Université de Lille UFR3S-Phamacie

Année Universitaire 2025/2026.

THESE

POUR LE DIPLOME D'ETAT

DE DOCTEUR EN PHARMACIE

Soutenue publiquement le 30/01/2026

Par Mr Ramdani Kamil

Étude des survivants exceptionnels du cancer : comment l'analyse approfondie de leurs données cliniques, génomiques et biologiques peut-elle orienter la découverte de nouveaux traitements anticancéreux ?

Membres du jury :

Président : Cazin, Jean Louis, PU pharmacologie / chef d'unité au centre Oscar Lambret

Directeur, conseiller de thèse : Pinçon, Claire, MCU Statistique à l'Université de Lille

Assesseur : BENHALIMA, Ilyès, Assistant Hospitalo-Universitaire CHU lyon

Université de Lille

Président	Régis BORDET
Premier Vice-président	Bertrand DÉCAUDIN
Vice-présidente Formation	Corinne ROBACZEWSKI
Vice-président Recherche	Olivier COLOT
Vice-président Ressources Humaine	Jean-Philippe TRICOIT
Directrice Générale des Services	Anne-Valérie CHIRIS-FABRE

UFR3S

Doyen	Dominique LACROIX
Premier Vice-Doyen, Vice-Doyen RH, SI et Qualité	Hervé HUBERT
Vice-Doyenne Recherche	Karine FAURE
Vice-Doyen Finances et Patrimoine	Emmanuelle LIPKA
Vice-Doyen International	Vincent DERAMECOURT
Vice-Doyen Coordination pluriprofessionnelle et Formations sanitaires	Sébastien D'HARANCY
Vice-Doyenne Formation tout au long de la vie	Caroline LANIER
Vice-Doyen Territoire-Partenariats	Thomas MORGENROTH
Vice-Doyen Santé numérique et Communication	Vincent SOBANSKI
Vice-Doyenne Vie de Campus	Anne-Laure BARBOTIN
Vice-Doyen étudiant	Victor HELENA

Faculté de Pharmacie

Vice - Doyen	Pascal ODOU
Premier Assesseur et Assesseur à la Santé et à l'Accompagnement	Anne GARAT
Assesseur à la Vie de la Faculté et Assesseur aux Ressources et Personnels	Emmanuelle LIPKA
Responsable de l'Administration et du Pilotage	Cyrille PORTA
Représentant étudiant	Honoré GUISE
Chargé de mission 1er cycle	Philippe GERVOIS
Chargée de mission 2eme cycle	Héloïse HENRY
Chargé de mission Accompagnement et Formation à la Recherche	Nicolas WILLAND
Chargé de mission Relations Internationales	Christophe FURMAN
Chargée de Mission Qualité	Marie-Françoise ODOU
Chargé de mission dossier HCERES	Réjane LESTRELIN

Professeurs des Universités - Praticiens Hospitaliers (PU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	ALLORGE	Delphine	Toxicologie et Santé publique	81
M.	BROUSSEAU	Thierry	Biochimie	82
M.	DÉCAUDIN	Bertrand	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	DINE	Thierry	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
Mme	DUPONT-PRADO	Annabelle	Hématologie	82
Mme	GOFFARD	Anne	Bactériologie - Virologie	82
M.	GRESSIER	Bernard	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	ODOU	Pascal	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	POULAIN	Stéphanie	Hématologie	82
M.	SIMON	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	STAELS	Bart	Biologie cellulaire	82

Professeurs des Universités (PU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	ALIOUAT	El Moukhtar	Parasitologie - Biologie animale	87
Mme	ALIOUAT	Cécile-Marie	Parasitologie - Biologie animale	87
Mme	AZAROUAL	Nathalie	Biophysique - RMN	85
M.	BERLARBI	Karim	Physiologie	86
M.	BERTIN	Benjamin	Immunologie	87
M.	BLANCHEMAIN	Nicolas	Pharmacotechnie industrielle	85
M.	CARNOY	Christophe	Immunologie	87
M.	CAZIN	Jean-Louis	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	CUNY	Damien	Sciences végétales et fongiques	87

Mme	DELBAERE	Stéphanie	Biophysique - RMN	85
Mme	DEPREZ	Rebecca	Chimie thérapeutique	86
M.	DEPREZ	Benoît	Chimie bio inorganique	85
Mme	DUMONT	Julie	Biologie cellulaire	87
M.	ELATI	Mohamed	Biomathématiques	27
M.	FOLIGNÉ	Benoît	Bactériologie - Virologie	87
Mme	FOULON	Catherine	Chimie analytique	85
M.	GARÇON	Guillaume	Toxicologie et Santé publique	86
M.	GOOSSENS	Jean-François	Chimie analytique	85
M.	HENNEBELLE	Thierry	Pharmacognosie	86
M.	LEBEGUE	Nicolas	Chimie thérapeutique	86
M.	LEMDANI	Mohamed	Biomathématiques	26
Mme	LESTAVEL	Sophie	Biologie cellulaire	87
Mme	LESTRELIN	Réjane	Biologie cellulaire	87
Mme	LIPKA	Emmanuelle	Chimie analytique	85
Mme	MELNYK	Patricia	Chimie physique	85
M.	MILLET	Régis	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	MOREAU	Pierre-Arthur	Sciences végétales et fongiques	87
Mme	MUHR-TAILLEUX	Anne	Biochimie	87
Mme	PERROY	Anne-Catherine	Droit et Economie pharmaceutique	86
Mme	RIVIÈRE	Céline	Pharmacognosie	86
Mme	ROMOND	Marie-Bénédicte	Bactériologie - Virologie	87
Mme	SAHPAZ	Sevser	Pharmacognosie	86
M.	SERGHERAERT	Éric	Droit et Economie pharmaceutique	86
M.	SIEPMANN	Juergen	Pharmacotechnie industrielle	85
Mme	SIEPMANN	Florence	Pharmacotechnie industrielle	85
M.	WILLAND	Nicolas	Chimie organique	86

