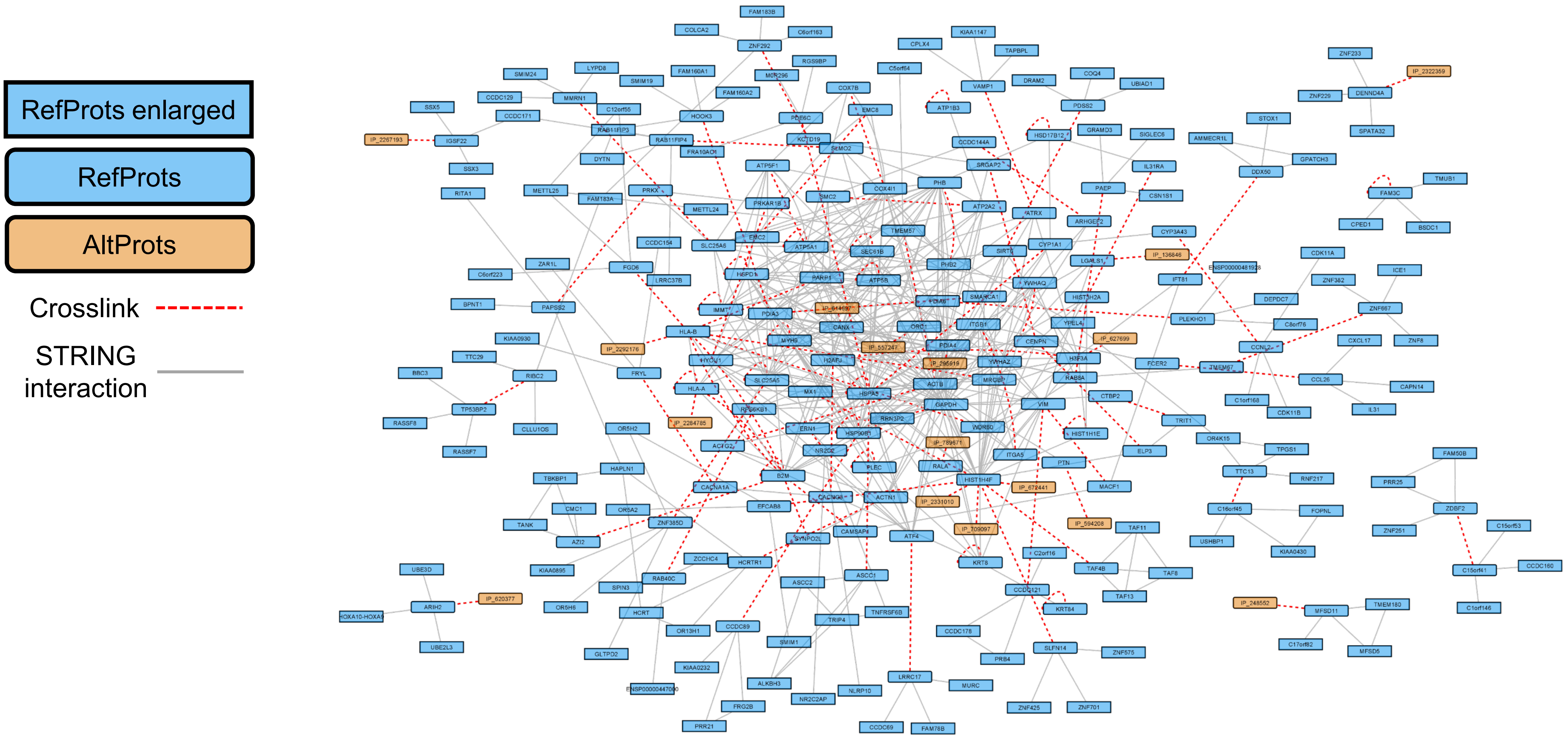
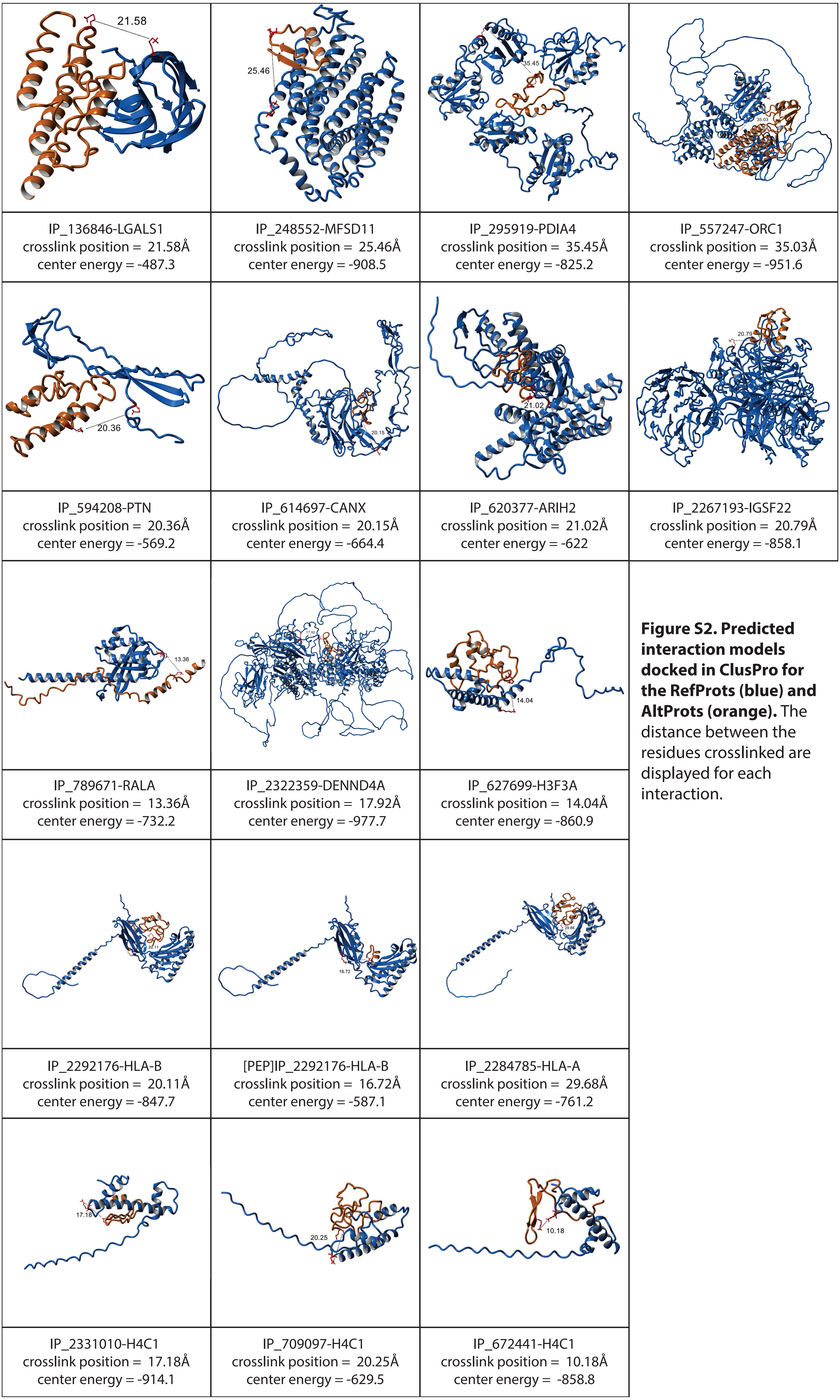


Figure S1. Crosslinked network enriched by the STRING interactions (gray lines) retrieved between these crosslinked (red dash lines) RefProts. For the RefProts that did not present referenced STRING interaction, an enrichment has been performed to expand the network.



IP_136846-LGALS1
crosslink position = 21.58Å
center energy = -487.3

IP_248552-MFSD11
crosslink position = 25.46Å
center energy = -908.5

IP_295919-PDIA4
crosslink position = 35.45Å
center energy = -825.2

IP_557247-ORC1
crosslink position = 35.03Å
center energy = -951.6

IP_594208-PTN
crosslink position = 20.36Å
center energy = -569.2

IP_614697-CANX
crosslink position = 20.15Å
center energy = -664.4

IP_620377-ARIH2
crosslink position = 21.02Å
center energy = -622

IP_2267193-IGSF22
crosslink position = 20.79Å
center energy = -858.1

IP_789671-RALA
crosslink position = 13.36Å
center energy = -732.2

IP_2322359-DENND4A
crosslink position = 17.92Å
center energy = -977.7

IP_627699-H3F3A
crosslink position = 14.04Å
center energy = -860.9

IP_2292176-HLA-B
crosslink position = 20.11Å
center energy = -847.7

[PEP]IP_2292176-HLA-B
crosslink position = 16.72Å
center energy = -587.1

IP_2284785-HLA-A
crosslink position = 29.68Å
center energy = -761.2

IP_2331010-H4C1
crosslink position = 17.18Å
center energy = -914.1

IP_709097-H4C1
crosslink position = 20.25Å
center energy = -629.5

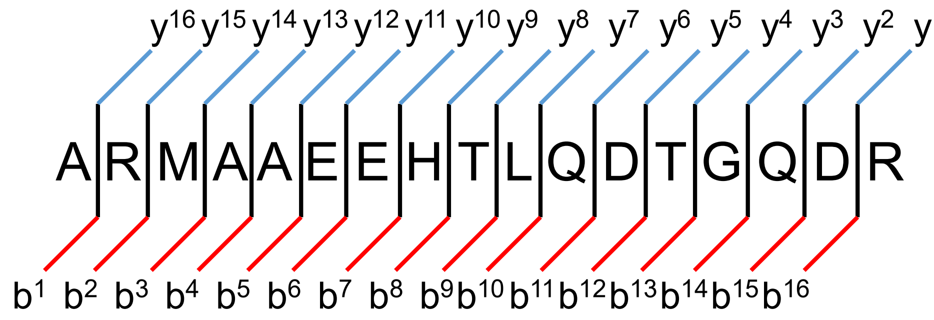
IP_672441-H4C1
crosslink position = 10.18Å
center energy = -858.8

Figure S2. Predicted interaction models docked in ClusPro for the RefProts (blue) and AltProts (orange). The distance between the residues crosslinked are displayed for each interaction.

IP_756980

MEKQLLEELE RQRQAELAAQ KARERKLARM
AAEEHTLQDT GQDRGRTCKT

Protein BLAST: No significant similarity was found NextProt Peptide uniqueness checker: Peptide not found


y16 y15 y14 y13 y12 y11 y10 y9 y8 y7 y6 y5 y4 y3 y2 y1
A R M A A E E H T L Q D T G Q D R
b1 b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12 b13 b14 b15 b16

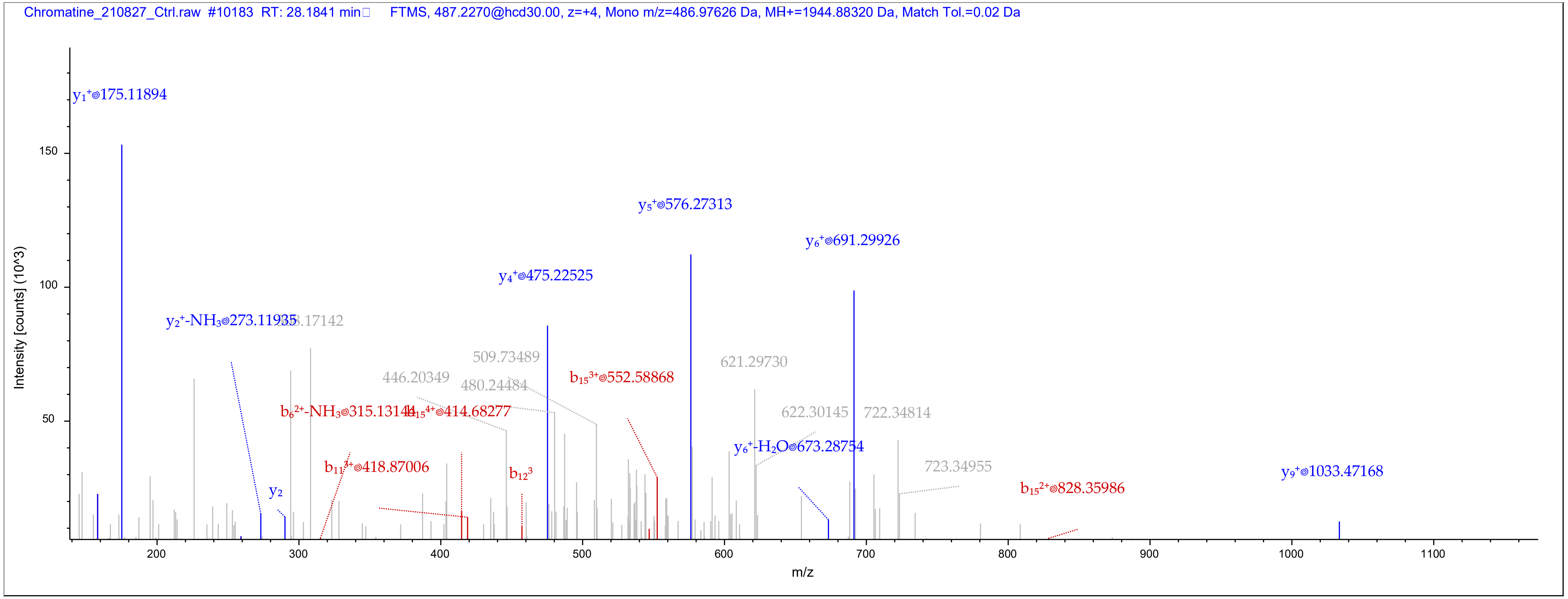
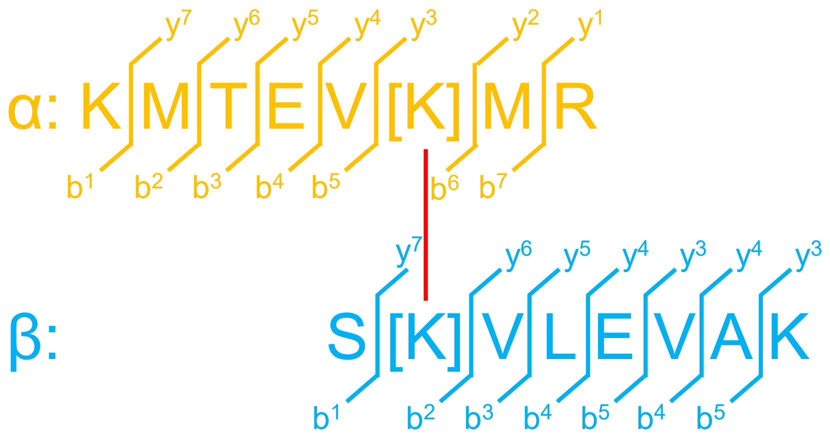


Figure S3. Tandem mass spectrum for the peptide ARMAAEEHTLQDT GQDRGRTCKT which was unique for the AltProt IP_756980. In red (b ions) and in blue (y ions) are observed in the spectra. No significant similarity was found for this AltProt in protein BLAST. Additionally, the peptide was not identified at NextProt Peptide uniqueness checker.

(α) IP_295919 — (β) PDIA4

Max. XlinkX Score: 46.35
CSM:1

NextProt Peptide uniqueness checker: Peptide not found



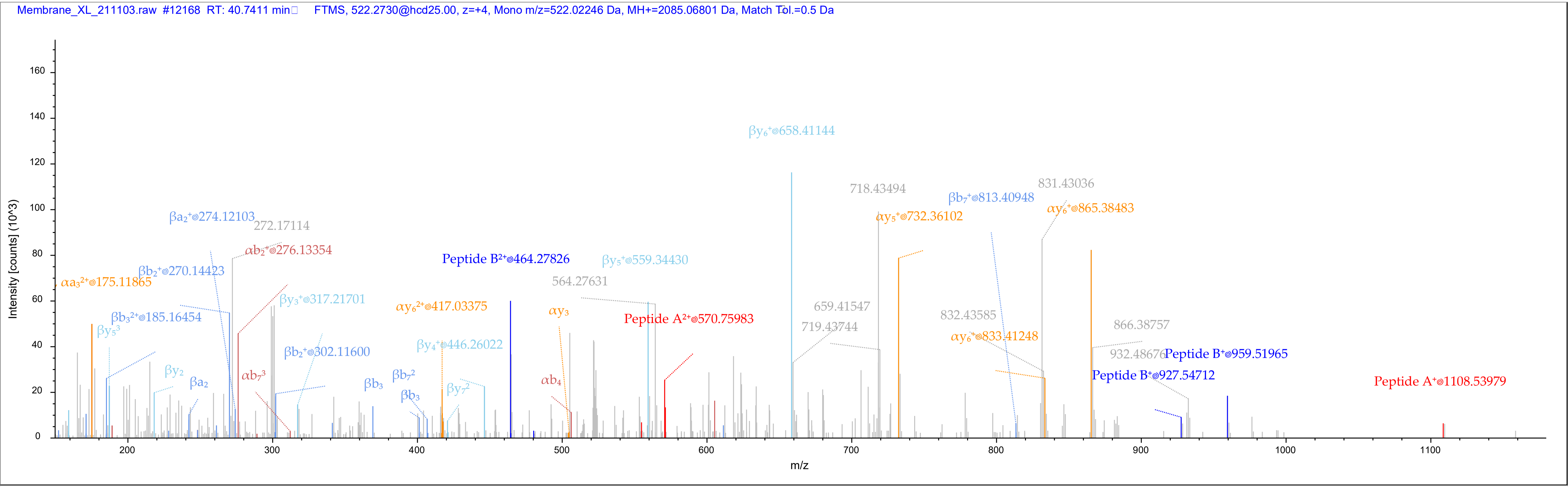


Figure S4. Crosslinking tandem mass spectrum which identified the interaction between IP_295919 (orange) and PDIA4 (blue). The b and y ions for the individual and crosslinking peptides are observed. The unique peptide of IP_295919 was not identified at NextProt.

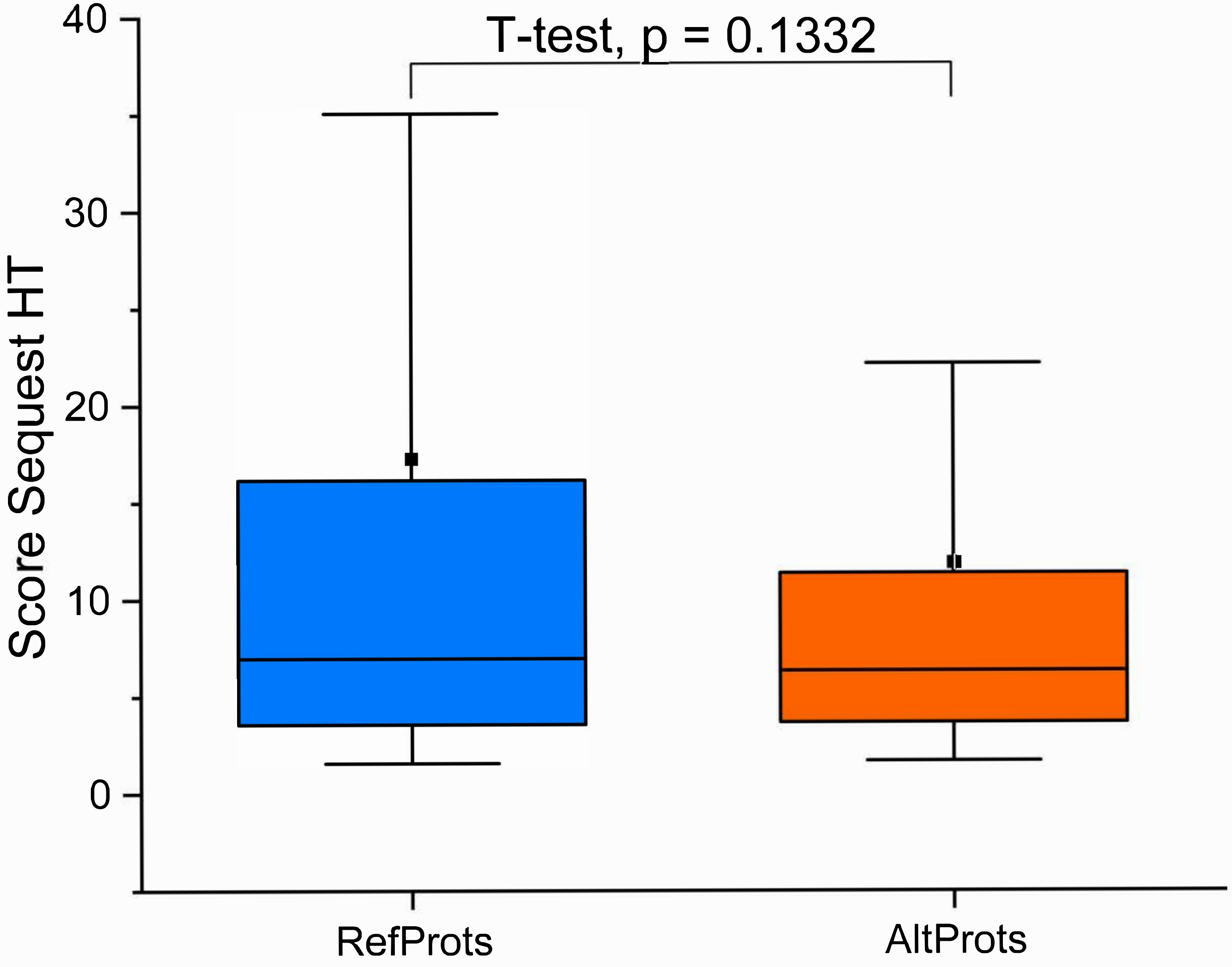


Figure S5. Boxplot of the Sequest HT Scores from the RefProts (blue) and AltProts (orange) identified using at least one peptide and OpenProt from the nuclear fraction. Between them there is no significant difference (T-test, p = 0.1332) between the Sequest HT scores

Supplemental Tables

Table S1. List of the alternative proteins (AltProts) identified in the subcellular fractionation experiments. For each AltProt the subcellular compartment in which it was found and the complete AltProts description is showed.

Table S2. List of protein-protein interactions (PPIs) identified. For each PPI, the XlinkX score, type of crosslink, number of crosslinking spectrum matches (CSMs), proteins involved, and the subcellular compartment in which it was found are displayed.

Table S3. List and complete description of the AltProts identified to be crosslinked to a reference protein (RefProts).

Compartment found	Protein accession	Protein length (a.a.)	Molecular weight (kDa)	Isoelectric point	Gene symbol	Annotation	Genomic coordinates	Strand	Transcript accession	RNA type	Localization	Frame	Transcript coordinates	Kozak motif	Blast	Unique peptide	Domains	Domain
Nuc	IP_169121	51	6.07	6.66	LINC01289	GRCh38.p12	8:63769683-63769838	+	ENST00000519550.1	ncRNA	-	2	254-410	-	No significant similarity found	MNVHVEVR	0	
									NR_038875.2	ncRNA	-	2	254-410					
Memb	IP_223184	446	50.34	5.23	POTEM	GRCh38.p12	14:18997148-18999995	+	NM_001145442.1	mRNA	CDS	3	1569-2910	-	alternative protein POTEM	SFTTMAEQEIVR	19	Actins and actin-related proteins signature.
Cyt	IP_2309391	74	8.15	10.46	LOC105369580	GRCh38.p12	11:132972946-132979674	+	XR_948206.2	ncRNA	-	2	425-650	-	No significant similarity found	SAQSLENIR	0	
Cyt	IP_274177	179	20.2	11.59	ZNF568	GRCh38.95	19:36997211-36997750	+	ENST00000455427.6	mRNA	CDS	1	1657-2197	-	No similarity > 80% found	VVAQNLPDIR	0	
									ENST00000591887.1	ncRNA	-	3	1689-2229					
									ENST00000617745.4	mRNA	CDS	1	1720-2260					
									NM_001204838.1	mRNA	CDS	1	2038-2578					
									NM_001204839.1	mRNA	CDS	1	1777-2317					
									XM_017026772.1	mRNA	CDS	1	2626-3166					
									XM_017026775.2	mRNA	CDS	2	1655-2195					
									XM_017026776.1	mRNA	CDS	2	1733-2273					
									XM_017026777.1	mRNA	CDS	3	1629-2169					
									XM_017026778.1	mRNA	CDS	1	1774-2314					
Cyt	IP_558777	167	20.01	4.26	KRT8P17	GRCh38.95	X:57985034-57985537	+	ENST00000453110.1	ncRNA	-	1	349-853	-	92 % KRT8 protein	QLETLGR	8	Intermediate filament protein
Chr	IP_561478	372	41.77	4.5	TUBBP5	GRCh38.95	9:138175651-138177264	+	ENST00000290377.9	ncRNA	-	3	276-1395	+	96% tubulin beta 8B isoform 1	TAAAIFQGRMPMR	32	Tubulin/FtsZ family, GTPase domain
									ENST00000508529.3	ncRNA	-	3	216-1335					
Cyt	IP_564603	175	19.07	7.5	RPSAP9	GRCh38.p12	9:76398699-76399226	+	ENST00000418207.1	ncRNA	-	1	1-529	-	96.5% RPSA	LAASAIVAIENPADVSVISSR	14	Ribosomal protein S2
									NR_026890.1	misc_RNA	-	2	101-629					
Memb	IP_566083	222	24.19	10.07	SLC25A6P2	GRCh38.95	9:31253901-31254569	+	ENST00000497974.1	ncRNA	-	1	1-670	-	85.96% ADP/ATP translocase 3	TAVASIKRVQL	19	Mitochondrial carrier protein
Cyt	IP_582251	158	17.2	4.24	HSPA8P8	GRCh38.95	7:10452409-10452885	+	ENST00000419827.2	ncRNA	-	1	1099-1576	-	92% heat shock protein family A (Hsp70) member 8	GIETASGVMTILINCNTTIPTKQTQTF	13	70kDa heat shock protein signature
Memb	IP_584241	176	20.45	11.03	RPL7AP38	GRCh38.95	7:26922161-26922691	+	ENST00000441433.1	ncRNA	-	1	52-583	-	No similarity > 80% found	KAQLVVTVHDVDPIK	13	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family
Cyt, Memb	IP_591993	74	8.55	4.94	PGAM1P10	GRCh38.95	6:73055097-73055321	+	ENST00000402426.2	ncRNA	-	1	1-226	+	80.52% PGAM1	IP_591993	5	Histidine phosphatase superfamily (branch 1)
Chr	IP_602534	203	23.9	4.56	KRT8P32	GRCh38.95	5:98392738-98393349	-	ENST00000512863.1	ncRNA	-	2	182-794	+	93% keratin 8, partial	MGGITAITFNQSL	13	Intermediate filament protein
Ske	IP_602541	179	19.46	7.95	AC233964.1	GRCh38.95	5:98338748-98339287	+	ENST00000512165.1	ncRNA	-	2	5-545	-	84% actin gamma 1	PASQVTMPPGHLPLCHGCQGVMVGMGQK	12	Actin signature
Memb	IP_603389	307	32.92	8.09	EEF1A1P19	GRCh38.95	5:43495073-43495996	-	ENST00000513637.1	ncRNA	-	3	459-1383	+	95% Elongation factor 1-alpha 1	KSGDAAIVDMVPGKPIVESF	10	Elongation factor Tu C-terminal domain
Chr	IP_604016	68	7.72	10.28	HSPD1P15	GRCh38.95	5:19233725-19233931	+	ENST00000505573.2	ncRNA	-	3	360-567	-	No similarity > 80% found	EETAQDATISANREK	2	Chaperonin Cpn60/TCP-1 family
Memb	IP_612331	64	6.68	9.21	HSPD1P5	GRCh38.95	4:144846059-144846253	+	ENST00000511127.1	ncRNA	-	3	435-630	-	87% heat shock protein family D (Hsp60) member 1	KVGSKGIITVNNGK	3	Chaperonin Cpn60/TCP-1 family
Chr	IP_613285	78	8.78	6.77	AC108941.1	GRCh38.95	4:69888943-69889179	+	ENST00000502853.2	ncRNA	-	1	67-304	-	No similarity > 80% found	RISGHMHILTR	1	Mitochondrial carrier domain
Memb, Cyt, Chr, Ske, Nuc	IP_623199	236	26.79	7.5	KRT8P25	GRCh38.95	3:87323268-87323978	+	ENST00000473150.1	ncRNA	-	1	1-712	-	83% keratin 8	KVVDDAYMNK	13	Intermediate filament protein
Nuc	IP_624445	137	15.52	4.85	KRT18P15	GRCh38.95	3:32259817-32260230	+	ENST00000458108.1	ncRNA	-	2	878-1292	-	94% cytokeratin 18	RVDGKVVSETNDTK	8	Intermediate filament protein
Memb, Nuc	IP_625997	293	31.62	8.09	HSPD1P6	GRCh38.95	3:36767831-36768712	-	ENST00000388967.3	ncRNA	-	1	88-970	-	89% 60 kDa heat shock protein, mitochondrial	EIGHILSDAVKK	8	TCP-1/cpn60 chaperonin family
Memb	IP_635611	327	37.79	6.22	EIF2S2P4	GRCh38.95	2:170751805-170752788	+	ENST00000461070.1	ncRNA	-	1	1-985	-	96.7% eukaryotic translation initiation factor 2 subunit 2 isoform 1	LQCETCHSRCSVAIIK	6	domain present in translation initiation factor eIF2B and Eif5
Chr, Nuc	IP_638540	251	26.77	8.29	EEF1A1P12	GRCh38.95	2:106697331-106698086	-	ENST00000443590.1	ncRNA	-	3	591-1347	+	94.96% EEF1A1 protein	IGGIGSVPVGR	14	Elongation factor Tu domain 2
Chr, Memb	IP_667299	258	28.91	7.55	EEF1A1P32	GRCh38.95	1:197688760-197690057	+	ENST00000418613.2	ncRNA	-	1	1-778	+	No similarity > 80% found	ASTLIVKQPTVHVNK	10	Elongation factor Tu GTP binding domain
Nuc, Ske	IP_667319	259	27.69	7.2	EEF1A1P14	GRCh38.95	1:194189559-194190338	+	ENST00000421195.1	ncRNA	-	2	593-1373	-	84.81% EEF1A1	KPGIVVTFAPVNITTEVK	8	Translation elongation factor EFTu/EF1A, domain 2
Nuc	IP_668634	165	18.79	10.45	RPL4P3	GRCh38.95	1:171683528-171684025	-	ENST00000428881.1	ncRNA	-	3	414-912	-	90.3% 60S ribosomal protein L4	YIIYNEDNGIIK	5	Ribosomal protein L4/L1 family
Memb	IP_669854	317	34.57	5.39	PKMP1	GRCh38.95	1:114535995-114536948	+	ENST00000455804.1	ncRNA	-	1	1-955	-	91% pyruvate kinase PKM	LEHMCHLDIHTPPIPAW	14	Pyruvate kinase, barrel domain
Chr, Nuc	IP_672475	287	33.17	4.71	KRT8P47	GRCh38.95	1:44103652-44104515	-	ENST00000604727.1	ncRNA	-	3	189-1053	-	85.42% keratin, type II cytoskeletal 8	TEMENQFVLIK	18	Intermediate filament protein
Nuc	IP_689114	394	44.7	6.33	AC078899.1	GRCh38.95	19:20257720-20259418	+	ENST00000521432.2	ncRNA	-	1	1-1186	+	93.91% actin-related protein	GLLTAVVVDSGDGVTHIY	16	Actin family
Memb, Nuc	IP_701684	48	5.7	5.68	AC090340.1	GRCh38.95	18:57431641-57431787	+	ENST00000620778.1	ncRNA	-	3	4509-4656	-	No significant similarity found	ISVEELHR	0	
Cyt, Memb, Nuc	IP_721455	187	21.62	4.32	TUBB8P7	GRCh38.95	16:90095626-90096189	+	ENST00000563927.1	ncRNA	-	3	960-1524	+	96.95% tubulin beta-8	TEMENQFVLIK	14	Beta-tubulin signature
									ENST00000564451.1	ncRNA	-	1	1387-1951					
Chr	IP_723351	287	31.98	4.55	KRT8P22	GRCh38.95	16:78048621-78049484	-	ENST00000562317.1	ncRNA	-	3	414-1278	-	No similarity > 80% found	TLDMDSIITEVR	11	Intermediate filament protein
Cyt, Memb	IP_746205	41	5.19	4.39	RPSAP3	GRCh38.95	14:76635018-76635143	-	ENST00000464073.1	ncRNA	-	3	519-645	-	92.68% 40S ribosomal protein	EVLRMHGTISR	2	Ribosomal protein S2, eukaryotic/archaeal
Memb	IP_746478	319	35.43	10.47	YBX1P1	GRCh38.95	14:66012830-66013789	-	ENST00000445822.1	ncRNA	-	1	1-961	-	95.67% Y-box-binding protein 1	GAKAANVTGPGGVPVQGSKY	9	'Cold-shock' DNA-binding domain
Cyt	IP_758382	189	20.27	4.93	EEF1A1P3	GRCh38.95	13:28214818-28215387	+	ENST00000417549.1	ncRNA	-	3	306-876	-	86-17% elongation factor 1-alpha 1	NPDTVAFVPFAGW	8	Elongation factor Tu GTP binding domain
Cyt	IP_761273	351	38.65	8.33	LDHAL6CP	GRCh38.95	12:63003553-63004608	+	ENST00000502352.1	ncRNA	-	1	1-1057	-	86.59% L-lactate dehydrogenase A-like	NVHKEVIASAYEIIEMK	15	lactate/malate dehydrogenase, NAD binding domain
Memb	IP_762241	295	32.74	4.19	RPSAP12	GRCh38.95	12:68552995-68553882	-	ENST00000435879.2	ncRNA	-	1	1-889	-	96.27% 40S ribosomal protein SA	TAIQPEVADWSEGVQVPSVPIQQFPTEDW	17	Ribosomal protein S2
Cyt	IP_762707	158	18.61	6.88	EIF4A1P4	GRCh38.95	12:53153763-53154239	+	ENST00000550578.1	ncRNA	-	3	492-969	-	94.50% eukaryotic initiation factor 4A-I isoform 2	LSAAMPSDVLEVTK	7	Superfamilies 1 and 2 helicase ATP-binding type-1 domain profile.
Chr, Cyt, Memb, Nuc, Ske	IP_774693	75	8.68	3.92	TUBAP2	GRCh38.95	11:90283776-90284003	+	ENST00000530835.1	ncRNA	-	2	1217-1445	-	98% tubulin alpha-4A chain isoform 1	MLSNTTAIAEAWAR	5	
Memb	IP_775141	183	21.27	5.29	FTH1P16	GRCh38.95	11:77734475-77735026	-	ENST00000435241.1	ncRNA	-	1	1-553	-	95.63% ferritin heavy chain	ESGLNVMECALHL	8	Ferritin-like domain
Cyt	IP_775646	318	34.31	5.84	RPLP0P2	GRCh38.p12	11:61636680-61637636	+	ENST00000490750.1	ncRNA	-	1	1-958	-	89.47% 60S acidic ribosomal protein	KSGDAAIVDMVPGKPIVESF	6	Ribosomal protein L10
									ENST00000496593.5	ncRNA	-	3	756-1713					
									NR_002775.2	misc_RNA	-	3	756-1713					
Cyt, Memb	IP_777545	188	21.95	4.46	KRT8P49	GRCh38.95	11:4723255-4723821	-	ENST00000604166.1	ncRNA	-	3	351-918	-	85.56% keratin, type II cytoskeletal 8	QMLAGKQEDDPQR	9	Intermediate filament protein
Memb, Nuc	IP_781261	171	18.37	10.67	AC139143.1	GRCh38.95	11:1803239-1803754	-	ENST00000441772.1	ncRNA	-	3	342-858	-	No similarity > 80% found	SKYTPGHTTGIMMNSSDGVTCTVSIY	6	Actin family
Chr, Cyt, Memb, Nuc, Ske	IP_790379	42	4.38	3.71	AL161932.1	GRCh38.95	10:32111661-32111789	-	ENST00000437408.1	ncRNA	-	1	1-130	+	88.10% ras-related protein Rab-18	ILIIGESGAGK	8	Ras family
Cyt, Memb	IP_091037	76	8.71	11.95	NEB	GRCh38.p12	2:151679807-151682666	-	ENST00000172853.14	mRNA	CDS	1	3142-3373	-	No significant similarity found	KNIANIQTPSSLPR	0	
									ENST00000397345.7	mRNA	CDS	1	3142-3373					
									ENST00000409198.5	mRNA	CDS	1	3142-3373					
									ENST00000427231.6	mRNA	CDS	1	3142-3373					
									ENST00000603639.5	mRNA	CDS	2	2939-3170					
									ENST00000604864.5	mRNA	CDS	2	2939-3170					
									ENST00000618972.4	mRNA	CDS	1	3142-3373					
									NM_001164507.1	mRNA	CDS	1	3142-3373					
									NM_001164508.1	mRNA	CDS	1	3142-3373					
									NM_001271208.1	mRNA	CDS	1	3142-3373					
									NM_004543.4	mRNA	CDS	1	3142-3373					
									XM_005246590.2	mRNA	CDS	3	3129-3360					
									XM_005246591.2	mRNA	CDS	3	3129-3360					
									XM_005246592.2	mRNA	CDS	3	3129-3360					
									XM_005246593.2	mRNA	CDS	3	3129-3360					
									XM_005246594.2	mRNA	CDS	3	3129-3360					
									XM_005246596.2	mRNA	CDS	3	3129-3360					
									XM_005246597.2	mRNA	CDS	3	3129-3360					
									XM_005246598.2	mRNA	CDS	3	3129-3360					
									XM_005246599.2	mRNA	CDS	3	3129-3360					
									XM_005246601.2	mRNA	CDS	3	3129-3360					
									XM_005246602.2	mRNA	CDS	3	3129-3360					
									XM_005246603.2	mRNA	CDS	3	3129-3360					
									XM_005246604.2	mRNA	CDS	3	3129-3360					
									XM_005246606.2	mRNA	CDS	3	3129-3360					

									XM_005246608.2	mRNA	CDS	3	3129-3360					
									XM_005246610.2	mRNA	CDS	3	3129-3360					
									XM_005246611.2	mRNA	CDS	3	3129-3360					
									XM_005246612.2	mRNA	CDS	3	3129-3360					
									XM_005246613.2	mRNA	CDS	3	3129-3360					
									XM_005246615.2	mRNA	CDS	3	3129-3360					
									XM_005246616.1	mRNA	CDS	3	3129-3360					
									XM_005246617.2	mRNA	CDS	3	3129-3360					
									XM_006712541.2	mRNA	CDS	3	3129-3360					
									XM_006712542.2	mRNA	CDS	3	3129-3360					
									XM_011511225.2	mRNA	CDS	3	3129-3360					
									XM_011511226.2	mRNA	CDS	3	3129-3360					
									XM_011511227.2	mRNA	CDS	3	3129-3360					
									XM_017004177.1	mRNA	CDS	3	3129-3360					
									XM_017004178.1	mRNA	CDS	3	3129-3360					
									XM_017004179.1	mRNA	CDS	3	3129-3360					
									XM_017004180.1	mRNA	CDS	3	3129-3360					
									XM_017004181.1	mRNA	CDS	3	3129-3360					
									XM_017004182.1	mRNA	CDS	3	3129-3360					
									XM_017004183.1	mRNA	CDS	3	3129-3360					
									XM_017004184.1	mRNA	CDS	3	3129-3360					
									XM_017004185.1	mRNA	CDS	3	3129-3360					
Nuc	IP_133955	73	8.55	8.08	FAM196B	GRCh38.p12	5:169980397-169980618	-	ENST00000377365.3	mRNA	5'UTR	3	123-345	-	No significant similarity found	EVFSFHMVAAEQF	0	
									NM_001129891.2	mRNA	5'UTR	2	140-362					
									NM_001346304.1	mRNA	5'UTR	2	140-362					
Memb	IP_237799	63	7.45	12.81	SYNM	GRCh38.p12	15:99130624-99130815	+	ENST00000336292.10	mRNA	CDS	2	2429-2621	-	No significant similarity found	VMEKAMSHSQL	0	
									ENST00000560674.5	mRNA	CDS	3	1875-2067					
									ENST00000594047.2	mRNA	CDS	2	2384-2576					
									NM_015286.5	mRNA	CDS	2	2384-2576					
									NM_145728.2	mRNA	CDS	2	2384-2576					
									XM_017022035.1	mRNA	CDS	1	2395-2587					
Chr	IP_274399	156	16.93	5.14	ZNF607	GRCh38.p12	19:37696932-37697402	-	ENST00000355202.8	mRNA	3'UTR	1	3325-3796	+	87.18% Prohibitin 1	AVAGGMVNSALCNVDAGHRAAIF	13	Prohibitin signature
									ENST00000395835.7	mRNA	3'UTR	1	2974-3445					
									NM_001172677.1	mRNA	3'UTR	1	2974-3445					
									NM_032689.4	mRNA	3'UTR	1	3325-3796					
									XM_006723435.4	mRNA	3'UTR	3	3954-4425					
									XM_006723436.3	mRNA	3'UTR	1	3280-3751					
Cyt	IP_304484	32	3.66	11.48	ZNF185	GRCh38.p12	X:152972198-152972296	+	ENST00000318504.11	mRNA	3'UTR	1	2866-2965	-	No significant similarity found	EVAIALR	0	
									ENST00000318529.12	mRNA	3'UTR	1	2596-2695					
									ENST00000324823.10	mRNA	3'UTR	2	2324-2423					
									ENST00000370268.8	mRNA	3'UTR	2	3032-3131					
									ENST00000426821.6	mRNA	3'UTR	2	2270-2369					
									ENST00000449285.6	mRNA	3'UTR	1	3046-3145					
									ENST00000454925.1	mRNA	3'UTR	3	1917-2016					
									ENST00000535861.5	mRNA	3'UTR	1	3139-3238					
									ENST00000539731.5	mRNA	3'UTR	1	3052-3151					
									NM_001178106.1	mRNA	3'UTR	1	3139-3238					
									NM_001178107.1	mRNA	3'UTR	1	3052-3151					
									NM_001178108.1	mRNA	3'UTR	1	3046-3145					
									NM_001178109.1	mRNA	3'UTR	1	2956-3055					
									NM_001178110.1	mRNA	3'UTR	1	2866-2965					
									NM_001178113.1	mRNA	3'UTR	1	2596-2695					
									NM_001178115.1	mRNA	3'UTR	3	1947-2046					
									NM_007150.3	mRNA	3'UTR	1	3043-3142					
									XM_005274730.2	mRNA	3'UTR	3	3246-3345					
									XM_005274731.2	mRNA	3'UTR	3	3243-3342					
									XM_005274735.2	mRNA	3'UTR	3	3156-3255					
									XM_005274737.2	mRNA	3'UTR	3	3150-3249					
									XM_005274738.2	mRNA	3'UTR	3	3147-3246					
									XM_005274740.2	mRNA	3'UTR	3	3060-3159					
									XM_005274741.2	mRNA	3'UTR	3	3060-3159					
									XM_005274742.2	mRNA	3'UTR	3	3057-3156					
									XM_005274744.2	mRNA	3'UTR	3	2973-3072					
									XM_005274745.2	mRNA	3'UTR	3	2970-3069					
									XM_005274746.3	mRNA	3'UTR	3	2889-2988					
									XM_011531194.1	mRNA	3'UTR	3	3051-3150					
									XM_011531195.1	mRNA	3'UTR	3	3048-3147					
									XM_017029821.1	mRNA	3'UTR	3	3378-3477					
									XM_017029822.2	mRNA	3'UTR	1	3397-3496					
									XM_017029823.1	mRNA	3'UTR	3	3375-3474					
									XM_017029824.1	mRNA	3'UTR	3	3294-3393					
									XM_017029825.1	mRNA	3'UTR	3	3291-3390					
									XM_017029826.1	mRNA	3'UTR	3	3288-3387					
									XM_017029827.1	mRNA	3'UTR	3	3285-3384					
									XM_017029828.1	mRNA	3'UTR	3	3282-3381					
									XM_017029829.1	mRNA	3'UTR	3	3201-3300					
									XM_017029830.1	mRNA	3'UTR	3	3195-3294					
									XM_017029831.1	mRNA	3'UTR	3	3192-3291					
									XM_017029832.1	mRNA	3'UTR	3	3183-3282					
									XM_017029833.1	mRNA	3'UTR	3	3117-3216					
									XM_017029834.1	mRNA	3'UTR	3	3105-3204					
									XM_017029835.1	mRNA	3'UTR	3	3066-3165					
Nuc	IP_3416431	445	51.34	8.24	XRCC6P5	GRCh38.p12	X:99719480-99785856	-	NR_024608.1	misc_RNA	-	1	529-1867	-	85.28% X-ray repair cross-complementing protein	ILMDSTSCPL	13	Ku70/Ku80 domain
Cyt, Memb	IP_556260	233	26.61	7.86	AL359263.1	GRCh38.95	X:93223652-93224353	-	ENST00000441188.2	ncRNA	-	3	663-1365	-	92.27% AP-1 complex subunit beta-1	VVQEVIVVIR	5	Adaptin like
Memb	IP_556266	137	15.31	7.58	KRT18P11	GRCh38.95	X:92459838-92460251	+	ENST00000435214.1	ncRNA	-	1	169-583	+	91.60% keratin 18	MGPGGLAAGMAGGLAGMGGIQNK	7	Intermediate filament protein
Memb	IP_556387	480	53.91	5.91	HK2P1	GRCh38.95	X:80572754-80574196	-	ENST00000447818.2	ncRNA	-	1	412-1855	-	94.79% hexokinase-2	GMDPTQEDCVATHRICQIVSTHSASL	19	Hexokinase
Cyt	IP_556624	267	30.3	5.1	KRT8P8	GRCh38.95	X:152479577-152480380	+	ENST00000434750.1	ncRNA	-	1	1-805	-	87.65% cytokeratin 8	DVDEADMNKVELESR	13	Intermediate filament protein
Memb	IP_557242	97	10.8	4.37	TUBB4AP1	GRCh38.95	X:123561724-123562017	+	ENST00000450755.1	ncRNA	-	2	83-377	+	86.46% tubulin beta 4A	VDLEPSTMDSILSGPF	11	Tubulin family, GTPase domain
Memb	IP_559479	205	23.36	10.79	KRT8P14	GRCh38.95	X:45633116-45633733	-	ENST00000414254.1	ncRNA	-	1	1-619	-	87% keratin 8	GGVSGMGGITTIMVKQSLL	5	Intermediate filament protein
Nuc	IP_561566	248	27.7	4.68	NCLP1	GRCh38.95	9:136813589-136815121	-	ENST00000445357.1	ncRNA	-	2	416-1163	-	No similarity > 80% found	MKAAAGAPASGDEDEDDEDDAEEDGSEAAPMETTPAK	6	RNA recognition motif domain
Nuc	IP_571469	295	32.71	4.45	RPSAP47	GRCh38.95	8:80558870-80559757	+	ENST00000479228.1	ncRNA	-	1	1-889	-	95.93% Ribosomal protein SA	RDPEEIEKEEQPAAEK	16	Ribosomal protein S2
Chr	IP_578661	93	10	11.11	AC009517.1	GRCh38.95	7:137164205-137164486	+	ENST00000454750.1	ncRNA	-	1	163-445	-	84.95% ribosomal protein L18	SLSWMIQNIK	3	Ribosomal protein L18e
Chr	IP_581070	181	20.04	11.06	AC091685.2	GRCh38.95	7:64141856-64142401	-	ENST00000454924.1	ncRNA	-	1	1-547	+	87.18% Ribosomal protein L6	IILTGCH	5	Ribosomal protein L6
Cyt, Memb, Nuc	IP_581916	154	16.64	8.84	SLC25A5P5	GRCh38.95	7:32471513-32471977	+	ENST00000435239.1	ncRNA	-	1	1-466	-	88.31% ADP/ATP translocase 2	MTDAAVSFTK	15	Mitochondrial carrier protein
Cyt	IP_586907	174	18.9	6.78	KRT8P44	GRCh38.95	6:162569204-162569728	+	ENST00000441609.1	ncRNA	-	1	826-1351	-	82.76% keratin, type II cytoskeletal 8	QQAMQDMAWQLHEY	7	Intermediate filament protein
Memb	IP_595191	285	32.75	10.74	RPL3P2	GRCh38.95	6:31280317-31281174	+	ENST00000413027.1	ncRNA	-	1	1-859	-	91.32% 60S ribosomal protein L3	EVVEAVTIVER	7	Ribosomal protein L3

Memb	IP_596887	229	24.77	5.35	SUCLA2P1	GRCh38.95	6:30469542-30470231	+	ENST00000425839.1	ncRNA	-	1	661-1351	+	86.34% succinate--CoA ligase [ADP-forming] subunit beta, mitochondrial precursor	EIKIPVVVQL	8	CoA-ligase
Memb	IP_596971	334	37.06	5.97	HLA-J	GRCh38.95	6:30006606-30009148	+	ENST00000494367.1	ncRNA	-	1	1-1006	+	No similarity > 80% found	GTEIHMPCAAQGAAQAPHPEMGALSPAHHPHCGYHCW	17	MHC class I signature
Cyt	IP_600567	118	12.43	7.97	GAPDHP40	GRCh38.95	5:159950833-159951189	-	ENST00000505218.1	ncRNA	-	1	310-667	+	81% glyceraldehyde-3-phosphate dehydrogenase	QNITPASTGTAKAVGRGHL	11	Glyceraldehyde 3-phosphate dehydrogenase
Memb	IP_602007	198	22.57	5.21	KRT8P33	GRCh38.95	5:123401212-123401808	-	ENST00000508125.1	ncRNA	-	3	537-1134	-	89.87% keratin, type II cytoskeletal 8	SLDMNSVIAEVK	6	Intermediate filament protein
Cyt	IP_603739	166	17.7	7	HSPD1P1	GRCh38.95	5:21882753-21883253	+	ENST00000503199.1	ncRNA	-	1	169-670	+	96.39% short heat shock protein 60 Hsp60s1	PVTTPEEIAQVAMISANGDKEIGNIISDAMK	6	TCP-1/cpn60 chaperonin family
Memb, Nuc	IP_606752	176	20	5.34	CCT6P2	GRCh38.95	5:14640288-14640818	-	ENST00000507234.1	ncRNA	-	3	201-732	-	89.20% T-complex protein 1 subunit zeta	TEAVVDSILAIKR	9	TCP-1/cpn60 chaperonin family
Cyt	IP_610367	362	40.57	4.73	TUBB7P	GRCh38.95	4:189982523-189984315	-	ENST00000428444.2	ncRNA	-	1	217-1306	+	93.80% tubulin beta-8 chain	TAAAIFQGRMPMR	32	Tubulin/FtsZ family, GTPase domain
Nuc	IP_612631	255	27.22	6.23	AC104619.3	GRCh38.95	4:135046062-135046829	+	ENST00000509116.1	ncRNA	-	2	599-1367	-	86% EEF1A1 protein	KIGGIGAVPAGR	7	Elongation factor Tu
Cyt	IP_612791	217	23.89	7.96	EEF1A1P9	GRCh38.95	4:105484836-105485489	+	ENST00000514975.1	ncRNA	-	1	139-793	+	95.39% elongation factor 1-alpha 1	MDSTEPPYSHKR	11	Elongation factor Tu GTP binding domain
Cyt	IP_612847	114	12.17	9.63	KRT8P46	GRCh38.95	4:102729827-102730171	-	ENST00000512037.2	ncRNA	-	1	1-346	-	84.21% keratin, type II cytoskeletal 8	GEGSSVGGITTITVNQSLLSPL	4	Intermediate filament protein
Memb	IP_614518	82	9.22	9.6	RPSAP39	GRCh38.95	4:80161129-80161377	+	ENST00000467439.1	ncRNA	-	1	1-250	-	95.18% 40S ribosomal protein SA	SDVIYIINLK	3	Ribosomal protein S2
Memb	IP_619671	184	21.34	10.74	AC144530.1	GRCh38.95	3:197850400-197850954	-	ENST00000452862.1	ncRNA	-	1	1-556	+	93.48% 60S ribosomal protein L17	KNAESDAELK	5	Ribosomal protein
Cyt	IP_620408	230	26.29	11.6	RPL4P4	GRCh38.95	3:185417870-185418562	-	ENST00000422486.1	ncRNA	-	1	217-910	+	97.39% ribosomal protein L4	NDIEKVYASQR	6	Ribosomal protein
Cyt	IP_623047	111	12.14	10.69	EEF1A1P8	GRCh38.95	3:184027222-184027557	-	ENST00000419025.1	ncRNA	-	2	200-536	-	86.52% elongation factor 1-alpha 1	TLGVKQPIVGVNK	9	Transcription factor, GTP-binding
Cyt	IP_624300	158	17.29	9.49	AC026877.1	GRCh38.95	3:76434832-76435308	+	ENST00000476047.1	ncRNA	-	2	815-1292	-	94.94% adenylyl cyclase-associated protein 1	VFDDMVGIVEIINSRDVK	8	Adenylate cyclase associated (CAP)
Cyt, Nuc	IP_635576	151	16.08	9.71	AC092573.1	GRCh38.95	2:173297961-173298416	+	ENST00000433054.1	ncRNA	-	3	333-789	-	84.83% 40S ribosomal protein S2	VAGCCGSVLVCLISVPR	9	Ribosomal protein
Memb	IP_636378	64	7.39	4.72	UBBP1	GRCh38.95	2:136329719-136329913	-	ENST00000392399.3	ncRNA	-	1	1-196	-	85.94% Chain X, Ubiquitin	TLTSKTIAL	9	Ubiquitin domain
Ske	IP_636787	304	33.85	7.82	LOC150776	GRCh38.p12	2:131517878-131520426	+	ENST00000438378.2 NR_026922.1	ncRNA misc_RNA	- -	2 2	1724-2639 1724-2639	-	85.30% sphingomyelin phosphodiesterase 4	LAQLITEAK	9	Mitochondrial-associated sphingomyelin phosphodiesterase
Chr	IP_637160	255	28.4	9.5	AL078621.1	GRCh38.95	2:113657908-113658675	+	ENST00000272567.4	ncRNA	-	1	1-769	+	97.65% U2 small nuclear ribonucleoprotein A'	QSGQIPGRERK	9	Leucine-rich repeat
Ske	IP_638654	228	26.18	4.72	KRT18P33	GRCh38.95	2:65666717-65667403	+	ENST00000398525.2	ncRNA	-	2	23-710	-	86.57% keratin, type I cytoskeletal 18	WSQQIEESITVVTMQSTEVGAAEMMLMEL	9	Intermediate filament protein
Chr	IP_641466	157	17.72	11.17	RPL23AP37	GRCh38.95	2:64347266-64347739	-	ENST00000479658.1	ncRNA	-	3	3-477	+	88.19% 60S ribosomal protein L23a	PLTTEFTMK	7	Ribosomal protein
Cyt	IP_644798	119	11.7	9	PSD4	GRCh38.95	2:64347266-64347739	-	ENST00000464559.1	ncRNA	-	3	3-477	-	88.19% 60S ribosomal protein L23a	PAGSTEEPAGAAGPW	0	
Cyt, Nuc	IP_654859	80	8.64	9.92	RPL3	GRCh38.95	22:39313502-39313744	-	ENST00000467105.1	ncRNA	-	1	745-988	-	100% 60S ribosomal protein L23a	VEGGSKATDAGGPVVGEER	4	Ribosomal protein L3
Cyt	IP_658154	164	17.75	10.43	HSPD1P7	GRCh38.95	21:28888659-28889153	-	ENST00000447985.1	ncRNA	-	3	120-615	-	86.50% 60 kDa heat shock protein, mitochondrial	TAVAITMGPKGK	7	TCP-1/cpn60 chaperonin family
Cyt, Nuc	IP_667318	106	11.83	6.09	EEF1A1P14	GRCh38.95	1:194189081-194189401	+	ENST00000421195.1	ncRNA	-	1	115-436	+	85.85% elongation factor 1-alpha 1	NMITGTSQADSAVLIVAAGVGEFEGSIPR	9	Elongation factor Tu GTP binding domain
Memb, Nuc	IP_667404	173	19.27	7.67	AL390728.1	GRCh38.95	1:247230199-247230720	+	ENST00000397642.3	ncRNA	-	2	260-782	-	89.40% heat shock cognate 70 kDa protein 8	EIAEAYLGNMVTNAVVTVPAY	10	Hsp70 protein
Nuc	IP_668974	118	13.26	4.37	AL365440.1	GRCh38.95	1:158525482-158525838	+	ENST00000436135.1	ncRNA	-	2	1811-2168	+	85.37% heat shock protein HSP 90-alpha	MIKLGVGIDENDPTADDTAEEMSPL	3	Core histone
Cyt	IP_669779	69	7.64	10.73	HIST2H2BA	GRCh38.p12	1:121108560-121117065	-	ENST00000430394.1 NR_027337.1	ncRNA misc_RNA	- -	1 1	193-403 187-397	+	97.01% histone H2B type 1-N	MGIMNSFVNDIFER	8	Heat shock protein Hsp90 family
Chr	IP_669823	238	25.8	7.53	HNRNPA1P43	GRCh38.95	1:115856970-115857686	-	ENST00000452680.3	ncRNA	-	2	134-851	-	90.07% heterogeneous nuclear ribonucleoprotein A1	CGKMELIEIITDR	11	RNA recognition motif
Chr	IP_670821	126	14.16	6.68	HNRNPA1P63	GRCh38.95	1:54536796-54537176	+	ENST00000427917.1	ncRNA	-	1	1-382	-	89.68% heterogeneous nuclear ribonucleoprotein A1	SIEATNESLRSHF	7	RNA recognition motif
Memb	IP_671626	426	47.24	4.69	SLC2A3P2	GRCh38.95	1:64984707-64985987	-	ENST00000331747.6	ncRNA	-	2	101-1382	-	84.04% solute carrier family 2, facilitated glucose transporter member 3	KDKANTPPSEML	45	Sugar (and other) transporter
Memb	IP_713380	194	22.11	10.79	RPS7P1	GRCh38.95	17:28467822-28468406	+	ENST00000571702.1	ncRNA	-	1	1-586	-	98.45% 40S ribosomal protein S7	NITAAKEIEVGGGQKAIIIF	4	Ribosomal protein
Nuc	IP_722050	108	12.45	4.3	KRT18P18	GRCh38.95	16:72729357-72729683	+	ENST00000565278.1	ncRNA	-	1	742-1069	+	92.59% keratin, type I cytoskeletal 18	VGAAETTLTELRY	8	Intermediate filament protein
Memb, Nuc	IP_724292	144	15.79	8.66	EEF1A1P38	GRCh38.95	16:27133839-27134273	-	ENST00000567658.1	ncRNA	-	2	599-1034	-	81.54% elongation factor 1-alpha 1	ALVSVTTEVNCVEKHHEAL	4	Translation protein
Chr	IP_736936	152	16.59	9.54	HNRNPCP3	GRCh38.95	15:79236332-79236790	+	ENST00000560385.2	ncRNA	-	1	1-460	+	81.05% heterogeneous nuclear ribonucleoproteins C-like 1	QGDYDDTMYSY	7	RNA recognition motif
Chr	IP_746496	192	20.8	9.54	PTBP1P	GRCh38.95	14:65279787-65280664	+	ENST00000554374.2	ncRNA	-	2	521-1100	+	No similarity > 80% found	FSASTGTCSA	6	RNA recognition motif
Chr	IP_747029	200	21.56	4.95	AL133163.1	GRCh38.95	14:35388305-35388907	-	ENST00000459899.1	ncRNA	-	1	163-766	-	87.13% 60S acidic ribosomal protein P0	VGASEATLLNTPNISPF	4	Ribosomal protein
Memb	IP_756776	153	17.13	7.71	RPS4XP16	GRCh38.95	13:51460784-51461245	+	ENST00000427198.2 ENST00000595905.1	ncRNA ncRNA	- -	2 3	260-722 279-741	-	89.54% 40S ribosomal protein S4, X isoform	AVHHITPEEAKYK	4	Ribosomal protein
Chr	IP_756980	50	5.85	8.54	ARGLU1	GRCh38.95	13:106557109-106559503	-	ENST00000360629.3 ENST00000375926.5	ncRNA ncRNA	- -	2 2	206-359 275-428	+	No significant similarity found	ARMAAEEHTLQDTGQDR	1	
Memb	IP_759877	351	39.3	4.67	KRT18P20	GRCh38.95	12:104976802-104977857	-	ENST00000548986.1	ncRNA	-	3	171-1227	-	87.39% keratin, type I cytoskeletal 18	NEEIAKVVTTQSAEVGAAEMTL	15	Intermediate filament protein
Memb	IP_761691	448	48.79	6.82	HSPD1P4	GRCh38.95	12:56511002-56512348	+	ENST00000547203.1	ncRNA	-	1	1-1348	-	95.50% 60 kDa heat shock protein, mitochondrial	PVTTPEEIAQVATISANGDKEIGNISDAMK	16	TCP-1/cpn60 chaperonin
Chr	IP_762232	129	15.08	4.38	KRT8P39	GRCh38.95	12:68706136-68706525	+	ENST00000396270.3	ncRNA	-	2	503-893	-	86.84% keratin, type II cytoskeletal 8	ESHLEGLTEEISF	5	Intermediate filament protein
Cyt	IP_762916	94	10.04	4.66	EEF1A1P16	GRCh38.95	12:16990973-16991257	+	ENST00000325349.6	ncRNA	-	1	304-589	-	90.43% elongation factor 1-alpha 1	MITGTSQADCAVL	7	Elongation factor Tu GTP binding
Cyt	IP_773785	133	15	4.68	KRT18P59 PKNOX2	GRCh38.95 GRCh38.p12	11:125113960-125114361 11:125113960-125114361	+ +	ENST00000524879.1 ENST00000534728.1 XM_017018110.1 XM_024448643.1	ncRNA ncRNA mRNA mRNA	- - 5'UTR 5'UTR	2 2 2 2	392-794 800-1202 1964-2366 1964-2366	-	No similarity > 80% found	IASSELTMEVNAPK	7	Intermediate filament protein
Cyt, Nuc, Ske	IP_773966	192	22.37	7.1	KRT8P7	GRCh38.95	11:119603375-119603953	-	ENST00000533003.1	ncRNA	-	3	354-933	-	91.04% keratin, type II cytoskeletal 8	LGSTGMTCGAQR	10	Intermediate filament protein
Cyt	IP_774695	374	41.01	5.68	TUBAP2	GRCh38.95	11:90282658-90283782	+	ENST00000530835.1	ncRNA	-	3	99-1224	-	93.85% tubulin alpha-1B chain	AVFVDLEPMVIDEVCTGTY	30	Tubulin signature
Chr	IP_788332	181	20.65	8.16	YWHAZP5	GRCh38.95	10:105686322-105686867	+	ENST00000419975.1	ncRNA	-	1	1-547	+	93.14% 14-3-3 protein zeta/delta	IEMELRDISNDVL	12	14-3-3 protein zeta signature
Nuc	IP_789717	150	16.61	10.04	KRT8P38	GRCh38.95	10:88727544-88727996	+	ENST00000441370.1	ncRNA	-	3	60-513	-	No similarity > 80% found	YRSIPSACIR	5	Intermediate filament protein
Chr	IP_790150	269	29.45	8.47	HNRNPA3P1	GRCh38.p12	10:43789578-43790387	-	ENST00000401429.1 NR_002726.2	ncRNA misc_RNA	- -	1 1	1-811 31-841	+	93.81% heterogeneous nuclear ribonucleoprotein A3	TDCLVMRDPQTK	12	RNA recognition motif
Cyt	IP_790880	294	33.89	4.57	KRT8P37	GRCh38.95	10:8513896-8514780	-	ENST00000451609.1	ncRNA	-	3	327-1212	-	85.71% keratin, type II cytoskeletal 8	SMDNSHSLDMESIIAEVK	14	Intermediate filament protein

Max. XlinkX Score	Crosslink Type	# CSMs	Protein A	Protein B	Compartment Identified
83.29	Inter	1	PHB2	PHB	Cytoplasm
82.28	Intra	1	SEC61B	SEC61B	Cytoplasm
79.93	Intra	1	HSPA5	HSPA5	Cytoplasm
61.14	Intra	1	MYH9	MYH9	Cytoplasm
57.5	Inter	1	VIM	CCDC121	Cytoplasm
54.89	Inter	1	IL31RA	HIST3H2A	Cytoplasm
50.31	Inter	1	HSP90B1	CCDC89	Cytoplasm
49.35	Intra	1	HSPA5	HSPA5	Cytoplasm
43.7	Inter	1	HIST1H4F	IP_672441	Cytoplasm
42.28	Intra	2	HSPA5	HSPA5	Cytoplasm
39.02	Inter	1	CYP1A1	HSPA5	Cytoplasm
31.93	Inter	1	KCTD19	ZNF292	Cytoplasm
29.98	Inter	1	IP_2331010	HIST1H4F	Cytoplasm
132.01	Inter	2	PHB2	PHB	Membrane
99.17	Intra	1	HSP90B1	HSP90B1	Membrane
93.6	Intra	2	HSPD1	HSPD1	Membrane
92.28	Intra	1	HSP90B1	HSP90B1	Membrane
86.12	Intra	1	HLA-A	HLA-A	Membrane
79.12	Intra	1	ATP5F1A	ATP5F1A	Membrane
77.94	Intra	1	IMMT	IMMT	Membrane
72.05	Inter	1	COX4I1	COX7B	Membrane
71.35	Intra	1	HSP90B1	HSP90B1	Membrane
70.22	Intra	2	B2M	B2M	Membrane
67.06	Intra	1	PLEC	PLEC	Membrane
65.47	Inter	1	ATP5F1A	ATP5F1	Membrane
63.56	Inter	1	EMC8	EMC2	Membrane
59.64	Intra	1	SEC61B	SEC61B	Membrane
56.12	Inter	2	SLFN14	HIST1H4F	Membrane
56.05	Intra	1	MYH9	MYH9	Membrane
54.67	Intra	1	HSD17B12	HSD17B12	Membrane
52.4	Intra	1	ATP5F1B	ATP5F1B	Membrane
51.79	Inter	1	CACNG3	H2AJ	Membrane
50.97	Intra	1	FAM3C	FAM3C	Membrane
47.82	Inter	1	VAMP1	ATRX	Membrane
47.82	Inter	1	VIM	MACF1	Membrane
47.75	Intra	1	HSP90B1	HSP90B1	Membrane
47.05	Inter	1	YWHAQ	YWHAZ	Membrane
46.35	Inter	1	IP_295919	PDIA4	Membrane
44.14	Inter	1	HIST1H4F	IP_672441	Membrane
44.12	Inter	1	IP_557247	ORC1	Membrane
43	Inter	1	SMC2	ATP2A2	Membrane
42.11	Inter	1	MMRN1	SLC25A6	Membrane
40.95	Inter	1	PDIA6	PLEKHO1	Membrane
40.36	Inter	1	HYOU1	PRKAR1B	Membrane

40.33	Inter	1	HOOK3	PDIA3	Membrane
36.2	Inter	1	ASCC1	HSPA5	Membrane
35.03	Inter	1	IP_614697	CANX	Membrane
34.23	Inter	1	PDIA3	SMARCA1	Membrane
33.69	Inter	1	IP_789671	RALA	Membrane
32.35	Inter	1	GAPDH	RRN3P2	Membrane
32.22	Inter	1	HIST1H4F	TAF4B	Membrane
32.22	Inter	1	C15orf41	ZDBF2	Membrane
30.33	Inter	1	PDE6C	HSPA5	Membrane
29.98	Inter	1	TP53BP2	RIBC2	Membrane
26.51	Inter	1	CENPN	ELP3	Membrane
132.01	Inter	2	PHB2	PHB	Nucleus
99.17	Intra	1	HSP90B1	HSP90B1	Nucleus
93.6	Intra	2	HSPD1	HSPD1	Nucleus
92.28	Intra	1	HSP90B1	HSP90B1	Nucleus
86.12	Intra	1	HLA-A	HLA-A	Nucleus
79.12	Intra	1	ATP5F1A	ATP5F1A	Nucleus
77.94	Intra	1	IMMT	IMMT	Nucleus
72.05	Inter	1	COX4I1	COX7B	Nucleus
71.35	Intra	1	HSP90B1	HSP90B1	Nucleus
70.22	Intra	2	B2M	B2M	Nucleus
67.06	Intra	1	PLEC	PLEC	Nucleus
65.47	Inter	1	ATP5F1A	ATP5F1	Nucleus
63.56	Inter	1	EMC8	EMC2	Nucleus
59.64	Intra	1	SEC61B	SEC61B	Nucleus
56.12	Inter	2	SLFN14	HIST1H4F	Nucleus
56.05	Intra	1	MYH9	MYH9	Nucleus
54.67	Intra	1	HSD17B12	HSD17B12	Nucleus
52.4	Intra	1	ATP5F1B	ATP5F1B	Nucleus
51.79	Inter	1	CACNG3	H2AJ	Nucleus
50.97	Intra	1	FAM3C	FAM3C	Nucleus
47.82	Inter	1	VAMP1	ATRX	Nucleus
47.82	Inter	1	VIM	MACF1	Nucleus
47.75	Intra	1	HSP90B1	HSP90B1	Nucleus
47.05	Inter	1	YWHAQ	YWHAZ	Nucleus
46.35	Inter	1	IP_295919	PDIA4	Nucleus
44.14	Inter	1	HIST1H4F	IP_672441	Nucleus
44.12	Inter	1	IP_557247	ORC1	Nucleus
43	Inter	1	SMC2	ATP2A2	Nucleus
42.11	Inter	1	MMRN1	SLC25A6	Nucleus
40.95	Inter	1	PDIA6	PLEKHO1	Nucleus
40.36	Inter	1	HYOU1	PRKAR1B	Nucleus
40.33	Inter	1	HOOK3	PDIA3	Nucleus
36.2	Inter	1	ASCC1	HSPA5	Nucleus
35.03	Inter	1	IP_614697	CANX	Nucleus
34.23	Inter	1	PDIA3	SMARCA1	Nucleus
33.69	Inter	1	IP_789671	RALA	Nucleus

32.35	Inter	1	GAPDH	RRN3P2	Nucleus
32.22	Inter	1	HIST1H4F	TAF4B	Nucleus
32.22	Inter	1	C15orf41	ZDBF2	Nucleus
30.33	Inter	1	PDE6C	HSPA5	Nucleus
29.98	Inter	1	TP53BP2	RIBC2	Nucleus
26.51	Inter	1	CENPN	ELP3	Nucleus
69.2	Intra	1	HLA-A	HLA-A	Cytoplasm
49.59	Inter	1	HLA-B	B2M	Cytoplasm
47.64	Intra	1	B2M	B2M	Cytoplasm
29.98	Inter	1	HLA-B	PARP1	Cytoplasm
28.13	Inter	1	CCDC144A	ARHGEF2	Cytoplasm
59.21	Inter	1	HLA-B	B2M	Membrane
48.85	Intra	1	B2M	B2M	Membrane
41.65	Inter	1	IP_2267193	IGSF22	Membrane
40.57	Intra	1	HLA-A	HLA-A	Membrane
39.84	Intra	1	B2M	B2M	Membrane
38.2	Inter	1	LRRC17	ATF4	Membrane
31.37	Inter	1	CAMSAP1	B2M	Membrane
28.13	Inter	1	CCDC144A	ARHGEF2	Membrane
25.67	Inter	1	FCER2	CCL26	Membrane
67.03	Inter	1	HLA-B	B2M	Nucleus
40.36	Intra	1	B2M	B2M	Nucleus
36.2	Inter	1	HLA-A	IP_2284785	Nucleus
77.3	Intra	2	HLA-A	HLA-A	Cytoplasm
39.44	Inter	1	ACTN1	HCRTR1	Cytoplasm
35.11	Inter	1	MFSD11	IP_248552	Cytoplasm
32.12	Inter	1	ITGA5	ITGB1	Membrane
55.44	Inter	1	HLA-B	B2M	Membrane
58.65	Intra	2	B2M	B2M	Membrane
43.63	Inter	1	B2M	HLA-B	Membrane
56.09	Intra	2	B2M	B2M	Membrane
100.82	Intra	2	HLA-A	HLA-A	Nucleus
55.6	Inter	2	HLA-B	B2M	Nucleus
45.08	Inter	2	IP_709097	HIST1H4F	Nucleus
44.53	Inter	1	ITGA5	ITGB1	Nucleus
40.1	Inter	1	HLA-A	B2M	Nucleus
37.41	Intra	1	B2M	B2M	Nucleus
34.81	Inter	1	H3F3A	IP_627699	Chromatin
60.29	Intra	1	HLA-A	HLA-A	Cytoplasm
43.7	Inter	1	HLA-B	B2M	Cytoplasm
32.94	Inter	1	IP_594208	PTN	Cytoplasm
49.42	Intra	1	B2M	B2M	Membrane
45.01	Inter	1	ACTB	ACTG2	Membrane
44.8	Intra	1	SLC25A5	SLC25A5	Membrane
44.14	Inter	1	HLA-B	B2M	Membrane
33.23	Inter	1	NR2C2	ERN1	Membrane
28.74	Inter	1	SLFN14	HIST1H4F	Membrane

30.19	Inter	1	TMEM67	ZNF667	Nucleus
60.36	Inter	1	HLA-B	B2M	Chromatin
34.82	Inter	1	EFCAB8	HIST1H4F	Chromatin
77.3	Intra	1	HLA-A	HLA-A	Chromatin
75.68	Intra	1	HIST1H1E	HIST1H1E	Chromatin
39.84	Inter	1	H3F3A	H2AFJ	Chromatin
35.3	Inter	1	IFT81	DDX50	Chromatin
27.32	Inter	1	CENPN	SRGAP2	Chromatin
77.29	Intra	1	HLA-A	HLA-A	Chromatin
43	Inter	1	TMEM57	ACTB	Chromatin
34.81	Inter	1	H3F3A	H2AFJ	Chromatin
50.63	Inter	1	SLFN14	HIST1H4F	Chromatin
30.1	Inter	1	H3F3A	H2AFJ	Chromatin
98.81	Intra	1	HLA-A	HLA-A	Cytoplasm
85.85	Inter	2	HLA-B	B2M	Cytoplasm
57.74	Intra	1	B2M	B2M	Cytoplasm
41.11	Inter	1	HLA-A	B2M	Cytoplasm
29.98	Inter	1	HLA-B	PARP1	Cytoplasm
27.62	Inter	1	IP_620377	ARIH2	Cytoplasm
85.52	Intra	1	HLA-A	HLA-A	Cytoplasm
59.56	Intra	1	B2M	B2M	Cytoplasm
46.75	Inter	1	B2M	SYNPO2L	Cytoplasm
36.56	Inter	1	HLA-B	WDR60	Cytoplasm
33.69	Inter	1	HLA-B	MRGBP	Cytoplasm
57.5	Intra	1	B2M	B2M	Cytoplasm
51.46	Inter	1	HLA-B	B2M	Cytoplasm
47.75	Intra	1	HLA-A	HLA-A	Cytoplasm
39.96	Inter	1	SLFN14	HIST1H4F	Cytoplasm
28.13	Inter	1	SLMO2	RAB11FIP4	Cytoplasm
89.3	Intra	1	HLA-A	HLA-A	Membrane
85.85	Inter	2	HLA-B	B2M	Membrane
66.99	Inter	1	SLFN14	HIST1H4F	Membrane
57.71	Intra	1	KRT8	KRT8	Membrane
49.86	Inter	1	B2M	HLA-B	Membrane
44.12	Intra	1	B2M	B2M	Membrane
44.12	Inter	1	LGALS1	IP_136846	Membrane
43.07	Inter	1	ITGA5	ITGB1	Membrane
38	Inter	1	SLC25A5	ZNF385D	Membrane
37.14	Inter	1	HLA-A	B2M	Membrane
89.3	Intra	3	HLA-A	HLA-A	Membrane
63.65	Intra	1	B2M	B2M	Membrane
54.79	Inter	1	HLA-B	B2M	Membrane
50.23	Intra	1	B2M	B2M	Membrane
47.14	Inter	1	B2M	HLA-B	Membrane
46.57	Inter	1	ITGA5	ITGB1	Membrane
40.36	Intra	1	ATP1B3	ATP1B3	Membrane
30.64	Inter	1	SIRT6	PDSS2	Membrane

107.18	Intra	3	HLA-A	HLA-A	Membrane
62.09	Intra	1	KRT84	KRT84	Membrane
57.5	Intra	1	B2M	B2M	Membrane
50.9	Inter	1	SLFN14	HIST1H4F	Membrane
47.56	Inter	1	B2M	HLA-B	Membrane
45.45	Inter	1	RPS6KB1	RAB40C	Membrane
31	Inter	1	HLA-B	B2M	Membrane
29.98	Inter	1	HLA-B	PARP1	Membrane
77.3	Intra	1	HLA-A	HLA-A	Nucleus
59.56	Intra	1	B2M	B2M	Nucleus
45.95	Inter	1	SLFN14	HIST1H4F	Nucleus
38.6	Inter	1	HLA-B	B2M	Nucleus
30.07	Inter	1	PAPSS2	PRKX	Nucleus
29.98	Inter	1	DENND4A	IP_2322359	Nucleus
27.64	Inter	1	AZI2	B2M	Nucleus
91.14	Intra	1	HLA-A	HLA-A	Nucleus
45.4	Intra	1	B2M	B2M	Nucleus
44.45	Inter	1	SLFN14	HIST1H4F	Nucleus
43.07	Inter	1	CACNA1A	FRYL	Nucleus
56.09	Intra	1	B2M	B2M	Nucleus
34.08	Inter	1	HLA-B	IP_2292176	Nucleus
93.82	Intra	1	HLA-A	HLA-A	Skeleton
57.5	Intra	2	B2M	B2M	Skeleton
53.58	Inter	1	HLA-B	B2M	Skeleton
52.33	Intra	1	B2M	B2M	Skeleton
50.97	Inter	1	MX1	HIST1H4F	Skeleton
49.37	Inter	1	TTC13	C16orf45	Skeleton
40.57	Inter	1	HLA-A	B2M	Skeleton
35.11	Inter	1	HLA-B	FGD6	Skeleton
31.96	Inter	1	CTBP2	TRIT1	Skeleton
70.2	Intra	1	HLA-A	HLA-A	Skeleton
29.3	Inter	1	PAEP	RAB8A	Skeleton
24.82	Inter	1	CCDC144A	ARHGEF2	Skeleton
77.37	Intra	1	HLA-A	HLA-A	Skeleton
44.12	Inter	1	CCNL2	CYP3A43	Skeleton

Potein accession	Protein length (a.a.)	Molecular weight (kDa)	Isoelectric point	Gene symbol	Annotation	Genomic coordinates	Strand	Transcript accession	RNA type	Localization	Frame	Transcript coordinates	Kozak motif	Domains	Domains
IP_2292176	67	7.82	8.5	FAM227B	GRCh38.p12	15:49618904-49619107	-	XM_011521319.2	mRNA	5'UTR	2	1019-1223	-	6	Signal peptide region and region of a membrane-bound protein in the extracellular region.
IP_789671	71	7.68	9.99	RPL23AP61	GRCh38.95	10:46063250-46063465	+	ENST00000623642.2	ncRNA	-	1	1-217	+	2	Ribosomal protein L23
IP_557247	262	29.22	10.53	MRRFP1	GRCh38.95	X:123116968-123117756	-	ENST00000435941.1	ncRNA	-	2	26-815	+	4	Ribosome recycling factor
IP_672441	42	4.8	10.68	RPS15AP10	GRCh38.p12	1:45645950-45646078	-	NR_026768.1	misc_RNA	-	2	608-737	-	4	Ribosomal protein S8
					GRCh38.95	1:45645950-45646078	-	ENST00000432472.1	ncRNA	-	3	120-249			
IP_2267193	52	6.21	11.88	SAMSN1	GRCh38.p12	21:14512522-14546239	-	XM_011529685.1	mRNA	CDS	2	77-236	-	4	No prediction
IP_594208	101	11.56	9.99	CNKSR3	GRCh38.95	6:154389002-154389307	-	ENST00000607772.5	mRNA	3'UTR	3	19260-19566	-	2	Reverse transcriptase
IP_136846	140	16.08	11.94	PRPF4B	GRCh38.95	6:4031936-4032358	+	ENST00000337659.10	mRNA	CDS	3	519-942	-	1	No prediction
					GRCh38.95	6:4031936-4032358	+	ENST00000480058.5	mRNA	CDS	2	578-1001			
					GRCh38.p12	6:4031936-4032358	+	NM_003913.4	mRNA	CDS	3	510-933			
					GRCh38.p12	6:4031936-4032358	+	NR_146783.1	misc_RNA	-	3	510-933			
					GRCh38.p12	6:4031936-4032358	+	NR_146784.1	misc_RNA	-	3	510-933			
					GRCh38.p12	6:4031936-4032358	+	NR_146785.1	misc_RNA	-	3	510-933			
					GRCh38.p12	6:4031936-4032358	+	XM_011514970.3	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011410.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011411.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011413.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011414.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011415.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011416.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011417.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011418.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011420.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011421.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011422.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011423.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011424.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011425.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011426.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011427.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011428.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011430.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011432.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011433.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011434.2	mRNA	CDS	3	516-939			
					GRCh38.p12	6:4031936-4032358	+	XM_017011435.2	mRNA	CDS	3	516-939			
IP_627699	87	9.98	9.27	SLC41A3	GRCh38.95	3:126022809-126033659	-	ENST00000507008.5	mRNA	CDS	3	261-525	-	1	No prediction
IP_248552	37	4.04	10.51	LONP2	GRCh38.p12	16:48361587-48361700	+	NM_001348078.1	mRNA	CDS	3	2934-3048	-	0	No prediction
					GRCh38.95	16:48361587-48361700	+	ENST00000564259.1	ncRNA	-	3	570-684			
					GRCh38.95	16:48361587-48361700	+	ENST00000566719.1	ncRNA	-	2	1136-1250			
					GRCh38.95	16:48361587-48361700	+	ENST00000565185.1	ncRNA	-	1	373-487			
IP_295919	60	6.59	6.22	BRD1	GRCh38.p12	22:49779394-49779576	-	XM_017028719.1	mRNA	3'UTR	3	7275-7458	-	0	No prediction
					GRCh38.p12	22:49779394-49779576	-	XM_017028721.1	mRNA	3'UTR	2	7193-7376			
					GRCh38.p12	22:49779394-49779576	-	XR_001755193.1	misc_RNA	-	3	7212-7395			
					GRCh38.p12	22:49779394-49779576	-	XR_001755194.1	misc_RNA	-	1	7126-7309			
IP_614697	36	4.23	9.39	LINC02484	GRCh38.95	4:34121087-34121316	-	ENST00000513843.5	ncRNA	-	3	243-354	-	0	No prediction
IP_620377	43	4.85	8.56	LINC02051	GRCh38.p12	3:186477112-186478197	+	XR_001741057.2	ncRNA	-	3	513-645	-	0	No prediction
					GRCh38.95	3:186477112-186478197	+	ENST00000456535.1	ncRNA	-	2	386-518			
IP_709097	64	7.11	10.91	AC123769.1	GRCh38.95	17:34005795-34005989	-	ENST00000624332.1	ncRNA	-	1	784-979	-	0	No prediction
IP_2284785	76	8.9	10.42	LOC105371077	GRCh38.p12	16:9789781-9790011	+	XR_933062.3	ncRNA	-	3	1035-1266	-	0	No prediction
IP_2322359	38	4.63	11.47	CAMK1D	GRCh38.p12	10:12758628-12760971	+	XM_011519594.3	mRNA	5'UTR	3	102-219	-	0	No prediction
IP_2331010	39	4.37	9.38	KDM4C	GRCh38.p12	9:7055130-7055249	+	XM_017014502.2	mRNA	3'UTR	2	6614-6734	+	0	No prediction

Conclusion

Our study pioneers a new integrative strategy that combines subcellular fractionation, XL-MS, structural modeling, docking and GO term enrichment to shed light on the alternative proteome in an unbiased manner. This multi-technique approach allowed us to gain several key insights into this overlooked dimension of the proteome.

Here, we identified 112 AltProts, most of which were derived from non-coding RNAs, supporting their translational potential. Localization analysis allocated several AltProts to specific compartments. Membrane and cytoplasm were the compartments in which more AltProts were identified. The study identified 220 protein-protein crosslinks only with subcellular fractionation enrichment. Among them, 16 were between AltProts and reference proteins. This interaction network connected AltProts to particular pathways and processes like antigen presentation and gene regulation. Several examples, such as AltFAM227B-HLA-B and multiple AltProt interactions with HIST1H4F and ORC1, suggest possible involvement of AltProts in MHC-mediated immunity and transcriptional control. Structural modeling supported the feasibility of several AltProt-protein crosslinking-derived interactions, lending orthogonal support. The unbiased strategy revealed new interactors and localized unstudied AltProts, providing a framework to propose functional hypotheses and roles in cell biology.

Overall, our study expanded the characterization of the hidden proteome, which may harbor unknown regulators and signaling molecules with impacts on cellular physiology that have been missed. The integrated omics workflow can be applied to determine AltProt involvement in diverse biological contexts. The results provide a foundation for future efforts to unravel AltProt functions and mechanisms now that an unbiased analysis platform has been established.

Several avenues of research could be pursued based on this study. Researchers could perform similar interactome/localization mapping in diseased versus normal cells to reveal disrupted AltProt networks and roles. Following up on individual AltProts like AltFAM227B to validate predicted interactions and functions using targeted methods could be another option. Extending this approach to explore AltProt roles in particular processes like immune cell function, development, and chromatin regulation could also be a fruitful path.

One option to validate the interactions found is PLA, as described in Part I. However, different host antibodies are needed, one for each protein. Since there are no antibodies available for AltProts, the sequence of the AltProt needs to be fused with a molecular tag. This will allow us to use an antibody that recognizes the molecular tag (e.g., Flag, HIS, or HA tags). Additionally, HLA-B targeted Co-IP experiments can be performed to identify the presence of the AltProt as an immunopeptide or in the surroundings of the MHC class I complex. To investigate possible interactions between DNA and the AltProts crosslinked with histones, chromatin immunoprecipitation (ChIP) can be used. ChIP is an antibody-based technology that selectively enriches specific DNA-binding proteins along with their DNA targets. It can be used to investigate a particular protein-DNA interaction, several protein-DNA interactions, or interactions across the whole genome or a subset of genes. Virotrap and BioID are targeted interactomic techniques that can help to validate the crosslinked PPIs identified. In addition, we could obtain a more comprehensive view of the interactions in which the bait protein is involved. For both techniques, the AltProt sequence needs to be fused in an expression system that contains the HIV-GAG proteins or the BirA* enzyme, respectively. Enhancing the workflow with deeper fractionation methods or more selective crosslinkers to boost detection, time course experiments to track AltProt dynamic responses and interactions, and integration with transcriptomics and genomics to link AltProt mechanisms to altered gene regulation are other possible avenues of research.

In conclusion, our study highlights the power of unbiased omics techniques to illuminate biology's "dark matter", the overlooked alternative proteome. We provided an exploratory framework to catalyze future efforts to elucidate the mechanistic contributions of AltProts in diverse cell processes and states. Characterizing this hidden dimension could uncover new regulators of signaling pathways and new biomarkers for precision medicine.

PART V PROTEOGENOMIC APPROACH TO IDENTIFY AND QUANTIFY THE GHOST PROTEOME IN OVARIAN CANCER CELL LINES

Proteogenomics

Proteogenomics is a rapidly growing field that combines proteomics, genomics and/or transcriptomics[361]. By integrating these three disciplines, proteogenomics aims to unravel the complex relationship between genes, transcripts and proteins. This innovative approach enhances our understanding of biological systems by facilitating the discovery and identification of peptides unique to specific proteins (e.g., mutated proteins). To achieve this, NGS genomic and/or transcriptomic data are utilized to create customized protein sequence databases[362]. These databases serve as the basis for interpretation of MS/MS data, allowing the accurate identification of peptides and proteins. In recent years, the field of proteogenomics has been growing due to advancements in NGS and MS-based proteomics. These advancements have resulted in enhanced depth and throughput, making proteogenomics more appealing and advantageous for studying proteomes on a system-wide level. By improving protein inference and database searching, proteogenomics has proven to be a valuable approach. Moreover, it enables the integration of datasets from various disciplines, eliminating the reliance on limited genomic derived models and facilitating the development of a comprehensive database of proteins or genetic markers[363]. This technique thus allows the discovery of novel peptides opening up new avenues for studying protein functions, physiological pathways and disease mechanisms. Moreover, proteogenomics plays a significant role in advancing precision medicine as it helps the development of targeted therapies and personalized medicine approaches.