Maîtres de Conférences - Praticiens Hospitaliers (MCU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	CUVELIER	Élodie	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
Mme	DANEL	Cécile	Chimie analytique	85
Mme	DEMARET	Julie	Immunologie	82
Mme	GARAT	Anne	Toxicologie et Santé publique	81
Mme	GENAY	Stéphanie	Biopharmacie, Pharmacie galénique et hospitalière	81
Mme	GILLIOT	Sixtine	Biopharmacie, Pharmacie galénique et hospitalière	80
M.	GRZYCH	Guillaume	Biochimie	82
Mme	HENRY	Héloïse	Biopharmacie, Pharmacie galénique et hospitalière	80
M.	LANNOY	Damien	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	MASSE	Morgane	Biopharmacie, Pharmacie galénique et hospitalière	81
Mme	ODOU	Marie-Françoise	Bactériologie - Virologie	82

Maîtres de Conférences des Universités (MCU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	ANTHÉRIEU	Sébastien	Toxicologie et Santé publique	86
M.	BANTUBUNGI-BLUM	Kadiombo	Biologie cellulaire	87
M.	BERTHET	Jérôme	Biophysique - RMN	85
M	BEDART	Corentin	ICPAL	86
M.	BOCHU	Christophe	Biophysique - RMN	85
M.	BORDAGE	Simon	Pharmacognosie	86

M.	BOSC	Damien	Chimie thérapeutique	86
Mme	BOU KARROUM	Nour	Chimie bioinorganique	
M.	BRIAND	Olivier	Biochimie	87
Mme	CARON-HOUDE	Sandrine	Biologie cellulaire	87
Mme	CARRIÉ	Hélène	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
Mme	CHABÉ	Magali	Parasitologie - Biologie animale	87
Mme	CHARTON	Julie	Chimie organique	86
M.	CHEVALIER	Dany	Toxicologie et Santé publique	86
Mme	DEMANCHE	Christine	Parasitologie - Biologie animale	87
Mme	DEMARQUILLY	Catherine	Biomathématiques	85
M.	DHIFLI	Wajdi	Biomathématiques	27
M.	EL BAKALI	Jamal	Chimie thérapeutique	86
M.	FARCE	Amaury	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	FLIPO	Marion	Chimie organique	86
M.	FRULEUX	Alexandre	Sciences végétales et fongiques	
M.	FURMAN	Christophe	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	GERVOIS	Philippe	Biochimie	87
Mme	GOOSSENS	Laurence	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	GRAVE	Béatrice	Toxicologie et Santé publique	86
M.	HAMONIER	Julien	Biomathématiques	26
Mme	HAMOUDI-BEN YELLES	Chérifa-Mounira	Pharmacotechnie industrielle	85
Mme	HANNOTHIAUX	Marie-Hélène	Toxicologie et Santé publique	86
Mme	HELLEBOID	Audrey	Physiologie	86
M.	HERMANN	Emmanuel	Immunologie	87
M.	KAMBIA KPAKPAGA	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	KARROUT	Younes	Pharmacotechnie industrielle	85
Mme	LALLOYER	Fanny	Biochimie	87

Mme	LECOEUR	Marie	Chimie analytique	85
Mme	LEHMANN	Hélène	Droit et Economie pharmaceutique	86
Mme	LELEU	Natascha	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	LIBERELLE	Maxime	Biophysique - RMN	
Mme	LOINGEVILLE	Florence	Biomathématiques	26
Mme	MARTIN	Françoise	Physiologie	86
M.	MARTIN MENA	Anthony	Biopharmacie, Pharmacie galénique et hospitalière	
M.	MENETREY	Quentin	Bactériologie - Virologie	87
M.	MORGENROTH	Thomas	Droit et Economie pharmaceutique	86
Mme	MUSCHERT	Susanne	Pharmacotechnie industrielle	85
Mme	NIKASINOVIC	Lydia	Toxicologie et Santé publique	86
Mme	PINÇON	Claire	Biomathématiques	85
M.	PIVA	Frank	Biochimie	85
Mme	PLATEL	Anne	Toxicologie et Santé publique	86
M.	POURCET	Benoît	Biochimie	87
M.	RAVAUX	Pierre	Biomathématiques / Innovations pédagogiques	85
Mme	RAVEZ	Séverine	Chimie thérapeutique	86
Mme	ROGEL	Anne	Immunologie	
M.	ROSA	Mickaël	Hématologie	87
M.	ROUMY	Vincent	Pharmacognosie	86
Mme	SEBTI	Yasmine	Biochimie	87
Mme	SINGER	Elisabeth	Bactériologie - Virologie	87
Mme	STANDAERT	Annie	Parasitologie - Biologie animale	87
M.	TAGZIRT	Madjid	Hématologie	87
M.	VILLEMAGNE	Baptiste	Chimie organique	86
M.	WELTI	Stéphane	Sciences végétales et fongiques	87
M.	YOUS	Saïd	Chimie thérapeutique	86