During the last years, it has been shown that among two patients with the same type of cancer, the tumors aren’t the same. Therefore, a patient’s response to treatment can be very different. Precision oncology assesses the molecular signatures of each patient to evaluate/predict the tumor response to certain treatments, the assumptions being that by matching the mechanism of action of a certain drug to the status of that drug´s target, the tumor response will be improved[364].

One of the major efforts in cancer characterization is conducted by the Cancer Genome Atlas (TCGA)[365]. The objective of this resource is to create a comprehensive catalog of genomic changes implicated in cancer. While genomic characterization has improved patient outcomes[366], various studies have highlighted the limitations of making therapeutic decisions solely based on mutational profiling[367]. Although genomics enables us to comprehend the blueprint of cancer, a thorough analysis of the resulting proteins is necessary to identify and understand the precise state of the tumor to treat the underlying molecular pathology. Additionally, at the proteomic level is where most therapeutic targets are located. Therefore, it is essential to bridge the gap between cancer genotype and cancer phenotype. The objective of proteogenomic analyses is to investigate the connections between proteins that result from altered genes and related biological processes. The aim of the combination of transcriptomics and proteomics is to determine whether incorporating additional molecular information could enhance the understanding of the molecular mechanisms underlying cancers, beyond what can be accomplished solely through genomics.

Some of the findings that proteogenomics has provided are that RNA expression levels often do not accurately predict protein levels[368]. It allows for a more comprehensive understanding of signaling and regulatory pathways, providing insights into which pathways are activated or deactivated in a specific tumor[369]. Proteogenomics also enables customized searches in proteomics databases to identify new proteins and prioritize potential neoantigens[370]. Additionally, it helps prioritize genomic alterations that may act as oncogenic drivers, such as copy-number alterations[371].

Therefore, by integrating proteomic and genomic data, proteogenomics has revolutionized our understanding of gene annotation, protein translation, post-translational modifications, and splice isoforms. Despite challenges, ongoing advancements in mass spectrometry and bioinformatics are addressing these limitations, making proteogenomics a crucial tool in cancer research, microbiology and other fields. As proteomic and genomic technologies continue to advance, proteogenomics is poised to play an increasingly important role in advancing molecular biology.

Cancer cell line research

Cancer cell lines are highly valuable and extensively used *in vitro* model systems that play a crucial role in advancing medical research in various fields, especially in basic cancer research and drug discovery[372]. These cell lines are essential tools in laboratories, providing a platform to thoroughly study the complex biology of cancer and evaluate the effectiveness of therapeutic drugs. By utilizing cancer cell lines, valuable insights into the intricate mechanisms involved in cancer development, progression and response to treatment can be gained. Additionally, these cell lines are crucial in validating cancer targets and assessing the efficacy of potential treatments[373].

1.1. The PEO-4 cell line

The PEO-4 cell line is a human ovarian cancer cell line that is part of the PE ovarian adenocarcinoma panel. It is an adherent cell line derived from a malignant effusion from the peritoneal cavity of a patient with ovarian cancer and the cells have a doubling time of around 27 hours. Particularly, PEO-4 cells have a high-grade serous histology and were collected after clinical resistance in a patient who previously received cisplatin, 5-fluorouracil and chlorambucil treatment[374]. PEO-4 cells were collected after clinical resistance developed to chemotherapy. Additionally, PEO-4 cells have been xenografted into immune-deprived mice and found to be tumorigenic[375]. The key genetic mutations in PEO-4 cells include p53, BRCA1 and PI3KCA mutations. The cells are negative for estrogen and progesterone receptors. The PEO-4 cell line is an important research tool for understanding and developing new ovarian cancer treatments. Its drug resistant nature makes it a good model for testing therapies that may be able to overcome resistance.

1.2. The SKOV-3 cell line

The SKOV-3 cell line is a clear cell carcinoma cell line. It has an epithelial-like morphology that closely resembles the characteristics of ovarian adenocarcinoma. This cell line was derived from the ascites of a 64-year-old Caucasian female diagnosed with adenocarcinoma of the ovary in 1973. One notable feature of SKOV-3 cells is their resistance to tumor necrosis factor and various cytotoxic drugs like diphtheria toxin, cisplatin and adriamycin. This resistance poses challenges in developing effective treatment strategies for ovarian cancer. In animal studies, injecting these cells intraperitoneally into immunocompromised mice led to the growth of tumors[376]. A recent study showed that UNBS5162, a potential therapeutic agent, inhibits SKOV-3 ovarian cancer cell proliferation by modulating the PI3K/AKT signaling pathway[377]. Another study revealed that SKOV-3 cells exhibit higher migratory and invasive potential compared to PEO-4[378].

Mass Spectrometry-Based Protein Quantification

The versatility of MS-based proteomics has facilitated the detection and development of protein quantitative workflows. These approaches allow for the investigation of variations in biological processes or pathways under different conditions, answering the question of "what and how much" variation can be identified[379]. By addressing this question, decision-making for new disease-related biomarkers can be improved, leading to better diagnostics, prediction and treatment.

MS offers various strategies for quantification. Untargeted or global quantification enables the profiling of thousands of proteins in a system or the comparison of different conditions. In contrast, targeted quantification focuses on the quantification of single or specific proteins. Quantification can be performed at the protein level using top-down approaches or at the peptide level using bottom-up workflows. Another classification is based on the use of labeling reagents. Label-based quantification incorporates stable isotope labels into peptides, while label-free quantification (LFQ) analyzes peptides or proteins in their natural state. Furthermore, relative quantification compares protein ratios between samples, while absolute quantification provides the exact concentration of proteins in a sample.

The peptide-centered approach takes advantage of the fact that peptides are easier to fragment than entire proteins. However, this approach generates a list of proteins based on the unique peptides derived from these proteins. Therefore, the peptides are first quantified and the data are then transferred to the protein level[380].

1.3. Label-free quantification

LFQ compares the variation of peptides and proteins in their natural state in consecutive experiments. Due to the variation that can arise from multiple sample analysis, a normalization step is required to make the data more comparable. Additionally, LFQ allows an unlimited number of samples to be prepared and analyzed without the need for labeling steps, which reduces the costs of analysis. As a result, this method is preferred for biomarker research as it provides the widest dynamic range and coverage. However, LFQ has lower quantification accuracy and reproducibility compared to label-based techniques[381]. Additionally, a larger number of technical replicates are needed to compensate for its poor reproducibility[382].

In order to perform LFQ, two main methods are available. For instance, spectral counting is based on the observation that more abundant peptides are more likely to be detected multiple times in a MS run. This method involves counting the identified peptides or fragment spectra observed for a particular protein. A ratio of the number of peptides observed against the total number of peptides which a protein could produce is calculated and named as protein abundance index. A linear correlations has been identified between the number of spectra and the relative protein abundance[383].

The second main method is called intensity-based LFQ. This method is based on the correlation of the signal intensities of the ions after ESI[384]. Therefore, the peak areas from the extracted ion chromatograms (XIC) can be used for relative quantification. When performing this method, the variation in LC retention time and/or m/z values of identical peptides between measurement runs should be considered. This is accomplished by aligning the individual ion chromatograms and feature detection[385].

In order to perform this alignment and feature detection, different algorithms have been developed. Among them is Minora Feature Detector, an algorithm used to detect and quantify chromatographic peaks in MS data. The algorithm works by detecting peaks in the MS data. It then aligns the peaks across different MS runs and matches them to the corresponding peptide sequences identified by MS2. This allows the algorithm to quantify the peptides across all the runs. Additionally, Minora can provide two different types of quantitative values, the height of the most abundant peak at the apex of the chromatographic profile (intensity) or the integrated peak area. Finally, the normalization method can be based on the total peptide intensity or on the abundance of an internal reference protein[386].

Objective

The main objective of this article is to utilize a proteogenomic approach to thoroughly characterize and compare the proteomes of two different ovarian cancer cell lines, namely PEO-4 and SKOV-3 cell, as well as an immortalized ovarian epithelial cell line known as

T1074. This will involve the implementation of the subcellular fractionation protocol, as detailed in Chapter III, which will greatly enhance the depth of characterization of both the alternative and reference proteomes across all three cell lines.

To begin, we will carry out RNA-seq experiments to accurately map the obtained reads to the reference transcriptome and subsequently identify any variants present within each cell's transcriptome. Once these variants have been successfully identified, we will generate cell-specific protein databases using OpenCustumDB. Furthermore, the RNA-seq experiments will enable us to analyze the differential expression of both transcripts and genes. These RNA-seq-derived databases will serve as the foundation for evaluating and comparing the different abundances of AltProts, RefProts and novel isoforms. Additionally, they will facilitate the identification of protein variants specific to each individual cell line.

In order to gain a deeper understanding of the disparities between high-grade serous and clear cell carcinoma, we will employ functional enrichment algorithms to analyze the specific variated genes and RefProts. This will allow us to map these features onto different pathways, ultimately identifying key differences that will enhance our overall comprehension of these two types of ovarian cancer.

Lastly, we will utilize the proposed workflow outlined in Chapter III, which includes XL-MS, molecular modeling, docking and GO enrichment, to gain valuable insights into the "hidden" proteome of both the ovarian cancer cells and the epithelial ovarian cell. This comprehensive approach will provide us with a more thorough understanding of the complex proteomic landscape in these cell lines.

MANUSCRIPT TITLE

Deciphering the ghost proteome in ovarian cancer cells by deep proteogenomic characterization

AUTHORS

Diego Fernando Garcia-del Rio[1, 5, 6], Mehdi Derhourhi[2], Amelie Bonnefond[2, 3], Sébastien Leblanc[4], Noé Guilloy[4], Xavier Roucou[4], Sven Eyckerman[5, 6], Kris Gevaert[5, 6], Michel Salzet[1†*], Tristan Cardon[1†*]

[1]Université de Lille, Univ. Lille, CHU Lille, Inserm U1192 - Protéomique Réponse Inflammatoire Spectrométrie de Masse - PRISM, F-59000 Lille, France, Lille, France

[2]Université de Lille, Inserm/CNRS UMR 1283/8199, Pasteur Institute of Lille, EGID, Lille, France University of Lille, Lille, France.

[3]Department of Metabolism, Digestion and Reproduction, Imperial College London, London, United Kingdom.

[4]Department of Biochemistry and Functional Genomics, Université de Sherbrooke, Sherbrooke, Québec J1E4K8, Canada

[5]VIB Center for Medical Biotechnology, VIB, Ghent, 9052, Belgium

[6]Department of Biomolecular Medicine, Ghent University, Ghent, 9052, Belgium

[†] Joint Authors

* To whom correspondence should be addressed.
Tel: (+33)0320434385; Email: tristan.cardon@univ-lille.fr Correspondence may also be addressed to michel.salzet@univ-lille.fr

GRAPHICAL ABSTRACT

ABSTRACT

Proteogenomics is becoming a powerful tool in personalized medicine by linking genomics, transcriptomics and mass spectrometry (MS)-based proteomics. Due to increasing evidence of alternative open reading frame-encoded proteins (AltProts), proteogenomics has a high potential to unravel the characteristics, variants and expression levels of the alternative proteome, in addition to already annotated proteins (RefProts). To obtain a broader view of the proteome of ovarian cancer cells compared to ovarian epithelial cells, cell-specific total RNA-sequencing profiles and customized protein databases were generated. In total, 128 RefProts and 30 AltProts were identified exclusively in SKOV-3 and PEO-4 cells. Among them, an AltProt variant of IP_715944, translated from *DHX8*, was found mutated (p.Leu44Pro). We show high variation in protein expression levels of RefProts and AltProts in different subcellular compartments. The presence of 117 RefProt and two AltProt variants was described, along with their possible implications in the different physiological/pathological characteristics. To identify the possible involvement of AltProts in cellular processes, crosslinking-MS (XL-MS) was performed in each cell line to identify AltProt-RefProt interactions. This approach revealed an interaction between POLD3 and the AltProt IP_183088, which after molecular docking, was placed between POLD3-POLD2 binding sites, highlighting its possibility of the involvement in DNA replication and repair.

INTRODUCTION

Historically, protein sequence databases have only considered proteins to originate from the coding regions of mRNA molecules (CDS) (1, 2). However, we now know that the sequences of many products of transcript translation are not stored in such databases (3). Such translated transcripts include small open reading frames (smORFs) (4–6), which translate to short encoding proteins (SEPs) (7, 8) with a length of less than 100 amino acids. Additionally, alternative proteins (AltProt) (9–11) are translated from alternative ORFs (AltORFs) present in non-coding regions, including the 5' and 3'UTRs, overlapping a CDS with a +1 or +2 reading, or present in non-coding RNAs (ncRNAs). In contrast to SEPs, AltProts are not limited to a maximum length of 100 amino acids. Synthesis of such proteins may result from leaky scanning and reinitiation of ribosomes as described by Marylin Kozak (12, 13). However, such underlying mechanisms remain poorly understood and, importantly, they were not considered when the first protein databases were built, explaining the absence of quite some protein sequences in the most-often used protein sequence databases such as Swiss-Prot. Nevertheless, an effort has been made to make such databases more comprehensive, notably by integrating predicted protein sequences (TrEmbl) (14) which increase the size of the (theoretical) proteome. Yet, the used prediction rules are restrictive and do not consider the concept of AltProts. To tackle this, databases holding predicted sequence for AltProts such as OpenProt (9, 15) have been created. With such databases AltProts can now be identified from bottom-up proteomic datasets. However, although such databases consider the presence of the "ghost proteome", they do not consider mutations and neither the transcriptomic expression of samples. To overcome these limitations, OpenCustomDB(16), is a new tool that uses RNA-seq data to generate sample-specific protein sequence databases incorporating AltProts and their genetic variants. Such a proteogenomic approach coupled with AltProt research, is therefore expected to provide more comprehensive views on cellular proteomes.

AltProts are ubiquitously expressed in cells and can carry physiological functions (17). Several AltProts have been linked to several pathways such as protein synthesis (18–20), DNA repair (8) and innate immunity (17). AltProts have also been linked to pathologies (21, 22) such as cancers (glioblastoma, breast, ovarian and colorectal cancer) (23–26) and amyotrophic lateral sclerosis (Alt-FUS) (27). Although their identification has been facilitated by specific enrichment and detection strategies (19, 28–30), for the overall majority of AltProts, their functions remains to be elucidated, yet targeted approaches have shed light on the function of a few AltProts (20, 29, 31–33). Recently, we have demonstrated the effectiveness of a protein crosslink strategy coupled to mass spectrometry (XL-MS) to annotate AltProt functions. XL-MS enabled us to identify interactions that are very close in space from 5.3 Å (34) to 30 Å (35), and by identifying crosslinked peptides between AltProts and known proteins, it completed our understanding of the function of these new proteins.

Ovarian cancer (OvCa) is considered a stealth killer due to its misdiagnosis and extended chemoresistance to treatment. In 2021, OvCa was the 8th most frequently diagnosed and source of fatal cancer in women (36). The high mortality rate of OvCa is related to its late detection. In the initial stages of the pathology, few unspecific symptoms are evident and diagnostic methods are not sufficiently effective (37). The current standard treatment is based on surgery or chemotherapy. For advanced stage tumours, debulking surgery and subsequent adjuvant chemotherapy is needed (carboplatin combined with paclitaxel is most commonly used). With this combination of treatments, up to 80% of patients will go into remission, but around 65% will relapse. Radical strategies such as oophorectomy and salpingectomy are recommended for avoiding recurrence (38).

Among the metabolic pathways involved in cancer. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (39) summarized different metabolic pathways. Among the central carbon metabolism in cancer (hsa05230) summarizes the metabolic changes that take place in cancer cells to facilitate their growth and survival (40). This pathway involves the conversion of glucose and glutamine into intermediate molecules, which are then used to synthesize the necessary macromolecules for the replication of cancer cell biomass and genome. The Warburg effect (41), a key feature of this pathway, is characterized by the heightened utilization of glucose and glutamine by cancer cells. This phenomenon describes the extensive glucose consumption, high rates of glycolysis, and conversion of a significant portion of glucose into lactic acid even in the presence of sufficient oxygen (42). More recently, it has been realized that the Warburg effect also encompasses an increased reliance on glutamine. Along the signalling pathways that regulate c-MYC, HIF-1, and p53, numerous other oncogenes and tumour suppressor genes are clustered(40).

We hypothesized that molecular characterization of OvCa at the proteomic level might help to improve patient care and treatment. In this context, studying AltProts may shed light on mechanisms that are not yet completely understood yet have an impact on OvCa pathology. Therefore, we here describe a proteogenomic approach to characterize the ghost proteome of two OvCa cell lines and an immortalized epithelial ovarian cell line. This approach allowed us to identify differential expression of RefProts, novel isoforms, AltProts and their transcripts. Additionally, the subcellular location, characteristics and interactors of several AltProts were mapped.

MATERIAL AND METHODS

Cell culture

Human PEO-4 ovarian cancer cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (Thermo Fisher Scientific), supplemented with 10% fetal bovine serum (Thermo Fisher Scientific), 2 mM L-glutamine (Thermo Fisher Scientific) and 100 U/mL penicillin-streptomycin (Thermo Fisher Scientific). Human SKOV-3 ovarian cancer cells were cultured in McCoy's 5A (modified) medium (Thermo Fisher Scientific), supplemented with 10% fetal bovine serum and 100 U/mL penicillin-streptomycin. Human immortalized ovarian epithelial cells SV-40 (T1074) were cultured in Prigrow I medium (Applied Biological Materials), supplemented with 10% fetal bovine serum and 100 U/mL penicillin-streptomycin. The three cell lines were grown in a humidified air incubator at 37 °C under an atmosphere of 5% CO_2. Aliquots of three million cells were harvested by trypsin-EDTA (0.05%, phenol red) (Thermo Fisher Scientific), centrifuged at 100 x g for 5 min at 20 °C and washed three times with DPBS (Thermo Fisher Scientific).

Cell line specific database creation

Total RNA sequencing (RNA-Seq). RNA was extracted from four replicates of three million cells from each cell line employing the NucleoSpin RNA Mini kit for RNA purification (MACHEREY-

NAGEL), following the vendor's protocol. 1 µg of RNA was utilized for library preparation using RiboNaut rRNA Depletion Kit and Rapid Directional RNAseq Kit 2.0 (PerkinElmer). Nine cycles of PCR were performed during this preparation. Library sequencing was carried out using the NovaSeq6000 sequencing platform (Illumina; SP flow cell) following a 2x75 paired-end mode. Demultiplexing was performed using bcl2fastq v2.20.0.422. Subsequent fastq trimming utilized trimmomatic v0.39 with parameters MINLEN:35 and AVGQUAL:20. The mapping and counting steps were executed using RSEM v1.3.1 along with STAR v2.7.3a, referencing genome version hg38 and GTF from Gencode v39. Differential analysis was conducted through DESeq2 v1.24.0, employing R v3.6.3.

Customized protein database generation with OpenCustomDB. RNA-Seq reads were aligned to the reference genome GRCh38.p12 using STAR version 2.7.3a with default parameters except for '–outSAMprimaryFlag: AllBestScore,–outFilterMismatchNmax: 5, –alignSJoverhangMin 10, –alignMatesGapMax 200 000, –alignIntronMax 200 000, –alignSJstitchMismatchNmax "5-1 5 5",–bamRemoveDuplicatesType UniqueIdenticalNotMulti'. Transcript expression was quantified in transcripts per million (tpm) with Kallisto version 0.46.0 with default parameters. Variant calling files (VCF) were generated from BAM files with FreeBayes version 1.3.1 with the setting "–min-alternate-count" set to 5. SNPs and Indels with FreeBayes quality of less than 20 were filtered out with an internal Python script. Variations were inserted in the corresponding transcripts with the variant annotator OpenVar. Next, the transcripts quantified by Kallisto were arranged in descending order based on their expression level (top 100,000 transcripts). Subsequently, OpenProt-annotated proteins linked to these transcripts were incorporated into the customized database until 100,000 entries (100K DB) were reached, as described by Guilloy *et al*. (16). Upon adding a protein variant to the database, the corresponding reference protein without any variation was simultaneously included to account for potential heterozygosity.

Chemical protein cross-linking and subcellular fractionation

In cellulo chemical cross-linking. The cross-linking methodology was described in Garcia-del Rio *et al.* (17, 30). To prepare a 50 mM stock solution of disuccinimidyl sulfoxide (DSSO, Thermo Fisher Scientific), dry DMSO (Sigma-Aldrich) was used. Three million cells of each cell line were resuspended in 200 µL of DPBS. The crosslinking reaction was carried out with 2 mM of DSSO (final concentration) at 37 °C with end-over-end stirring. After one hour, the reaction was quenched by adding 10 µL of 500 mM Tris-HCl pH 8.5 and gently stirring for 30 min.

Protein subcellular fractionation. The subcellular fractionation methodology was also described in our previous work (17, 30). In brief, three replicates of three million cells that underwent crosslinking were pelleted and the supernatant was removed. The Subcellular Protein Fractionation Kit for Cultured Cells (Thermo Fisher Scientific) was used to isolate five distinct protein cell compartments: cytoplasmic, membrane, nuclear, chromatin-bound and cytoskeletal proteins. Each fraction was extracted following the manufacturer's instructions and stored at -80 °C until use.

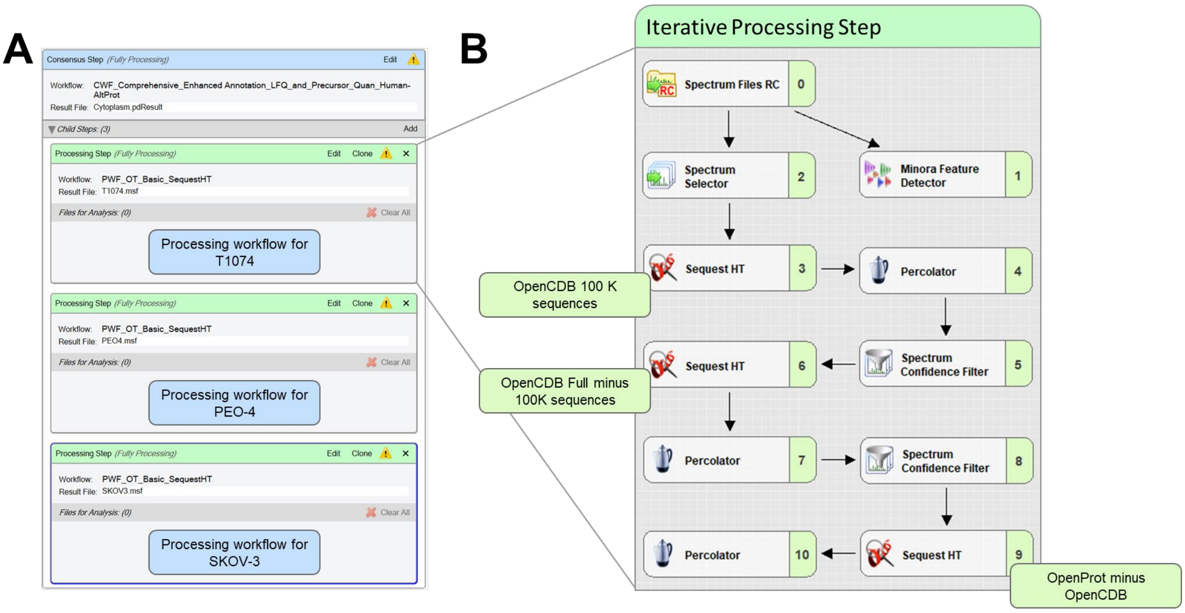
Filter Aided Sample Preparation (FASP) and digestion. Each subcellular fraction was transferred to a 50 kDa molecular weight cut-off Amicon filter (Merck) and concentrated by centrifugation (14,000 g x 15 min at 4 °C). Proteins were denatured by adding 100 mL of a denaturing buffer (8 M urea (Euromedex), 100 mM Tris-HCl (Interchim), pH 8.5). Reduction was performed by adding 100 mL of 100 mM dithiothreitol (VWR Life Science) in the denaturing buffer and incubating at 56 °C for 40 min. Alkylation was then done by adding 100 mL of 50 mM iodoacetamide (Sigma-Aldrich) in the denaturing buffer at room temperature (RT) for 30 min in the dark. After alkylation, three washes with 200 µL of 50 mM ammonium bicarbonate buffer were performed. Sequential digestion was performed in each fraction by adding 40 µL of 40 ng/µL trypsin/Lys-C Mix, Mass Spec Grade (Promega) to the Amicon filter and incubating at 37 °C overnight, followed by 25 µL of 40 ng/µL chymotrypsin, Sequencing Grade (Promega) at room temperature for 4 h. Finally, the resulting peptides were recuperated by adding 50 µL of ammonium bicarbonate buffer and centrifugating for 15 min at 14,000 x g. Finally, this flowthrough was acidified with 0.1% TFA (Sigma-Aldrich) and vacuum dried.

Nano LC-MS/MS analysis

The peptides of each replicate were suspended in 20 µL of 0.1% TFA and desalted using a ZipTip with C18 resin (Merck), following the manufacturer's instructions. Afterwards, the samples were vacuum-dried and resuspended in 20 µL of a solution containing acetonitrile (ACN, Carlo Erba Reagents) and 0.1% formic acid (2:98 v/v, TCI America). Five microliters of the resulting peptide solution were analysed on a nanoAcquity (Waters) coupled to a Q Exactive mass spectrometer (Thermo Fisher Scientific), as described in (24).

Label-free quantification (LFQ) data analysis

Processing workflow. The raw data obtained by nanoLC-MS/MS analysis were analysed using Proteome Discoverer V2.5 (Thermo Fisher Scientific). For each subcellular compartment, a different LFQ analysis was performed. Here, three processing steps (for each cell line's replicates) were employed using Minor Feature Detector and three iterative Sequest HT nodes (Figure 1A). The detailed parameters of the Sequest HT node are described in (30). In the first Sequest HT node, the top 100,000 sequences derived from RNA-seq experiments (100K DB) were utilized. Next, a percolator with a relaxed 0.05 FDR and strict 0.01 FDR was applied. A spectrum confidence filter was applied before moving on to the next Sequest HT node, discarding any spectra with a confidence rating worse than high. In the second Sequest HT node, the full transcript-derived database (Full DB) from OpenCustomDB was used, minus the sequences contained in the 100K DB. The same parameters were used for a second percolator and spectrum confidence filter. Finally, in the third Sequest HT node, OpenProt was used to interrogate the sequences not found in the two previous databases (Figure 1B).


Consensus Step *(Fully Processing)*

Workflow: CWF_Comprehensive_Enhanced Annotation_LFQ_and_Precursor_Quan_Human-AltProt
Result File: Cytoplasm.pdResult

Child Steps: (3)

Processing Step *(Fully Processing)*
Workflow: PWF_OT_Basic_SequestHT
Result File: T1074.msf
Files for Analysis: (0)

Processing Step *(Fully Processing)*
Workflow: PWF_OT_Basic_SequestHT
Result File: PEO4.msf
Files for Analysis: (0)

Processing Step *(Fully Processing)*
Workflow: PWF_OT_Basic_SequestHT
Result File: SKOV3.msf
Files for Analysis: (0)

A
Edit
Add
Edit Clone
Clear All
Processing workflow for T1074
Processing workflow for PEO-4
Processing workflow for SKOV-3
B
Iterative Processing Step
Spectrum Files RC 0
Minora Feature Detector 1
Spectrum Selector 2
Sequest HT 3
Percolator 4
Spectrum Confidence Filter 5
Sequest HT 6
Percolator 7
Spectrum Confidence Filter 8
Sequest HT 9
Percolator 10
OpenCDB 100 K sequences
OpenCDB Full minus 100K sequences
OpenProt minus OpenCDB

Figure 1. LFQ analysis workflow. (A) Illustration of the Proteome Discoverer analysis steps used. Each child processing step corresponds to the interrogation using the cell-specific database. (B) Workflow nodes present in each processing child step.

Consensus workflow. The five different subcellular fractionation MSF files were subjected to independent consensus workflows. At the feature mapper node, chromatographic alignment was performed with a maximum retention time shift of 10 min, a 10 ppm mass tolerance and coarse tuning. Unique and razor peptides were used at the precursor ions quantifier node. Protein groups were considered for peptide uniqueness and shared quant results were used. Precursor abundance was based on intensity without any threshold. Total peptide amount was used for normalization mode without scaling mode. All peptides were used for normalization and protein roll-up. Modified peptides (methionine oxidation, N-terminus acetylation and cysteine carbamidomethylation) were excluded for pairwise ratios. At the PSM grouper node, the site probability threshold was set to 75. The strict and relaxed FDRs were set at 0.01 and 0.05, respectively, at the peptide validator node. Validation was based on the q-value, and automatic target/decoy selection was used for PSM level FDR calculation based on score. At the peptides and protein filter node, the peptide confidence was set to medium with six amino acids per peptide. Additionally, a minimum of one peptide was set. A strict (0.01) and relaxed (0.05) FDR confidence threshold were set at the protein FDR validator. The results were filtered for RefProts, AltProts and novel isoforms (9). Briefly, a RefProt is a protein matching a NCBI RefSeq, Ensembl or UniProt protein entry. A novel isoform is a protein encoded by the same gene as a RefProt with a significant level of identity (over 80% of protein sequence identity with the RefProt over 50% of the length). An AltProt does not have any significant similarity with a RefProt.

Protein identification. The master protein files were uploaded as a text file to Perseus v.1.6.10.43. The abundance matrix was annotated into three categories based on the cell lines used: SKOV-3, PEO-4 and T1074. Next to count an identification, proteins needed to be identified in 70% of the replicates from at least one cell line and the groups were averaged. A numeric Venn diagram was used to identify the unique RefProts, AltProts and novel isoforms in each compartment for each cell line.

Statistical analysis workflow. The master protein files were uploaded as a text file to Perseus v.1.6.10.43. As a first step, log2 transformation and categorical annotation were performed on the normalized abundance values matrix, with cell lines SKOV-3, PEO-4 and T1074. To consider a valid identification, proteins needed to be identified in 70% of the replicates from each cell line. Moreover, missing values were replaced with low values of the normal distribution. An ANOVA multiple sample test was performed using a Benjamini-Hochberg FDR q-value cutoff of 0.05. Non-significant values were filtered out, and a Z-score processing was applied without grouping. To ensure quality control, a principal component analysis (PCA) was conducted with a Benjamini-Hochberg FDR cutoff of 0.05. Finally, hierarchical clustering employing Pearson correlation was applied to the averaged Z-scores to identify the different protein clusters.

Crosslinking data analysis

Processing workflow. The RAW data obtained by nano LC-MS/MS analysis were analysed using Proteome Discoverer V2.5 (Thermo Fisher Scientific). The detailed parameters for the Sequest HT and XlinkX nodes are described in (24). The triple Sequest HT nodes mentioned earlier were utilized. Instead of a percolator, a target decoy PSM validator was used after each Sequest HT node. A concatenated target decoy strategy was employed, with strict (0.01) and relaxed (0.05) FDR targets.

Consensus workflow. The resulting crosslinking MSF files were submitted to a consensus workflow of which the parameters are described in detail in (24).

RESULTS

For this study, we selected three cell line models to investigate differences in the reference proteome, novel isoforms and the alternative proteome. Two of these cell lines (PEO-4 and SKOV-3 cells) are derived from ascitic fluid from ovarian adenocarcinomas. Particularly, PEO-4 cells have a high-grade serous histology and were collected after clinical resistance from a patient who previously received cisplatin, 5-fluorouracil and chlorambucil treatment (43). PEO-4 cells have been xenografted into immune-deprived mice and found to be tumorigenic (44). SKOV-3 cells are clear cell carcinoma cells and resistant to tumour necrosis factor, diphtheria toxin, cisplatin and adriamycin (45). According to Hernandez *et al.* (46) and Hallas-Potts *et al.* (47), PEO-4 cells have a lower tumorigenicity than SKOV-3 cells when injected in nude mice. The T1074 ovarian cancer cell line was immortalized by SV40 virus and originally derived from normal human ovarian surface epithelial cells.

Differential gene expression analysis

In order to generate custom databases using OpenCustomDB, RNA-Seq data is required. From these reads, the assessment of differential gene expression can be performed. Mapping the RNA-Seq reads to the genome using RSEM and STAR enabled the identification of 117,636 transcripts expressed in 70% of four replicates between cell lines. Of these, 96,442 transcripts were shared by the three cell lines. Additionally, 1567, 2391, and 1780 transcripts were only identified in T1074, PEO-4 and SKOV-3 cells respectively (Figure 2A). Total RNA-seq data analysis showed that 37,197 transcripts were differentially expressed (DESeq2, FDR <0.05). Hierarchical clustering (Figure 2B and Supplemental Table 1) indicated six main transcript clusters: upregulation in PEO-4 (cluster 1, 3117) in SKOV-3 (cluster 2, 3220), or in both PEO-4 and SKOV-3 (cluster 3, 1138 transcripts); and downregulation in SKOV-3 (cluster 4), in PEO-4 (cluster 5), and in both cancerous cells (cluster 6, 12,129 transcripts).

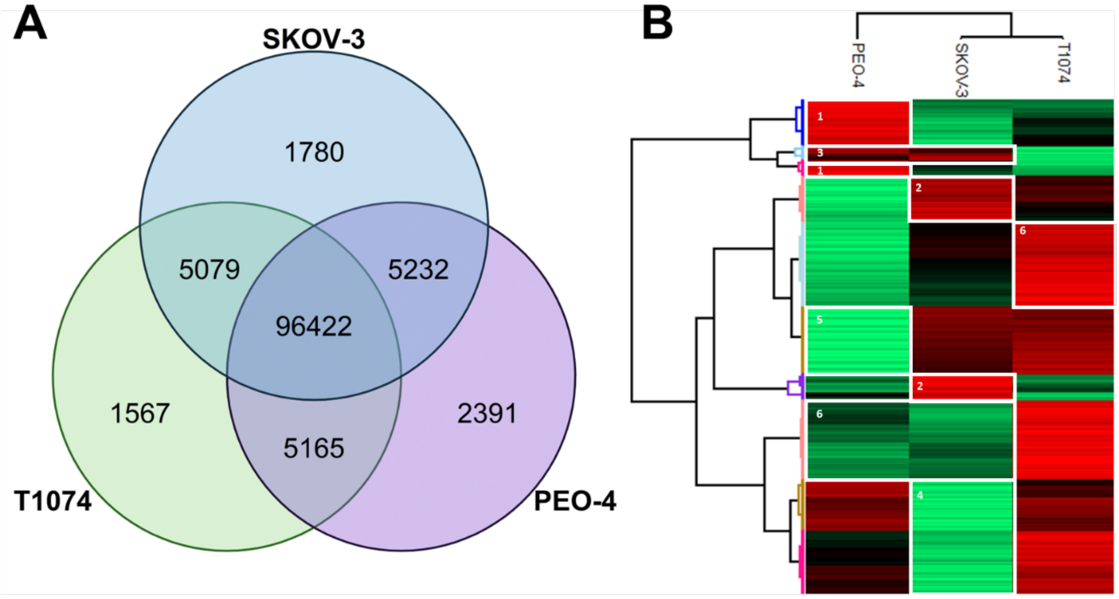


Figure 2. DESeq2 transcripts analysis. (A) Venn diagram displaying the number of exclusive and shared transcripts between the three cell lines. (B) Hierarchical clustering heatmap showing the different transcript clusters that can be observed among the three cell lines. Z-score range from -1.3509 (green) to 1.3523 (red).

Mapping RNA-Seq reads to the human genome Hg38 allowed us to find 29,245 expressed genes among the three cell lines. Among these expressed genes, 420, 407 and 540 were identified to be specific for T1074, SKOV-3 and PEO-4 cells respectively (Figure. 3A). Figure 3B displays the different categories of genes annotated and the major category of these genes were annotated as non-coding (pseudogenes and lncRNAs, 60.9%), while approximately 37% of the genes were annotated as coding genes. Hierarchical clustering was performed on the expression values obtained from the DESeq2 workflow. A total of 17,368 genes were identified as significantly differentially expressed between the three cell lines (Figure 3C and Supplemental Table 2), and of these, 2142 and 1949 genes were upregulated in PEO-4 and SKOV-3 cells respectively. On the other hand, 3345 and 2692 genes were downregulated in

PEO-4 and SKOV-3 cells respectively. Between the two cancerous cell lines, 632 genes were identified as upregulated and 6608 as downregulated.

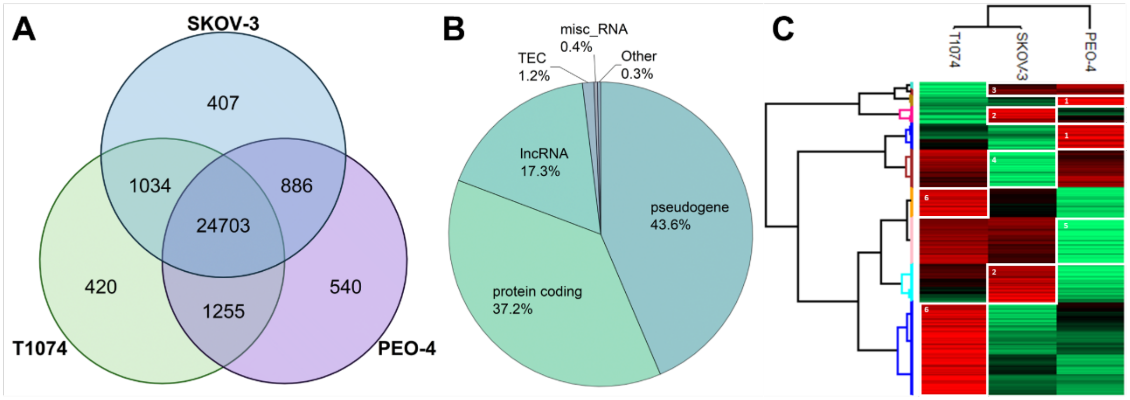


Figure 3. DESeq2 gene analysis. (A) Venn diagram displaying the number of exclusive and shared genes between the three cell lines. (B) Pie chart displaying the ratios of the different types of RNAs sequenced. (C) Hierarchical clustering heatmap showing the different gene clusters that can be observed among the three cell lines. Z-score range from -1.351 (green) to 1.3496 (red).

RNA-Seq based databases

We used RNA-Seq data from the ovarian epithelial cell (T1074) and the OvCa cell lines (PEO-4 and SKOV-3) to generate two cell-specific protein databases for each cell line. Figure 4 summarizes the protein types of the sequences stored in these databases. The distribution is similar for the three cell lines used and the custom 100K DB contained around 15% of wild-type (WT) RefProts, 2% of variant RefProts, 5% of WT novel isoforms, less than 1% of variant novel isoforms, 73% of WT AltProts and 5% of variant AltProts (Figure 4).

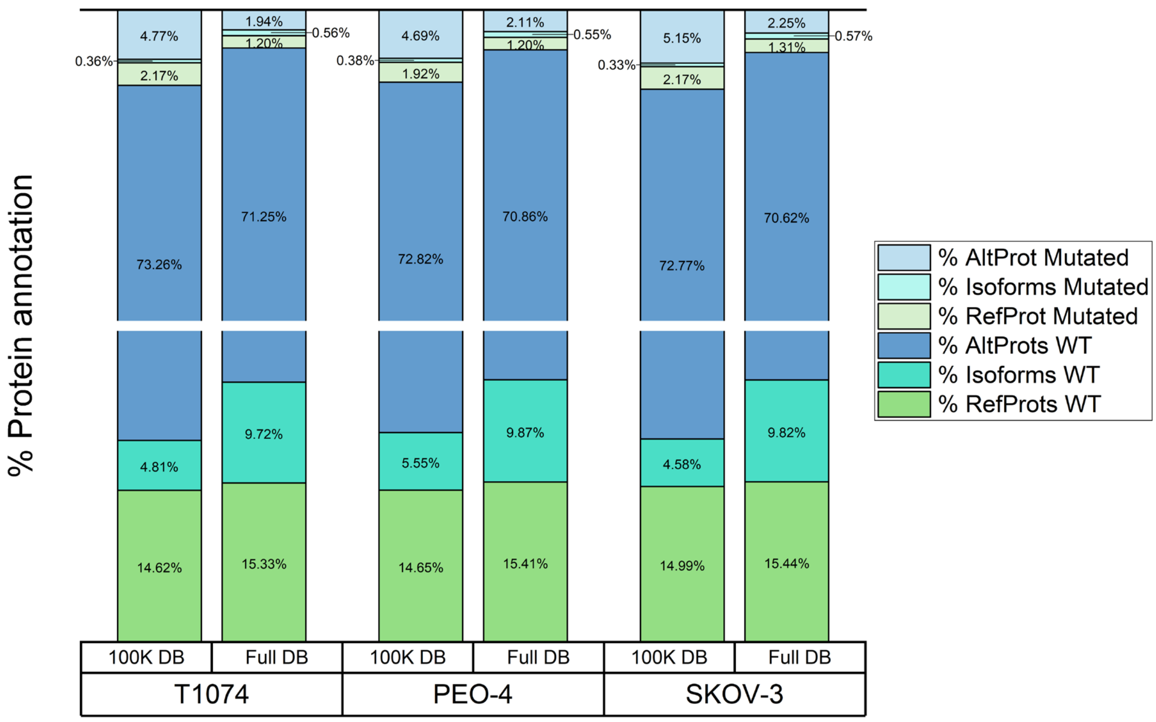


Figure 4. WT and variant proteins predicted by OpenCustomDB. For each cell line and database, the fractions of AltProts, RefProts, novel isoforms and their variants are displayed.

The OpenCustomDB workflow was used to generate comprehensive transcript databases (Full DB) without limiting the maximum number of entries to 100,000. These databases included 448,569, 443,177 and 437,568 entries for T1074, PEO-4 and SKOV-3 cells, respectively. For example, for T1074 cells, 68,759 WT RefProts (15.33%), 5366 variant RefProts (1.2%), 43,609 WT novel isoforms (9.7%), 2529 variant isoforms (0.6%), 319,612 WT AltProts (71.3%) and 8694 variant AltProts (1.9%) were stored in the database. Similar ratios were observed for PEO-4 and SKOV-3 cells (Figure 4).

Of the AltProts predicted, we mapped their transcriptomic origin by extracting information from OpenProt (Figure 5). AltORFs overlapping a CDS in a shifted reading frame, or in 3’UTRs and ncRNA were found to be the main sources of predicted AltProts.

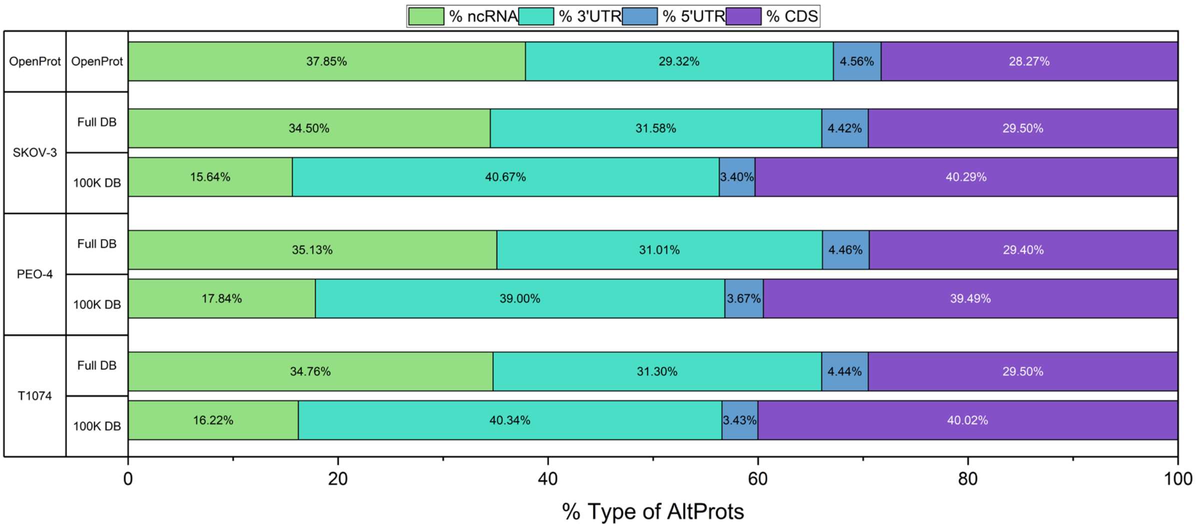


Figure 5. Types of AltProts predicted by OpenCustomDB. The percentages of ncRNA, CDS frameshifts, 3' and 5'UTR derived AltProts are displayed for each database and cell line.