M.	ZITOUNI	Djamel	Biomathématiques	85

Professeurs certifiés

Civ.	Nom	Prénom	Service d'enseignement
Mme	FAUQUANT	Soline	Anglais
M.	HUGES	Dominique	Anglais
Mme	KUBIK	Laurence	Anglais
M.	OSTYN	Gaël	Anglais

Professeurs Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	BAILLY	Christian	ICPAL	86
M.	DAO PHAN	Haï Pascal	Chimie thérapeutique	86
M.	DHANANI	Alban	Droit et Economie pharmaceutique	86

Maîtres de Conférences Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M	AYED	Elya	Pharmacie officinale	
M.	COUSEIN	Etienne	Biopharmacie, Pharmacie galénique et hospitalière	
Mme	CUCCHI	Malgorzata	Biomathématiques	85
Mme	DANICOURT	Frédérique	Pharmacie officinale	
Mme	DUPIRE	Fanny	Pharmacie officinale	
M.	DUFOSSEZ	François	Biomathématiques	85

M.	FRIMAT	Bruno	Pharmacologie, Pharmacocinétique et Pharmacie clinique	85
Mme	GEILER	Isabelle	Pharmacie officinale	
M.	GILLOT	François	Droit et Economie pharmaceutique	86
M.	MITOUMBA	Fabrice	Biopharmacie, Pharmacie galénique et hospitalière	86
M.	PELLETIER	Franck	Droit et Economie pharmaceutique	86
M	POTHIER	Jean-Claude	Pharmacie officinale	
Mme	ROGNON	Carole	Pharmacie officinale	

Assistants Hospitalo-Universitaire (AHU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	BOUDRY	Augustin	Biomathématiques	
Mme	DERAMOUDT	Laure	Pharmacologie, Pharmacocinétique et Pharmacie clinique	
M.	GISH	Alexandr	Toxicologie et Santé publique	
Mme	NEGRIER	Laura	Chimie analytique	

Hospitalo-Universitaire (PHU)

	Nom	Prénom	Service d'enseignement	Section CNU
M.	DESVAGES	Maximilien	Hématologie	
Mme	LENSKI	Marie	Toxicologie et Santé publique	

Attachés Temporaires d'Enseignement et de Recherche (ATER)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	BERNARD	Lucie	Physiologie	
Mme	BARBIER	Emeline	Toxicologie	
Mme	COMPAGNE	Nina	Chimie Organique	
Mme	COULON	Audrey	Pharmacologie, Pharmacocinétique et Pharmacie clinique	
M.	DUFOSSEZ	Robin	Chimie physique	
Mme	FERRY	Lise	Biochimie	
M	HASYEOUI	Mohamed	Chimie Organique	
Mme	HENRY	Doriane	Biochimie	
Mme	KOUAGOU	Yolène	Sciences végétales et fongiques	
M	LAURENT	Arthur	Chimie-Physique	
M.	MACKIN MOHAMOUR	Synthia	Biopharmacie, Pharmacie galénique et hospitalière	
Mme	RAAB	Sadia	Physiologie	

Enseignant contractuel

Civ.	Nom	Prénom	Service d'enseignement
Mme	DELOBEAU	Iris	Pharmacie officinale
M	RIVART	Simon	Pharmacie officinale
Mme	SERGEANT	Sophie	Pharmacie officinale
M.	ZANETTI	Sébastien	Biomathématiques

LRU / MAST

Civ.	Nom	Prénom	Service d'enseignement
Mme	FRAPPE	Jade	Pharmacie officinale
M	LATRON-FREMEAU	Pierre-Manuel	Pharmacie officinale
M.	MASCAUT	Daniel	Pharmacologie, Pharmacocinétique et Pharmacie clinique

Remerciements :

Monsieur le Professeur André Tartar

Professeur émérite des Universités

Faculté de Pharmacie de Lille

Vous avez été le professeur qui a eu le plus grand impact sur ma perception de la pharmacie, et plus particulièrement de l'industrie pharmaceutique et de ses débouchés, ce qui m'a donné l'envie de poursuivre dans cette voie.

Madame le Docteur Mounira Hamoudi

Faculté de Pharmacie de Lille

Je vous remercie pour l'enseignement que vous nous avez dispensé à la faculté de pharmacie, mais aussi pour votre accompagnement à la fois personnel et professionnel, notamment lors de mon aventure entrepreneuriale, au cours de laquelle vous avez été d'un grand soutien.

Madame le Docteur Claire Pinçon

Faculté de Pharmacie de Lille

Je vous remercie pour l'enseignement que vous nous avez dispensé à la Faculté de Pharmacie, mais aussi pour avoir accepté de m'accompagner dans la direction de ma thèse, de me conseiller sur l'écriture de celle-ci et sur la manière de la composer. Je vous remercie également pour votre patience.

Monsieur le Professeur jean louis cazin

PU-PH

Centre oscar lambre

Je vous remercie très sincèrement d'avoir accepté d'évaluer mon travail. Connaissant votre expertise dans le domaine de l'oncologie, votre présence au sein de mon jury représente pour moi un grand honneur.

À ma famille :

À mes parents, je ne vous remercierai jamais assez pour tout l'amour et l'accompagnement que vous m'avez apportés. Lorsque aucun de mes enseignants ne croyait en moi, vous avez toujours été d'un soutien sans nul égal.

Mon père, je te remercie pour la confiance que tu me portes, pour avoir toujours cru en moi et pour m'avoir permis de vivre des aventures qui m'ont appris la vie.

Ma mère, mon pilier, celle qui me soutient quoi que je fasse, qui me supporte sans jamais hausser la voix, mais qui sait toujours trouver les mots justes pour me conseiller. Tu es mon modèle et mon plus grand repère. Les mots ne sont pas assez grands pour exprimer mes remerciements les plus profonds.

mon grand frère Redouane, l'aîné, celui qui nous a ouvert la voie. Il est pour moi une véritable source d'inspiration, à la fois en tant que personne et dans sa vie d'entrepreneur. Sa dextérité et sa persévérance sont comme une boussole qui me permet d'avancer avec confiance vers l'avenir. Ayant suivi l'une des écoles de commerce les plus prestigieuses au monde, il a su me guider et me conseiller tout au long de mon parcours. Il est aussi une boussole personnelle. Nous avons beaucoup en commun et, même si tu habites loin, nous communiquons chaque jour.

Mon grand frère Mehdi, mon cadet, mon colocataire durant mes études de pharmacie. Pharmacien comme moi, il a suivi le parcours de l'internat et de la biologie. Ses conseils et son aide ont été pour moi inestimables. Je n'oublierai jamais l'accompagnement qu'il m'a offert durant mes deux années de PACES : il me préparait des plats chaque jour et me les apportait à la bibliothèque, me conseillait et m'a ouvert la voie de la pharmacie. Il a toujours été présent, tant pour des conseils professionnels que personnels. J'espère vivre encore de nombreuses aventures et voyages avec toi, Mehdi.

À mes grands-parents, Hadoum et Ahmed, ainsi qu'à ma grand-mère Nana, qui veille sur moi d'en haut et regarde cette thèse. Je vous remercie pour tout l'amour que vous m'avez donné. Vous avez toujours été un repère de générosité, de bonté et de sagesse. Vous êtes des exemples pour moi.

À mes tantes et oncles, et plus particulièrement à mes tantes Mama et Hafida, qui nous ont malheureusement quittés ces dernières années, emportées par le cancer, je dédie cette thèse. Vous avez toujours été une source de joie et de réconfort durant mon enfance, et je vous en remercie du fond du cœur. Le temps passe, mais les souvenirs partagés restent gravés à jamais. J'aurais tant aimé que vous soyez présentes aujourd'hui, mais je suis certain que vous l'êtes de là-haut en ce jour.

À tous mes cousins et cousines, merci pour votre soutien et pour tout l'amour que vous m'avez toujours témoigné. Votre présence et vos encouragements ont compté énormément pour moi tout au long de ce parcours.

À mes amis, Clément, Alexandre, Juliette, Hugo et Adam, je vous remercie pour tous les rires et les moments passés, présents et à venir. Vous êtes bien plus que de simples amis : vous êtes une partie intégrante de ma vie et ma deuxième famille.

A mes collègue de travail :

Je remercie Nicolas Wolikow, CEO de Cure51, la biotech qui mène actuellement des recherches sur les survivants exceptionnels. Depuis mon stage de cinquième année, tu as été pour moi une source précieuse de connaissances et une véritable inspiration, tant par ton parcours professionnel que par ta personnalité.

À Simon Istolainen, co-fondateur de Cure51, je te remercie pour ton aide précieuse dans la structuration des thématiques abordées au cours de cette thèse, ainsi que pour ta bienveillance et ta bonne humeur constantes.

Je remercie également Anne et Omar pour les moments partagés au travail, ainsi que pour votre encadrement et votre accompagnement au sein de Cure51.

Je tiens aussi à remercier l'ensemble des chercheurs et cliniciens qui ont accepté de m'accorder de leur temps et de partager leur expertise : le Dr Naouel Zerrouk, le Dr Rémy Nicolle, le chercheur Samuel Blanck, le Dr Paloma Cejas, ainsi que le Dr Yaovi Eric Amela. Je vous remercie sincèrement pour votre disponibilité, pour la qualité de vos échanges, et pour les éclairages scientifiques et méthodologiques que vous m'avez apportés tout au long de l'analyse de cette thèse.

Table des matières

Liste des abréviations

ADN : Acide désoxyribonucléique
ALK : Anaplastic Lymphoma Kinase
BH : Benjamini–Hochberg
CAF : Cancer-Associated Fibroblasts
CCLE : Cancer Cell Line Encyclopedia
CNV : Copy Number Variation
CR : Complete Response
CRISPR : Clustered Regularly Interspaced Short Palindromic Repeats
CXCL9 / CXCL10 : C-X-C motif chemokine ligand 9 / 10
DDR : DNA Damage Response
DESeq2 : Differential Expression analysis for Sequence count data
EGFR : Epidermal Growth Factor Receptor
FDR : False Discovery Rate
FFPE : Formalin-Fixed Paraffin-Embedded
GDSC : Genomics of Drug Sensitivity in Cancer
GSEA : Gene Set Enrichment Analysis
GO : Gene Ontology
HRD : Homologous Recombination Deficiency
ICGC : International Cancer Genome Consortium
IFNG : Interferon Gamma
KEGG : Kyoto Encyclopedia of Genes and Genomes
LASSO : Least Absolute Shrinkage and Selection Operator
MA plot : Mean–Average plot
MOFA : Multi-Omics Factor Analysis
MSI-H : Microsatellite Instability – High
MSigDB : Molecular Signatures Database
NCI : National Cancer Institute
NTRK : Neurotrophic Tyrosine Receptor Kinase
OS : Overall Survival
PCA : Principal Component Analysis
PD-1 : Programmed cell death protein 1
PDX : Patient-Derived Xenograft
PR : Partial Response
RNA-seq : RNA sequencing
scRNA-seq : single-cell RNA sequencing
siRNA : small interfering RNA
sPLS : sparse Partial Least Squares
TCGA : The Cancer Genome Atlas
TLR : Toll-Like Receptor