Additionally, a comparison was performed between the databases across the three cell lines (see Supplemental Figure 1). In total, 282,287 AltProts were found to overlap across the three cell lines and, 15,109, 11,026 and 8897 unique AltProts were predicted in T1074, PEO-4 and SKOV-3 cells, respectively. Among the cancerous cell lines, 8055 AltProts were found to overlap. Approximately 39,000 sequences of novel isoforms were predicted to be shared across the three cell lines, with specific novel isoforms also identified in each cell line and in both cancerous cells. Almost 60,000 RefProts were found to overlap across all cell lines, with approximately 6000 being specific for each cell line. The same analysis was performed on the 100K DB, with 52,483 AltProts, 3116 novel isoforms and 10,346 RefProts being predicted to overlap across all three cell lines. A main advantage of these databases is that they contain predicted AltProt variants specific of each sample; for instance, 4321 specific AltProt variants were predicted for PEO-4 cells and, 4355 for SKOV-3 and 3540 for T1074 cells. This also shows that both cancerous cells have an increased number of transcript variants, which may be translated into mutated AltProts.

Proteome analysis of subcellular compartments

To evaluate the deeper differences in the proteome of these three different cell lines. The MS/MS data sets obtained from analysing each subcellular proteome of the three cell lines were analysed using Proteome Discoverer V2.5. Three different child processing workflows that contained three sequential Sequest HT (48) nodes were used with the databases as described in the material and methods section. We considered a protein as identified when it was present in at least one subcellular compartment in 70% of the replicates of at least one cell line. Figure 6A displays the distributions of all identified proteins. 6301 RefProts were identified in T1074 cells, 6268 in PEO-4 cells and 6319 in SKOV-3 cells. Among the identified RefProts, 234 (T1074 cells), 224 (PEO-4 cells) and 233 (SKOV-3 cells) were variants of RefProts. In addition, 137 novel isoforms were identified in T1074 cells, and 136 in PEO-4 and SKOV-3 cells. A total of 8 variants of novel isoforms were annotated in T1074 cells, and 9 in SKOV-3 and PEO-4 cells. Finally, over 500 AltProts were identified in each cell line with similar numbers of AltProts identified in SKOV-3 cells (577), T1074 (556) and PEO-4 cells (549). The number of AltProt variants identified was 12 for PEO-4 cells, and 13 for T1074 and SKOV-3 cells. Additionally, the distribution of WT and variant proteins is shown in Figure 6B.

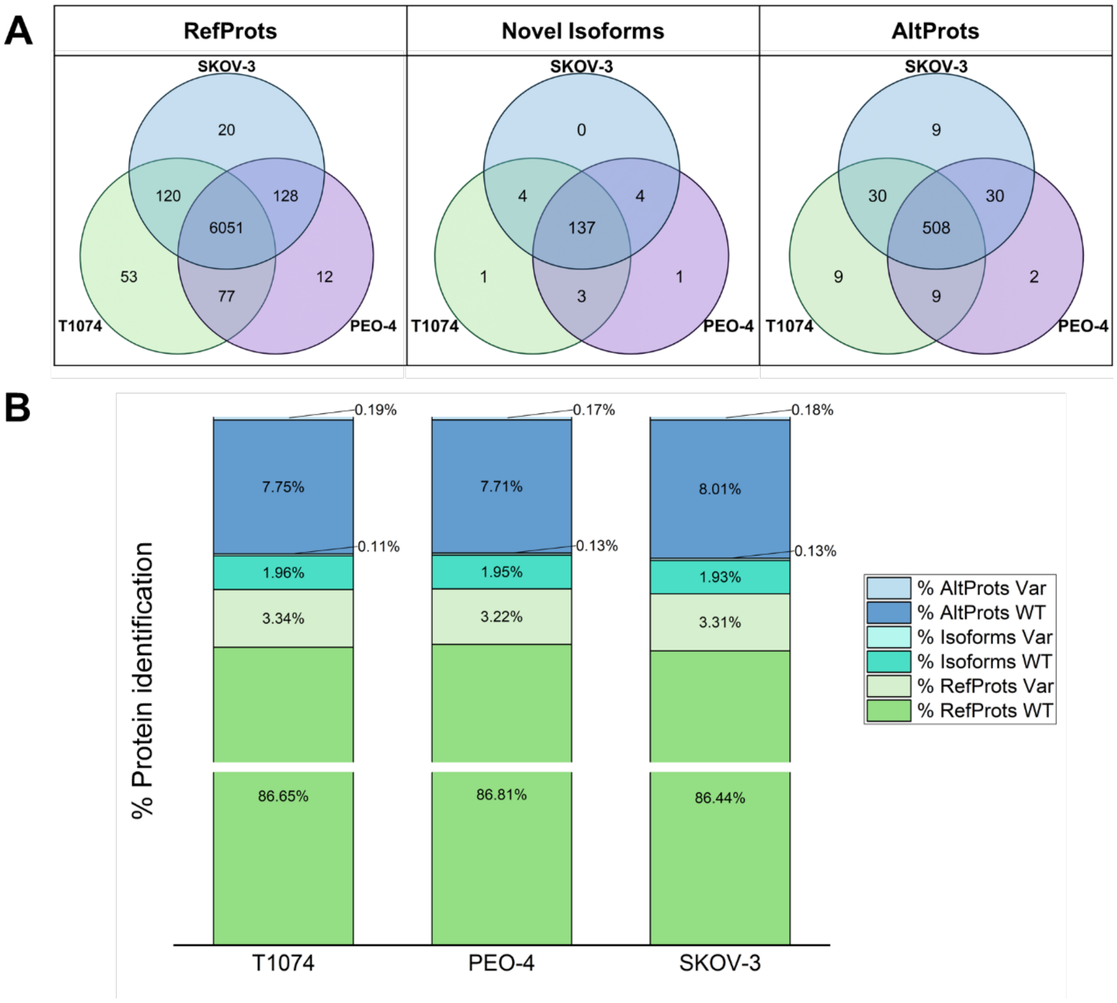


Figure 6. Analysis of the identified proteins. (A) Venn diagrams displaying the number of exclusive and shared proteins identified between the three cell lines. (B) Bar plot displaying the fractions of WT and variant RefProts, novel isoforms and AltProts identified in each cell line.

Subcellular fractionation was used to link (a) cellular compartment(s) to identified AltProts (Figure 7A). The membrane-bound fraction of all three cell lines contained the highest number of identified AltProts. In Figures 7 B and C, some general descriptions of the identified AltProts are displayed. Here, the majority of the AltProts identified possess a 3'UTR origin. Additionally, the vast majority (80.9%) have a molecular weight less than 10 KDa.

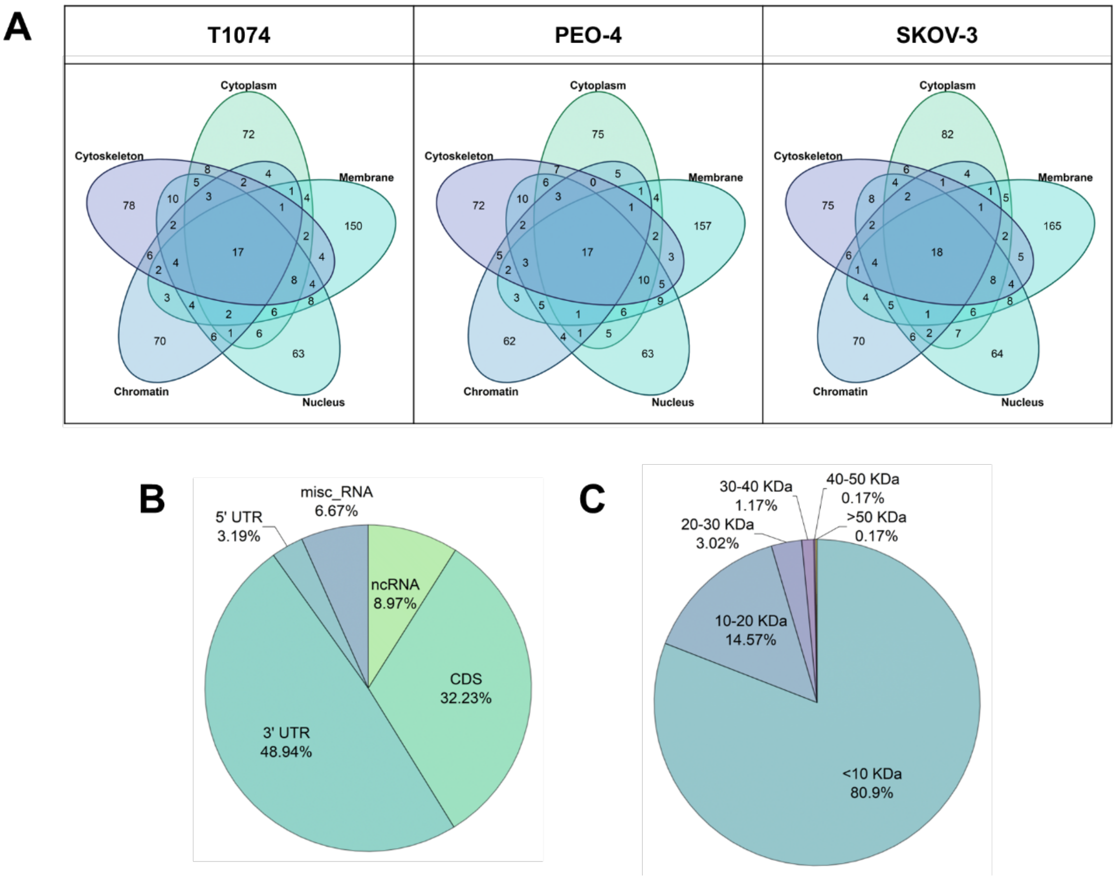


Figure 7. Subcellular compartment distribution and characteristics of identified AltProts. (A) Venn diagram displaying the distribution of AltProts identified in the different subcellular fractions. (B) RNA origin and (C) molecular weight distribution of the identified AltProts.

In addition, we identified cell line-specific RefProts, novel isoforms and AltProts. In T1074 cells, nine specific AltProts were identified, including the variant AltProt IP_290059@Asp99fs, which was found in the cytoskeletal fraction. SKOV-3 cells also had nine cell-specific AltProts, but without any variants, and PEO-4 cells had two specific wild-type AltProts identified. The characteristics of the cell line-specific AltProts are described in Supplemental Table 3. Overall, 508 AltProts were identified shared by all three cell lines, including 11 variants.

Among the identifications, 30 AltProts were identified in both cancerous cell lines. The variant IP_715944@Leu44Pro was identified in the cytoskeletal fraction of both cell lines. The WT

AltProt IP_715944 is a 4.82 KDa protein composed by 47 amino acids. It is coded in the *DHX8* gene. The variant of this AltProt is the result of a base substitution (c.131T>C) observed in the transcript ENST00000587574, which changed the proline at position 44 to a leucine. To verify the impact of the mutation, the sequence was analysed using protein BLAST (49), InterProScan (50) and Phobious (51). No significant similarity or any change in the predicted domains were identified.

Next, we performed a label-free quantitative analysis on the subcellular proteomes (n=4), which led to the identification of 1,022 RefProts with significantly altered levels (ANOVA, q-value <0.05) in the cytoplasmic fraction, 995 in the membrane-bound fraction, 561 in the nuclear fraction, and 159 in the chromatin and 590 in cytoskeletal fractions. The used RNA-Seq derived databases allowed us to identify and quantify variant proteins, and 88 RefProt variants were found at significantly different levels in the three cell lines. Of these variants, 39 were found in the cytoplasm, 39 in membrane-bound structures, 15 in the nucleus, 6 in the chromatin fraction and 23 in the cytoskeleton. Note, that 22 of the 88 RefProt variants were found in more than one cellular fraction.

Hierarchical clustering (Supplemental Figure 2A and Supplemental Table 4) pointed to six main groups of proteins: up-regulation in (1) PEO-4 cells, (2) SKOV-3 cells, and (3) in both cancerous cells; and down-regulation in (4) SKOV-3 cells, (5) PEO-4 cells, and (6) in both cancerous cells. Table 1 displays the number of significantly deregulated WT and RefProt variants quantified in the three cell lines.

Table 1. Wild-type and variant RefProts significantly varied (ANOVA, q-value <0.05). The number of WT and variant RefProts is displayed for the six main clusters identified upon LFQ proteomics.

Cluster		WT RefProts	RefProt variants
Upregulated	PEO-4 cells	482	10
	SKOV-3 cells	383	6
	Both cancerous cells	666	29
Downregulated	PEO-4 cells	195	4
	SKOV-3 cells	328	16
	Both cancerous cells	1154	54

An identical hierarchical clustering was performed on novel isoforms, resulting in the identification of 53 wild-type novel isoforms and three novel isoform variants that were significantly varied (ANOVA, q-value<0.05) between the three cell lines (Supplemental Figure

2B and Supplemental Table 5). One of these novel isoform variants, II_587587@Asn359Asp, was found upregulated in both cancerous cell lines in the cytoplasm and membrane-bound fractions. This protein is a novel isoform expressed from the *PMPCB* gene. A second variant, II_702738@Ala184Thr[Leu79LeuAsn72Asn], was found to be downregulated in SKOV-3 cells in the nuclear fraction. This novel isoform is encoded by the *WDR18* gene and possesses a substitution in position 184 and three silent mutations. II_597059@Glu65GlnAsn139AspAla57ValLys122ArgIle6ValGlu80Lys[Val118Val] was identified as upregulated in SKOV-3 cells in the chromatin-bound fraction. This protein is a novel isoform of HLA-H, which possesses seven mutations, one of which is a silent mutation.

The same workflow was used to compare the AltProt profiles between the three studied cell lines. In total, 73 AltProts were found at significantly altered levels and 41 of these were upregulated in the ovarian cancer cells, with 12 upregulated only in PEO-4 cells, nine in SKOV-3 cells, and 20 in both. Four AltProts were found to be downregulated only in PEO-4 cells or only in SKOV-3 cells, while 36 AltProts were downregulated in both cells (Supplemental tables 6 and 7). Figure 8 shows the distribution of the significantly altered AltProts over the five different subcellular fractions. We found 11 AltProts to be significantly regulated in more than one unique compartment. IP_067626, IP_070304, IP_108778, IP_147518, IP_178464, IP_213023, IP_246003 and IP_282949 were downregulated in both cancerous cells. Interestingly, IP_582685 (translated from a ncRNA transcript of the pseudogene *GDI2P1*) was identified upregulated at the membrane-bound fraction of both cancerous cells. Moreover, it was also found upregulated in the cytoplasmic and nuclear fractions of SKOV-3 cells. IP_062385 (translated from the 3’UTR part of the transcript ENST00000457946.1 coded by *ZMYM4* gene) was found upregulated in both cancerous cells’ cytoplasmic fractions, while it was downregulated in the cytoskeletal fraction of these cells. A similar observation was made for IP_774693 (translated from an ncRNA of *TUBAP2*): this AltProt was upregulated in the membrane-bound fractions of the cancerous cells yet, downregulated in their cytoplasmic fractions.

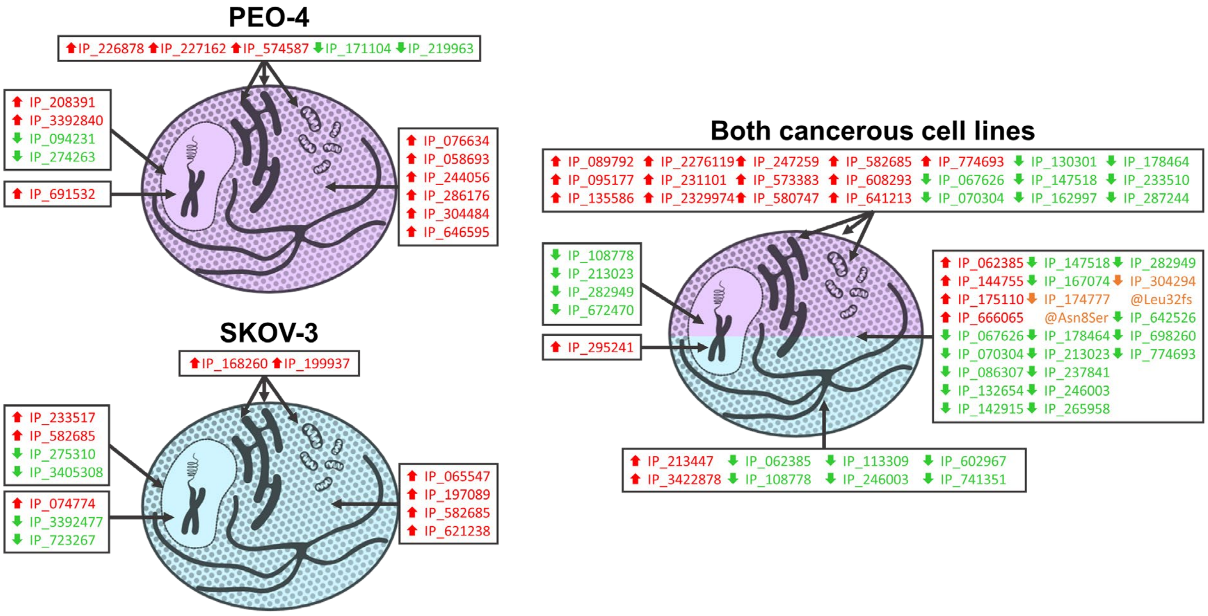

PEO-4
↑IP_226878 ↑IP_227162 ↑IP_574587 ↓IP_171104 ↓IP_219963
↑ IP_208391
↑ IP_3392840
↓ IP_094231
↓ IP_274263
↑ IP_691532
↑ IP_076634
↑ IP_058693
↑ IP_244056
↑ IP_286176
↑ IP_304484
↑ IP_646595
SKOV-3
↑IP_168260 ↑IP_199937
↑ IP_233517
↑ IP_582685
↓ IP_275310
↓ IP_3405308
↑ IP_074774
↓ IP_3392477
↓ IP_723267
↑ IP_065547
↑ IP_197089
↑ IP_582685
↑ IP_621238
Both cancerous cell lines
↑ IP_089792 ↑ IP_2276119 ↑ IP_247259 ↑ IP_582685 ↑ IP_774693 ↓ IP_130301 ↓ IP_178464
↑ IP_095177 ↑ IP_231101 ↑ IP_573383 ↑ IP_608293 ↓ IP_067626 ↓ IP_147518 ↓ IP_233510
↑ IP_135586 ↑ IP_2329974 ↑ IP_580747 ↑ IP_641213 ↓ IP_070304 ↓ IP_162997 ↓ IP_287244
↓ IP_108778
↓ IP_213023
↓ IP_282949
↓ IP_672470
↑ IP_295241
↑ IP_062385 ↓ IP_147518 ↓ IP_282949
↑ IP_144755 ↓ IP_167074 ↓ IP_304294
↑ IP_175110 ↓ IP_174777 @Leu32fs
↑ IP_666065 @Asn8Ser ↓ IP_642526
↓ IP_067626 ↓ IP_178464 ↓ IP_698260
↓ IP_070304 ↓ IP_213023 ↓ IP_774693
↓ IP_086307 ↓ IP_237841
↓ IP_132654 ↓ IP_246003
↓ IP_142915 ↓ IP_265958
↑ IP_213447 ↓ IP_062385 ↓ IP_113309 ↓ IP_602967
↑ IP_3422878 ↓ IP_108778 ↓ IP_246003 ↓ IP_741351

Figure 8. AltProts with significantly changed levels exclusively in one of two cancerous cell lines or common in both (ANOVA, FDR <0.05). For each cell line, the subcellular compartment, the AltProts upregulated (red) and downregulated (green) are shown.

Note that only two AltProt variants were found at significantly different levels. IP_174777 is a 53-amino acid AltProt encoded from the 3'UTR RNA of the *TMEM245* gene. During the creation of our databases, a single base substitution (23A>G) in transcript ENST00000374586 led to the prediction of the variant IP_174777@Asn8Ser. This mutant AltProt was identified as significantly downregulated in both cancerous cells, compared to the epithelial ovarian cell line. The second AltProt variant identified as downregulated in the cancerous ovarian cell lines was IP_304294@Leu32fs. The WT AltProt, IP_304294, is a 57-amino acid protein coded by the *MTMR1* gene and is translated from the 3'UTR of the transcript ENST00000445323. A guanine deletion at position 93 results in a reading frame shift at leucine 32. This shortens the protein to 44 amino acids and substituted the last 13 amino acids. For both proteins, a cytoplasmic domain was predicted by Phobius, and this prediction remained unchanged after the frame shift.

Proteome and transcriptome functional annotation

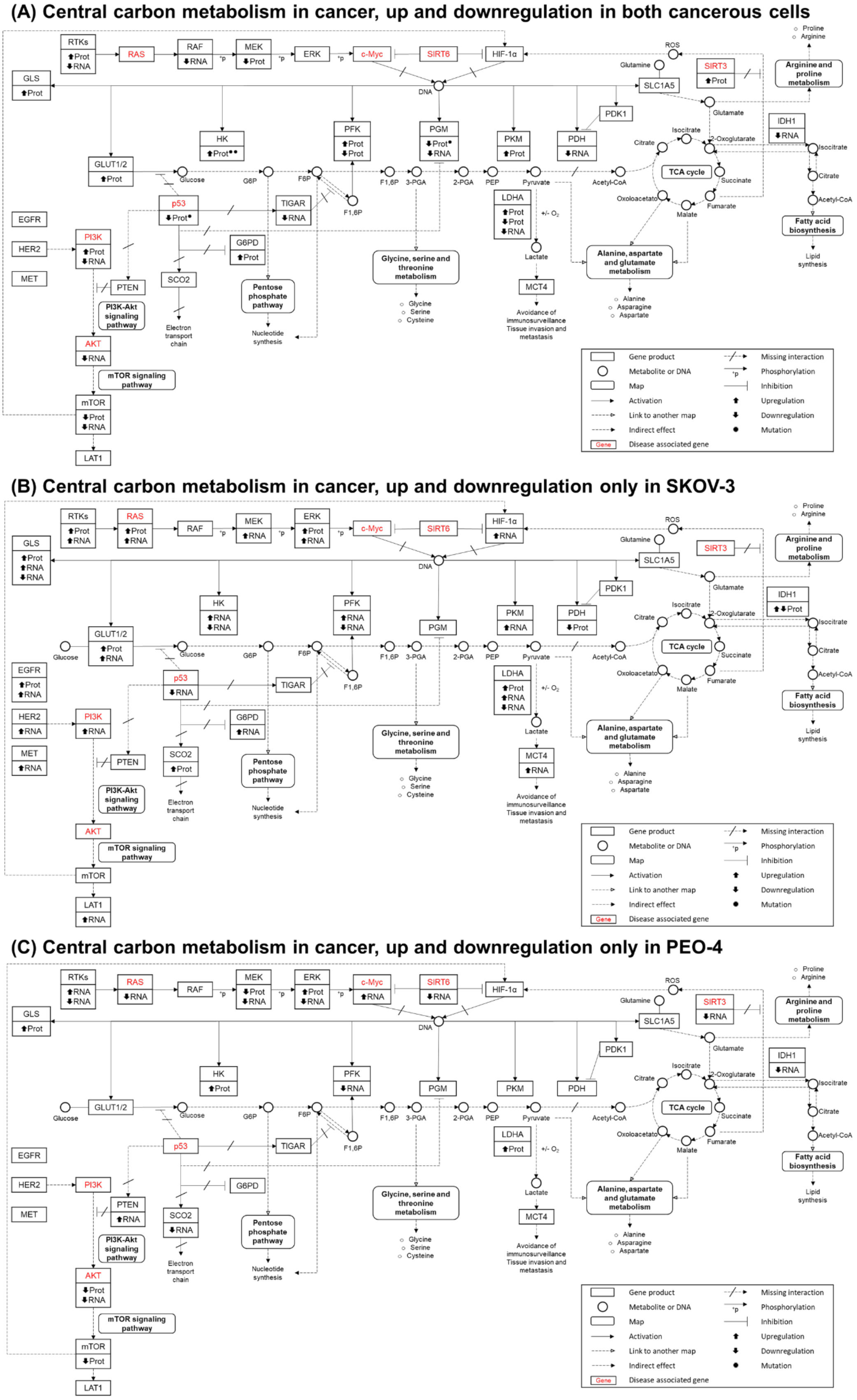
To integrate and interpret the data obtained from the differentially expressed reference proteome and transcriptome, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID) (52). This online tool allows users to perform GO term enrichment, cluster redundant enriched terms, visualize enriched pathway maps and extract gene functionality and literature.

The RefProts identified as upregulated in cancerous cells were submitted to DAVID and showed that two major cancer-related KEGG pathways (39) were significantly enriched: central carbon metabolism in cancer (hsa05230; p-value: 1.90E-04) and chemical carcinogenesis - reactive oxygen species (hsa05208; p-value: 5.26E-06). The KEGG pathway proteoglycans in cancer (hsa05205; p-value: 0.026) was significantly enriched among the downregulated cancer RefProts.

Regulated protein clusters in SKOV-3 cells were found significantly enriched for the central carbon metabolism in cancer pathways (p-value: 7.3E-5). On the contrary, no significant enrichment was identified in PEO-4 cells. Based on this difference we presented the protein and transcript expression profiles on an adapted central carbon metabolism pathway in a cancer pathway map (Figure 9). The complete list of genes and proteins enriched for this pathway can be found in Supplemental Table 8. One observes a significant upregulation of the NRAS protein in the RAS/RAF/MEK/ERK/c-Myc pathway in SKOV-3 cells (ANOVA q-value: 0.017). On the other hand, its transcript levels were significantly downregulated in PEO-4 cells (ANOVA q-value: 0.0004). Moreover, for the other two members of the oncogene RAS family, no significant variation was found at the proteome level whereas on the transcript level, *HRAS* was downregulated in PEO-4 cells (ANOVA q-value: 3.7E-6) and *KRAS* upregulated in SKOV-3 cells (ANOVA q-value: 5.58E-5). Other differences were observed for the MEK kinases MAP2K1 and MAP2K2; for instance, MAP2K2 was significantly downregulated in both cancerous cells' membrane-bound fraction (ANOVA q-value: 0.005) and downregulated in the PEO-4 cytoskeletal fraction (ANOVA q-value: 0.028). MAP2K1 was found downregulated in PEO-4 cells (ANOVA q-value: 2.29E-6) while its transcript level was found upregulated in SKOV-3 cells (ANOVA q-value: 1.49E-5).

In another part of the central carbon metabolism in cancer pathway, SIRT6 and SIRT3 are considered as cancer associated genes (53–55). It has been found that downregulation of SIRT6 increased ovarian cancer cells growth (55). The transcript levels of SIRT6, a tumour suppressor gene, were found downregulated in PEO-4 cells (ANOVA q-value: 4.65E-6), while the transcript levels of c-Myc, an oncogene, were upregulated in these cells (ANOVA q-value:

3.88E-5). Protein levels of SIRT3, another tumour suppressor gene, were upregulated in both cancerous cells (ANOVA q-value: 0.005), while its transcript levels were found downregulated in PEO-4 cells (ANOVA q-value: 0.0001). The expression of the oncogenic PI3K family was also found significantly regulated among the three cell lines. PIK3R1 was upregulated in both cancerous cells' cytoplasmic fraction (ANOVA q-value: 0.037), while its transcript was only upregulated in SKOV-3 cells (ANOVA q-value: 2.31E-5). Additionally, the transcripts of *PIK3CB* (ANOVA q-value: 0.0001) and *PIK3R2* (ANOVA q-value: 0.005) were also only upregulated in these cells. On the contrary, the *PIK3CA* (ANOVA q-value: 0.001) and *PIK3CD* (ANOVA q-value: 0.0001) transcripts were found downregulated in both cancerous cells.



(A) Central carbon metabolism in cancer, up and downregulation in both cancerous cells

(B) Central carbon metabolism in cancer, up and downregulation only in SKOV-3

(C) Central carbon metabolism in cancer, up and downregulation only in PEO-4

Figure 9. RefProts and genes significantly varied (ANOVA, FDR <0.05) in the central carbon metabolism in the cancer pathway. (A) Central carbon metabolism in cancer, up and downregulation in both cancerous cells. (B) Central carbon metabolism in cancer, up and downregulation only in SKOV-3. (C) Central carbon metabolism in cancer, up and downregulation only in PEO-4.

Other oncogenes in the central carbon metabolism cancer pathway are members of the *AKT* family. AKT1 protein (ANOVA q-value: 0.0002) and transcript levels (ANOVA q-value: 2E-5) were downregulated in PEO-4 cells. For AKT2 and AKT3, no significant variation in protein expression was found, while their transcript levels were significantly downregulated in both cancerous cells (ANOVA q-value: 0.02 and 3.6E-6).

With our proteogenomic workflow, we could identify a variant form of p53 (ENSP00000269305.8: p.Pro72Arg), an amino acid substitution that stems from the c.215C>G variant in *TP53*. This p53 mutant was significantly downregulated in both cancerous cells' cytoplasmic (ANOVA q-value: 0.0036) and cytoskeletal (ANOVA q-value: 0.0096) fractions, while its transcript levels were only significantly downregulated in SKOV-3 cells (ANOVA p-value: 1.17E-10). Three other RefProt variants were identified in this pathway. ENSP00000359991.5: p.Thr238Met, a mutant of PGAM1 was downregulated in both cancerous cells (ANOVA q-value: 0.0013), while two mutants of HKDC1 were upregulated in both cancerous cells; ENSP00000346643.5: p.Thr124Ile, p.Asn917Lys, p.Arg827Trp, p.Trp721Arg, [p.Phe601Phe] (ANOVA q-value: 0.008) and ENSP00000346643.5: p.Thr124Ile, p.Asn917Lys, p.Trp721Arg, [p.Phe601Phe] (ANOVA q-value: 0.023).

Crosslinking network analysis

The computational analysis of the crosslinked samples was carried out as described in (30), which allowed us to generate a protein interaction map in Cytoscape (56) (Supplemental Figure 3). A total of 90 crosslinks were identified (Supplemental table 9), among them 20 intra-crosslinks were identified, which do not give interactome information, but might be useful for structural studies. In this protein network (Supplemental Figure 3), 28 protein-protein interactions (PPIs) were found in PEO-4 cells (marked in purple), 27 in SKOV-3 cells (marked in blue) and 35 in T1074 cells (marked in green). From these pairs, 12 crosslink interactions were identified in at least two cell lines. Among all the crosslinked pairs, 20 involved AltProts, four crosslinks were AltProt-AltProt interactions, and 13 AltProt-RefProt crosslinks were identified. The latter were considered most important for our study as they provide hints to an AltProt's physiological or pathological involvement.

To attribute functions to an AltProt from this set of PPIs, we retrieved the known interactions from the STRING (57, 58), BioGrid (59) and IntAct (60) databases and included the identified crosslinked interactions (Supplemental Figure 4). Additionally, for the RefProts that did not present a referenced STRING interaction within the crosslinked network, the addition of three STRING interactors has been performed to expand the network. We observed that seven PPIs had already been described (pink lines): B2M-HLA-B, B2M-HLA-A, ITGA5-ITGA1, TUBA1C-TUBB, HIST3H2A-HIST2H3D, PRC1-ORC1 and VP39-VPS13C. Using this network, a molecular function GO term and KEGG pathway enrichment analysis was performed with the ClueGO App(61) from Cytoscape. The interactions between AltProts and RefProts were displayed along with the enriched GO terms (Figure 10). Four direct AltProt-RefProt-GO-term interactions were detected. The AltProt IP_192190 was crosslinked to KIF13A in PEO-4 cells and linked to the vesicle-mediated transport of plasma membrane (GO:0098876), Golgi to plasma membrane protein transport (GO:0043001), protein localization to plasma membrane (GO:0072659) and post-Golgi vesicle mediated transport (GO:0006892). The AltProt IP_136846 was identified as crosslinked to LGALS1 in T1074 cells, which is linked to the GO terms viral entry into host cell (GO:0046718) and biological process involved in interaction with host (GO:0051701). Similarly, IP_235241, crosslinked to ITGA5 in T1074 cells, was linked to the phagosome KEGG pathway (KEGG:04145) and the GO terms virus receptor activity (GO:0001618), biological process involved in interaction with host (GO:0051701) and viral entry into host cell (GO:0046718). Finally, IP_183088 was crosslinked to POLD3 in T1074 and PEO-4 cells. POLD3 is part of the DNA polymerase involved in the replication and reparation of DNA and linked to the UV-damage excision repair (GO:0070914) and response to UV (GO:0009411) GO terms.

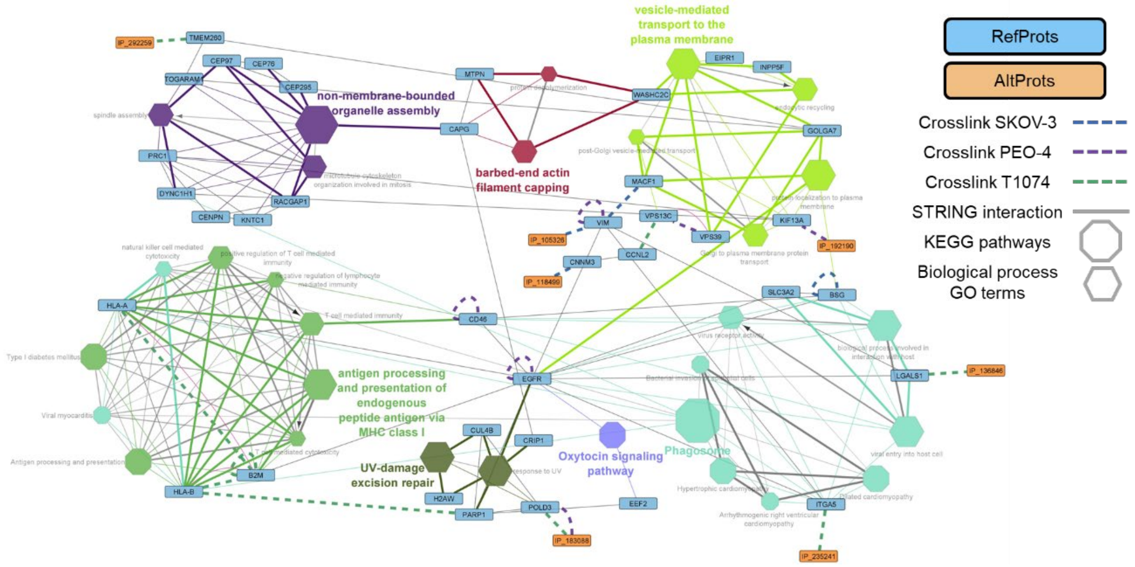

IP_292259
TMEM260
CEP97
CEP76
TOGARAM1
CEP295
non-membrane-bounded organelle assembly
spindle assembly
PRC1
microtubule cytoskeleton organization involved in mitosis
DYNC1H1
RACGAP1
CENPN
KNTC1
MTPN
protein depolymerization
CAPG
barbed-end actin filament capping
vesicle-mediated transport to the plasma membrane
EIPR1
INPP5F
WASHC2C
endocytic recycling
GOLGA7
post-Golgi vesicle-mediated transport
MACF1
protein localization to plasma membrane
VIM
VPS13C
KIF13A
VPS39
IP_105326
IP_192190
CCNL2
CNNM3
Golgi to plasma membrane protein transport
IP_118499
SLC3A2
BSG
natural killer cell mediated cytotoxicity
positive regulation of T cell mediated immunity
negative regulation of lymphocyte mediated immunity
HLA-A
T cell mediated immunity
CD46
virus receptor activity
biological process involved in interaction with host
Type I diabetes mellitus
antigen processing and presentation of endogenous peptide antigen via MHC class I
EGFR
Bacterial invasion of epithelial cells
LGALS1
IP_136846
Viral myocarditis
T cell mediated cytotoxicity
CUL4B
CRIP1
Oxytocin signaling pathway
Phagosome
viral entry into host cell
Antigen processing and presentation
UV-damage excision repair
response to UV
B2M
HLA-B
H2AW
Hypertrophic cardiomyopathy
Dilated cardiomyopathy
PARP1
POLD3
EEF2
ITGA5
Arrhythmogenic right ventricular cardiomyopathy
IP_183088
IP_235241
RefProts
AltProts
Crosslink SKOV-3
Crosslink PEO-4
Crosslink T1074
STRING interaction
KEGG pathways
Biological process GO terms

Figure 10: GO molecular function enrichment network generated with ClueGO in Cytoscape. GO enrichment was generated from the accession numbers of Supplemental figure 4. AltProts are marked in orange and RefProts in blue. Enriched GO terms

are displayed as hexagons. KEGG pathways are displayed as octagons and crosslinks are marked in blue (SKOV-3 cells), purple (PEO-4 cells) and green (T1074 cells) dashed lines.

Three AltProt-GO-term/KEGG pathways indirect links were identified. IP_292259, crosslinked to TMEM260 in T1074 cells, and TMEM260 possesses a STRING interaction with TOGARAM, which is linked to the non-membrane-bounded organelle assembly (GO:0140694), spindle assembly (GO:0051225) and microtubule cytoskeleton organization involved in mitosis (GO:1902850). Additionally, TMEM260 interacts with GOLGA7, which is linked to GO terms related to the vesicle-mediated transport to the plasma membrane. In addition, two AltProts were also identified to be related to these GO terms: IP_105326 and IP_118499. The former was crosslinked to VIM in SKOV-3 cells, and VIM was crosslinked to MACF1, which is linked to vesicle-mediated transport GO terms. IP_118499 was found crosslinked to CNNM3 in SKOV-3 cells, which processes a STRING interaction with CCNL2, which was crosslinked to VPS13C, which is linked to vesicle-mediated transport GO terms.

To confirm the probability of the observed interactions, we analysed 3D models of RefProt-AltProts using unguided interaction docking between the two partners (as described in (30)). The structures of the AltProts were predicted using I-Tasser(62), while those of the interactors were predicted using ClusPro (63). The RefProt, for which the structure was predicted by AlphaFold(64), was used as a receptor of the AltProt, which was smaller in structure. By measuring the distance of the predicted interactions, we confirmed the observed interactions from XL-MS with a mean of 23.467 Å (Supplemental Figure 5), which is consistent with the distances described in the literature for DSSO, ranging from 5.3 (34) to 30 Å (35).

DISCUSSION

Proteogenomics establishes a direct connection between the genome blueprint and the constructed proteome. We utilized this approach to explore potential implications of AltProts in ovarian cancer. We selected the PEO-4 cell line possessing a high-grade serous histology, the SKOV-3 clear cell carcinoma cell line, and the T1074 ovarian epithelial cell line, originally derived from normal human ovarian surface epithelial cells, serving as a non-tumorous control.

The transcriptome as a source of information for the proteomic perspective

The transcriptomic analysis employing DESeq2 to analyse the RNA-seq data enabled us to identify clusters of regulated genes in the cancer cell models. Each cell line showed about 500 uniquely expressed genes. Among the 540 genes uniquely expressed in PEO-4 cells, proto-oncogenes SSX1, SSX2 and SSX2B were found, along with an additional 24 genes related to cancer according to the Gene-Disease Associations Dataset (GAD) (65). Among the 406 genes uniquely expressed in SKOV-3 cells, 23 were related to cancer according to GAD. While transcriptomic analysis provided cell specificity information, the strength of this approach lies in the custom creation of cell-specific databases using OpenCustomDB. These databases contain a larger number of AltProt variants due to a high number of predicted AltProts. The ratio of variant RefProts to WT RefProts was greater than the ratio of variant AltProts to WT AltProts, which can be attributed to differences in sequence length. Longer genomic sequences have higher mutation rates and replication errors. Additionally, predicted AltProts mostly originate from ncRNAs, but mRNA CDS frame shifts and 3'UTRs also contributed significantly to the top 100,000 most abundant transcripts. This suggests a greater potential for ncRNAs to code for AltProts, although there is a larger abundance of mRNAs capable of coding for AltProts.

The proteogenomic approach of constructing a custom database, combined with reading frame prediction for AltProt generation, presents analytical challenges. However, our iterative triple SEQUEST HT processing workflow using the 100,000-abundance cut-off database in the first node overcomes the FDR limitations of a 400,000-sequences database (full database) search, which may increase the number of false positives and false negative identifications (16). To not lose possible identifications, such iterative workflows provide a stepwise increase in possible protein identifications by expanding the search space, until the last step with OpenProtDB, where proteins translated from ncRNAs not detected by RNA-Seq can be recovered. Finally, using Percolator, we removed false positive identifications by this semi-supervised machine learning algorithm (66). Percolator effectively estimates the statistical significance of peptide-spectrum matches and assigns confidence scores to identified peptides in a fast and accurate way. It enhances the rate of confident peptide identifications from a collection of tandem mass spectra (67).

A larger view on the proteomic landscape

Subcellular fractionation is a validated approach to decrease sample complexity and to maximize resolution in LC-MS/MS analysis. In our previous works (17, 30), such subcellular fractionation was proven beneficial for XL-MS workflows and provided better coverage of the proteome compared to analysing whole cell lysates (68). This enhanced the detection of low-abundant proteins (AltProts and crosslinked proteins). Furthermore, subcellular fractionation helps to determine the subcellular localization of AltProts and monitors changes under different cellular conditions (69). For instance, IP_062385 was found to be located in the cytoplasm and upregulated in cancerous cells, while downregulated in their cytoskeleton fractions. This may reflect a functional change linked to cancer, yet targeted studies will be necessary to prove such links between tumour development and AltProts re-localising over different cellular compartments. However, it is important to be note that subcellular fractionation based on the use of protein extraction using different detergents can lead to potential cross-contamination and inaccuracies in downstream data interpretation.

Subcellular fractionation led to the identification of ~6,000 common RefProts among the three cell lines. Over 3% of all identified proteins in each cell line were RefProt variants (Figure 6B). However, these ~180 RefProt variants require deeper characterization to understand their (pathological) role. Cell line-specific AltProts were also found in all three cell lines, AltProts in SKOV-3 and PEO-4 cells are of interest as potential new protein markers for OvCa. Among them, IP_715944@Leu44Pro (Figure 11) caught our attention as it is a variant AltProt not predicted in the T1074 RNA-Seq database. Moreover, six additional AltProts from this group were also not predicted, which highlights the importance of a cell-specific analysis to identify new biomarkers.

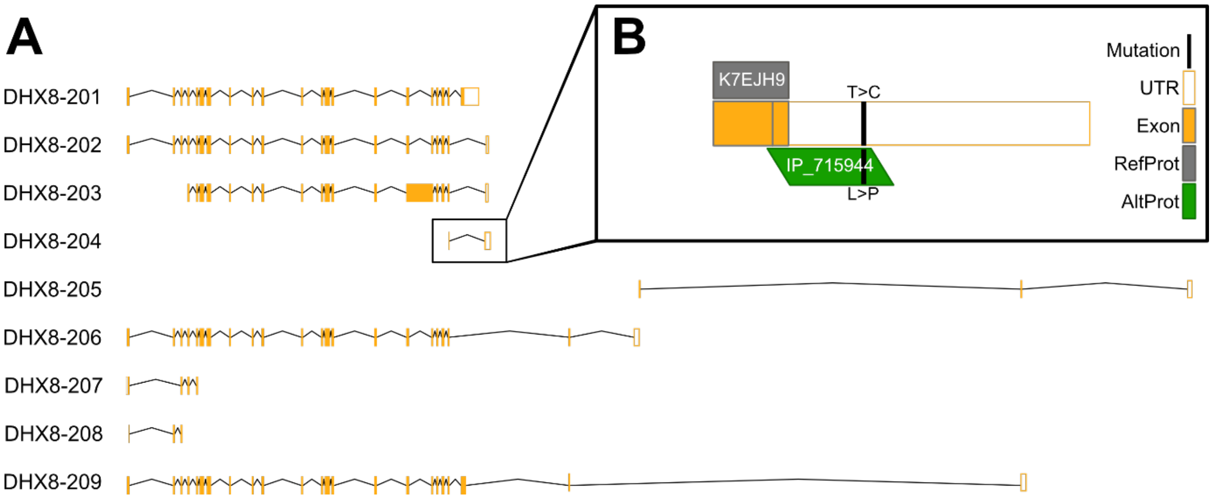


Figure 11. Synthesis of AltProts from the DHX8 gene (A) List of transcripts referenced in Ensembl. (B) Zoom on DHX8-204 described to translate to "K7EJH9", a predicted protein from TrEMBL without the 5'UTR part or methionine as the first amino acid, when IP_715944 is described from the overlap between the CDS and the 3'UTR. As a result, the mutation is only

observable by the proteogenomic construction, as it would be considered a silent mutation due to its position in the UTR part of K7EJH9.