TRIPOD : Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis
TRIPOD-AI : Extension of TRIPOD for Artificial Intelligence
UMAP : Uniform Manifold Approximation and Projection
WES : Whole-Exome Sequencing
WGS : Whole-Genome Sequencing
ANLN : Anillin
BRCA1/2 : Breast Cancer gene 1 / 2
ComBat : Empirical Bayes method for batch effect correction
CRISPR-Cas9 : Clustered Regularly Interspaced Short Palindromic Repeats – CRISPR associated protein 9
CTHRC1 : Collagen Triple Helix Repeat Containing 1
cGAS : Cyclic GMP–AMP Synthase
EGFR-TKI : Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitor
HR : Homologous Recombination
HR-proficient : Homologous Recombination Proficient
ICB : Immune Checkpoint Blockade
IFN : Interferon
JAK-STAT : Janus Kinase – Signal Transducer and Activator of Transcription
MKI67 : Marker of Proliferation Ki-67
NF-κB : Nuclear Factor kappa-light-chain-enhancer of activated B cells
NT5E : 5'-Nucleotidase Ecto (CD73)
PARP : Poly(ADP-ribose) Polymerase
PD-L1 : Programmed Death-Ligand 1
RNA-seq (bulk) : Bulk RNA sequencing
SASP : Senescence-Associated Secretory Phenotype
STING : Stimulator of Interferon Genes
TCGA-STAD : The Cancer Genome Atlas – Stomach Adenocarcinoma
TNF-α : Tumor Necrosis Factor alpha
UMI : Unique Molecular Identifier
voom : Variance modelling at the observational level (limma extension for RNA-seq)

Introduction

La notion de survivant exceptionnel désigne des patients présentant une survie ou une réponse thérapeutique très largement supérieure à ce qui est attendu au regard du pronostic habituel de leur pathologie. Ces trajectoires cliniques atypiques, bien que rares, interrogent directement les mécanismes biologiques de la maladie et de la réponse aux traitements. Le concept de survivant exceptionnel, ses définitions opérationnelles et les implications méthodologiques associées seront développés plus en détail dans la Partie 1.

Les survivants exceptionnels constituent une population extrêmement minoritaire, représentant le plus souvent moins de 5 % des patients. Cette rareté explique en grande partie la faible abondance de travaux dédiés dans la littérature. Leur étude se heurte à plusieurs limitations majeures, parmi lesquelles la disponibilité limitée de tissu tumoral exploitable pour des analyses moléculaires approfondies, l'absence de définitions harmonisées du concept de survie ou de réponse exceptionnelle, ainsi que la taille réduite des cohortes pouvant être constituées. À cela s'ajoute le manque de programmes structurés et systématiques permettant d'identifier ces patients à grande échelle, ce qui freine la reproductibilité des observations et la généralisation des résultats.

L'intérêt scientifique d'étudier ces patients exceptionnels repose sur l'hypothèse de pouvoir identifier des biomarqueurs spécifiques et prédictifs, ou encore des signatures moléculaires, susceptibles d'expliquer leur survie prolongée. De nombreuses recherches, dont l'**Exceptional Responders Initiative** menée par le *National Cancer Institute (NCI)(1)*, ont montré que l'analyse approfondie du génome de ces patients pouvait mettre en évidence des altérations rares et distinctives. Ces dernières, bien que peu fréquentes dans la population générale, pourraient être directement associées à une sensibilité particulière à certaines thérapies, ouvrant ainsi la voie à des approches thérapeutiques plus personnalisées.Sur le plan clinique, l'étude des survivants exceptionnels présente un double intérêt. D'une part, elle permet de mieux comprendre les mécanismes de résistance ou, au contraire, de sensibilité marquée aux traitements, afin d'affiner la stratification des patients et d'anticiper leur réponse thérapeutique. D'autre part, elle ouvre la possibilité de repositionner certains médicaments ou de concevoir de nouvelles stratégies combinatoires, en s'appuyant sur les profils génétiques ou immunologiques identifiés chez ces patients. Enfin, les connaissances issues de ces analyses peuvent contribuer à la conception d'essais cliniques plus ciblés, où les patients sont sélectionnés sur la base de biomarqueurs prédictifs, augmentant ainsi les chances de succès et réduisant l'exposition inutile à des traitements inefficaces.

Cette thèse ne vise pas à identifier de nouveaux biomarqueurs ou à établir une causalité biologique définitive entre certaines altérations moléculaires et la survie prolongée.

L'objectif principal est d'explorer la méthodologie et la valeur scientifique de l'étude des *survivants exceptionnels du cancer*, en particulier dans le cadre de petites cohortes.

Le travail s'inscrit donc dans une perspective méthodologique et translationnelle, cherchant à comprendre :

- comment les approches *multi-omiques* (génomique, transcriptomique, immunologique, etc.) peuvent être mobilisées pour caractériser ces patients,
- dans quelle mesure les données issues d'un faible effectif peuvent être exploitées de manière robuste,
- et enfin, comment les signatures issues de ces analyses pourraient, à terme, être intégrées dans la pratique clinique pour orienter la décision thérapeutique.

Cette réflexion se fonde à la fois sur une analyse bibliographique, sur des entretiens menés auprès de chercheurs, bio-informaticiens et oncologues, ainsi que sur une immersion au sein d'une entreprise de biotechnologie spécialisée dans l'étude des survivants exceptionnels (Cure51).

Ainsi, le périmètre du travail exclut volontairement :

- la validation expérimentale de biomarqueurs spécifiques,
- la conduite ou l'analyse statistique d'un essai clinique,
- et toute évaluation médico-économique des approches proposées.

Le propos de cette thèse est donc d'établir un cadre conceptuel et méthodologique permettant de mieux comprendre comment les données rares et multi-omiques issues de patients exceptionnels peuvent contribuer, à terme, à enrichir la médecine de précision en oncologie.

Afin de mieux comprendre les objectifs et les différentes parties de cette thèse, un schéma a été élaboré afin de visualiser clairement le rôle et la finalité de chaque section. Cette thèse est structurée en cinq parties complémentaires, qui suivent la progression logique d'un programme de recherche visant à comprendre les mécanismes biologiques associés aux survivants exceptionnels du cancer.

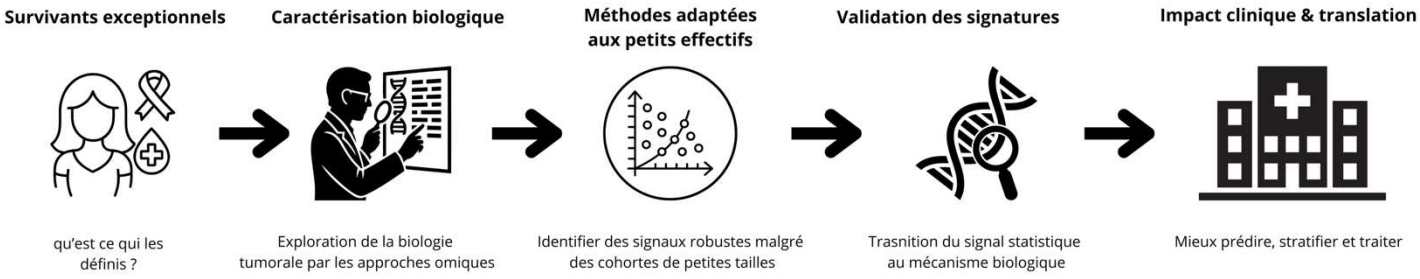

Figure 1 Organisation générale de la thèse et progression méthodologique de l'étude des survivants exceptionnels

La première partie est consacrée à la définition et à la compréhension du concept de survivant exceptionnel. Elle s'appuie sur les données de la littérature et sur des initiatives internationales, telles que l'Exceptional Responders Initiative du National Cancer Institute (NCI), afin de préciser les critères utilisés pour caractériser ces patients et les enjeux liés à leur identification. Elle intègre également un rappel des données épidémiologiques et biologiques des principales pathologies concernées, permettant de contextualiser la rareté de ces trajectoires cliniques. Cette partie vise ainsi à clarifier la notion de survie ou de réponse exceptionnelle et à montrer en quoi ces situations atypiques constituent un modèle pertinent pour explorer les mécanismes biologiques associés à une survie prolongée.

La deuxième partie présente les principales stratégies de caractérisation biologique mobilisées dans l'étude des survivants exceptionnels. Elle introduit, en des termes accessibles, les différentes approches omiques (génomique, transcriptomique, épigénomique, protéomique, immunologique) ainsi que les technologies associées (WES, WGS, single-cell RNA-seq, transcriptomique spatiale). Cette section établit le socle conceptuel et technologique nécessaire à l'interprétation des données générées chez ces patients. Elle s'appuie à la fois sur la littérature scientifique récente et sur les échanges menés avec **le Dr Naouel Zerrouk**, dont l'expertise a permis d'éclairer les choix technologiques, leurs apports respectifs et leurs limites, en particulier dans le contexte de petites cohortes et de phénotypes rares.

La troisième partie est consacrée aux outils biostatistiques et méthodologiques adaptés aux petites cohortes. Elle décrit les modèles utilisables (case–control, analyses différentielles, réduction de dimension, intégration multi-omique) et discute leurs forces et leurs limites. Cette partie s'appuie sur les entretiens menés avec deux experts Rémy Nicolle et Samuel Blanck afin d'ancrer la réflexion dans des pratiques méthodologiques concrètes.

La quatrième partie traite de l'interprétation biologique et de la validation des signatures identifiées. Elle détaille les différentes étapes nécessaires pour établir qu'un signal est robuste sur le plan analytique, qu'il correspond à un mécanisme biologique réel et qu'il possède une pertinence clinique potentielle. Cette section s'appuie sur des exemples issus de la littérature multi-omique en oncologie, mais également sur les échanges avec des experts de la biologie translationnelle, en particulier **le Dr Paloma Cejas**, afin d'illustrer comment relier signatures moléculaires, mécanismes biologiques et observations tissulaires.