Based on the LFQ proteome analysis data, AltProts were found to be upregulated in all compartments except the cytoskeleton in PEO-4 and SKOV-3 cells, while downregulation of AltProts was only observed in the membrane-bound and nuclear fractions in PEO-4 cells, and in the nuclear and chromatin fractions in SKOV-3 cells. Such differentially expressed AltProts can be important for distinguishing between cancer cell lines. When comparing both cancerous cell lines to T1074 cells, significant downregulation of AltProts was observed in all five compartments. AltProts upregulated in both cancerous cells were present in all compartments except the nucleus. These findings provide some insights into the specific expression of AltProts in high grade serous and non-serous OvCa. Functional domains were predicted for 23 out of 73 AltProts, which can help us understand their potential roles in interactions. Future targeted interactomic approaches such as Virotrap (70), BioID (71) and proximity ligation assays (72) could be used to identify the interaction partners of these AltProts, which may shed light on their involvement in the pathogenic development of OvCa or drug resistance.

Interpretation of the major protein and transcript fluctuations from the three cell-line highlights cancer-related KEGG pathways

NRAS, a member of the RAS oncogene family, is involved in cell signalling, regulation of cell growth, differentiation and angiogenesis. In ovarian clear cell carcinoma, no NRAS mutations were found in our SKOV-3 cell transcriptome data (73). Overexpression of NRAS was shown to increase tumor aggressiveness in mice (74). KRAS, another member of the RAS oncogene family, was found to be upregulated in SKOV-3 cells and in metastatic lesions in endometrial cancer (75), which is associated with adverse prognosis (76). Downregulation of HRAS has been linked to lower aggressiveness and reduced cell proliferation in certain types of cancer (46, 77, 78). Another branch of the pathway also shows MEK (mitogen-activated extracellular signal-regulated kinase) which is a kinase cascade pathway that plays a central role in carcinogenesis and the maintenance of several cancers. We found downregulation of MAP2K1 and MAP2K2 in both cancerous cell lines, as also evident from data in The Human Protein Atlas (79). In parallel, related to cancer metabolism, we observed *SIRT6* downregulation and *c-Myc* upregulation in PEO-4 cells. Lower levels of *SIRT6* are associated with poorer prognosis and increased tumour aggressiveness (54, 55). *SIRT6* also regulates ribosome metabolism by repressing *c-Myc* activity. As a result, higher levels of *c-Myc*, resulting from downregulation of *SIRT6*, promote energy production and biomolecule synthesis for rapid cell proliferation. On the other hand, *SIRT3* is described as a tumor suppressor gene in OvCa (80) and its expression increases in detached cells and tumor cells from malignant ascites, indicating its pro-metastatic role in OvCa (53). Our proteomic data show upregulation of SIRT3 in both cancerous cells, while *SIRT3* transcripts are downregulated in PEO-4 cells. Discordance between mRNA and protein levels has been observed in various studies (81–84), attributed to post-transcriptional regulation, transcript isoform switching and DNA variants (82, 85). We found that PIK3R1 (p85α) was upregulated in the tumoral cells, which also corresponds to the identified overexpression of PIK3R1 in an OvCa cohort of 98 patients (86). However, contrary to literature findings (87), transcript levels of PIK3CD were downregulated in both cancerous cell lines. Stronach *et al*. (88) and Liu *et al*. (89) have studied the role of the AKT kinase signalling pathway in OvCa cell proliferation, cell cycle regulation and anti-apoptosis. They discovered that SKOV-3 cells rely on AKT1 for cisplatin resistance, while PEO-4 cells depend on AKT3. In line with this study, in our dataset, both protein and transcript levels of AKT1 were found to be overexpressed in SKOV-3 cells.

On the importance of identifying variants

Among the significantly deregulated RefProts identified in our study, P53 rs1042522 was found downregulated in both cancer cell lines. The corresponding Pro72Arg substitution in the canonical P53 sequence (UniProtKB: P04637-1) occurs in a proline-rich, intrinsically disordered region (residues 64–92) (90). This region is described as rigid (91) and a substitution of one of the prolines in this region might decrease its stiffness. Moreover, position 72 is part of the binding site of P53 with the oncogenic protein MDM2 (92). Even though there is evidence suggesting that there may be an association between this mutation and OvCa risk, a meta-analysis by Schildkraut *et al.* could not confirm an association with OvCa (93). Additionally, using our proteogenomic approach we were able to confirm the observations of Yaginuma *et al.* (94) describing SKOV-3 as a null-WT-P53 cell line.

HKDC1 variants were found upregulated for both cancerous cells. Three (rs906219, rs1111335 and rs874556) of the four single nucleotide variations (SNVs) are reported as natural variants of HKDC1 (UniProtKB: Q2TB90). The last, SNV rs138235256 is not reported in UniProt and does not possess any clinical significance so far. Additionally, the variant ENSP00000359991.5@Thr238Met (PGAM1) was identified downregulated in both cancer cells and results from rs202055965 SNV (C>T).

XL-MS reveals clues about AltProt functions based on AltProt-RefProt PPIs

IP_183088, a 38-amino acid AltProt, is encoded by *MAPK8* and was found to interact with POLD3 in T1074 and PEO-4 cell lines. Figure 12A displays the model of the human polymerase delta holoenzyme complex (PDB: 6s1m). Herein, the four subunits of the complex are shown (POLD1 turquoise, POLD2 green, POLD3 blue and POLD4 yellow), additionally, the proliferating cell nuclear antigen is displayed in light blue and the AltProt IP_183088 in red, together with its crosslinks. Figure 12B zooms in on the crosslinked region of POLD3-IP_183088, revealing that this interaction occurs in the region where POLD2 and POLD3 interact. Our transcriptomic data point to POLD3 downregulation in both cancerous cells. This correlates with the findings of Willes *et al*. who described that POLD3 downregulation is correlated with a poor cancer outcome (95) and those of Weberpals *et al.* who showed that *POLD3* is overexpressed in patients with high grade serous ovarian carcinoma and with good response to carboplatin/paclitaxel (96). On the other hand, the inhibition of the interaction between POLD3 and POLD2 driven by IP_183088 can reflect two effects. (i) An increase of the mutagenesis in the cells upon reduced activity of the POLD complex and, therefore, errors in DNA replication are more likely to occur and go unrepaired, which can be expected in PEO-4 cells. (ii) A regulatory system of the POLD complex, where the POLD3-IP183088 interaction in T1074 cells could lead to cell apoptosis; Murga *et al.* (97) showed that POLD3 stabilizes the POLD complex and in its absence, the cell is driven to apoptosis. The difficulty of detecting interactions by XL-MS means that we cannot claim that the observed interactions are cell-type specific, but they do provide information about potential protein functions for unreferenced proteins. The use of this approach for studying AltProt thus makes sense, and in the case of IP_183088, allowed us to hypothesize a regulatory function of POLD3-POLD2 interaction, the stability of the POLD complex and therefore an effect in the regulation of DNA replication error repair.

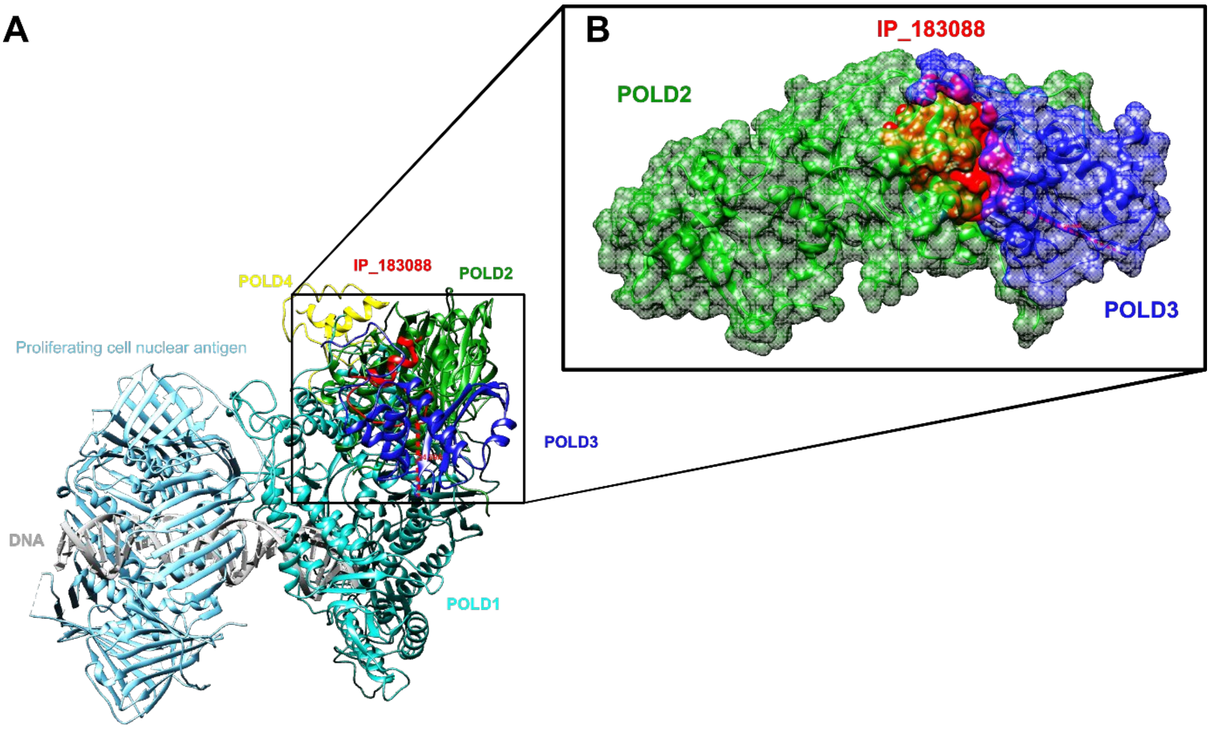


Figure 12. IP_183088 (AltMAPK8) predicted models docked to the human polymerase delta holoenzyme complex. (A) The interaction of IP_183088 and the full POLD complex is shown. The distance between the two lysines involved in the crosslink is 24.59 Å. (B) Zoom of the interaction of IP_183088 and POLD3. The surface representation shows the possible placement of IP_183088 at the interaction site of POLD3 and POLD2.

To conclude, one main advantage of the databases generated by OpenCustomDB is the possibility of predicting and identifying cell-specific proteins in cell lines and, in the future, in patient samples, resulting in a big step forward towards personalized medicine. Subcellular fractionation allowed us to study differences in the reference, alternative and novel isoforms proteome of OvCa cell lines compared to a non-tumoral ovarian epithelial cell. Additionally, it allowed us to identify RefProts variants and understudied AltProts and their variants. The versatility of these databases allowed us to identify AltProt-RefProts PPIs and gave some clue about the function of AltProts, which however need to be validated. In summary, our large-scale characterization study revealed other research targets and demonstrated the complexity of the cell proteome and its largely unmapped ghost proteome.

DATA AVAILABILITY

"The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE(98) *partner repository with the dataset identifier PXD045689".*

Reviewer account details:

Username: reviewer_pxd045689@ebi.ac.uk

Password: nI04XiIQ

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

AUTHOR CONTRIBUTIONS

Diego Fernando Garcia-del Rio: Conceptualization, Formal analysis, Methodology, Validation, Writing & editing—original draft. Mehdi Derhourhi: Formal analysis, Methodology, Writing. Amelie Bonnefond: Methodology, Writing—review & editing. Sébastien Leblanc: Formal analysis, Methodology, Writing—review & editing. Noé Guilloy: Methodology, Writing—review & editing. Xavier Roucou: Methodology, Writing—review & editing. Kris Gevaert: review & editing, Funding. Sven Eyckerman: review & editing, Funding. Michel Salzet: Conceptualization, review & editing, Funding. Tristan Cardon: Conceptualization, Methodology, Validation, Writing & editing—original draft.

FUNDING

This research was supported by funding from I-SITE, Institut National de la Santé et de la Recherche Médicale (Inserm) and Université de Lille and by The Research Foundation - Flanders (FWO), project number G008018N.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

1. The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Research*, **43**, D204–D212.

2. Breuza,L., Poux,S., Estreicher,A., Famiglietti,M.L., Magrane,M., Tognolli,M., Bridge,A., Baratin,D., Redaschi,N., and UniProt Consortium (2016) The UniProtKB guide to the human proteome. *Database (Oxford)*, **2016**, bav120.

3. Mouilleron,H., Delcourt,V. and Roucou,X. (2016) Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res*, **44**, 14–23.

4. Hao,Y., Zhang,L., Niu,Y., Cai,T., Luo,J., He,S., Zhang,B., Zhang,D., Qin,Y., Yang,F., *et al.* (2018) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Briefings in Bioinformatics*, **19**, 636–643.

5. Galindo,M.I., Pueyo,J.I., Fouix,S., Bishop,S.A. and Couso,J.P. (2007) Peptides Encoded by Short ORFs Control Development and Define a New Eukaryotic Gene Family. *PLOS Biology*, **5**, e106.

6. Albuquerque,J.P., Tobias-Santos,V., Rodrigues,A.C., Mury,F.B. and Fonseca,R.N. da (2015) small ORFs: A new class of essential genes for development. *Genet. Mol. Biol.*, **38**, 278–283.

7. Ruiz-Orera,J., Messeguer,X., Subirana,J.A. and Alba,M.M. (2014) Long non-coding RNAs as a source of new peptides. *eLife*, **3**, e03523.

8. Slavoff,S.A., Heo,J., Budnik,B.A., Hanakahi,L.A. and Saghatelian,A. (2014) A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J Biol Chem*, **289**, 10950–10957.

9. Brunet,M.A., Brunelle,M., Lucier,J.-F., Delcourt,V., Levesque,M., Grenier,F., Samandi,S., Leblanc,S., Aguilar,J.-D., Dufour,P., *et al.* (2018) OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Research*, 10.1093/nar/gky936.

10. Cardon,T., Fournier,I. and Salzet,M. (2021) Shedding Light on the Ghost Proteome. *Trends in Biochemical Sciences*, **46**, 239–250.

11. Brunet,M.A. and Roucou,X. (2019) Mass Spectrometry-Based Proteomics Analyses Using the OpenProt Database to Unveil Novel Proteins Translated from Non-Canonical Open Reading Frames. *JoVE (Journal of Visualized Experiments)*, 10.3791/59589.

12. Kozak,M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**, 187–208.

13. Kozak,M. (2006) Rethinking some mechanisms invoked to explain translational regulation in eukaryotes. *Gene*, **382**, 1–11.

14. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, **31**, 365–370.

15. Brunet,M.A., Lucier,J.-F., Levesque,M., Leblanc,S., Jacques,J.-F., Al-Saedi,H.R.H., Guilloy,N., Grenier,F., Avino,M., Fournier,I., *et al.* (2021) OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Research*, **49**, D380–D388.

16. Guilloy,N., Brunet,M.A., Leblanc,S., Jacques,J.-F., Hardy,M.-P., Ehx,G., Lanoix,J., Thibault,P., Perreault,C. and Roucou,X. (2023) OpenCustomDB: Integration of Unannotated Open

Reading Frames and Genetic Variants to Generate More Comprehensive Customized Protein Databases. *J. Proteome Res.*, **22**, 1492–1500.

17. Garcia-del Rio,D.F., Cardon,T., Eyckerman,S., Fournier,I., Bonnefond,A., Gevaert,K. and Salzet,M. (2023) Employing non-targeted interactomics approach and subcellular fractionation to increase our understanding of the ghost proteome. *iScience*, **26**.

18. Cao,X., Khitun,A., Harold,C.M., Bryant,C.J., Zheng,S.-J., Baserga,S.J. and Slavoff,S.A. (2022) Nascent alt-protein chemoproteomics reveals a pre-60S assembly checkpoint inhibitor. *Nat Chem Biol*, **18**, 643–651.

19. Cardon,T., Salzet,M., Franck,J. and Fournier,I. (2019) Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation. *Biochimica et Biophysica Acta (BBA) - General Subjects*, **1863**, 1458–1470.

20. D'Lima,N.G., Ma,J., Winkler,L., Chu,Q., Loh,K.H., Corpuz,E.O., Budnik,B.A., Lykke-Andersen,J., Saghatelian,A. and Slavoff,S.A. (2017) A human microprotein that interacts with the mRNA decapping complex. *Nat Chem Biol*, **13**, 174–180.

21. Matsumoto,A., Pasut,A., Matsumoto,M., Yamashita,R., Fung,J., Monteleone,E., Saghatelian,A., Nakayama,K.I., Clohessy,J.G. and Pandolfi,P.P. (2017) mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*, **541**, 228–232.

22. Stein,C.S., Jadiya,P., Zhang,X., McLendon,J.M., Abouassaly,G.M., Witmer,N.H., Anderson,E.J., Elrod,J.W. and Boudreau,R.L. (2018) Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Rep*, **23**, 3710-3720.e8.

23. Cardon,T., Ozcan,B., Aboulouard,S., Kobeissy,F., Duhamel,M., Rodet,F., Fournier,I. and Salzet,M. (2020) Epigenetic Studies Revealed a Ghost Proteome in PC1/3 KD Macrophages under Antitumoral Resistance Induced by IL-10. *ACS Omega*, 10.1021/acsomega.0c02530.

24. Delcourt,V., Franck,J., Leblanc,E., Narducci,F., Robin,Y.-M., Gimeno,J.-P., Quanico,J., Wisztorski,M., Kobeissy,F., Jacques,J.-F., *et al.* (2017) Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer. *EBioMedicine*, **21**, 55–64.

25. Huang,J.-Z., Chen,M., Chen,D., Gao,X.-C., Zhu,S., Huang,H., Hu,M., Zhu,H. and Yan,G.-R. (2017) A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Molecular Cell*, **68**, 171-184.e6.

26. Polycarpou-Schwarz,M., Groß,M., Mestdagh,P., Schott,J., Grund,S.E., Hildenbrand,C., Rom,J., Aulmann,S., Sinn,H.-P., Vandesompele,J., *et al.* (2018) The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene*, **37**, 4750–4768.

27. Brunet,M.A., Jacques,J.-F., Nassari,S., Tyzack,G.E., McGoldrick,P., Zinman,L., Jean,S., Robertson,J., Patani,R. and Roucou,X. (2020) The FUS gene is dual-coding with both proteins contributing to FUS-mediated toxicity. *EMBO reports*, 10.15252/embr.202050640.

28. Cao,X., Chen,Y., Khitun,A. and Slavoff,S.A. (2023) BONCAT-based Profiling of Nascent Small and Alternative Open Reading Frame-encoded Proteins. *Bio Protoc*, **13**, e4585.

29. Slavoff,S.A., Mitchell,A.J., Schwaid,A.G., Cabili,M.N., Ma,J., Levin,J.Z., Karger,A.D., Budnik,B.A., Rinn,J.L. and Saghatelian,A. (2013) Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat Chem Biol*, **9**, 59–64.

30. Garcia-del Rio,D.F., Fournier,I., Cardon,T. and Salzet,M. (2023) Protocol to identify human subcellular alternative protein interactions using cross-linking mass spectrometry. *STAR Protocols*, **4**, 102380.

31. Vanderperre,B., Staskevicius,A.B., Tremblay,G., McCoy,M., O'Neill,M.A., Cashman,N.R. and Roucou,X. (2011) An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *The FASEB Journal*, **25**, 2373–2386.

32. Zhang,Q., Vashisht,A.A., O'Rourke,J., Corbel,S.Y., Moran,R., Romero,A., Miraglia,L., Zhang,J., Durrant,E., Schmedt,C., *et al.* (2017) The microprotein Minion controls cell fusion and muscle formation. *Nat Commun*, **8**, 15664.

33. Yosten,G.L.C., Liu,J., Ji,H., Sandberg,K., Speth,R. and Samson,W.K. (2016) A 5′-upstream short open reading frame encoded peptide regulates angiotensin type 1a receptor production and signalling via the β-arrestin pathway. *The Journal of Physiology*, **594**, 1601–1605.

34. Kao,A., Chiu,C., Vellucci,D., Yang,Y., Patel,V.R., Guan,S., Randall,A., Baldi,P., Rychnovsky,S.D. and Huang,L. (2011) Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes. *Mol Cell Proteomics*, **10**, M110.002212.

35. Hevler,J.F., Lukassen,M.V., Cabrera-Orefice,A., Arnold,S., Pronker,M.F., Franc,V. and Heck,A.J.R. (2021) Selective cross-linking of coinciding protein assemblies by in-gel cross-linking mass spectrometry. *The EMBO Journal*, **40**, e106174.

36. Berek,J.S., Renz,M., Kehoe,S., Kumar,L. and Friedlander,M. (2021) Cancer of the ovary, fallopian tube, and peritoneum: 2021 update. *International Journal of Gynecology & Obstetrics*, **155**, 61–85.

37. Wentzensen,N., Poole,E.M., Trabert,B., White,E., Arslan,A.A., Patel,A.V., Setiawan,V.W., Visvanathan,K., Weiderpass,E., Adami,H.-O., *et al.* (2016) Ovarian Cancer Risk Factors by Histologic Subtype: An Analysis From the Ovarian Cancer Cohort Consortium. *J Clin Oncol*, **34**, 2888–2898.

38. Stewart,C., Ralyea,C. and Lockwood,S. (2019) Ovarian Cancer: An Integrated Review. *Seminars in Oncology Nursing*, **35**, 151–156.

39. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27–30.

40. Soga,T. (2013) Cancer metabolism: Key players in metabolic reprogramming. *Cancer Science*, **104**, 275–281.

41. Warburg,O. (1925) The Metabolism of Carcinoma Cells1. *The Journal of Cancer Research*, **9**, 148–163.

42. Vander Heiden,M.G., Cantley,L.C. and Thompson,C.B. (2009) Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science*, **324**, 1029–1033.

43. Wolf,C.R., Hayward,I.P., Lawrie,S.S., Buckton,K., McIntyre,M.A., Adams,D.J., Lewis,A.D., Scott,A.R.R. and Smyth,J.F. (1987) Cellular heterogeneity and drug resistance in two ovarian adenocarcinoma cell lines derived from a single patient. *International Journal of Cancer*, **39**, 695–702.

44. Langdon,S.P., Lawrie,S.S., Hay,F.G., Hawkes,M.M., McDonald,A., Hayward,I.P., Schol,D.J., Hilgers,J., Leonard,R.C.F. and Smyth,J.F. Characterization and Properties of Nine Human Ovarian Adenocarcinoma Cell Lines.

45. Fogh,J., Fogh,J.M. and Orfeo,T. (1977) One hundred and twenty-seven cultured human tumor cell lines producing tumors in nude mice. *J Natl Cancer Inst*, **59**, 221–226.

46. Hernandez,L., Kim,M.K., Lyle,L.T., Bunch,K.P., House,C.D., Ning,F., Noonan,A.M. and Annunziata,C.M. (2016) Characterization of ovarian cancer cell lines as in vivo models for preclinical studies. *Gynecol Oncol*, **142**, 332–340.

47. Hallas-Potts,A., Dawson,J.C. and Herrington,C.S. (2019) Ovarian cancer cell lines derived from non-serous carcinomas migrate and invade more aggressively than those derived from high-grade serous carcinomas. *Sci Rep*, **9**, 5515.

48. Tabb,D.L., Eng,J.K. and Yates,J.R. (2001) Protein Identification by SEQUEST. In James,P. (ed), *Proteome Research: Mass Spectrometry*, Principles and Practice. Springer, Berlin, Heidelberg, pp. 125–142.

49. McGinnis,S. and Madden,T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, **32**, W20-25.

50. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

51. Käll,L., Krogh,A. and Sonnhammer,E.L.L. (2004) A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, **338**, 1027–1036.

52. Sherman,B.T., Hao,M., Qiu,J., Jiao,X., Baseler,M.W., Lane,H.C., Imamichi,T. and Chang,W. (2022) DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Research*, **50**, W216–W221.

53. Dong,X.-C., Jing,L.-M., Wang,W.-X. and Gao,Y.-X. (2016) Down-regulation of SIRT3 promotes ovarian carcinoma metastasis. *Biochem Biophys Res Commun*, **475**, 245–250.

54. Sebastián,C., Zwaans,B.M.M., Silberman,D.M., Gymrek,M., Goren,A., Zhong,L., Ram,O., Truelove,J., Guimaraes,A.R., Toiber,D., *et al.* (2012) The histone deacetylase SIRT6 is a tumor suppressor that controls cancer metabolism. *Cell*, **151**, 1185–1199.

55. Zhang,J., Yin,X.-J., Xu,C.-J., Ning,Y.-X., Chen,M., Zhang,H., Chen,S.-F. and Yao,L.-Q. (2015) The histone deacetylase SIRT6 inhibits ovarian cancer cell proliferation via down-regulation of Notch 3 expression. *Eur Rev Med Pharmacol Sci*, **19**, 818–824.

56. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.*, **13**, 2498–2504.

57. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M., *et al.* (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*, **37**, D412-416.

58. Doncheva,N.T., Morris,J.H., Gorodkin,J. and Jensen,L.J. (2019) Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J Proteome Res*, **18**, 623–632.

59. Oughtred,R., Rust,J., Chang,C., Breitkreutz,B.-J., Stark,C., Willems,A., Boucher,L., Leung,G., Kolas,N., Zhang,F., *et al.* (2021) The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci*, **30**, 187–200.

60. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N., *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, **42**, D358–D363.

61. Bindea,G., Mlecnik,B., Hackl,H., Charoentong,P., Tosolini,M., Kirilovsky,A., Fridman,W.-H., Pagès,F., Trajanoski,Z. and Galon,J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**, 1091–1093.

62. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J. and Zhang,Y. (2015) The I-TASSER Suite: protein structure and function prediction. *Nat Methods*, **12**, 7–8.

63. Kozakov,D., Hall,D.R., Xia,B., Porter,K.A., Padhorny,D., Yueh,C., Beglov,D. and Vajda,S. (2017) The ClusPro web server for protein–protein docking. *Nat Protoc*, **12**, 255–278.

64. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

65. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat Genet*, **36**, 431–432.

66. Käll,L., Canterbury,J.D., Weston,J., Noble,W.S. and MacCoss,M.J. (2007) Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods*, **4**, 923–925.

67. The,M., MacCoss,M.J., Noble,W.S. and Käll,L. (2016) Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.*, **27**, 1719–1727.

68. Paulo,J.A., Gaun,A., Kadiyala,V., Ghoulidi,A., Banks,P.A., Conwell,D.L. and Steen,H. (2013) Subcellular Fractionation Enhances Proteome Coverage of Pancreatic Duct Cells. *Biochim Biophys Acta*, **1834**, 791–797.

69. Na,Z., Dai,X., Zheng,S.-J., Bryant,C.J., Loh,K.H., Su,H., Luo,Y., Buhagiar,A.F., Cao,X., Baserga,S.J., *et al.* (2022) Mapping subcellular localizations of unannotated microproteins and alternative proteins with MicroID. *Molecular Cell*, **82**, 2900-2911.e7.

70. Eyckerman,S., Titeca,K., Van Quickelberghe,E., Cloots,E., Verhee,A., Samyn,N., De Ceuninck,L., Timmerman,E., De Sutter,D., Lievens,S., *et al.* (2016) Trapping mammalian protein complexes in viral particles. *Nat Commun*, **7**, 11416.

71. Roux,K.J., Kim,D.I., Burke,B. and May,D.G. (2018) BioID: A Screen for Protein-Protein Interactions. *Curr Protoc Protein Sci*, **91**, 19.23.1-19.23.15.

72. Alam,M.S. (2018) Proximity Ligation Assay (PLA). *Curr Protoc Immunol*, **123**, e58.

73. Therachiyil,L., Anand,A., Azmi,A., Bhat,A., Korashy,H.M. and Uddin,S. (2022) Role of RAS signaling in ovarian cancer. *F1000Res*, **11**, 1253.

74. Zheng,Z.-Y., Elsarraj,H., Lei,J.T., Hong,Y., Anurag,M., Feng,L., Kennedy,H., Shen,Y., Lo,F., Zhao,Z., *et al.* (2022) Elevated NRAS expression during DCIS is a potential driver for progression to basal-like properties and local invasiveness. *Breast Cancer Research*, **24**, 68.

75. Birkeland,E., Wik,E., Mjøs,S., Hoivik,E.A., Trovik,J., Werner,H.M.J., Kusonmano,K., Petersen,K., Raeder,M.B., Holst,F., *et al.* (2012) KRAS gene amplification and overexpression but not mutation associates with aggressive and metastatic endometrial cancer. *Br J Cancer*, **107**, 1997–2004.

76. Zhou,J.-D., Yao,D.-M., Li,X.-X., Zhang,T.-J., Zhang,W., Ma,J.-C., Guo,H., Deng,Z.-Q., Lin,J. and Qian,J. (2017) KRAS overexpression independent of RAS mutations confers an adverse prognosis in cytogenetically normal acute myeloid leukemia. *Oncotarget*, **8**, 66087–66097.

77. Jung,J., Cho,K.-J., Naji,A.K., Clemons,K.N., Wong,C.O., Villanueva,M., Gregory,S., Karagas,N.E., Tan,L., Liang,H., *et al.* (2019) HRAS-driven cancer cells are vulnerable to TRPML1 inhibition. *EMBO reports*, **20**, e46685.

78. Miglietta,G., Gouda,A.S., Cogoi,S., Pedersen,E.B. and Xodo,L.E. (2015) Nucleic Acid Targeted Therapy: G4 Oligonucleotides Downregulate HRAS in Bladder Cancer Cells through a Decoy Mechanism. *ACS Med. Chem. Lett.*, **6**, 1179–1183.

79. November 2020,19 (2020) The Human Protein Atlas: A 20-year journey into the body. *Science | AAAS*.

80. Ouyang,S., Zhang,Q., Lou,L., Zhu,K., Li,Z., Liu,P. and Zhang,X. (2022) The Double-Edged Sword of SIRT3 in Cancer and Its Therapeutic Applications. *Frontiers in Pharmacology*, **13**.

81. Chen,G., Gharib,T.G., Huang,C.-C., Taylor,J.M.G., Misek,D.E., Kardia,S.L.R., Giordano,T.J., Iannettoni,M.D., Orringer,M.B., Hanash,S.M., *et al.* (2002) Discordant Protein and mRNA Expression in Lung Adenocarcinomas *. *Molecular & Cellular Proteomics*, **1**, 304–313.

82. Bauernfeind,A.L. and Babbitt,C.C. (2017) The predictive nature of transcript expression levels on protein expression in adult human brain. *BMC Genomics*, **18**, 322.

83. Perl,K., Ushakov,K., Pozniak,Y., Yizhar-Barnea,O., Bhonker,Y., Shivatzki,S., Geiger,T., Avraham,K.B. and Shamir,R. (2017) Reduced changes in protein compared to mRNA levels across non-proliferating tissues. *BMC Genomics*, **18**, 305.

84. Fukao,Y. (2015) Discordance between protein and transcript levels detected by selected reaction monitoring. *Plant Signal Behav*, **10**, e1017697.

85. Brion,C., Lutz,S.M. and Albert,F.W. (2020) Simultaneous quantification of mRNA and protein in single cells reveals post-transcriptional effects of genetic variation. *eLife*, **9**, e60645.

86. De Marco,C., Rinaldo,N., Bruni,P., Malzoni,C., Zullo,F., Fabiani,F., Losito,S., Scrima,M., Marino,F.Z., Franco,R., *et al.* (2013) Multiple genetic alterations within the PI3K pathway are responsible for AKT activation in patients with ovarian carcinoma. *PLoS One*, **8**, e55362.

87. Wang,G., Yang,X., Li,C., Cao,X., Luo,X. and Hu,J. (2014) PIK3R3 Induces Epithelial-to-Mesenchymal Transition and Promotes Metastasis in Colorectal Cancer. *Molecular Cancer Therapeutics*, **13**, 1837–1847.

88. Stronach,E.A., Chen,M., Maginn,E.N., Agarwal,R., Mills,G.B., Wasan,H. and Gabra,H. (2011) DNA-PK mediates AKT activation and apoptosis inhibition in clinically acquired platinum resistance. *Neoplasia*, **13**, 1069–1080.

89. Liu,Q., Turner,K.M., Alfred Yung,W.K., Chen,K. and Zhang,W. (2014) Role of AKT signaling in DNA repair and clinical response to cancer therapy. *Neuro Oncol*, **16**, 1313–1323.

90. Arlt,C., Ihling,C.H. and Sinz,A. (2015) Structure of full-length p53 tumor suppressor probed by chemical cross-linking and mass spectrometry. *PROTEOMICS*, **15**, 2746–2755.

91. Wells,M., Tidow,H., Rutherford,T.J., Markwick,P., Jensen,M.R., Mylonas,E., Svergun,D.I., Blackledge,M. and Fersht,A.R. (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proceedings of the National Academy of Sciences*, **105**, 5762–5767.

92. Hoyos,D., Greenbaum,B. and Levine,A.J. (2022) The genotypes and phenotypes of missense mutations in the proline domain of the p53 protein. *Cell Death Differ*, **29**, 938–945.

93. Schildkraut,J.M., Goode,E.L., Clyde,M.A., Iversen,E.S., Moorman,P.G., Berchuck,A., Marks,J.R., Lissowska,J., Brinton,L., Peplonska,B., *et al.* (2009) Single Nucleotide Polymorphisms in the TP53 Region and Susceptibility to Invasive Epithelial Ovarian Cancer. *Cancer Research*, **69**, 2349–2357.

94. Yaginuma,Y. and Westphal,H. (1992) Abnormal structure and expression of the p53 gene in human ovarian carcinoma cell lines. *Cancer Res*, **52**, 4196–4199.

95. Willis,S., Villalobos,V.M., Gevaert,O., Abramovitz,M., Williams,C., Sikic,B.I. and Leyland-Jones,B. (2016) Single Gene Prognostic Biomarkers in Ovarian Cancer: A Meta-Analysis. *PLoS One*, **11**, e0149183.

96. Weberpals,J.I., Pugh,T.J., Marco-Casanova,P., Goss,G.D., Andrews Wright,N., Rath,P., Torchia,J., Fortuna,A., Jones,G.N., Roudier,M.P., *et al.* (2021) Tumor genomic, transcriptomic, and immune profiling characterizes differential response to first-line platinum chemotherapy in high grade serous ovarian cancer. *Cancer Med*, **10**, 3045–3058.

97. Murga,M., Lecona,E., Kamileri,I., Díaz,M., Lugli,N., Sotiriou,S.K., Anton,M.E., Méndez,J., Halazonetis,T.D. and Fernandez-Capetillo,O. (2016) POLD3 Is Haploinsufficient for DNA Replication in Mice. *Molecular Cell*, **63**, 877–883.

98. Perez-Riverol,Y., Bai,J., Bandla,C., García-Seisdedos,D., Hewapathirana,S., Kamatchinathan,S., Kundu,D.J., Prakash,A., Frericks-Zipper,A., Eisenacher,M., *et al.* (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research*, **50**, D543–D552.

TABLE AND FIGURES LEGENDS

Table 1. Wild-type and variant RefProts significantly varied (ANOVA, q-value <0.05). The number of WT and variant RefProts is displayed for the six main clusters identified upon LFQ proteomics.

Figure 1. LFQ analysis workflow. (A) Illustration of the Proteome Discoverer analysis steps used. Each child processing step corresponds to the interrogation using the cell-specific database. (B) Workflow nodes present in each processing child step.

Figure 2. DESeq2 transcripts analysis. (A) Venn diagram displaying the number of exclusive and shared transcripts between the three cell lines. (B) Hierarchical clustering heatmap showing the different transcript clusters that can be observed among the three cell lines. Z-score range from -1.3509 (green) to 1.3523 (red).

Figure 3. DESeq2 gene analysis. (A) Venn diagram displaying the number of exclusive and shared genes between the three cell lines. (B) Pie chart displaying the ratios of the different types of RNAs sequenced. (C) Hierarchical clustering heatmap showing the different gene clusters that can be observed among the three cell lines. Z-score range from -1.351 (green) to 1.3496 (red).

Figure 4. WT and variant proteins predicted by OpenCustomDB. For each cell line and database, the fractions of AltProts, RefProts, novel isoforms and their variants are displayed.

Figure 5. Types of AltProts predicted by OpenCustomDB. The percentages of ncRNA, CDS frameshifts, 3' and 5'UTR derived AltProts are displayed for each database and cell line.

Figure 6. Analysis of the identified proteins. (A) Venn diagrams displaying the number of exclusive and shared proteins identified between the three cell lines. (B) Bar plot displaying the fractions of WT and variant RefProts, novel isoforms and AltProts identified in each cell line.

Figure 7. Subcellular compartment distribution and characteristics of identified AltProts. (A) Venn diagram displaying the distribution of AltProts identified in the different subcellular fractions. (B) RNA origin and (C) molecular weight distribution of the identified AltProts.

Figure 8. AltProts with significantly changed levels exclusively in one of two cancerous cell lines or common in both (ANOVA, FDR <0.05). For each cell line, the subcellular compartment, the AltProts upregulated (red) and downregulated (green) are shown.

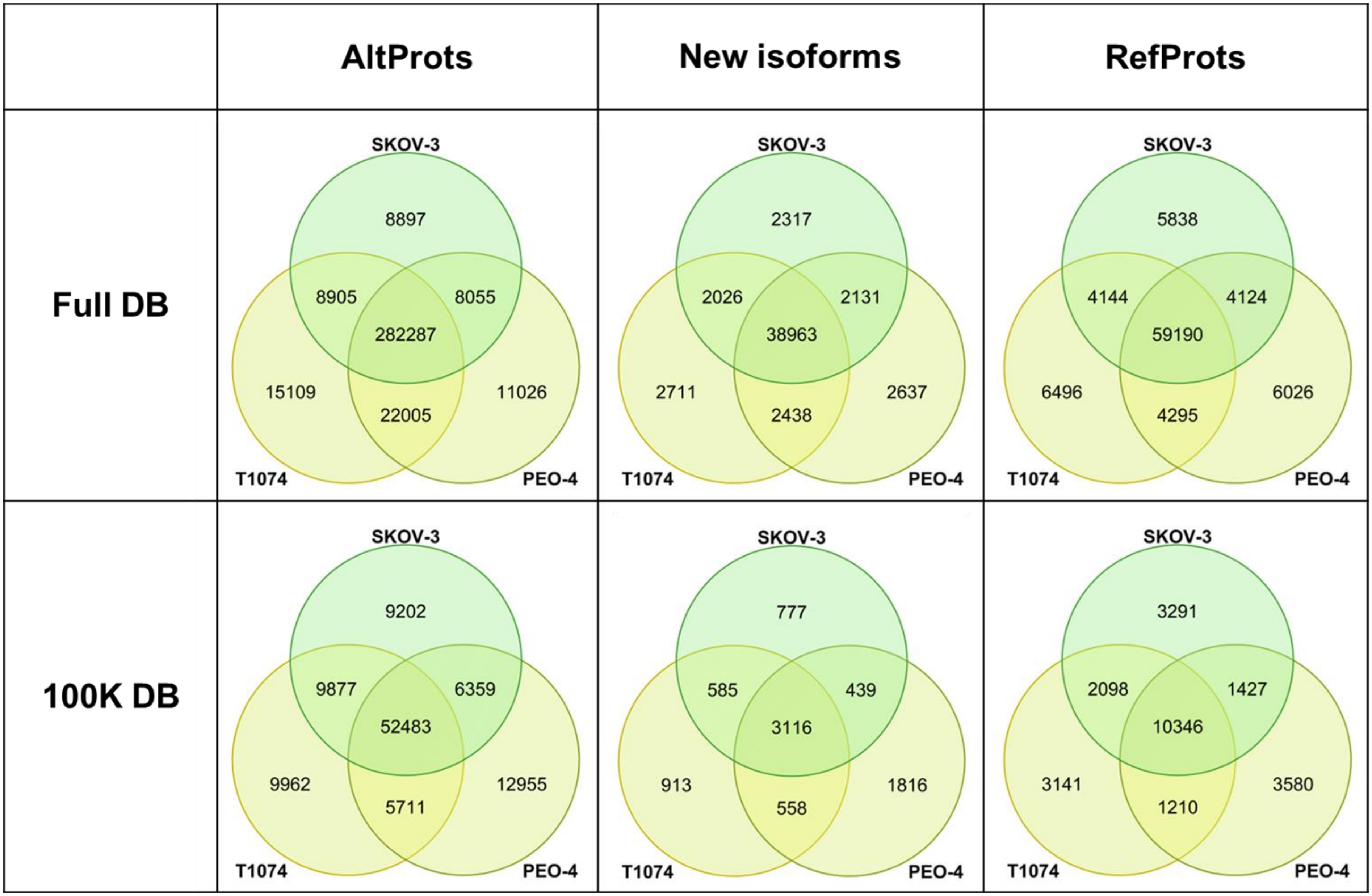
Figure 9. RefProts and genes significantly varied (ANOVA, FDR <0.05) in the central carbon metabolism in the cancer pathway. (A) Central carbon metabolism in cancer, up and downregulation in both cancerous cells. (B) Central carbon metabolism in cancer, up and downregulation only in SKOV-3. (C) Central carbon metabolism in cancer, up and downregulation only in PEO-4.

Figure 10: GO molecular function enrichment network generated with ClueGO in Cytoscape. GO enrichment was generated from the accession numbers of Supplemental figure 4. AltProts are marked in orange and RefProts in blue. Enriched GO terms are displayed as hexagons. KEGG pathways are displayed as octagons and crosslinks are marked in blue (SKOV-3 cells), purple (PEO-4 cells) and green (T1074 cells) dashed lines.

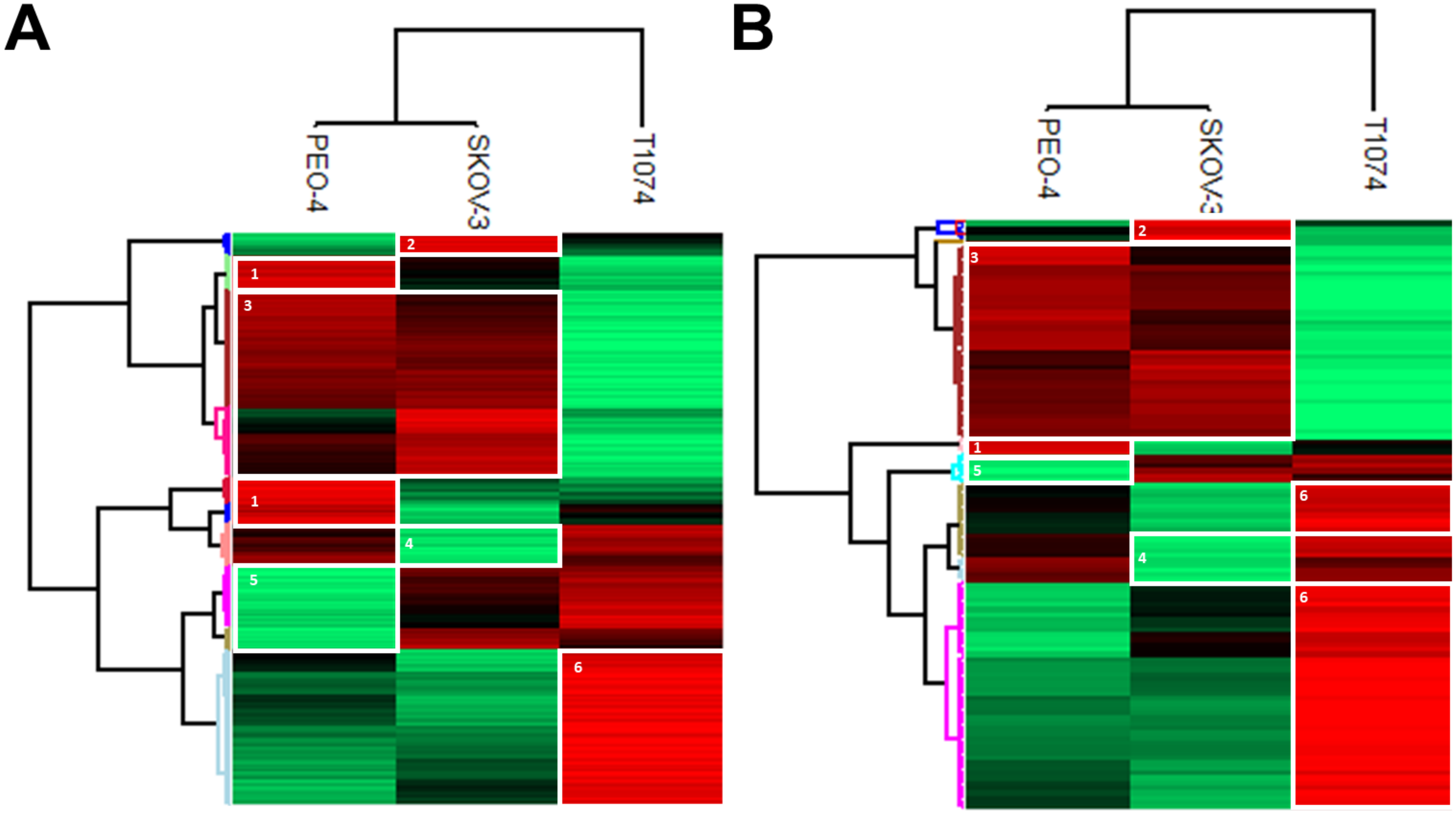
Figure 11. Synthesis of AltProts from the DHX8 gene (A) List of transcripts referenced in Ensembl. (B) Zoom on DHX8-204 described to translate to “K7EJH9”, a predicted protein from TrEMBL without the 5’UTR part or methionine as the first amino acid, when IP_715944 is described from the overlap between the CDS and the 3’UTR. As a result, the mutation is only observable by the proteogenomic construction, as it would be considered a silent mutation due to its position in the UTR part of K7EJH9.