La cinquième partie explore les apports cliniques et translationnels possibles des signatures issues des survivants exceptionnels. Elle s'intéresse à leur potentiel pour la prédiction de la réponse aux traitements, la stratification thérapeutique des patients, le repositionnement de médicaments et la conception d'essais cliniques adaptés à l'étude de sous-groupes très rares. Cette partie est éclairée par les entretiens menés avec **le Dr Yaovi Eric Amela**, clinicien oncologue, et **le Dr Paloma Cejas**

Enfin, la sixième partie propose une synthèse des résultats. Elle met en perspective les enseignements issus de l'ensemble des sections, discute les apports principaux de ce travail ainsi que ses limites et les points de vigilance méthodologiques, et ouvre des perspectives pour l'étude future des survivants exceptionnels.

1. Épidémiologie et définitions

1.1rappels épidémiologiques et biologiques(2)

Le cancer constitue aujourd'hui un enjeu majeur de santé publique. Il se définit comme une maladie résultant de la prolifération incontrôlée de cellules qui échappent aux mécanismes physiologiques régulant l'homéostasie tissulaire, notamment la prolifération, la différenciation et la survie cellulaire. Ces cellules acquièrent progressivement la capacité d'envahir les tissus adjacents et de disséminer à distance sous forme de métastases. Le développement tumoral est un processus lent et progressif, s'étendant le plus souvent sur plusieurs années, voire plusieurs décennies, mais dont l'évolution clinique est très variable selon le type de cancer, son origine tissulaire et ses caractéristiques biologiques.

1.1.1 Épidémiologie générale des cancers

D'un point de vue épidémiologique, trois indicateurs principaux permettent de décrire le poids des cancers dans une population : l'incidence, correspondant au nombre de nouveaux cas diagnostiqués sur une période donnée, la mortalité, qui représente le nombre de décès attribuables au cancer, et la prévalence, définie comme le nombre de personnes vivant avec un diagnostic de cancer, qu'elles soient en cours de traitement ou considérées comme guéries. En France Le cancer constitue actuellement la première cause de mortalité. En 2023, le nombre de nouveaux cas de cancer en France métropolitaine est estimé à environ 433 000, (2) dont 57 % chez les hommes. Bien que l'incidence globale des cancers diminue chez l'homme depuis 2005, elle continue d'augmenter chez la femme. La mortalité par cancer, en revanche, diminue régulièrement chez les deux sexes. Le cancer demeure une maladie majoritairement masculine et affecte préférentiellement la seconde moitié de la vie. L'âge médian au diagnostic est de 70 ans chez l'homme et de 68 ans chez la femme, tandis que l'âge médian au décès est respectivement de 73 et 77 ans. Cette évolution démographique explique l'importance croissante de la prise en charge oncogériatrique. Quatre localisations tumorales concentrent à elles seules environ la moitié des nouveaux cas de cancer : la prostate, le sein, le poumon et le côlon-rectum. En France, les cancers responsables du plus grand nombre de décès diffèrent selon le sexe. Chez les hommes, les principales causes de mortalité par cancer sont le cancer du poumon (environ 20 500 décès), suivi des cancers colorectaux (environ 9 000 décès) et du cancer de la prostate (environ 9 200 décès). Chez les femmes, les cancers les plus meurtriers sont le cancer du sein (environ 12 600 décès), le cancer du poumon (environ 9 900 décès) et les cancers colorectaux (environ 8 000 décès).(3)

1.1.2. Épidémiologie des cancers

Le cancer du sein est le cancer le plus fréquent chez la femme, représentant environ un tiers des nouveaux cas de cancers féminins. Son incidence augmente légèrement, avec un âge médian au diagnostic de 64 ans. Bien qu'il reste la première cause de mortalité par cancer chez la femme, sa mortalité diminue de manière régulière.

Le cancer de la prostate est le cancer le plus fréquent chez l'homme, représentant près d'un quart des nouveaux cas. Son incidence est en diminution, l'âge médian au diagnostic étant de 69 ans. Il constitue la troisième cause de décès par cancer chez l'homme, avec une mortalité également en baisse.

Le cancer colorectal est le quatrième cancer le plus fréquent en France. Son incidence diminue chez l'homme et augmente légèrement chez la femme. Il constitue la deuxième cause de mortalité par cancer, soulignant son impact majeur en santé publique.

Le cancer du poumon occupe une place particulière par sa gravité. Il est le troisième cancer le plus fréquent mais demeure la première cause de décès par cancer. Son incidence diminue lentement chez l'homme mais augmente fortement chez la femme, traduisant l'évolution des comportements à risque, notamment le tabagisme.

Parmi les cancers plus rares mais de pronostic particulièrement défavorable, les tumeurs primitives du système nerveux central occupent une place spécifique. Leur incidence est estimée à environ 15 cas pour 100 000 habitants, représentant près de 1,2 % de l'ensemble des cancers, avec environ 5 910 nouveaux cas diagnostiqués en France en 2023. Chez l'adulte, les gliomes constituent les tumeurs primitives les plus fréquentes, et parmi eux, le glioblastome représente la forme majoritaire, avec environ 2 500 nouveaux cas par an en France. Ces tumeurs présentent un sex-ratio en faveur des hommes, compris entre 1,3 et 1,8 selon l'histologie, et sont associées à un pronostic particulièrement sombre. Chez l'enfant, les tumeurs cérébrales constituent la deuxième cause de cancer après les leucémies et regroupent des entités biologiquement distinctes, telles que les gliomes diffus du tronc cérébral, les astrocytomes pilocytiques ou les médulloblastomes.