Figure 12. IP_183088 (AltMAPK8) predicted models docked to the human polymerase delta holoenzyme complex. (A) The interaction of IP_183088 and the full POLD complex is shown. The distance between the two lysines involved in the crosslink is 24.59 Å. (B) Zoom of the interaction of IP_183088 and POLD3. The surface representation shows the possible placement of IP_183088 at the interaction site of POLD3 and POLD2.

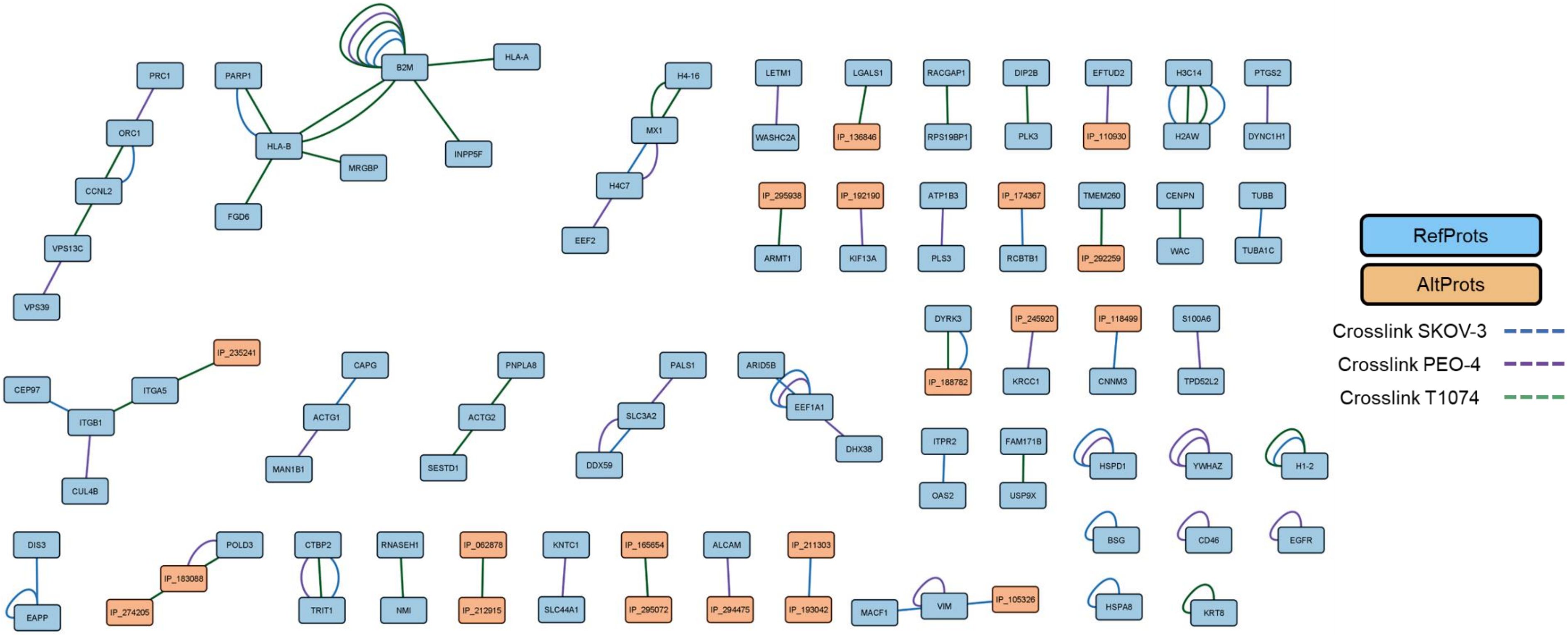
Supplemental figure 1. Venn diagrams describing the RNA-seq derived databases. The number of specific and common.



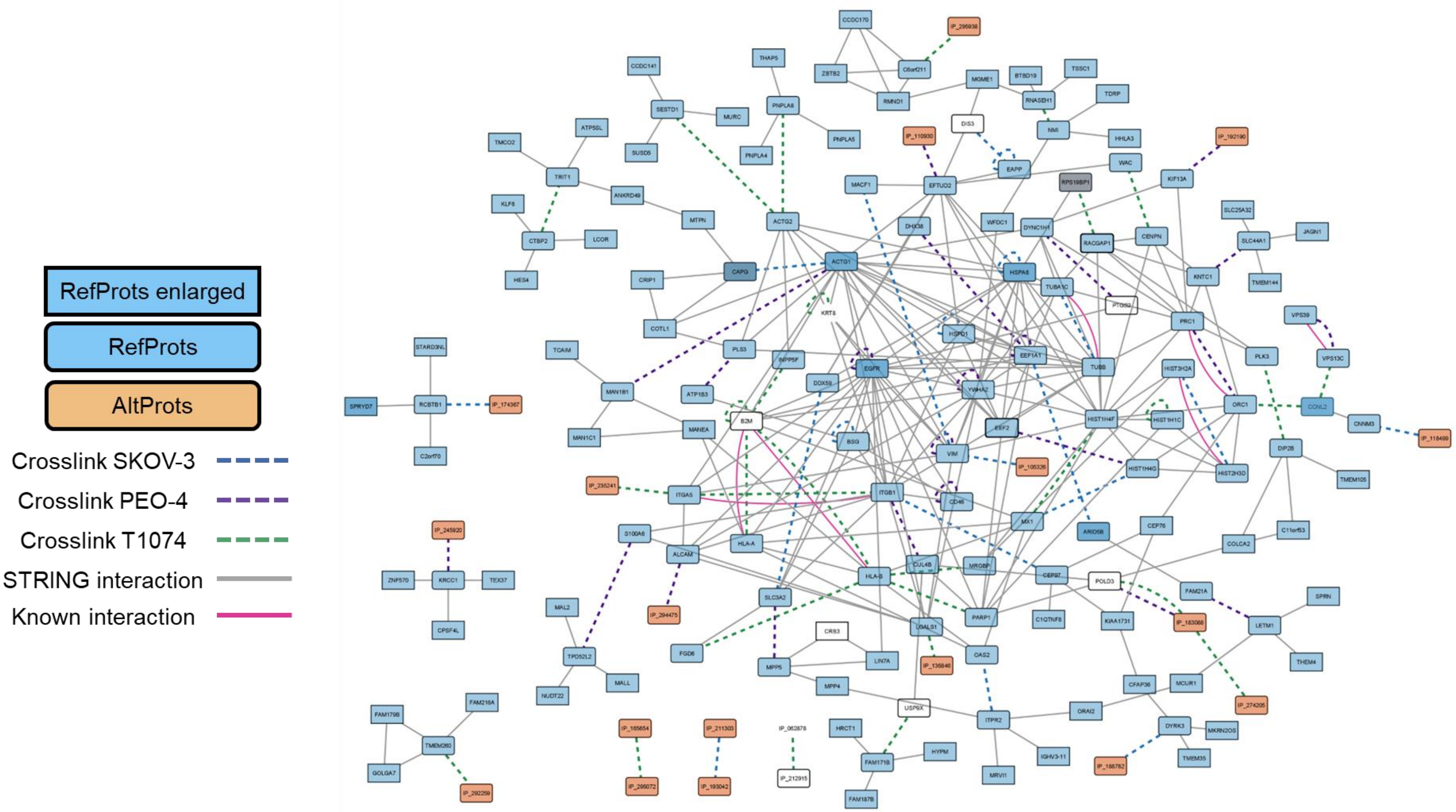
Supplemental figure 2. (A) Hierarchical clustering heatmap showing the different RefProt clusters that can be observed among the three cell lines. Z-score range from -1.349 (green) to 1.307 (red).(B) Hierarchical clustering heatmap showing the different novel isoforms clusters that can be observed among the three cell lines. Z-score range from -1.348 (green) to 1.273 (red).



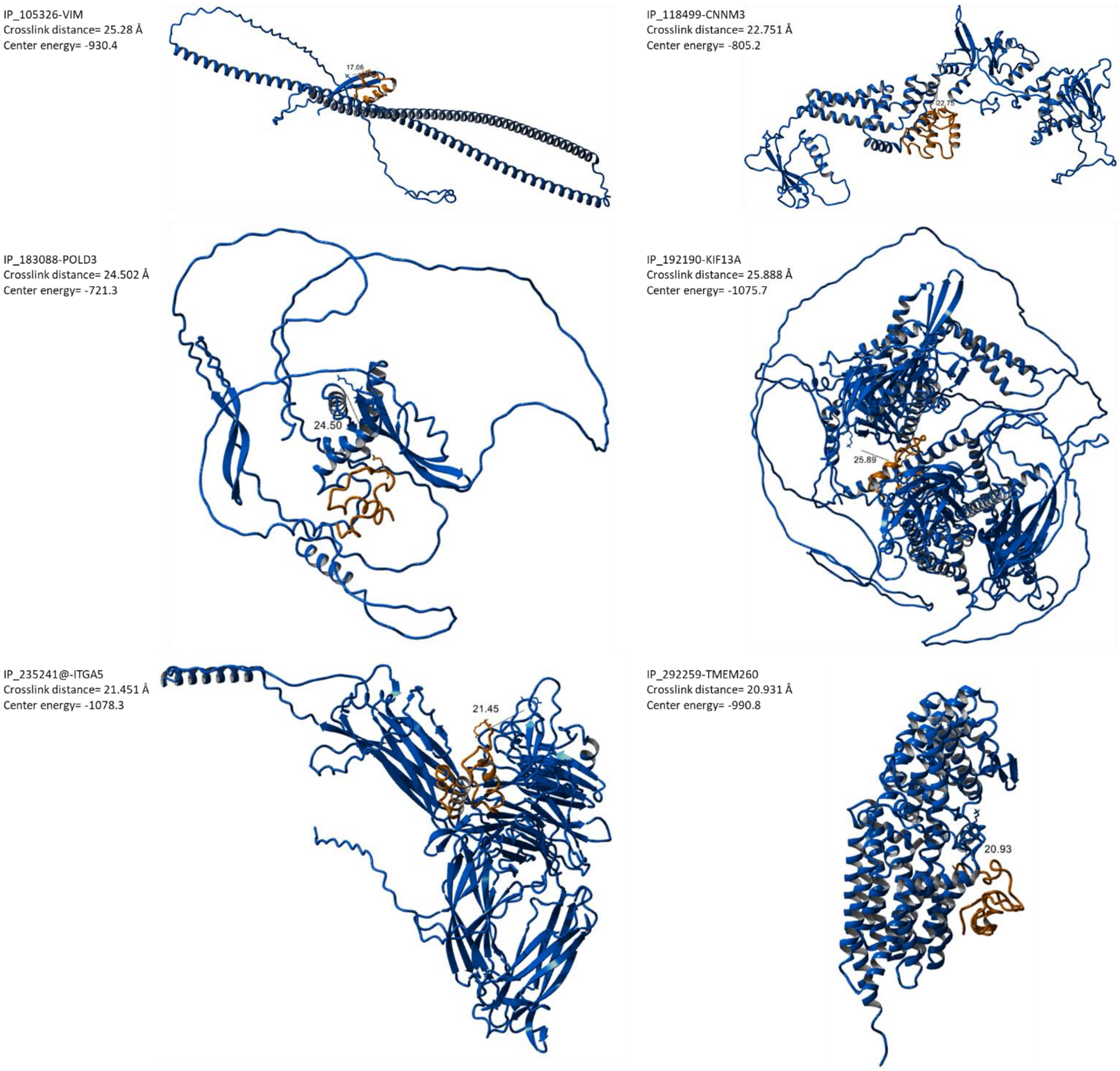
Supplemental figure 3. Raw crosslink network in which AltProts are marked in orange and RefProts are marked in blue. Crosslinks are marked in dark blue (SKOV-3 cells), purple (PEO-4 cells) and green (T1074 cells) dashed lines.



Supplemental figure 4. Crosslink network in which AltProts are marked in orange and RefProts are in blue. Crosslinks are marked in dark blue (SKOV-3 cells), purple (PEO-4 cells) and green (T1074 cells) dashed lines. Enriched by the STRING interactions (gray lines) retrieved. For the RefProts that did not present a referenced STRING interaction, an enrichment was performed to expand the network.



Supplemental figure 5. Predicted interaction models docked in ClusPro for the RefProts (blue) and AltProts (orange). The distance between the residues crosslinked are given for each interaction.



IP_105326-VIM
Crosslink distance= 25.28 Å
Center energy= -930.4

IP_118499-CNNM3
Crosslink distance= 22.751 Å
Center energy= -805.2

IP_183088-POLD3
Crosslink distance= 24.502 Å
Center energy= -721.3

IP_192190-KIF13A
Crosslink distance= 25.888 Å
Center energy= -1075.7

IP_235241@-ITGA5
Crosslink distance= 21.451 Å
Center energy= -1078.3

IP_292259-TMEM260
Crosslink distance= 20.931 Å
Center energy= -990.8

Conclusions

In recent years, personalized medicine has gained significant attention and recognition as a valuable approach for diagnosing and treating various diseases. This innovative approach offers numerous advantages, with one key advantage being the acquisition of detailed information about a patient's biomolecular profile. By understanding this profile, healthcare professionals can make more informed and precise decisions when selecting the most appropriate treatment for each patient. This tailored approach thus ensures that patients receive the most effective and personalized treatment, resulting in improved health outcomes and overall patient satisfaction.

Here, proteogenomics aids to link the genomic/transcriptomic profile with the proteome. As technology and algorithms improve, novel tools have been developed to make this approach available for researchers and ultimately for clinicians. For this, the team of Prof. Xavier Roucou developed a tool called OpenCustomDB, which allowed us to generate RNA-seq-derived databases to interrogate the proteome of three cell lines. A key feature of this tool is the capability of generating databases from variant call files. Therefore, in these databases we are able to import genomic mutations and then assess if they are translated in the proteome. This is important for deeply characterizing the molecular composition of a disease or different subgroups of a disease. Additionally, with the increasing interest in alternative proteomes and their demonstrated potential involvement in diseases, this tool allows us to identify AltProt variants. Since there are no databases for AltProt variants, proteogenomics is the only way to identify them on a large scale and assess their levels. This can potentially make them a target for focused experiments.

Given the significance of OvCa in women's health, we investigated the potential impact of AltProts on it. To do this, we selected different cell lines, including PEO-4 cells with high-grade serous histology, SKOV-3 cells with clear cell carcinoma, and T1074 cells derived from normal human ovarian surface epithelial cells, which served as a non-tumorous control. In this context, we were able to identify AltProts that were common to all three cell lines. More importantly, we also identified AltProts that were exclusively found in each cancerous cell line or in both cancerous cell lines.

The example of the AltProt variant (IP_715944@Leu44Pro) identified in both cancerous cell lines demonstrates the power of a proteogenomic approach. Indeed, using RNA-seq, the transcript corresponding to this mutation was not annotated and the predicted sequence was not generated. Therefore, we can infer that the variant is not present in the genomic information of the non-tumor cell line. Additionally, six other wild-type AltProts were only predicted and identified in the OvCa cell lines. These findings highlight that the proteogenomic approach establishes a direct connection between a genome blueprint and the resulting proteome. Moreover, it provides different possible targets which can have a role in pathology, setting a ground base for future research. Moreover, based on these predicted databases, data reuse can be done considering the subtypes analyzed. Doing so, we can try to identify if these AltProts are identified in high-grade serous carcinoma and clear cell OvCa patient data.

One advantage of combining NGS data and proteomics data is the ability to assess the differential expression of both types of data simultaneously. Based on this, we decided to map the identified RefProts and novel isoforms variations in connected datasets to pathways using functional annotation enrichment strategies, which are useful for identifying differences in large gene/protein datasets. We found a significant enrichment in the central carbon metabolism pathway associated with cancer according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. Using this pathway, we were able to map expression differences between the three cell lines used in this study, revealing such differences in key genes related to cancer. Further research is needed to determine whether these differences are related to the different subtype, morphology, or resistance between the two cancerous cells.

As described in Chapters III and IV, a methodology based on XL-MS, subcellular fractionation, 3D structure simulation and docking was developed. The application of this methodology to the three cell lines was done with the aim of identifying AltProt-RefProt interactions. In order to confirm possible interactions between proteins, molecular docking was performed. From the interactions identified, POLD3-IP183088 caught our attention in the OvCa context. POLD3 is part of the POLδ complex, which has exonuclease and 3’ to 5’ polymerase activity. For instance, it is involved in high fidelity replication (lagging strand synthesis)[387] and nucleotide excision repair (NER) synthesis following UV irradiation[388]. Particularly, POLD3 is an accessory component of POLδ, which plays a major role stabilizing the complex[389]. POLD3 was shown to possess better efficient proofreading activity than other components of the complex[390]. Therefore, due to the possible position of this interaction in PEO-4 and T1074 cells, a disruption of the POLδ complex may occur. A hypothesis is that this disruption increases mutagenesis in these tumor cells, leading to a higher likelihood of unrepaired errors in DNA replication, as expected in PEO-4 cells. Additionally, there could be a regulatory system involving the POLD complex, where this interaction may trigger the apoptosis of the cell.

In summary, the utilization of a proteogenomic approach provides several advantages. One of the key ones being the ability to predict and identify proteins that are specific to individual cells. This holds true for both cell lines and potentially as well as for samples taken from patients. The significance of this advantage cannot be overstated, as it greatly contributes to the progress of personalized medicine.

PART VI DISCUSSION AND PERSPECTIVES

During my three-year thesis, I successfully applied a novel workflow to study the cellular localization and functions of AltProts through large-scale investigations. This research challenged a fundamental principle in biology, particularly the long-believed idea that a mRNA-molecule could only encode for a single functional protein. My main emphasis was on developing a crosslinking mass spectrometry strategy in conjunction with subcellular fractionation, and this approach was combined with the development of cell-specific protein databases. Additionally, I integrated this workflow for the analysis of the protein crosslink data in network analysis software to understand the interactions, making significant novel contributions.

The development of RNA-seq derived cell line-specific databases allowed the addition of another level of information to the study of AltProts as it provided us a direct link between the genomic information stored in the cell lines and the produced cellular proteome. Additionally, the NGS studies allowed us to identify significant variations in the expression of the transcriptome between the studied cell lines. Coupled to the analysis of the proteome, I was able to generate a large set of data, which allowed us to point out some physio-pathological differences between the cell lines studied.

One of the main challenges for AltProt research is the lack of antibodies specifically targeting these proteins. As a consequence, the majority of functional studies have been based on targeted approaches in which an AltProt is expressed via an expression vector to then perform interactomic experiments. The other main technique used for their study is silencing (e.g. by short harping RNA, shRNA) the transcripts that encode AltProts. Usually, these silencing methods are costly and (very) time consuming, as well as raising antibodies that recognize AltProts. On the other hand, XL-MS experiments aim to identify AltProt-RefProt PPIs. These identified PPIs can then be selected as AltProts of interest for targeted experiments. The identification of AltProt-interacting proteins provides an opportunity to reconstruct the network of interactions between these proteins and RefProts, and thereby the integration of such new proteins into known signaling pathways, potentially involved in crucial cellular mechanisms. By this approach we were able to put forward some hypotheses on the possible function or pathway involvement of some AltProts.

Identification and characterization of AltProts

In shotgun MS-based proteomics, the identification process involves matching predicted peptide fragmentation spectra to fragmentation spectra obtained upon LC-MS/MS analysis of enzymatic digestion of proteomes. This allows to determine the presence of a protein in the analyzed sample by identifying peptides from this protein. Generally, one relies on at least two unique peptides of a protein for its unambiguous identification. However, for AltProts, which are small in size and low in abundance, one typically needs to lower the threshold to just one unique peptide. To ensure the accuracy of protein assignments, manual inspection of such unique peptide spectra, pairwise alignment[264], and querying the peptide in NextProt[391] are necessary steps. These additional measures increase the validity of the identification of an AltProt.

In order to gain a comprehensive understanding of the identified AltProts, one can extract various types of information from the OpenProt database. This includes the molecular weight, number of amino acids, isoelectric point, gene annotation, transcript annotation(s), transcript type, genomic and proteomic sequences, as well as predicted domains. Such information is crucial for planning future experiments and can guide in designing PCR primers and cloning vectors by retrieving the genomic sequence. Additionally, knowing the different transcripts from which an AltProt can be coded aids in selecting the appropriate transcript for silencing experiments. The protein sequence can be also used to predict the protein 3D structure of the AltProt in I-Tasser[340] or AlphaFold[341].

For instance, based on our proteomic experiments aimed at characterizing the PEO-4, SKOV-3 and T1074 cell lines, we identified two interesting AltProts (**Figure 13A**). The first one, IP_642002 (LncRNA, *IGFBP2*), was found to be upregulated in SKOV-3 cells, while IP_3424589 is derived from the 3’UTR part of a transcript from the *LCOR* gene and found upregulated in PEO-4 cells. Both AltProts were studied by RT-PCR, cloning and confocal microscopy. We preliminary tested the transfection efficiency using the vector pcDNA3 EGFP-IP_3424589 (**Figure 13B**) and PolyJet as transfection reagent, in the three cell lines. **Figure 13C** shows the favorable expression of an EGFP tagged-AltProt by confocal microscopy in SKOV-3 and PEO-4 cells. This, allow me to identify the correct conditions for transfection of the cell lines and facilitate the pre-IF treatment of the cells. Due to the easy detachment of T1074 cells from the flasks and slides, confocal microscopy was not performed in this cell line.

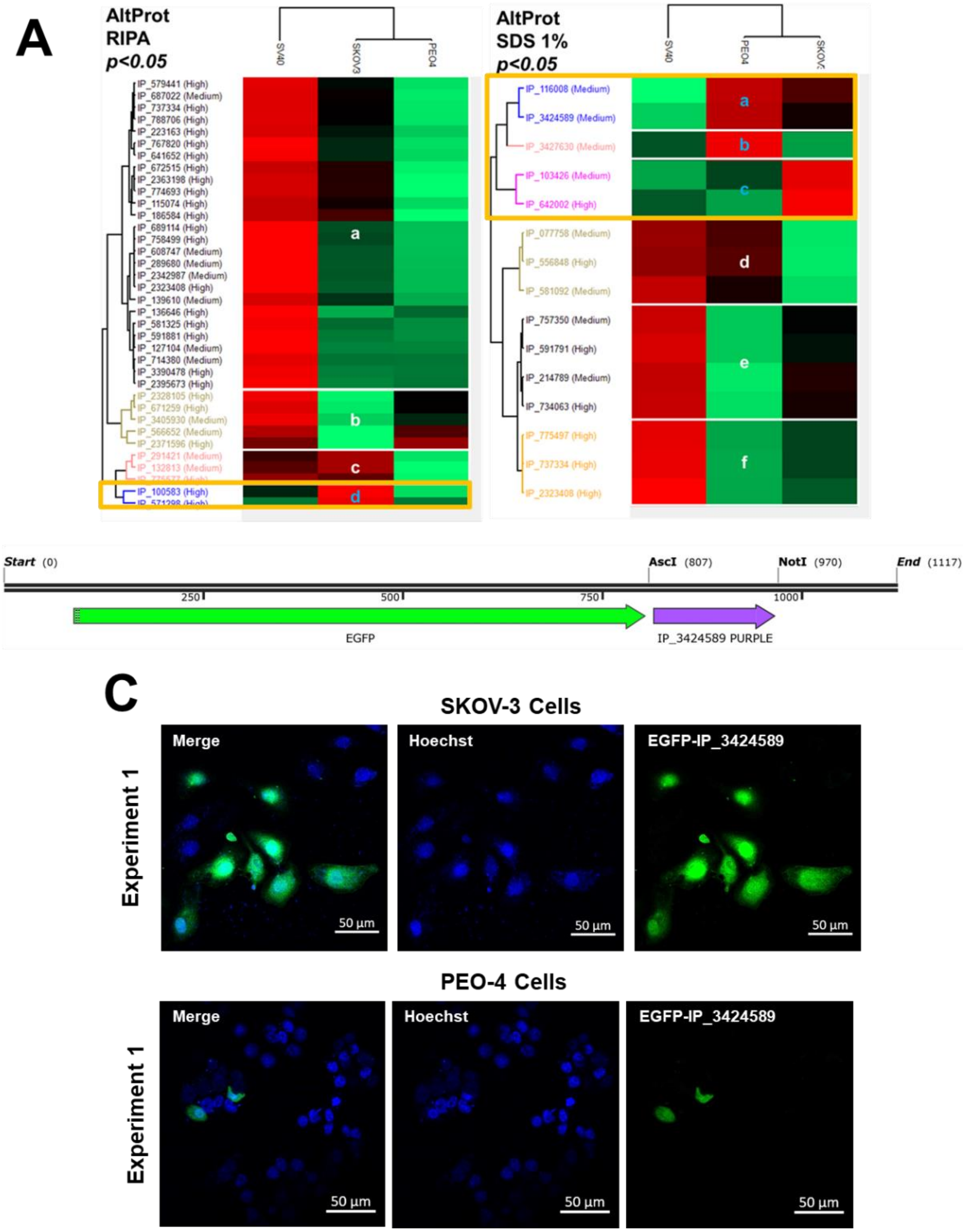


Figure 13. Total proteome cell line analysis and confocal microscopy images of transfected cells. *(A) Hierarchical clustering of significant regulated (ANOVA, p-value <0.05) AltProts in the 1% SDS and RIPA extractions of each cell line. (B) Sanger sequencing construct of the pcDNA3 EGFP plasmid in which IP_3424589 is incorporated. (C) Confocal microscopy images of transfected EGFP-IP_3424589 SKOV-3 (upper panel)*

and PEO-4 (lower panel) cells are displayed. Cellular nuclei are stained win Hoechst reagent.

Additionally, I started evaluating the subcellular localization of these AltProts. First, the transfection of a vector containing EGFP and IP_642002 was done (**Figure 14A**). **Figure 14B** and **C** display the confocal microscopy images from preliminary fluorescence and IF experiments. Hoechst reagent was used as a nuclei counterstain. Anti-RAB5C was used as an exosomes stain and anti-KRT as cytoplasmic stain.

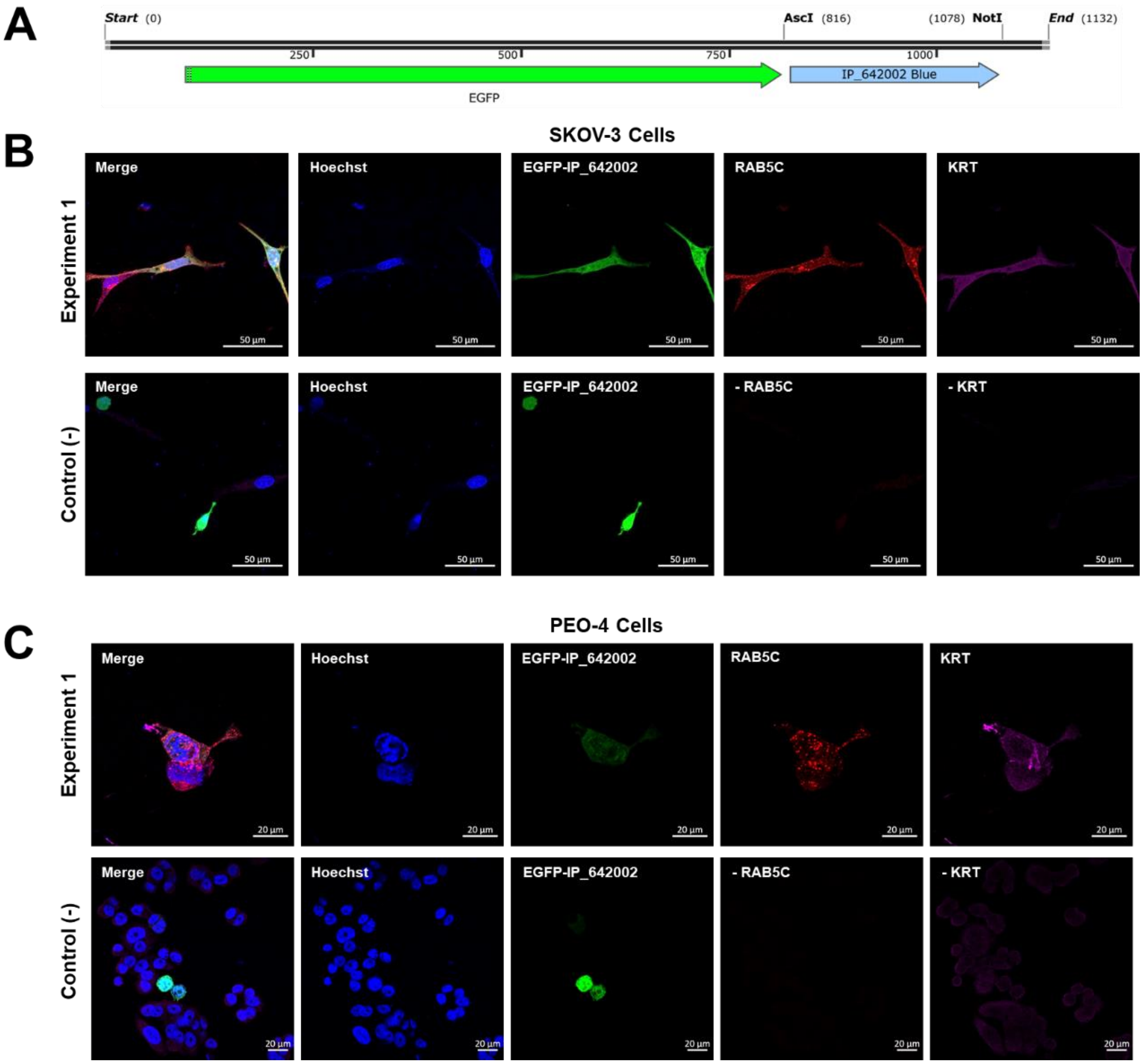


Figure 14. Confocal microscopy of IP_642002. *(A) Sanger sequencing construct of the pcDNA3 EGFP plasmid in which IP_642002 is incorporated. (B) Confocal microscopy images of SKOV-3 cells and PEO-4 cells (C) transfected with EGFP- IP_642002. For both, cellular nuclei are counterstained with Hoechst reagent. Exosomes are labelled with anti-RAB5 and the cytoplasm by anti-KRT.*

For both cell lines, the expression EGFP-IP_642002 (green) overlaps with the signal of anti-KRT, hinting to a possible cytoplasmic localization of this AltProt. However, these experiments need to be done with other organelle markers such as anti-cadherin for cellular membrane staining, anti-calreticulin for endoplasmic reticulum, anti-RCAS1 for Golgi apparatus and anti-COX4 for mitochondria.

shRNA experiments have been designed with the aid of Dr. Maheul Ploton. The primary objective of these experiments will be to inhibit the expression of the specific transcripts that give rise to AltProts of interest. By doing so, we can then proceed to re-express the various elements of the loci through the process of transient transfection. This allows for a comprehensive analysis of the phenotypical and molecular changes in cells, both before and after silencing the expression of AltProt.

By utilizing subcellular fractionation, we can deeper explore AltProts as it allows to perform a cellular fractionation before protein digestion. In our case, it permitted us to assign AltProts to specific subcellular compartments, providing further insights into AltProts. However, it is important to note that subcellular fractionation using different detergents has limitations, and a main one is the potential cross-contamination or carryover from one subcellular fraction to another, which may impact the accuracy of the results. Another approach to identify the subcellular compartment of AltProts is by differential ultracentrifugation, and yet another approach combines localization of organelle proteins by isotope tagging (LOPIT) and differential ultracentrifugation[392]. Here, cells are lysed and fractionated by differential centrifugation (3,000 to 120,000 x g). Then, each fraction is labelled with a tandem mass tag (TMT) 10-plex set post digestion. This allows the analysis of multiple subcellular components in the same run. Assigning a subcellular localization is done by comparing the protein level of well-defined subcellular organelle markers using machine learning algorithms. LOPIT would thus allow a higher resolved subcellular location of AltProts. Additionally, if the goal is to increase the number of AltProts identifications, protein enrichment methods can be used to deplete larger proteins which could interfere with AltProt signals[210,245,270,393]. However, the identification of larger AltProts and the information that can be obtained from the RefProts will then be lost.

Large-scale functional determination of AltProts

It is evident that OpenProt holds a wide range of information regarding AltProts. The prediction of functional protein domains present in AltProt sequences serves as an initial indicator of molecular functions of these proteins. However, it is important to note that such valuable information may not be available for all AltProts, mainly due to its size, thus limiting our ability to infer their possible involvement in physiological or pathological mechanisms.

In our large-scale studies, our primary objective was to link a potential function to an AltProt through the concept of "guilt by association". Indeed, by investigating the connections and relationships between AltProts and RefProts, we aimed to shed light on their potential roles and contributions in a broader biological context. During the last three years, some identified interactions caught our attention due to their possible importance in physiological processes.

In Part III, we described some of the interactions that occur. One of the interesting interactions to us is the interaction between IP_2292176 (*FAM227B*) and HLA-B. HLA-B is a component of the MHC I complex and plays a crucial role in the immune response by facilitating the presentation of antigenic peptides. These peptides, which are composed of 8-13 residues, are recognized by $CD8^+$ T cells, thereby initiating an antigen-specific immune response. This system is particularly important for tumor-derived antigens. Initially, our hypothesis was that a peptide from this AltProt could be presented by HLA-B or that the AltProt could inactivate the presentation site. However, after further analysis of the generated models, we found that the position of the AltProt aligns better with the position of B2M. Based on this new finding, we hypothesize that IP_2292176 could either disrupt the MHC I complex (B2M-HLA-B) or be located next to B2M. To test this hypothesis, a PLA (duolink) experiment can be conducted. This experiment should first validate the interaction between B2M and HLA-B. A preliminary experiment was done to assess the co-localization of HLA-B and B2M (**Figure 15A**). In the merged confocal microscopy image, orange coloration can be observed, meaning that HLA-B and B2M co-localize, and the antibodies are validated. In future experiments, the AltProt needs be fused with an epitope tag (e.g., EGFP, **Figure 15B**), which will serve as the target for the antibody-PLA probe. This approach will allow us to determine if IP_2292176 is in close proximity to HLA-B and if B2M is present in this interaction using IF. Once this interaction is confirmed by PLA, other interactomic techniques such as co-IP or BioID can be used to gain a broader understanding of the interactome of this AltProt. Additionally, a CRISPR-Cas9 construct has been prepared to perform a knockout of B2M. This will allow us to evaluate whether the expression of IP_2292176 rescues the MHC class I complex from depletion.

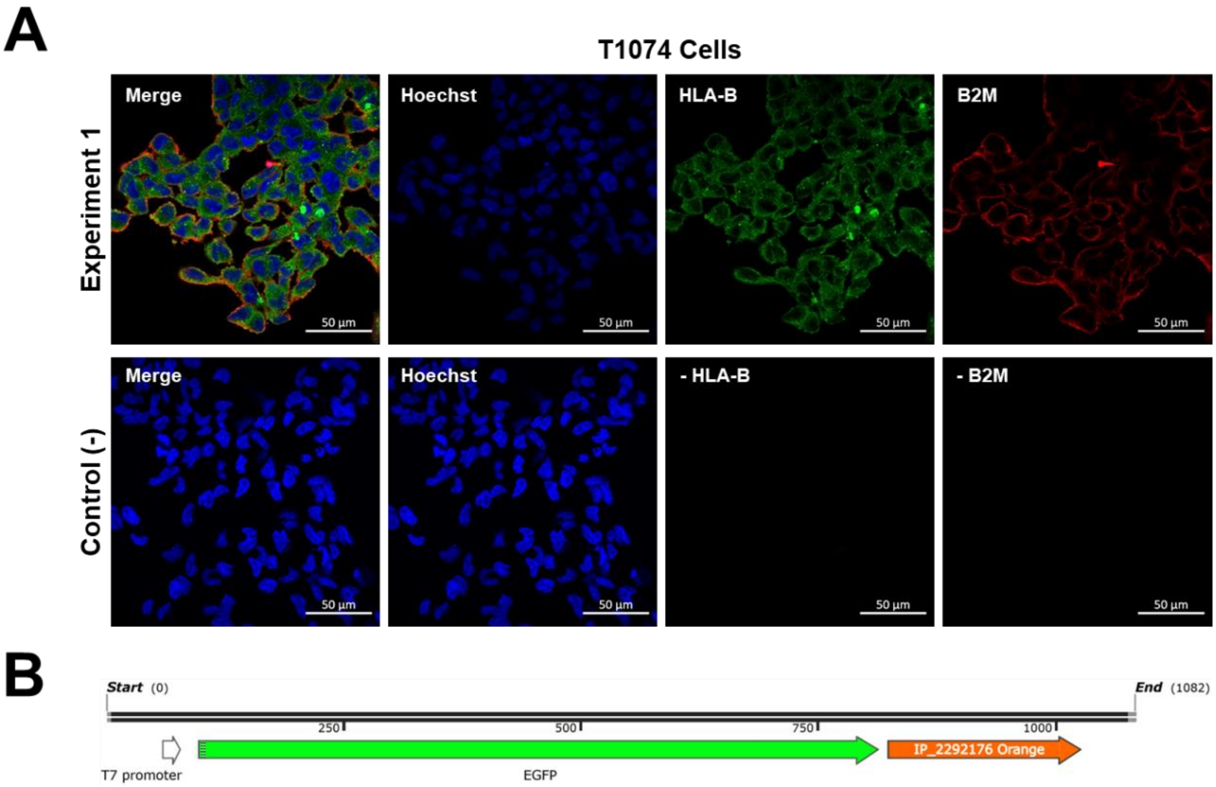


Figure 15. Confocal microscopy results validating the of the co-localization of B2M and HLA-B. *(A) IF confocal microscopy of T1074 cells. HLA-B is stained in red and B2M in green. Cellular nuclei are counterstained with Hoechst reagent. Orange coloration shows the co-localization of these proteins (B) Sanger sequencing map of the constructed pcDNA3 EGFP plasmid in which IP_2292176 is incorporated.*

The description of the next targeted phase demonstrates the complexity and variety of approaches that can be utilized to determine the function of an AltProt. In this study, we identified and examined at least 20 AltProts that are crosslinked to a RefProt and subsequently linked to a GO term or KEGG pathway. By employing the various experimental approaches described above, we can initiate multiple research projects for the coming years.

Secretome analysis

Recent studies have shown that AltProts can be secreted by cells[394,395]. Capuz *et al*. demonstrated that an AltProt named Heimdall is secreted by astrocytes under inflammatory conditions. Their findings suggest that this AltProt regulates the switch from neurons to astrocytes[394]. Additionally, Martinez *et al*. identified different expression levels of AltProts secreted between lean and obese mice. Moreover, they investigated the pro-appetite function of an AltProt derived from FAM237B in obese mice[395]. To further explore the secretion of AltProts by cells, a preliminary study of the secretome of PEO-4, SKOV-3 and T1074 cells was conducted. This study involved a FASP (10 kDa) enrichment, tryptic digestion and high-pH peptide fractionation (**Figure 16A**). The LC-MS/MS spectra were analyzed using Thermo Scientific Proteome Discoverer with Sequest HT (OpenProtDB) and Minora feature detector for LFQ analysis.

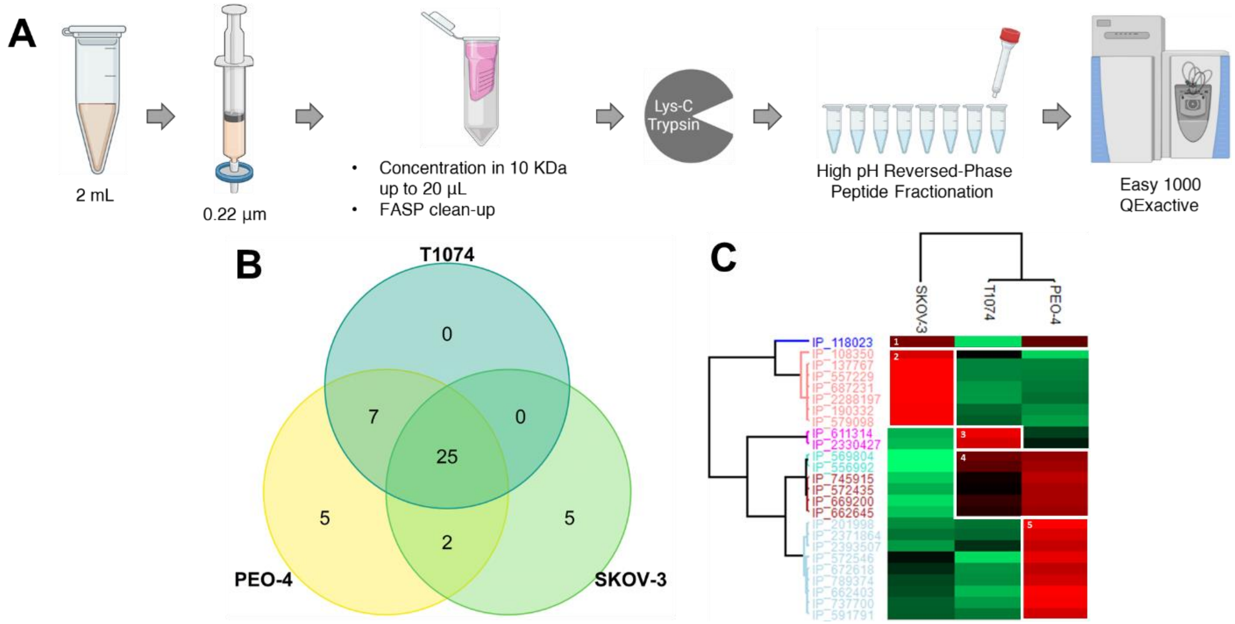


Figure 16. Secretome preliminary analysis. *(A) Workflow describing the steps applied to the secretome samples. (B) Venn diagram displaying the number of AltProts identified in each cell line. (C) Hierarchical clustering of significant regulated (ANOVA, FDR <0.05) AltProts in the secretome of each cell line.*

By this approach 841 RefProts were identified among the three cell lines. Moreover, as shown in **Figure 16B**, 44 AltProts were identified by this approach. Five AltProts were identified specifically in PEO-4 and SKOV-3 cells, and two of them were identified in both cancerous cells secretomes. Additionally, **Figure 16C** displays the hierarchical clustering of the significantly regulated AltProts. In this heatmap five different expression clusters were found. Up regulation in: (1) both cancerous cells, (2) SKOV-3 and (5) PEO-4. Downregulation in: (3) both cancerous cells and (4) SKOV-3.

These preliminary results showed that by using high pH fractionation, the identification of AltProts can be done in samples with low protein abundance levels. Furthermore, this technique provides the first glimpse into the presence of AltProts secreted in OvCa cells. There is also potential for further research to explore the roles of cell-specific secreted AltProts and their regulation mechanisms. By conducting additional studies, we can gain a deeper understanding of these proteins and their significance in OvCa.

Proteogenomic approach at data reuse for tissue and primary tumors

Proteomic data reuse emphasizes the importance of maximizing the value of existing proteomic datasets. This approach could be particularly relevant for the identification of AltProts in tissue and primary tumors, as it offers an economical and sustainable method. By reevaluating and reanalyzing previously generated proteomic data, there is the potential to uncover new insights and discoveries without conducting additional time-consuming and expensive wet-lab experiments[396]. This highlights the significance of data sharing and open-access repositories, where proteomic data from different studies can be deposited and accessed by the scientific community. Reusing proteomic data should enable the discovery of AltProts that may not have been initially identified in the original studies, thus expanding our understanding of the proteomic landscape in tissue biopsies and primary tumors.

In the context of proteomic data reuse, open search algorithms play an important role. These algorithms provide a systematic and unbiased approach for analyzing proteomic data, allowing for the identification of AltProts. Open search algorithms enable researchers to explore proteomic datasets extensively, uncovering hidden patterns and connections that may have been overlooked by traditional targeted approaches.

Additionally, to evaluate the presence and levels of the AltProts described in this work, in clinical samples, a workflow comprising parallel reaction monitoring (PRM) can be developed. PRM utilizes high-resolution mass spectrometry to detect and analyze all product ions generated from a specific precursor ion. This enables the accurate identification and quantification of analytes with enhanced sensitivity and specificity[397,398]. By capturing a broader range of product ions, PRM offers the possibility to correctly identify AltProts. This approach has shown potential to identify specific biomarkers in breast cancer derived FFPE tissue samples[399].

To sum up, the work done in this project has opened new research perspectives and expanded the horizons of AltProt research. Unexplored AltProt interactions hold potential to more deeply understand deeply the role of AltProts. The wide range of techniques to investigate AltProts is truly exciting. Our project demonstrated innovation and forward-thinking by venturing into new targets, pushing the boundaries of knowledge. The opportunities for collaboration and interdisciplinary work resulting from these targets are instrumental and will enhance our understanding of AltProts.

References

1. The Union for International Cancer Control (UICC). *World Cancer Day 2020: International Public Opinion Survey on Cancer 2020*. https://www.uicc.org/resources/world-cancer-day-2020-international-public-opinion-survey-cancer-2020 (2022).

2. Ferlay, J. *et al.* Global Cancer Observatory: Cancer Today. https://gco.iarc.fr/today (2020).

3. American Society of Clinical Oncology (ASCO). Ovarian, Fallopian Tube, and Peritoneal Cancer - Statistics. *Cancer.Net* https://www.cancer.net/cancer-types/ovarian-fallopian-tube-and-peritoneal-cancer/statistics (2012).

4. Prat, J. & FIGO Committee on Gynecologic Oncology. Staging classification for cancer of the ovary, fallopian tube, and peritoneum. *Int. J. Gynaecol. Obstet. Off. Organ Int. Fed. Gynaecol. Obstet.* **124**, 1–5 (2014).

5. Guo, T. *et al.* Cellular Mechanism of Gene Mutations and Potential Therapeutic Targets in Ovarian Cancer. *Cancer Manag. Res.* **13**, 3081–3100 (2021).

6. Berek, J. S., Renz, M., Kehoe, S., Kumar, L. & Friedlander, M. Cancer of the ovary, fallopian tube, and peritoneum: 2021 update. *Int. J. Gynecol. Obstet.* **155**, 61–85 (2021).

7. Wentzensen, N. *et al.* Ovarian Cancer Risk Factors by Histologic Subtype: An Analysis From the Ovarian Cancer Cohort Consortium. *J. Clin. Oncol.* **34**, 2888–2898 (2016).

8. Howlader, N. *et al.* SEER cancer statistics review, 1975–2010. *Bethesda MD Natl. Cancer Inst.* **21**, 12 (2013).

9. Olson, S. H. *et al.* Symptoms of ovarian cancer. *Obstet. Gynecol.* **98**, 212–217 (2001).

10. McGrogan, B. T., Gilmartin, B., Carney, D. N. & McCann, A. Taxanes, microtubules and chemoresistant breast cancer. *Biochim. Biophys. Acta* **1785**, 96–132 (2008).

11. Soslow, R. A. Histologic Subtypes of Ovarian Carcinoma: An Overview. *Int. J. Gynecol. Pathol.* **PAP**, (2008).

12. Lewin, S. *et al.* Paraneoplastic hypercalcemia in clear cell ovarian adenocarcinoma. http://ecancer.org/en/journal/article/271-paraneoplastic-hypercalcemia-in-clear-cell-ovarian-adenocarcinoma (2012) doi:10.3332/ecancer.2012.271.

13. Genestie, C. *et al.* Histological classification of mucinous ovarian tumors: inter-observer reproducibility, clinical relevance, and role of genetic biomarkers. *Virchows Arch.* **478**, 885–891 (2021).

14. Knox, R. J., Friedlos, F., Lydall, D. A. & Roberts, J. J. Mechanism of cytotoxicity of anticancer platinum drugs: evidence that cis-diamminedichloroplatinum(II) and cis-diammine-(1,1-cyclobutanedicarboxylato)platinum(II) differ only in the kinetics of their interaction with DNA. *Cancer Res.* **46**, 1972–1979 (1986).

15. Stewart, C., Ralyea, C. & Lockwood, S. Ovarian Cancer: An Integrated Review. *Semin. Oncol. Nurs.* **35**, 151–156 (2019).

16. Akter, S. *et al.* Recent Advances in Ovarian Cancer: Therapeutic Strategies, Potential Biomarkers, and Technological Improvements. *Cells* **11**, 650 (2022).

17. Doo, D. W., Norian, L. A. & Arend, R. C. Checkpoint inhibitors in ovarian cancer: A review of preclinical data. *Gynecol. Oncol. Rep.* **29**, 48–54 (2019).

18. Zhang, X.-W., Wu, Y.-S., Xu, T.-M. & Cui, M.-H. CAR-T Cells in the Treatment of Ovarian Cancer: A Promising Cell Therapy. *Biomolecules* **13**, 465 (2023).

19. Sanvictores, T. & Farci, F. Biochemistry, Primary Protein Structure. in *StatPearls* (StatPearls Publishing, 2023).

20. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).

21. Wang, D. & Farhana, A. Biochemistry, RNA Structure. in *StatPearls* (StatPearls Publishing, 2023).

22. Morris, R., Black, K. A. & Stollar, E. J. Uncovering protein function: from classification to complexes. *Essays Biochem.* **66**, 255–285 (2022).

23. Crick, F. H. On protein synthesis. in *Symp Soc Exp Biol* vol. 12 8 (1958).

24. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol.* **15**, e2003243 (2017).

25. Dever, T. E., Kinzy, T. G. & Pavitt, G. D. Mechanism and Regulation of Protein Synthesis in Saccharomyces cerevisiae. *Genetics* **203**, 65–107 (2016).

26. Lareau, L. F., Hite, D. H., Hogan, G. J. & Brown, P. O. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife* **3**, e01257 (2014).

27. Lawson, M. R. *et al.* Mechanisms that ensure speed and fidelity in eukaryotic translation termination. *Science* **373**, 876–882 (2021).

28. Young, D. J. & Guydosh, N. R. Hcr1/eIF3j Is a 60S Ribosomal Subunit Recycling Accessory Factor In Vivo. *Cell Rep.* **28**, 39-50.e4 (2019).

29. Pisarev, A. V., Hellen, C. U. T. & Pestova, T. V. Recycling of Eukaryotic Posttermination Ribosomal Complexes. *Cell* **131**, 286–299 (2007).

30. Jackson, R. J., Hellen, C. U. T. & Pestova, T. V. THE MECHANISM OF EUKARYOTIC TRANSLATION INITIATION AND PRINCIPLES OF ITS REGULATION. *Nat. Rev. Mol. Cell Biol.* **11**, 113–127 (2010).

31. Passmore, L. A. *et al.* The Eukaryotic Translation Initiation Factors eIF1 and eIF1A Induce an Open Conformation of the 40S Ribosome. *Mol. Cell* **26**, 41–50 (2007).

32. Yu, Y. *et al.* Position of eukaryotic translation initiation factor eIF1A on the 40S ribosomal subunit mapped by directed hydroxyl radical probing. *Nucleic Acids Res.* **37**, 5167–5182 (2009).

33. Gingras, A. C., Raught, B. & Sonenberg, N. eIF4 initiation factors: effectors of mRNA recruitment to ribosomes and regulators of translation. *Annu. Rev. Biochem.* **68**, 913–963 (1999).

34. Grüner, S. *et al.* The Structures of eIF4E-eIF4G Complexes Reveal an Extended Interface to Regulate Translation Initiation. *Mol. Cell* **64**, 467–479 (2016).

35. Pisarev, A. V., Kolupaeva, V. G., Yusupov, M. M., Hellen, C. U. & Pestova, T. V. Ribosomal position and contacts of mRNA in eukaryotic translation initiation complexes. *EMBO J.* **27**, 1609–1621 (2008).

36. Jackson, R. J. The ATP requirement for initiation of eukaryotic translation varies according to the mRNA species. *Eur. J. Biochem.* **200**, 285–294 (1991).

37. Svitkin, Y. V. *et al.* The requirement for eukaryotic initiation factor 4A (eIF4A) in translation is in direct proportion to the degree of mRNA 5’ secondary structure. *RNA* **7**, 382–394 (2001).

38. Siridechadilok, B., Fraser, C. S., Hall, R. J., Doudna, J. A. & Nogales, E. Structural Roles for Human Translation Factor eIF3 in Initiation of Protein Synthesis. *Science* **310**, 1513–1515 (2005).

39. Krause, L., Willing, F., Andreou, A. Z. & Klostermeier, D. The domains of yeast eIF4G, eIF4E and the cap fine-tune eIF4A activities through an intricate network of stimulatory and inhibitory effects. *Nucleic Acids Res.* **50**, 6497–6510 (2022).

40. Marintchev, A. *et al.* Topology and regulation of the human eIF4A/4G/4H helicase complex in translation initiation. *Cell* **136**, 447–460 (2009).

41. Berthelot, K., Muldoon, M., Rajkowitsch, L., Hughes, J. & McCarthy, J. E. G. Dynamics and processivity of 40S ribosome scanning on mRNA in yeast. *Mol. Microbiol.* **51**, 987–1001 (2004).

42. Kozak, M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.* **196**, 947–950 (1987).

43. Donahue, T., Sonenberg, N., Hershey, J. W. B. & Mathews, M. B. *Translational control of gene expression*. (Cold Spring Harbor Laboratory Press Cold Spring Harbor, 2000).

44. Pestova, T. V., Borukhov, S. I. & Hellen, C. U. Eukaryotic ribosomes require initiation factors 1 and 1A to locate initiation codons. *Nature* **394**, 854–859 (1998).

45. Pisarev, A. V. *et al.* Specific functional interactions of nucleotides at key -3 and +4 positions flanking the initiation codon with components of the mammalian 48S translation initiation complex. *Genes Dev.* **20**, 624–636 (2006).

46. Hernández, G., Osnaya, V. G. & Pérez-Martínez, X. Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. *Trends Biochem. Sci.* **44**, 1009–1021 (2019).

47. Mitchell, S. F. & Lorsch, J. R. Should I Stay or Should I Go? Eukaryotic Translation Initiation Factors 1 and 1A Control Start Codon Recognition. *J. Biol. Chem.* **283**, 27345–27349 (2008).

48. Kapp, L. D. & Lorsch, J. R. GTP-dependent recognition of the methionine moiety on initiator tRNA by translation factor eIF2. *J. Mol. Biol.* **335**, 923–936 (2004).

49. Pestova, T. V. *et al.* The joining of ribosomal subunits in eukaryotes requires eIF5B. *Nature* **403**, 332–335 (2000).

50. Unbehaun, A. *et al.* Position of eukaryotic initiation factor eIF5B on the 80S ribosome mapped by directed hydroxyl radical probing. *EMBO J.* **26**, 3109–3123 (2007).

51. Allen, G. S., Zavialov, A., Gursky, R., Ehrenberg, M. & Frank, J. The cryo-EM structure of a translation initiation complex from Escherichia coli. *Cell* **121**, 703–712 (2005).

52. Blumenthal, T. Operons in eukaryotes. *Brief. Funct. Genomic. Proteomic.* **3**, 199–211 (2004).

53. Vanderperre, B., Lucier, J.-F. & Roucou, X. HAltORF: a database of predicted out-of-frame alternative open reading frames in human. *Database* **2012**, bas025–bas025 (2012).

54. Lee, S.-J. Expression of growth/differentiation factor 1 in the nervous system: conservation of a bicistronic structure. *Proc. Natl. Acad. Sci.* **88**, 4250–4254 (1991).

55. Kochetov, A. V. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *BioEssays* **30**, 683–691 (2008).

56. Gould, P. S., Dyer, N. P., Croft, W., Ott, S. & Easton, A. J. Cellular mRNAs access second ORFs using a novel amino acid sequence-dependent coupled translation termination–reinitiation mechanism. *RNA* **20**, 373–381 (2014).

57. Kozak, M. The scanning model for translation: an update. *J. Cell Biol.* **108**, 229–241 (1989).

58. Peabody, D. S. & Berg, P. Termination-reinitiation occurs in the translation of mammalian cell mRNAs. *Mol. Cell. Biol.* **6**, 2695–2703 (1986).

59. Pöyry, T. A. A., Kaminski, A. & Jackson, R. J. What determines whether mammalian ribosomes resume scanning after translation of a short upstream open reading frame? *Genes Dev.* **18**, 62–75 (2004).

60. Tautz, D. Polycistronic peptide coding genes in eukaryotes--how widespread are they? *Brief. Funct. Genomic. Proteomic.* **8**, 68–74 (2009).

61. Vilela, C., Linz, B., Rodrigues-Pousada, C. & McCarthy, J. E. G. The yeast transcription factor genes YAP1 and YAP2 are subject to differential control at the levels of both translation and mRNA stability. *Nucleic Acids Res.* **26**, 1150–1159 (1998).

62. Rossi, A., Pisani, F., Nicchia, G. P., Svelto, M. & Frigeri, A. Evidences for a Leaky Scanning Mechanism for the Synthesis of the Shorter M23 Protein Isoform of Aquaporin-4. *J. Biol. Chem.* **285**, 4562–4569 (2010).

63. Shestakova, E. D., Smirnova, V. V., Shatsky, I. N. & Terenin, I. M. Specific mechanisms of translation initiation in higher eukaryotes: the eIF4G2 story. *RNA* **29**, 282–299 (2023).

64. Smirnova, V. V. *et al.* Ribosomal leaky scanning through a translated uORF requires eIF4G2. *Nucleic Acids Res.* **50**, 1111–1127 (2022).

65. Bohlen, J., Roiuk, M., Neff, M. & Teleman, A. A. PRRC2 proteins impact translation initiation by promoting leaky scanning. *Nucleic Acids Res.* **51**, 3391–3409 (2023).

66. Andreev, D. E. *et al.* Non-AUG translation initiation in mammals. *Genome Biol.* **23**, 111 (2022).

67. Basu, I., Gorai, B., Chandran, T., Maiti, P. K. & Hussain, T. Selection of start codon during mRNA scanning in eukaryotic translation initiation. *Commun. Biol.* **5**, 1–10 (2022).

68. Kearse, M. G. & Wilusz, J. E. Non-AUG translation: a new start for protein synthesis in eukaryotes. *Genes Dev.* **31**, 1717–1731 (2017).

69. Diaz de Arce, A. J., Noderer, W. L. & Wang, C. L. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res.* **46**, 985–994 (2018).

70. Anderson, D. M. *et al.* A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* **160**, 595–606 (2015).

71. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* **324**, 218–223 (2009).

72. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).

73. Gong, Z. *et al.* Long non-coding RNAs in cancer. *Sci. China Life Sci.* **55**, 1120–1124 (2012).

74. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **24**, 430–447 (2023).

75. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).

76. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).

77. Cabili, M. N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).

78. Gingeras, T. R. Origin of phenotypes: genes and transcripts. *Genome Res.* **17**, 682–690 (2007).

79. Kelley, D. & Rinn, J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.* **13**, R107 (2012).

80. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155–159 (2009).

81. Oh, H. J. *et al.* Jpx RNA regulates CTCF anchor site selection and formation of chromosome loops. *Cell* **184**, 6157-6173.e24 (2021).

82. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510–514 (2019).

83. Djupedal, I. & Ekwall, K. Epigenetics: heterochromatin meets RNAi. *Cell Res.* **19**, 282–295 (2009).

84. Garcia-Jove Navarro, M. *et al.* RNA is a critical element for the sizing and the composition of phase-separated RNA-protein condensates. *Nat. Commun.* **10**, 3230 (2019).

85. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11667–11672 (2009).

86. Chen, H. & Liang, H. A High-Resolution Map of Human Enhancer RNA Loci Characterizes Super-enhancer Activities in Cancer. *Cancer Cell* **38**, 701-715.e5 (2020).

87. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147**, 789–802 (2011).

88. van Heesch, S. *et al.* Extensive localization of long noncoding RNAs to the cytosol and mono- and polyribosomal complexes. *Genome Biol.* **15**, R6 (2014).

89. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).

90. Juntawong, P., Girke, T., Bazin, J. & Bailey-Serres, J. Translational dynamics revealed by genome-wide profiling of ribosome footprints in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E203-212 (2014).

91. Brent, M. R. Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res.* **15**, 1777–1786 (2005).

92. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012).

93. Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14**, 942–950 (2004).

94. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).

95. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).

96. Garcia-del Rio, D. F. *et al.* Employing non-targeted interactomics approach and subcellular fractionation to increase our understanding of the ghost proteome. *iScience* **26**, (2023).

97. Griss, J. *et al.* Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets. *Nat. Methods* **13**, 651–656 (2016).

98. Crappé, J. *et al.* PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.* **43**, e29 (2015).

99. Smith, L. M. & Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nat. Methods* **10**, 186–187 (2013).

100. Vanderperre, B. *et al.* Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLOS ONE* **8**, e70698 (2013).

101. Chung, W.-Y., Wadhawan, S., Szklarczyk, R., Pond, S. K. & Nekrutenko, A. A first look at ARFome: dual-coding genes in mammalian genomes. *PLoS Comput. Biol.* **3**, e91 (2007).

102. Ribrioux, S., Brüngger, A., Baumgarten, B., Seuwen, K. & John, M. R. Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts. *BMC Genomics* **9**, 122 (2008).

103. Mouilleron, H., Delcourt, V. & Roucou, X. Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res.* **44**, 14–23 (2016).

104. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame–encoded peptides in human cells. *Nat. Chem. Biol.* **9**, 59–64 (2013).

105. Li, Y. *et al.* SmProt: A Reliable Repository with Comprehensive Annotation of Small Proteins Identified from Ribosome Profiling. *Genomics Proteomics Bioinformatics* **19**, 602–610 (2021).

106. Murgoci, A.-N. *et al.* Reference and Ghost Proteins Identification in Rat C6 Glioma Extracellular Vesicles. *iScience* **23**, 101045 (2020).

107. Xu, G. *et al.* Global translational reprogramming is a fundamental layer of immune regulation in plants. *Nature* **545**, 487–490 (2017).

108. Xu, Q. *et al.* Histone deacetylases control lysine acetylation of ribosomal proteins in rice. *Nucleic Acids Res.* **49**, 4613–4628 (2021).

109. Hanada, K. *et al.* Small open reading frames associated with morphogenesis are hidden in plant genomes. *Proc. Natl. Acad. Sci.* **110**, 2395–2400 (2013).

110. Smith, J. E. *et al.* Translation of Small Open Reading Frames within Unannotated RNA Transcripts in Saccharomyces cerevisiae. *Cell Rep.* **7**, 1858–1866 (2014).

111. Yang, Y. *et al.* An Optimized Proteomics Approach Reveals Novel Alternative Proteins in Mouse Liver Development. *Mol. Cell. Proteomics* **22**, (2023).

112. Fabre, B., Combier, J.-P. & Plaza, S. Recent advances in mass spectrometry–based peptidomics workflows to identify short-open-reading-frame-encoded peptides and explore their functions. *Curr. Opin. Chem. Biol.* **60**, 122–130 (2021).

113. Zhang, S. *et al.* Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nat. Commun.* **11**, 1312 (2020).

114. Samandi, S. *et al.* Deep transcriptome annotation enables the discovery and functional characterization of cryptic small proteins. *eLife* **6**, e27860 (2017).

115. Schlötterer, C. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).

116. Klemke, M., Kehlenbach, R. H. & Huttner, W. B. Two overlapping reading frames in a single exon encode interacting proteins--a novel way of gene usage. *EMBO J.* **20**, 3849–3860 (2001).

117. Freson, K. *et al.* Functional polymorphisms in the paternally expressed XLalphas and its cofactor ALEX decrease their mutual interaction and enhance receptor-mediated cAMP formation. *Hum. Mol. Genet.* **12**, 1121–1130 (2003).

118. Bergeron, D. *et al.* An out-of-frame overlapping reading frame in the ataxin-1 coding sequence encodes a novel ataxin-1 interacting protein. *J. Biol. Chem.* **288**, 21824–21835 (2013).

119. Lee, C., Lai, H.-L., Lee, Y.-C., Chien, C.-L. & Chern, Y. The A2A adenosine receptor is a dual coding gene: a novel mechanism of gene usage and signal transduction. *J. Biol. Chem.* **289**, 1257–1270 (2014).

120. Akimoto, C. *et al.* Translational repression of the McKusick-Kaufman syndrome transcript by unique upstream open reading frames encoding mitochondrial proteins with alternative polyadenylation sites. *Biochim. Biophys. Acta* **1830**, 2728–2738 (2013).

121. Cao, X. *et al.* Nascent alt-protein chemoproteomics reveals a pre-60S assembly checkpoint inhibitor. *Nat. Chem. Biol.* **18**, 643–651 (2022).

122. Yosten, G. L. C. *et al.* A 5′-upstream short open reading frame encoded peptide regulates angiotensin type 1a receptor production and signalling via the β-arrestin pathway. *J. Physiol.* **594**, 1601–1605 (2016).

123. Nelson, B. R. *et al.* A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271–275 (2016).

124. Zhang, Q. *et al.* The microprotein Minion controls cell fusion and muscle formation. *Nat. Commun.* **8**, 15664 (2017).

125. Stein, C. S. *et al.* Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Rep.* **23**, 3710-3720.e8 (2018).

126. Magny, E. G. *et al.* Conserved Regulation of Cardiac Calcium Uptake by Peptides Encoded in Small Open Reading Frames. *Science* **341**, 1116–1120 (2013).

127. Makarewich, C. A. *et al.* MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. *Cell Rep.* **23**, 3701–3709 (2018).

128. Slavoff, S. A., Heo, J., Budnik, B. A., Hanakahi, L. A. & Saghatelian, A. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* **289**, 10950–10957 (2014).

129. D'Lima, N. G. *et al.* A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **13**, 174–180 (2017).

130. Cardon, T., Salzet, M., Franck, J. & Fournier, I. Nuclei of HeLa cells interactomes unravel a network of ghost proteins involved in proteins translation. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1863**, 1458–1470 (2019).

131. Pauli, A. *et al.* Toddler: An Embryonic Signal That Promotes Cell Movement via Apelin Receptors. *Science* **343**, 1248636 (2014).

132. Matsumoto, A. *et al.* mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature* **541**, 228–232 (2017).

133. Rion, N. & Rüegg, M. A. LncRNA-encoded peptides: More than translational noise? *Cell Res.* **27**, 604–605 (2017).

134. Cardon, T. *et al.* Alternative proteins are functional regulators in cell reprogramming by PKA activation. *Nucleic Acids Res.* gkaa277 (2020) doi:10.1093/nar/gkaa277.

135. Meng, N. *et al.* Small Protein Hidden in lncRNA LOC90024 Promotes "Cancerous" RNA Splicing and Tumorigenesis. *Adv. Sci.* **7**, 1903233 (2020).

136. Polycarpou-Schwarz, M. *et al.* The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene* **37**, 4750–4768 (2018).

137. Delcourt, V. *et al.* Combined Mass Spectrometry Imaging and Top-down Microproteomics Reveals Evidence of a Hidden Proteome in Ovarian Cancer. *EBioMedicine* **21**, 55–64 (2017).

138. Huang, J.-Z. *et al.* A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Mol. Cell* **68**, 171-184.e6 (2017).

139. Wang, Y. *et al.* LncRNA-encoded polypeptide ASRPS inhibits triple-negative breast cancer angiogenesis. *J. Exp. Med.* **217**, jem.20190950 (2020).

140. Zhang, M. *et al.* A novel protein encoded by the circular form of the SHPRH gene suppresses glioma tumorigenesis. *Oncogene* **37**, 1805–1814 (2018).

141. Yang, Y. *et al.* Novel Role of FBXW7 Circular RNA in Repressing Glioma Tumorigenesis. *J. Natl. Cancer Inst.* **110**, (2018).

142. Zhang, M. *et al.* A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat. Commun.* **9**, 4475 (2018).

143. Nagalakshmi, U. *et al.* The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344–1349 (2008).

144. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).

145. Rio, D. C., Ares, M., Hannon, G. J. & Nilsen, T. W. Purification of RNA Using TRIzol (TRI Reagent). *Cold Spring Harb. Protoc.* **2010**, pdb.prot5439 (2010).

146. Escobar, M. D. & Hunt, J. L. A cost-effective RNA extraction technique from animal cells and tissue using silica columns. *J. Biol. Methods* **4**, e72 (2017).

147. He, H. *et al.* Integrated DNA and RNA extraction using magnetic beads from viral pathogens causing acute respiratory infections. *Sci. Rep.* **7**, 45199 (2017).

148. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).

149. Smale, G. & Sasse, J. RNA isolation from cartilage using density gradient centrifugation in cesium trifluoroacetate: an RNA preparation technique effective in the presence of high proteoglycan content. *Anal. Biochem.* **203**, 352–356 (1992).

150. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).

151. Nolan, T. & Bustin, S. Chapter 9. Procedures for Quality Control of RNA Samples for Use in Quantitative Reverse Transcription PCR. in (eds. Keer, J. T. & Birch, L.) 189–207 (Royal Society of Chemistry, 2008). doi:10.1039/9781847558213-00189.

152. Rio, D. C., Ares, M., Hannon, G. J. & Nilsen, T. W. Enrichment of poly(A)+ mRNA using immobilized oligo(dT). *Cold Spring Harb. Protoc.* **2010**, pdb.prot5454 (2010).

153. Herbert, Z. T. *et al.* Cross-site comparison of ribosomal depletion kits for Illumina RNAseq library construction. *BMC Genomics* **19**, 199 (2018).

154. Morlan, J. D., Qu, K. & Sinicropi, D. V. Selective Depletion of rRNA Enables Whole Transcriptome Profiling of Archival Fixed Tissue. *PLOS ONE* **7**, e42882 (2012).

155. Menzel, U. *et al.* Comprehensive Evaluation and Optimization of Amplicon Library Preparation Methods for High-Throughput Antibody Sequencing. *PLoS ONE* **9**, e96727 (2014).

156. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).

157. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

158. Patel, R. K. & Jain, M. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLOS ONE* **7**, e30619 (2012).

159. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

160. Chen, G., Wang, C. & Shi, T. Overview of available methods for diverse RNA-Seq data analyses. *Sci. China Life Sci.* **54**, 1121–1128 (2011).

161. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: A Fast Search Method for Large DNA Databases. *Genome Res.* **11**, 1725–1729 (2001).

162. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).

163. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).

164. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

165. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).

166. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

167. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

168. Delhomme, N., Padioleau, I., Furlong, E. E. & Steinmetz, L. M. easyRNASeq: a bioconductor package for processing RNA-Seq data. *Bioinformatics* **28**, 2532–2533 (2012).

169. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

170. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

171. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

172. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

173. Abrams, Z. B., Johnson, T. S., Huang, K., Payne, P. R. O. & Coombes, K. A protocol to evaluate RNA sequencing normalization methods. *BMC Bioinformatics* **20**, 679 (2019).

174. Filloux, C. *et al.* An integrative method to normalize RNA-Seq data. *BMC Bioinformatics* **15**, 188 (2014).

175. Zhao, S., Ye, Z. & Stanton, R. Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* **26**, 903–909 (2020).

176. Zhao, Y. *et al.* TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *J. Transl. Med.* **19**, 269 (2021).

177. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

178. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).

179. Smyth, G. K. limma: Linear Models for Microarray Data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer, 2005). doi:10.1007/0-387-29362-0_23.

180. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).

181. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. 201178 Preprint at https://doi.org/10.1101/201178 (2018).

182. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinforma. Oxf. Engl.* **27**, 2987–2993 (2011).

183. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at https://doi.org/10.48550/arXiv.1207.3907 (2012).

184. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).

185. Brunet, M. A., Leblanc, S. & Roucou, X. OpenVar: functional annotation of variants in non-canonical open reading frames. *Cell Biosci.* **12**, 130 (2022).

186. Schmidt, B., Cmero, M., Ekert, P., Davidson, N. & Oshlack, A. Slinker: Visualising novel splicing events in RNA-Seq data. *F1000Research* **10**, 1255 (2021).

187. Fenton, D. A., Kiniry, S. J., Yordanova, M. M., Baranov, P. V. & Morrissey, J. P. Development of a ribosome profiling protocol to study translation in Kluyveromyces marxianus. *FEMS Yeast Res.* **22**, foac024 (2022).

188. Legrand, C., Duc, K. D. & Tuorto, F. Analysis of Ribosome Profiling Data. *Methods Mol. Biol. Clifton NJ* **2428**, 133–156 (2022).

189. Meindl, A. *et al.* A rapid protocol for ribosome profiling of low input samples. *Nucleic Acids Res.* gkad459 (2023) doi:10.1093/nar/gkad459.

190. Legrand, C. & Tuorto, F. RiboVIEW: a computational framework for visualization, quality control and statistical analysis of ribosome profiling data. *Nucleic Acids Res.* **48**, e7 (2020).

191. Kiniry, S. J., O'Connor, P. B. F., Michel, A. M. & Baranov, P. V. Trips-Viz: a transcriptome browser for exploring Ribo-Seq data. *Nucleic Acids Res.* **47**, D847–D852 (2019).

192. Kiniry, S. J., Judge, C. E., Michel, A. M. & Baranov, P. V. Trips-Viz: an environment for the analysis of public and user-generated ribosome profiling data. *Nucleic Acids Res.* **49**, W662–W670 (2021).

193. Olexiouk, V. *et al.* sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **44**, D324–D329 (2016).

194. Pan, J. *et al.* Functional Micropeptides Encoded by Long Non-Coding RNAs: A Comprehensive Review. *Front. Mol. Biosci.* **9**, 817517 (2022).

195. Cui, M., Cheng, C. & Zhang, L. High-throughput proteomics: a methodological mini-review. *Lab. Invest.* **102**, 1170–1181 (2022).

196. Zhang, Z., Wu, S., Stenoien, D. L. & Paša-Tolić, L. High-throughput proteomics. *Annu. Rev. Anal. Chem. Palo Alto Calif* **7**, 427–454 (2014).

197. Tran, J. C. *et al.* Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* **480**, 254–258 (2011).

198. Valeja, S. G. *et al.* Unit mass baseline resolution for an intact 148 kDa therapeutic monoclonal antibody by Fourier transform ion cyclotron resonance mass spectrometry. *Anal. Chem.* **83**, 8391–8395 (2011).

199. Cai, W. *et al.* Top-Down Proteomics of Large Proteins up to 223 kDa Enabled by Serial Size Exclusion Chromatography Strategy. *Anal. Chem.* **89**, 5467–5475 (2017).

200. Fornelli, L. *et al.* Advancing Top-down Analysis of the Human Proteome Using a Benchtop Quadrupole-Orbitrap Mass Spectrometer. *J. Proteome Res.* **16**, 609–618 (2017).

201. Fort, K. L. *et al.* Expanding the structural analysis capabilities on an Orbitrap-based mass spectrometer for large macromolecular complexes. *The Analyst* **143**, 100–105 (2017).

202. Cleland, T. P. *et al.* High-Throughput Analysis of Intact Human Proteins Using UVPD and HCD on an Orbitrap Mass Spectrometer. *J. Proteome Res.* **16**, 2072–2079 (2017).

203. Ferguson, J. T., Wenger, C. D., Metcalf, W. W. & Kelleher, N. L. Top-down proteomics reveals novel protein forms expressed in Methanosarcina acetivorans. *J. Am. Soc. Mass Spectrom.* **20**, 1743–1750 (2009).

204. Le Rhun, E. *et al.* Evaluation of non-supervised MALDI mass spectrometry imaging combined with microproteomics for glioma grade III classification. *Biochim. Biophys. Acta Proteins Proteomics* **1865**, 875–890 (2017).

205. Mann, M., Hendrickson, R. C. & Pandey, A. Analysis of Proteins and Proteomes by Mass Spectrometry. *Annu. Rev. Biochem.* **70**, 437–473 (2001).

206. Varnavides, G. *et al.* In Search of a Universal Method: A Comparative Survey of Bottom-Up Proteomics Sample Preparation Methods. *J. Proteome Res.* **21**, 2397–2411 (2022).

207. Cleland, W. W. DITHIOTHREITOL, A NEW PROTECTIVE REAGENT FOR SH GROUPS. *Biochemistry* **3**, 480–482 (1964).

208. Boja, E. S. & Fales, H. M. Overalkylation of a protein digest with iodoacetamide. *Anal. Chem.* **73**, 3576–3582 (2001).

209. Giansanti, P., Tsiatsiani, L., Low, T. Y. & Heck, A. J. R. Six alternative proteases for mass spectrometry-based proteomics beyond trypsin. *Nat. Protoc.* **11**, 993–1006 (2016).

210. Glatter, T., Ahrné, E. & Schmidt, A. Comparison of Different Sample Preparation Protocols Reveals Lysis Buffer-Specific Extraction Biases in Gram-Negative Bacteria and Human Cells. *J. Proteome Res.* **14**, 4472–4485 (2015).

211. Doellinger, J., Schneider, A., Hoeller, M. & Lasch, P. Sample Preparation by Easy Extraction and Digestion (SPEED) - A Universal, Rapid, and Detergent-free Protocol for Proteomics Based on Acid Extraction. *Mol. Cell. Proteomics MCP* **19**, 209–222 (2020).

212. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–362 (2009).

213. HaileMariam, M. *et al.* S-Trap, an Ultrafast Sample-Preparation Approach for Shotgun Proteomics. *J. Proteome Res.* **17**, 2917–2924 (2018).

214. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).

215. Johnston, H. E. *et al.* Solvent Precipitation SP3 (SP4) Enhances Recovery for Proteomics Sample Preparation without Magnetic Beads. *Anal. Chem.* **94**, 10320–10328 (2022).

216. Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N. & Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **11**, 319–324 (2014).

217. Mant, C. T. *et al.* HPLC Analysis and Purification of Peptides. *Pept. Charact. Appl. Protoc.* **386**, 3–55 (2007).

218. Irvine, G. B. High-performance size-exclusion chromatography of peptides. *J. Biochem. Biophys. Methods* **56**, 233–242 (2003).

219. Edelmann, M. J. Strong cation exchange chromatography in analysis of posttranslational modifications: innovations and perspectives. *J. Biomed. Biotechnol.* **2011**, 936508 (2011).

220. Alpert, A. J., Hudecz, O. & Mechtler, K. Anion-Exchange Chromatography of Phosphopeptides: Weak Anion Exchange versus Strong Anion Exchange and Anion-Exchange Chromatography versus Electrostatic Repulsion–Hydrophilic Interaction Chromatography. *Anal. Chem.* **87**, 4704–4711 (2015).

221. Boersema, P. J., Mohammed, S. & Heck, A. J. R. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. *Anal. Bioanal. Chem.* **391**, 151–159 (2008).

222. Žuvela, P. *et al.* Column Characterization and Selection Systems in Reversed-Phase High-Performance Liquid Chromatography. *Chem. Rev.* **119**, 3674–3729 (2019).

223. Wilm, M. & Mann, M. Analytical Properties of the Nanoelectrospray Ion Source. *Anal. Chem.* **68**, 1–8 (1996).

224. Brown, S. L., Zenaidee, M. A., Loo, J. A., Loo, R. R. O. & Donald, W. A. On the Mechanism of Theta Capillary Nanoelectrospray Ionization for the Formation of Highly Charged Protein Ions Directly from Native Solutions. *Anal. Chem.* **94**, 13010–13018 (2022).

225. Michalski, A. *et al.* Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics MCP* **10**, M111.011015 (2011).

226. Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **72**, 1156–1162 (2000).

227. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom. JMS* **40**, 430–443 (2005).

228. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712 (2007).

229. Wells, J. M. & McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol.* **402**, 148–185 (2005).

230. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **113**, 2343–2394 (2013).

231. Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nat. Methods* **1**, 39–45 (2004).

232. Verheggen, K. *et al.* Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. *Mass Spectrom. Rev.* **39**, 292–306 (2020).

233. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).

234. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).

235. Craig, R. & Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinforma. Oxf. Engl.* **20**, 1466–1467 (2004).

236. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

237. Dorfer, V. *et al.* MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J. Proteome Res.* **13**, 3679–3684 (2014).

238. Kim, S. & Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **5**, 5277 (2014).

239. Degroeve, S. *et al.* ionbot: a novel, innovative and sensitive machine learning approach to LC-MS/MS peptide identification. 2021.07.02.450686 Preprint at https://doi.org/10.1101/2021.07.02.450686 (2022).

240. Tabb, D. L., Eng, J. K. & Yates, J. R. Protein Identification by SEQUEST. in *Proteome Research: Mass Spectrometry* (ed. James, P.) 125–142 (Springer, 2001). doi:10.1007/978-3-642-56895-4_7.

241. Orsburn, B. C. Proteome Discoverer—A Community Enhanced Data Processing Suite for Protein Informatics. *Proteomes* **9**, 15 (2021).

242. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).

243. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).

244. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

245. Cardon, T. *et al.* Optimized Sample Preparation Workflow for Improved Identification of Ghost Proteins. *Anal. Chem.* **92**, 1122–1129 (2020).

246. Meyer, J. G. Deep learning neural network tools for proteomics. *Cell Rep. Methods* **1**, 100003 (2021).

247. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36-42 (2013).

248. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

249. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

250. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

251. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).

252. Sjöstedt, E. *et al.* An atlas of the protein-coding genes in the human, pig, and mouse brain. *Science* **367**, eaay5947 (2020).

253. Bairoch, A. & Boeckmann, B. The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19**, 2247–2249 (1991).

254. Junker, V. *et al.* The role SWISS-PROT and TrEMBL play in the genome research environment. *J. Biotechnol.* **78**, 221–234 (2000).

255. Wu, C. H. The Protein Information Resource. *Nucleic Acids Res.* **31**, 345–347 (2003).

256. Hao, Y. *et al.* SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.* **19**, 636–643 (2018).

257. Brunet, M. A. *et al.* OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky936.

258. Olexiouk, V., Van Criekinge, W. & Menschaert, G. An update on sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **46**, D497–D502 (2018).

259. Brunet, M. A. *et al.* OpenProt 2021: deeper functional annotation of the coding potential of eukaryotic genomes. *Nucleic Acids Res.* **49**, D380–D388 (2021).

260. Pruitt, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* **19**, 1316–1323 (2009).

261. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).

262. Risk, B. A., Spitzer, W. J. & Giddings, M. C. Peppy: proteogenomic search software. *J. Proteome Res.* **12**, 3019–3025 (2013).

263. Zhao, Y. *et al.* NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **44**, D203-208 (2016).

264. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* **32**, W20-25 (2004).

265. Guilloy, N. *et al.* OpenCustomDB: Integration of Unannotated Open Reading Frames and Genetic Variants to Generate More Comprehensive Customized Protein Databases. *J. Proteome Res.* **22**, 1492–1500 (2023).

266. Ma, J. *et al.* Discovery of Human sORF-Encoded Polypeptides (SEPs) in Cell Lines and Tissue. *J. Proteome Res.* **13**, 1757–1765 (2014).

267. Wang, B. *et al.* Identification and analysis of small proteins and short open reading frame encoded peptides in Hep3B cell. *J. Proteomics* **230**, 103965 (2021).

268. Cassidy, L., Prasse, D., Linke, D., Schmitz, R. A. & Tholey, A. Combination of Bottom-up 2D-LC-MS and Semi-top-down GelFree-LC-MS Enhances Coverage of Proteome and Low Molecular Weight Short Open Reading Frame Encoded Peptides of the Archaeon *Methanosarcina mazei*. *J. Proteome Res.* **15**, 3773–3783 (2016).

269. Cao, X. *et al.* Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines. *J. Proteome Res.* **19**, 3418–3426 (2020).

270. Cassidy, L., Kaulich, P. T. & Tholey, A. Depletion of High-Molecular-Mass Proteins for the Identification of Small Proteins and Short Open Reading Frame Encoded Peptides in Cellular Proteomes. *J. Proteome Res.* **18**, 1725–1734 (2019).

271. Ma, J. *et al.* Improved Identification and Analysis of Small Open Reading Frame Encoded Polypeptides. *Anal. Chem.* **88**, 3967–3975 (2016).

272. McDonald, L., Robertson, D. H. L., Hurst, J. L. & Beynon, R. J. Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat. Methods* **2**, 955–957 (2005).

273. Bogaert, A. *et al.* Limited Evidence for Protein Products of Noncoding Transcripts in the HEK293T Cellular Cytosol. *Mol. Cell. Proteomics MCP* **21**, 100264 (2022).

274. Staes, A. *et al.* Protease Substrate Profiling by N-Terminal COFRADIC. in *Protein Terminal Profiling: Methods and Protocols* (ed. Schilling, O.) 51–76 (Springer, 2017). doi:10.1007/978-1-4939-6850-3_5.

275. Letovsky, S. & Kasif, S. Predicting protein function from protein/proteininteraction data: a probabilistic approach. *Bioinformatics* **19**, i197–i204 (2003).

276. Van Criekinge, W. & Beyaert, R. Yeast two-hybrid: State of the art. *Biol. Proced. Online* **2**, 1–38 (1999).

277. Ha, T. Single-Molecule Fluorescence Resonance Energy Transfer. *Methods* **25**, 78–86 (2001).

278. Söderberg, O. *et al.* Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. *Methods* **45**, 227–232 (2008).

279. Dunham, W. H., Mullin, M. & Gingras, A.-C. Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *PROTEOMICS* **12**, 1576–1590 (2012).

280. Martell, J. D. *et al.* Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat. Biotechnol.* **30**, 1143–1148 (2012).

281. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* **196**, 801–810 (2012).

282. Eyckerman, S. *et al.* Trapping mammalian protein complexes in viral particles. *Nat. Commun.* **7**, 11416 (2016).

283. Liu, F., Rijkers, D. T. S., Post, H. & Heck, A. J. R. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* **12**, 1179–1184 (2015).

284. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

285. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).

286. Jensen, L. J. *et al.* STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* **37**, D412-416 (2009).

287. Doncheva, N. T., Morris, J. H., Gorodkin, J. & Jensen, L. J. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* **18**, 623–632 (2019).

288. Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358–D363 (2014).

289. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. Publ. Protein Soc.* **30**, 187–200 (2021).

290. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).

291. Brückner, A., Polge, C., Lentze, N., Auerbach, D. & Schlattner, U. Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *Int. J. Mol. Sci.* **10**, 2763–2788 (2009).

292. Inchingolo, M. A. *et al.* TP53BP1, a dual-coding gene, uses promoter switching and translational reinitiation to express a smORF protein. *iScience* **26**, 106757 (2023).

293. Alam, M. S. Proximity Ligation Assay (PLA). *Curr. Protoc. Immunol.* **123**, e58 (2018).

294. Sandmann, C.-L. *et al.* Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol. Cell* **83**, 994-1011.e18 (2023).

295. Kwan, J. H. M. & Emili, A. Simple and Effective Affinity Purification Procedures for Mass Spectrometry-Based Identification of Protein-Protein Interactions in Cell Signaling Pathways. *Methods Mol. Biol. Clifton NJ* **1394**, 181–187 (2016).

296. November 2020, 19. The Human Protein Atlas: A 20-year journey into the body. *Science | AAAS* https://www.sciencemag.org/collections/human-protein-atlas-20-year-journey-body?utm_source=3p-hl&utm_medium=email&utm_content=hpa-bklt&utm_campaign=cp2020&et_rid=38470108&et_cid=3573031 (2020).

297. Nguyen, T. M. T., Kim, J., Doan, T. T., Lee, M.-W. & Lee, M. APEX Proximity Labeling as a Versatile Tool for Biological Research. *Biochemistry* **59**, 260–269 (2020).

298. Singer-Krüger, B. *et al.* APEX2-mediated proximity labeling resolves protein networks in Saccharomyces cerevisiae cells. *FEBS J.* **287**, 325–344 (2020).

299. Chu, Q. *et al.* Identification of Microprotein–Protein Interactions via APEX Tagging. *Biochemistry* **56**, 3299–3306 (2017).

300. Roux, K. J., Kim, D. I., Burke, B. & May, D. G. BioID: A Screen for Protein-Protein Interactions. *Curr. Protoc. Protein Sci.* **91**, 19.23.1-19.23.15 (2018).

301. Go, C. D. *et al.* A proximity-dependent biotinylation map of a human cell. *Nature* 1–5 (2021) doi:10.1038/s41586-021-03592-2.

302. Cho, K. F. *et al.* Proximity labeling in mammalian cells with TurboID and split-TurboID. *Nat. Protoc.* **15**, 3971–3999 (2020).

303. Na, Z. *et al.* Mapping subcellular localizations of unannotated microproteins and alternative proteins with MicroID. *Mol. Cell* **82**, 2900-2911.e7 (2022).

304. Titeca, K. *et al.* Analyzing trapped protein complexes by Virotrap and SFINX. *Nat. Protoc.* **12**, 881–898 (2017).

305. De Meyer, M. *et al.* Capturing Salmonella SspH2 Host Targets in Virus-Like Particles. *Front. Med.* **8**, (2021).

306. Titeca, K. *et al.* SFINX: Straightforward Filtering Index for Affinity Purification–Mass Spectrometry Data Analysis. *J. Proteome Res.* **15**, 332–338 (2016).

307. Bogaert, A., Van de Steene, T., Vuylsteke, M., Eyckerman, S. & Gevaert, K. A decoupled Virotrap approach to study the interactomes of N-terminal proteoforms. *Methods Enzymol.* **684**, 253–287 (2023).

308. Piersimoni, L., Kastritis, P. L., Arlt, C. & Sinz, A. Cross-Linking Mass Spectrometry for Investigating Protein Conformations and Protein–Protein Interactions─A Method for All Seasons. *Chem. Rev.* (2021) doi:10.1021/acs.chemrev.1c00786.

309. Yu, C. & Huang, L. Cross-Linking Mass Spectrometry (XL-MS): an Emerging Technology for Interactomics and Structural Biology. *Anal. Chem.* **90**, 144–165 (2018).

310. Gaucher, S. P., Hadi, M. Z. & Young, M. M. Influence of crosslinker identity and position on gas-phase dissociation of lys-lys crosslinked peptides. *J. Am. Soc. Mass Spectrom.* **17**, 395–405 (2006).

311. Rappsilber, J., Siniossoglou, S., Hurt, E. C. & Mann, M. A Generic Strategy To Analyze the Spatial Organization of Multi-Protein Complexes by Cross-Linking and Mass Spectrometry. *Anal. Chem.* **72**, 267–275 (2000).

312. Sinz, A. Chemical cross-linking and mass spectrometry for mapping three-dimensional structures of proteins and protein complexes. *J. Mass Spectrom.* **38**, 1225–1237 (2003).

313. Arlt, C., Ihling, C. H. & Sinz, A. Structure of full-length p53 tumor suppressor probed by chemical cross-linking and mass spectrometry. *PROTEOMICS* **15**, 2746–2755 (2015).

314. Ryl, P. S. J. *et al.* In Situ Structural Restraints from Cross-Linking Mass Spectrometry in Human Mitochondria. *J. Proteome Res.* **19**, 327–336 (2020).

315. Stahl, K., Graziadei, A., Dau, T., Brock, O. & Rappsilber, J. Protein structure prediction with in-cell photo-crosslinking mass spectrometry and deep learning. *Nat. Biotechnol.* 1–10 (2023) doi:10.1038/s41587-023-01704-z.

316. Liu, F. & Heck, A. J. Interrogating the architecture of protein assemblies and protein interaction networks by cross-linking mass spectrometry. *Curr. Opin. Struct. Biol.* **35**, 100–108 (2015).

317. Götze, M. *et al.* Automated assignment of MS/MS cleavable cross-links in protein 3D-structure analysis. *J. Am. Soc. Mass Spectrom.* **26**, 83–97 (2015).

318. Müller, M. Q., Dreiocker, F., Ihling, C. H., Schäfer, M. & Sinz, A. Cleavable Cross-Linker for Protein Structure Analysis: Reliable Identification of Cross-Linking Products by Tandem MS. *Anal. Chem.* **82**, 6958–6968 (2010).

319. Kao, A. *et al.* Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes. *Mol. Cell. Proteomics MCP* **10**, M110.002212 (2011).

320. Petrotchenko, E. V., Serpa, J. J. & Borchers, C. H. An Isotopically Coded CID-cleavable Biotinylated Cross-linker for Structural Proteomics. *Mol. Cell. Proteomics MCP* **10**, (2011).

321. Matzinger, M. & Mechtler, K. Cleavable Cross-Linkers and Mass Spectrometry for the Ultimate Task of Profiling Protein–Protein Interaction Networks *in Vivo*. *J. Proteome Res.* **20**, 78–93 (2021).

322. Nury, C. *et al.* A Novel Bio-Orthogonal Cross-Linker for Improved Protein/Protein Interaction Analysis. *Anal. Chem.* **87**, 1853–1860 (2015).

323. Rey, M. *et al.* Advanced In Vivo Cross-Linking Mass Spectrometry Platform to Characterize Proteome-Wide Protein Interactions. *Anal. Chem.* (2021) doi:10.1021/acs.analchem.0c04430.

324. Steigenberger, B., Pieters, R. J., Heck, A. J. R. & Scheltema, R. A. PhoX: An IMAC-Enrichable Cross-Linking Reagent. *ACS Cent. Sci.* **5**, 1514–1522 (2019).

325. Jiang, P.-L. *et al.* A Membrane-Permeable and Immobilized Metal Affinity Chromatography (IMAC) Enrichable Cross-Linking Reagent to Advance In Vivo Cross-Linking Mass Spectrometry. *Angew. Chem. Int. Ed.* **61**, e202113937 (2022).

326. Yılmaz, Ş., Busch, F., Nagaraj, N. & Cox, J. Accurate and Automated High-Coverage Identification of Chemically Cross-Linked Peptides with MaxLynx. *Anal. Chem.* **94**, 1608–1617 (2022).

327. Götze, M. *et al.* StavroX--a software for analyzing crosslinked products in protein interaction studies. *J. Am. Soc. Mass Spectrom.* **23**, 76–87 (2012).

328. Netz, E. *et al.* OpenPepXL: An Open-Source Tool for Sensitive Identification of Cross-Linked Peptides in XL-MS. *Mol. Cell. Proteomics* **19**, 2157–2168 (2020).

329. Fischer, L. & Rappsilber, J. Quirks of Error Estimation in Cross-Linking/Mass Spectrometry. *Anal. Chem.* **89**, 3829–3833 (2017).

330. Heil, L. R. *et al. Evaluating the performance of the Astral mass analyzer for quantitative proteomics using data independent acquisition*. http://biorxiv.org/lookup/doi/10.1101/2023.06.03.543570 (2023) doi:10.1101/2023.06.03.543570.

331. Vanderperre, B. *et al.* An overlapping reading frame in the PRNP gene encodes a novel polypeptide distinct from the prion protein. *FASEB J.* **25**, 2373–2386 (2011).

332. Weirick, T. *et al.* The identification and characterization of novel transcripts from RNA-seq data. *Brief. Bioinform.* **17**, 678–685 (2016).

333. Schweppe, D. K. *et al.* Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1732–1737 (2017).

334. Yugandhar, K., Wang, T.-Y., Wierbowski, S. D., Shayhidin, E. E. & Yu, H. Structure-based validation can drastically underestimate error rate in proteome-wide cross-linking mass spectrometry studies. *Nat. Methods* **17**, 985–988 (2020).

335. Mendes, M. L. *et al.* An integrated workflow for crosslinking mass spectrometry. *Mol. Syst. Biol.* **15**, e8994 (2019).

336. Stieger, C. E., Doppler, P. & Mechtler, K. Optimized Fragmentation Improves the Identification of Peptides Cross-Linked by MS-Cleavable Reagents. *J. Proteome Res.* **18**, 1363–1370 (2019).

337. Paulo, J. A. *et al.* Subcellular Fractionation Enhances Proteome Coverage of Pancreatic Duct Cells. *Biochim. Biophys. Acta* **1834**, 791–797 (2013).

338. Ramsby, M. L. & Makowski, G. S. Differential Detergent Fractionation of Eukaryotic Cells. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 37–48 (Humana Press, 2005). doi:10.1385/1-59259-890-0:037.

339. Gao, H. *et al.* In-Depth In Vivo Crosslinking in Minutes by a Compact, Membrane-Permeable, and Alkynyl-Enrichable Crosslinker. *Anal. Chem.* **94**, 7551–7558 (2022).

340. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).

341. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

342. Kozakov, D. *et al.* The ClusPro web server for protein–protein docking. *Nat. Protoc.* **12**, 255–278 (2017).

343. Kim, M. *et al.* The lymphotactin receptor is expressed in epithelial ovarian carcinoma and contributes to cell migration and proliferation. *Mol. Cancer Res. MCR* **10**, 1419–1429 (2012).

344. Rahman, M. A. *et al.* Artonin E Induces Apoptosis via Mitochondrial Dysregulation in SKOV-3 Ovarian Cancer Cells. *PloS One* **11**, e0151466 (2016).

345. Liu, Q. *et al.* The role of R-spondin 1 through activating Wnt/β-catenin in the growth, survival and migration of ovarian cancer cells. *Gene* **689**, 124–130 (2019).

346. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

347. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2019).

348. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

349. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* **302**, 249–255 (2003).

350. Haubrich, B. A. & Swinney, D. C. Enzyme Activity Assays for Protein Kinases: Strategies to Identify Active Substrates. *Curr. Drug Discov. Technol.* **13**, 2–15 (2016).

351. Nakato, R. & Sakata, T. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods* **187**, 44–53 (2021).

352. Vartiainen, S. *et al.* Phenotypic rescue of a Drosophila model of mitochondrial ANT1 disease. *Dis. Model. Mech.* **7**, 635–648 (2014).

353. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

354. Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022).

355. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).

356. Andreatta, M. & Nielsen, M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* **32**, 511–517 (2016).

357. Wieczorek, M. *et al.* Major Histocompatibility Complex (MHC) Class I and MHC Class II Proteins: Conformational Plasticity in Antigen Presentation. *Front. Immunol.* **8**, (2017).

358. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).

359. Torres-Perez, J. V., Irfan, J., Febrianto, M. R., Giovanni, S. D. & Nagy, I. Histone post-translational modifications as potential therapeutic targets for pain management. *Trends Pharmacol. Sci.* **42**, 897–911 (2021).

360. Le, S. N. *et al.* The TAFs of TFIID Bind and Rearrange the Topology of the TATA-Less RPS5 Promoter. *Int. J. Mol. Sci.* **20**, 3290 (2019).

361. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).

362. Armengaud, J. Chapter Twelve - Reannotation of Genomes by Means of Proteomics Data. in *Methods in Enzymology* (ed. Shukla, A. K.) vol. 585 201–216 (Academic Press, 2017).

363. Fancello, L. & Burger, T. An analysis of proteogenomics and how and when transcriptome-informed reduction of protein databases can enhance eukaryotic proteomics. *Genome Biol.* **23**, 132 (2022).

364. Rodriguez, H., Zenklusen, J. C., Staudt, L. M., Doroshow, J. H. & Lowy, D. R. The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment. *Cell* **184**, 1661–1670 (2021).

365. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* **45**, 1113–1120 (2013).

366. Flaherty, K. T. *et al.* The Molecular Analysis for Therapy Choice (NCI-MATCH) Trial: Lessons for Genomic Trial Design. *JNCI J. Natl. Cancer Inst.* **112**, 1021–1029 (2020).

367. Saad, E. D., Paoletti, X., Burzykowski, T. & Buyse, M. Precision medicine needs randomized clinical trials. *Nat. Rev. Clin. Oncol.* **14**, 317–323 (2017).

368. Rogers, S. *et al.* Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics* **24**, 2894–2900 (2008).

369. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).

370. Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* **11**, 1759 (2020).

371. Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755–765 (2016).

372. Mirabelli, P., Coppola, L. & Salvatore, M. Cancer Cell Lines Are Useful Model Systems for Medical Research. *Cancers* **11**, 1098 (2019).

373. Gillet, J.-P., Varma, S. & Gottesman, M. M. The Clinical Relevance of Cancer Cell Lines. *JNCI J. Natl. Cancer Inst.* **105**, 452–458 (2013).

374. Wolf, C. R. *et al.* Cellular heterogeneity and drug resistance in two ovarian adenocarcinoma cell lines derived from a single patient. *Int. J. Cancer* **39**, 695–702 (1987).

375. Langdon, S. P. *et al.* Characterization and Properties of Nine Human Ovarian Adenocarcinoma Cell Lines. 8.

376. Fogh, J., Fogh, J. M. & Orfeo, T. One hundred and twenty-seven cultured human tumor cell lines producing tumors in nude mice. *J. Natl. Cancer Inst.* **59**, 221–226 (1977).

377. Wang, Q. & Shi, W. UNBS5162 inhibits SKOV3 ovarian cancer cell proliferation by regulating the PI3K/AKT signalling pathway. *Oncol. Lett.* **17**, 2976–2982 (2019).

378. Hernandez, L. *et al.* Characterization of ovarian cancer cell lines as in vivo models for preclinical studies. *Gynecol. Oncol.* **142**, 332–340 (2016).

379. Rozanova, S. *et al.* Quantitative Mass Spectrometry-Based Proteomics: An Overview. in *Quantitative Methods in Proteomics* (eds. Marcus, K., Eisenacher, M. & Sitek, B.) 85–116 (Springer US, 2021). doi:10.1007/978-1-0716-1024-4_8.

380. Perez-Riverol, Y. *et al.* In silico analysis of accurate proteomics, complemented by selective isolation of peptides. *J. Proteomics* **74**, 2071–2082 (2011).

381. Li, Z. *et al.* Systematic Comparison of Label-Free, Metabolic Labeling, and Isobaric Chemical Labeling for Quantitative Proteomics on LTQ Orbitrap Velos. *J. Proteome Res.* **11**, 1582–1590 (2012).

382. Bantscheff, M., Lemeer, S., Savitski, M. M. & Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **404**, 939–965 (2012).

383. Liu, H., Sadygov, R. G. & Yates, J. R. A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Anal. Chem.* **76**, 4193–4201 (2004).

384. Voyksner, R. D. & Lee, H. Investigating the use of an octupole ion guide for ion storage and high-pass mass filtering to improve the quantitative performance of electrospray ion trap mass spectrometry. *Rapid Commun. Mass Spectrom.* **13**, 1427–1437 (1999).

385. Wilm, M. Quantitative proteomics in biological research. *PROTEOMICS* **9**, 4590–4605 (2009).

386. Palomba, A. *et al.* Comparative Evaluation of MaxQuant and Proteome Discoverer MS1-Based Protein Quantification Tools. *J. Proteome Res.* **20**, 3497–3507 (2021).

387. Tumini, E., Barroso, S., -Calero, C. P. & Aguilera, A. Roles of human POLD1 and POLD3 in genome stability. *Sci. Rep.* **6**, 38873 (2016).

388. Ogi, T. *et al.* Three DNA polymerases, recruited by different mechanisms, carry out NER repair synthesis in human cells. *Mol. Cell* **37**, 714–727 (2010).

389. Murga, M. *et al.* POLD3 Is Haploinsufficient for DNA Replication in Mice. *Mol. Cell* **63**, 877–883 (2016).

390. Meng, X., Zhou, Y., Lee, E. Y. C., Lee, M. Y. W. T. & Frick, D. N. The p12 subunit of human polymerase delta modulates the rate and fidelity of DNA synthesis. *Biochemistry* **49**, 3545–3554 (2010).

391. Schaeffer, M. *et al.* The neXtProt peptide uniqueness checker: a tool for the proteomics community. *Bioinforma. Oxf. Engl.* **33**, 3471–3472 (2017).

392. Geladaki, A. *et al.* Combining LOPIT with differential ultracentrifugation for high-resolution spatial proteomics. *Nat. Commun.* **10**, 331 (2019).

393. Cao, X., Chen, Y., Khitun, A. & Slavoff, S. A. BONCAT-based Profiling of Nascent Small and Alternative Open Reading Frame-encoded Proteins. *Bio-Protoc.* **13**, e4585 (2023).

394. Capuz, A. *et al.* Heimdall, an alternative protein issued from a ncRNA related to kappa light chain variable region of immunoglobulins from astrocytes: a new player in neural proteome. *Cell Death Dis.* **14**, 1–23 (2023).

395. Martinez, T. F. *et al.* Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. *Cell Metab.* **35**, 166-183.e11 (2023).

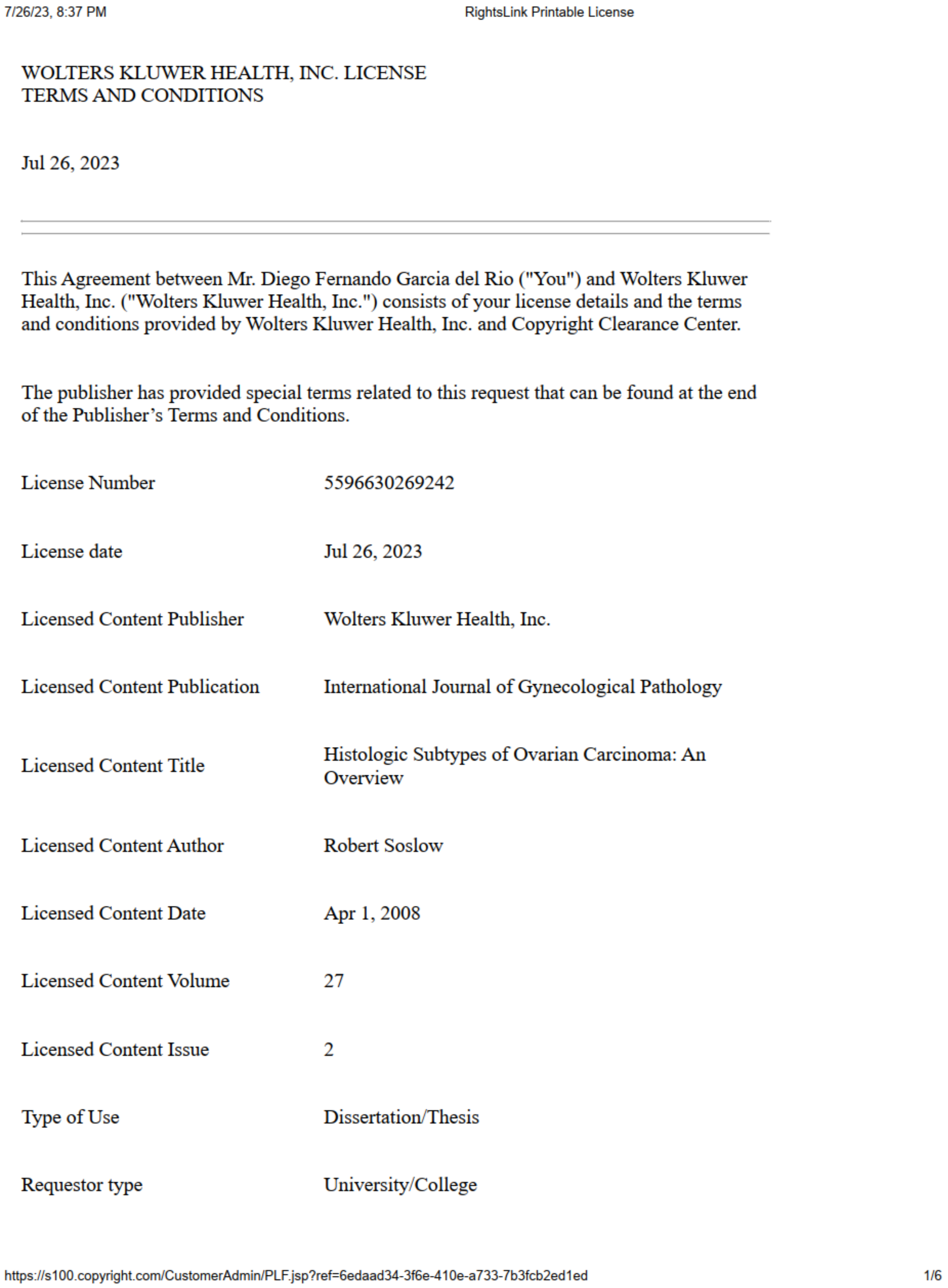
396. Cardon, T., Fournier, I. & Salzet, M. SARS-Cov-2 Interactome with Human Ghost Proteome: A Neglected World Encompassing a Wealth of Biological Data. *Microorganisms* **8**, 2036 (2020).

397. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S. & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol. Cell. Proteomics MCP* **11**, 1475–1488 (2012).

398. Bourmaud, A., Gallien, S. & Domon, B. Parallel reaction monitoring using quadrupole-Orbitrap mass spectrometer: Principle and applications. *PROTEOMICS* **16**, 2146–2159 (2016).

399. Park, J. *et al.* Parallel Reaction Monitoring-Mass Spectrometry (PRM-MS)-Based Targeted Proteomic Surrogates for Intrinsic Subtypes in Breast Cancer: Comparative Analysis with Immunohistochemical Phenotypes. *J. Proteome Res.* **19**, 2643–2653 (2020).

Figure Rights



7/26/23, 8:37 PM RightsLink Printable License

WOLTERS KLUWER HEALTH, INC. LICENSE
TERMS AND CONDITIONS

Jul 26, 2023

This Agreement between Mr. Diego Fernando Garcia del Rio ("You") and Wolters Kluwer Health, Inc. ("Wolters Kluwer Health, Inc.") consists of your license details and the terms and conditions provided by Wolters Kluwer Health, Inc. and Copyright Clearance Center.

The publisher has provided special terms related to this request that can be found at the end of the Publisher's Terms and Conditions.

License Number	5596630269242
License date	Jul 26, 2023
Licensed Content Publisher	Wolters Kluwer Health, Inc.
Licensed Content Publication	International Journal of Gynecological Pathology
Licensed Content Title	Histologic Subtypes of Ovarian Carcinoma: An Overview
Licensed Content Author	Robert Soslow
Licensed Content Date	Apr 1, 2008
Licensed Content Volume	27
Licensed Content Issue	2
Type of Use	Dissertation/Thesis
Requestor type	University/College

https://s100.copyright.com/CustomerAdmin/PLF.jsp?ref=6edaad34-3f6e-410e-a733-7b3fcb2ed1ed 1/6

Sponsorship	No Sponsorship
Format	Print and electronic
Will this be posted online?	Yes, on a secure website
Portion	Figures/tables/illustrations
Number of figures/tables/illustrations	3
Author of this Wolters Kluwer article	No
Will you be translating?	No
Intend to modify/change the content	No
Title	Studying Protein Complexes for Assessing the Function of Ghost Proteins (Ghost in the Cell)
Institution name	University of Lille
Expected presentation date	Oct 2023
Order reference number	1
Portions	Figures 1, 3, and 9
Requestor Location	Mr. Diego Fernando Garcia del Rio Avenue Paul Langevin - Bâtiment SN3 - 1e Université de Lille Villeneuve d'Ascq, 59655 France Attn: Mr. Diego Fernando Garcia del Rio
Publisher Tax ID	EU826013006

Billing Type	Invoice
Billing Address	Mr. Diego Fernando Garcia del Rio Avenue Paul Langevin - Bâtiment SN3 - 1e Université de Lille Villeneuve d'Ascq, France 59655 Attn: Mr. Diego Fernando Garcia del Rio
Total	0.00 EUR

Terms and Conditions

Wolters Kluwer Health Inc. Terms and Conditions

1. **Duration of License:** Permission is granted for a one time use only. Rights herein do not apply to future reproductions, editions, revisions, or other derivative works. This permission shall be effective as of the date of execution by the parties for the maximum period of 12 months and should be renewed after the term expires.
    i. When content is to be republished in a book or journal the validity of this agreement should be the life of the book edition or journal issue.
    ii. When content is licensed for use on a website, internet, intranet, or any publicly accessible site (not including a journal or book), you agree to remove the material from such site after 12 months, or request to renew your permission license
2. **Credit Line:** A credit line must be prominently placed and include: For book content: the author(s), title of book, edition, copyright holder, year of publication; For journal content: the author(s), titles of article, title of journal, volume number, issue number, inclusive pages and website URL to the journal page; If a journal is published by a learned society the credit line must include the details of that society.
3. **Warranties:** The requestor warrants that the material shall not be used in any manner which may be considered derogatory to the title, content, authors of the material, or to Wolters Kluwer Health, Inc.
4. **Indemnity:** You hereby indemnify and hold harmless Wolters Kluwer Health, Inc. and its respective officers, directors, employees and agents, from and against any and all claims, costs, proceeding or demands arising out of your unauthorized use of the Licensed Material
5. **Geographical Scope:** Permission granted is non-exclusive and is valid throughout the world in the English language and the languages specified in the license.
6. **Copy of Content:** Wolters Kluwer Health, Inc. cannot supply the requestor with the original artwork, high-resolution images, electronic files or a clean copy of content.
7. **Validity:** Permission is valid if the borrowed material is original to a Wolters Kluwer Health, Inc. imprint (J.B Lippincott, Lippincott-Raven Publishers, Williams & Wilkins, Lea & Febiger, Harwal, Rapid Science, Little Brown & Company, Harper & Row Medical, American Journal of Nursing Co, and Urban & Schwarzenberg - English Language, Raven Press, Paul Hoeber, Springhouse, Ovid), and the Anatomical Chart Company

8. **Third Party Material:** This permission does not apply to content that is credited to publications other than Wolters Kluwer Health, Inc. or its Societies. For images credited to non-Wolters Kluwer Health, Inc. books or journals, you must obtain permission from the source referenced in the figure or table legend or credit line before making any use of the image(s), table(s) or other content.
9. **Adaptations:** Adaptations are protected by copyright. For images that have been adapted, permission must be sought from the rightsholder of the original material and the rightsholder of the adapted material.
10. **Modifications:** Wolters Kluwer Health, Inc. material is not permitted to be modified or adapted without written approval from Wolters Kluwer Health, Inc. with the exception of text size or color. The adaptation should be credited as follows: Adapted with permission from Wolters Kluwer Health, Inc.: [the author(s), title of book, edition, copyright holder, year of publication] or [the author(s), titles of article, title of journal, volume number, issue number, inclusive pages and website URL to the journal page].
11. **Full Text Articles:** Republication of full articles in English is prohibited.
12. **Branding and Marketing:** No drug name, trade name, drug logo, or trade logo can be included on the same page as material borrowed from *Diseases of the Colon & Rectum, Plastic Reconstructive Surgery, Obstetrics & Gynecology (The Green Journal), Critical Care Medicine, Pediatric Critical Care Medicine, the American Heart Association publications and the American Academy of Neurology publications.*
13. **Open Access:** Unless you are publishing content under the same Creative Commons license, the following statement must be added when reprinting material in Open Access journals: "The Creative Commons license does not apply to this content. Use of the material in any format is prohibited without written permission from the publisher, Wolters Kluwer Health, Inc. Please contact permissions@lww.com for further information."
14. **Translations:** The following disclaimer must appear on all translated copies: Wolters Kluwer Health, Inc. and its Societies take no responsibility for the accuracy of the translation from the published English original and are not liable for any errors which may occur.
15. **Published Ahead of Print (PAP):** Articles in the PAP stage of publication can be cited using the online publication date and the unique DOI number.
    i. Disclaimer: Articles appearing in the PAP section have been peer-reviewed and accepted for publication in the relevant journal and posted online before print publication. Articles appearing as PAP may contain statements, opinions, and information that have errors in facts, figures, or interpretation. Any final changes in manuscripts will be made at the time of print publication and will be reflected in the final electronic version of the issue. Accordingly, Wolters Kluwer Health, Inc., the editors, authors and their respective employees are not responsible or liable for the use of any such inaccurate or misleading data, opinion or information contained in the articles in this section.
16. **Termination of Contract:** Wolters Kluwer Health, Inc. must be notified within 90 days of the original license date if you opt not to use the requested material.
17. **Waived Permission Fee:** Permission fees that have been waived are not subject to future waivers, including similar requests or renewing a license.
18. **Contingent on payment:** You may exercise these rights licensed immediately upon issuance of the license, however until full payment is received either by the publisher or our authorized vendor, this license is not valid. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of Wolters Kluwer Health, Inc.'s other billing and payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any

use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

19. **STM Signatories Only:** Any permission granted for a particular edition will apply to subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and do not involve the separate exploitation of the permitted illustrations or excerpts. Please view: STM Permissions Guidelines
20. **Warranties and Obligations:** LICENSOR further represents and warrants that, to the best of its knowledge and belief, LICENSEE's contemplated use of the Content as represented to LICENSOR does not infringe any valid rights to any third party.
21. **Breach:** If LICENSEE fails to comply with any provisions of this agreement, LICENSOR may serve written notice of breach of LICENSEE and, unless such breach is fully cured within fifteen (15) days from the receipt of notice by LICENSEE, LICENSOR may thereupon, at its option, serve notice of cancellation on LICENSEE, whereupon this Agreement shall immediately terminate.
22. **Assignment:** License conveyed hereunder by the LICENSOR shall not be assigned or granted in any manner conveyed to any third party by the LICENSEE without the consent in writing to the LICENSOR.
23. **Governing Law:** The laws of The State of New York shall govern interpretation of this Agreement and all rights and liabilities arising hereunder.
24. **Unlawful:** If any provision of this Agreement shall be found unlawful or otherwise legally unenforceable, all other conditions and provisions of this Agreement shall remain in full force and effect.

For Copyright Clearance Center / RightsLink Only:

1. **Service Description for Content Services:** Subject to these terms of use, any terms set forth on the particular order, and payment of the applicable fee, you may make the following uses of the ordered materials:
   i. **Content Rental:** You may access and view a single electronic copy of the materials ordered for the time period designated at the time the order is placed. Access to the materials will be provided through a dedicated content viewer or other portal, and access will be discontinued upon expiration of the designated time period. An order for Content Rental does not include any rights to print, download, save, create additional copies, to distribute or to reuse in any way the full text or parts of the materials.
   ii. **Content Purchase:** You may access and download a single electronic copy of the materials ordered. Copies will be provided by email or by such other means as publisher may make available from time to time. An order for Content Purchase does not include any rights to create additional copies or to distribute copies of the materials

Other Terms and Conditions:

If you are posting your thesis/dissertation online, the website on which you are posting must be a password protected website. Posting of our content to commercial/social media websites, such as ProQuest, YouTube, ResearchGate, Facebook is strictly prohibited.

v1.18

Questions? customercare@copyright.com.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Jul 26, 2023

This Agreement between Mr. Diego Fernando Garcia del Rio ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5596541114065
License date	Jul 26, 2023
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Virchows Archiv
Licensed Content Title	Histological classification of mucinous ovarian tumors: inter-observer reproducibility, clinical relevance, and role of genetic biomarkers
Licensed Content Author	Catherine Genestie et al
Licensed Content Date	Oct 3, 2020
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1

Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	Studying Protein Complexes for Assessing the Function of Ghost Proteins (Ghost in the Cell)
Institution name	University of Lille
Expected presentation date	Oct 2023
Order reference number	2
Portions	Figure 2
Requestor Location	Mr. Diego Fernando Garcia del Rio Avenue Paul Langevin - Bâtiment SN3 - 1e Université de Lille Villeneuve d'Ascq, 59655 France Attn: Mr. Diego Fernando Garcia del Rio
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable), Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature'*.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR

ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

May 31, 2023

This Agreement between Mr. Diego Fernando Garcia del Rio ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5559420391130
License date	May 31, 2023
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Molecular Cell Biology
Licensed Content Title	The mechanism of eukaryotic translation initiation and principles of its regulation
Licensed Content Author	Richard J. Jackson et al
Licensed Content Date	Dec 31, 1969
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Will you be translating?	no

Circulation/distribution	200 - 499
Author of this Springer Nature content	no
Title	Studying Protein Complexes for Assessing the Function of Ghost Proteins (Ghost in the Cell)
Institution name	University of Lille
Expected presentation date	Oct 2023
Portions	Figure 1
Requestor Location	Mr. Diego Fernando Garcia del Rio Avenue Paul Langevin - Bâtiment SN3 - 1e Université de Lille Villeneuve d'Ascq, 59655 France Attn: Mr. Diego Fernando Garcia del Rio
Total	0.00 EUR

Terms and Conditions

Springer Nature Customer Service Centre GmbH Terms and Conditions

The following terms and conditions ("Terms and Conditions") together with the terms specified in your [RightsLink] constitute the License ("License") between you as Licensee and Springer Nature Customer Service Centre GmbH as Licensor. By clicking 'accept' and completing the transaction for your use of the material ("Licensed Material"), you confirm your acceptance of and obligation to be bound by these Terms and Conditions.

1. Grant and Scope of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, non-sublicensable, revocable, world-wide License to reproduce, distribute, communicate to the public, make available, broadcast, electronically transmit or create derivative works using the Licensed Material for the purpose(s) specified in your RightsLink Licence Details only. Licenses are granted for the specific use requested in the order and for no other use, subject to these Terms and Conditions. You acknowledge and agree that the rights granted to you under this License do not include the right to modify, edit, translate, include in collective works, or create derivative works of the

Licensed Material in whole or in part unless expressly stated in your RightsLink Licence Details. You may use the Licensed Material only as permitted under this Agreement and will not reproduce, distribute, display, perform, or otherwise use or exploit any Licensed Material in any way, in whole or in part, except as expressly permitted by this License.

1. 2. You may only use the Licensed Content in the manner and to the extent permitted by these Terms and Conditions, by your RightsLink Licence Details and by any applicable laws.

1. 3. A separate license may be required for any additional use of the Licensed Material, e.g. where a license has been purchased for print use only, separate permission must be obtained for electronic re-use. Similarly, a License is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the License.

1. 4. Any content within the Licensed Material that is owned by third parties is expressly excluded from the License.

1. 5. Rights for additional reuses such as custom editions, computer/mobile applications, film or TV reuses and/or any other derivative rights requests require additional permission and may be subject to an additional fee. Please apply to journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

2. Reservation of Rights

Licensor reserves all rights not expressly granted to you under this License. You acknowledge and agree that nothing in this License limits or restricts Licensor's rights in or use of the Licensed Material in any way. Neither this License, nor any act, omission, or statement by Licensor or you, conveys any ownership right to you in any Licensed Material, or to any element or portion thereof. As between Licensor and you, Licensor owns and retains all right, title, and interest in and to the Licensed Material subject to the license granted in Section 1.1. Your permission to use the Licensed Material is expressly conditioned on you not impairing Licensor's or the applicable copyright owner's rights in the Licensed Material in any way.

3. Restrictions on use

3. 1. Minor editing privileges are allowed for adaptations for stylistic purposes or formatting purposes provided such alterations do not alter the original meaning or intention of the Licensed Material and the new figure(s) are still accurate and representative of the Licensed Material. Any other changes including but not limited to, cropping, adapting, and/or omitting material that affect the meaning, intention or moral rights of the author(s) are strictly prohibited.

3. 2. You must not use any Licensed Material as part of any design or trademark.

3. 3. Licensed Material may be used in Open Access Publications (OAP), but any such reuse must include a clear acknowledgment of this permission visible at the same time as the figures/tables/illustration or abstract and which must indicate that the Licensed Material is not part of the governing OA license but has been reproduced with permission. This may be indicated according to any standard referencing system but must include at a minimum 'Book/Journal title, Author, Journal Name (if applicable),

Volume (if applicable), Publisher, Year, reproduced with permission from SNCSC'.

4. STM Permission Guidelines

4. 1. An alternative scope of license may apply to signatories of the STM Permissions Guidelines ("STM PG") as amended from time to time and made available at https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/.

4. 2. For content reuse requests that qualify for permission under the STM PG, and which may be updated from time to time, the STM PG supersede the terms and conditions contained in this License.

4. 3. If a License has been granted under the STM PG, but the STM PG no longer apply at the time of publication, further permission must be sought from the Rightsholder. Contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

5. Duration of License

5. 1. Unless otherwise indicated on your License, a License is valid from the date of purchase ("License Date") until the end of the relevant period in the below table:

Reuse in a medical communications project	Reuse up to distribution or time period indicated in License
Reuse in a dissertation/thesis	Lifetime of thesis
Reuse in a journal/magazine	Lifetime of journal/magazine
Reuse in a book/textbook	Lifetime of edition
Reuse on a website	1 year unless otherwise specified in the License
Reuse in a presentation/slide kit/poster	Lifetime of presentation/slide kit/poster. Note: publication whether electronic or in print of presentation/slide kit/poster may require further permission.
Reuse in conference proceedings	Lifetime of conference proceedings
Reuse in an annual report	Lifetime of annual report
Reuse in training/CME materials	Reuse up to distribution or time period indicated in License
Reuse in newsmedia	Lifetime of newsmedia
Reuse in coursepack/classroom materials	Reuse up to distribution and/or time period indicated in license

6. Acknowledgement

6. 1. The Licensor's permission must be acknowledged next to the Licensed Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract and must be hyperlinked to the journal/book's homepage.

6. 2. Acknowledgement may be provided according to any standard referencing system and at a minimum should include "Author, Article/Book Title, Journal name/Book imprint, volume, page number, year, Springer Nature".

7. Reuse in a dissertation or thesis

7. 1. Where 'reuse in a dissertation/thesis' has been selected, the following terms apply: Print rights of the Version of Record are provided for; electronic rights for use only on institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/) and only up to what is required by the awarding institution.

7. 2. For theses published under an ISBN or ISSN, separate permission is required. Please contact journalpermissions@springernature.com or bookpermissions@springernature.com for these rights.

7. 3. Authors must properly cite the published manuscript in their thesis according to current citation standards and include the following acknowledgement: '*Reproduced with permission from Springer Nature'*.

8. License Fee

You must pay the fee set forth in the License Agreement (the "License Fees"). All amounts payable by you under this License are exclusive of any sales, use, withholding, value added or similar taxes, government fees or levies or other assessments. Collection and/or remittance of such taxes to the relevant tax authority shall be the responsibility of the party who has the legal obligation to do so.

9. Warranty

9. 1. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. **You are solely responsible for ensuring that the material you wish to license is original to the Licensor and does not carry the copyright of another entity or third party (as credited in the published version).** If the credit line on any part of the Licensed Material indicates that it was reprinted or adapted with permission from another source, then you should seek additional permission from that source to reuse the material.

9. 2. EXCEPT FOR THE EXPRESS WARRANTY STATED HEREIN AND TO THE EXTENT PERMITTED BY APPLICABLE LAW, LICENSOR PROVIDES THE LICENSED MATERIAL "AS IS" AND MAKES NO OTHER REPRESENTATION OR WARRANTY. LICENSOR EXPRESSLY DISCLAIMS ANY LIABILITY FOR ANY CLAIM ARISING FROM OR OUT OF THE CONTENT, INCLUDING BUT NOT LIMITED TO ANY ERRORS, INACCURACIES, OMISSIONS, OR DEFECTS CONTAINED THEREIN, AND ANY IMPLIED OR EXPRESS WARRANTY AS TO MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, PUNITIVE, OR EXEMPLARY DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE LICENSED MATERIAL REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION,

DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION APPLIES NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

10. Termination and Cancellation

10. 1. The License and all rights granted hereunder will continue until the end of the applicable period shown in Clause 5.1 above. Thereafter, this license will be terminated and all rights granted hereunder will cease.

10. 2. Licensor reserves the right to terminate the License in the event that payment is not received in full or if you breach the terms of this License.

11. General

11. 1. The License and the rights and obligations of the parties hereto shall be construed, interpreted and determined in accordance with the laws of the Federal Republic of Germany without reference to the stipulations of the CISG (United Nations Convention on Contracts for the International Sale of Goods) or to Germany´s choice-of-law principle.

11. 2. The parties acknowledge and agree that any controversies and disputes arising out of this License shall be decided exclusively by the courts of or having jurisdiction for Heidelberg, Germany, as far as legally permissible.

11. 3. This License is solely for Licensor's and Licensee's benefit. It is not for the benefit of any other person or entity.

Questions? For questions on Copyright Clearance Center accounts or website issues please contact springernaturesupport@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777. For questions on Springer Nature licensing please visit https://www.springernature.com/gp/partners/rights-permissions-third-party-distribution

Other Conditions:

Version 1.4 - Dec 2022

Questions? customercare@copyright.com.

RightsLink

? Help ⌄ Live Chat

Development of a Novel Cross-linking Strategy for Fast and Accurate Identification of Cross-linked Peptides of Protein Complexes*

Author:
Athit Kao,Chi-li Chiu,Danielle Vellucci,Yingying Yang,Vishal R. Patel,Shenheng Guan,Arlo Randall,Pierre Baldi,Scott D. Rychnovsky,Lan Huang

Publication: Molecular & Cellular Proteomics

Publisher: Elsevier

Date: January 2011

Creative Commons

This is an open access article distributed under the terms of the Creative Commons CC-BY license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

You are not required to obtain permission to reuse this article.

To request permission for a type of use not listed, please contact Elsevier Global Rights Department.

Are you the author of this Elsevier journal article?

Summary

Ovarian cancer (OvCa) has the highest mortality rate among female reproductive cancers worldwide. OvCa is often diagnosed late or misdiagnosed, and chemotherapy resistance poses a significant challenge. To overcome this, new targets and therapeutic strategies are urgently needed. The ghost proteome, consisting of proteins translated from alternative open reading frames (AltORFs), is a potential source of biomarkers. However, studying AltProts is complex and limited. "Guilty by association" approaches, such as assessing AltProts' protein-protein interactions (PPIs) with reference proteins (RefProts), can help identify their function. Crosslinking mass spectrometry (XL-MS) and bioinformatic tools are useful for this purpose. A methodology combining XL-MS and subcellular fractionation was developed, reducing sample complexity. Molecular modeling and network analysis provided insights into AltProts' roles. Proteogenomic analysis of ovarian cancer cell lines revealed differential expression of proteins, including AltProts, and their association with cancer-related pathways. This work uncovers new aspects of ovarian cancer biology, identifying previously unknown proteins and variants with functional significance from the "ghost proteome."

Resumé

Le cancer de l'ovaire (OvCa) est le cancer féminin le plus mortel, souvent diagnostiqué tardivement et résistant à la chimiothérapie. Pour surmonter ces défis, de nouvelles cibles et stratégies thérapeutiques sont nécessaires. Le protéome fantôme, composé de protéines traduites à partir de cadres de lecture ouverts alternatifs (AltORFs), est une source potentielle de biomarqueurs. Les études sur les protéines alternatives (AltProts) sont complexes mais peuvent être évaluées en identifiant leurs interactions protéine-protéine (PPI) avec des protéines de référence (RefProts). L'utilisation de la spectrométrie de masse en liaison chimique (XL-MS) et d'outils bioinformatiques permet d'analyser les AltProts. Une méthode combinant XL-MS et fractionnement subcellulaire a été développée pour réduire la complexité des échantillons. L'analyse a révélé des rôles des AltProts dans la réparation de l'ADN et la présentation d'antigènes. La protéogénomique a été utilisée pour étudier les protéomes de lignées cellulaires de cancer de l'ovaire, révélant des protéines différentiellement exprimées et associées à des voies de signalisation du cancer. Ce travail met en évidence le potentiel de l'approche protéogénomique pour comprendre le cancer de l'ovaire, en identifiant des protéines et des variants jusqu'alors inconnus du "protéome fantôme".