L'adénocarcinome pancréatique exocrine constitue quant à lui un problème majeur de santé publique en raison de son extrême gravité. À l'échelle mondiale, environ 496 000 nouveaux cas ont été estimés en 2020, pour 466 000 décès, faisant de ce cancer la septième cause de mortalité par cancer. En France, il représente le septième cancer le plus fréquent en 2023 et figure parmi les principales causes de décès par cancer, occupant la quatrième place chez la femme et la cinquième chez l'homme. Son incidence a fortement augmenté au cours des dernières décennies, avec une hausse estimée à +250 % entre 1980 et 2012, et une progression annuelle persistante depuis, tant chez l'homme que chez la femme. L'âge médian au diagnostic est élevé, autour de 71 ans chez l'homme et 74 ans chez la femme. Le pronostic demeure extrêmement

défavorable, avec une survie nette à cinq ans estimée à environ 9 %, en raison notamment du diagnostic tardif, de l'absence de dépistage, de la localisation profonde du pancréas et du potentiel métastatique précoce de la maladie.

1.1.3. Facteurs de risque et carcinogenèse

Le principal facteur de risque du cancer est l'âge. D'autres facteurs de risque correspondent à des expositions ou caractéristiques augmentant la probabilité de développer un cancer, comme les expositions professionnelles, environnementales ou les pathologies prédisposantes. La notion de risque attribuable permet d'estimer la proportion de cas directement imputables à un facteur donné.

Sur le plan biologique, la carcinogenèse est un processus multi-étapes classiquement décrit en trois phases : l'initiation, correspondant à une lésion irréversible de l'ADN induite par un agent carcinogène ; la promotion, liée à des stimuli favorisant l'expansion clonale des cellules initiées ; et la progression, marquée par l'acquisition de capacités de prolifération incontrôlée, de résistance à l'apoptose, d'invasion tissulaire et de dissémination métastatique comme l'illustre la figure 1 issu du collège de cancérologie **(2)**

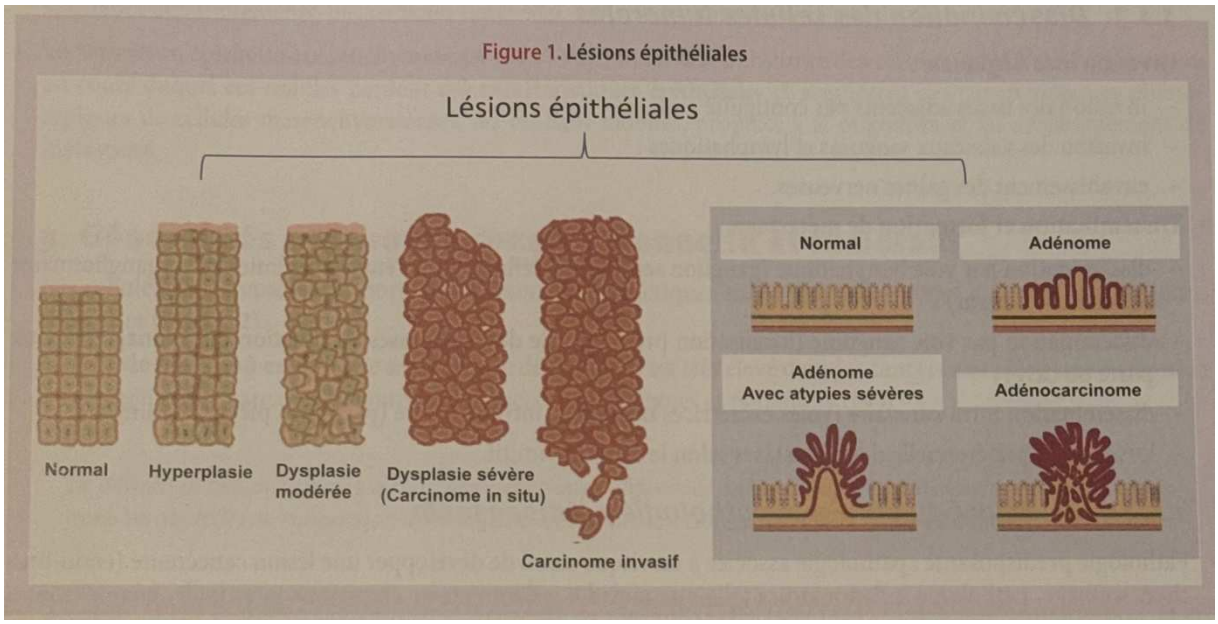

Figure 2: Progression des lésions épithéliales au cours de la carcinogenèse, de l'épithélium normal au carcinome invasif (2)

L'évolution histologique des cancers épithéliaux suit généralement une succession de lésions allant de l'épithélium normal à l'hyperplasie, la dysplasie, le carcinome in situ, puis le carcinome invasif. Cette progression s'accompagne de modifications architecturales et cytologiques de plus en plus marquées.

1.1.4. Biologie des cellules cancéreuses

Les cellules cancéreuses se caractérisent par un ensemble de propriétés fonctionnelles acquises au cours du processus tumoral. Ces caractéristiques incluent notamment l'auto-suffisance en signaux de prolifération, l'insensibilité aux signaux inhibiteurs de croissance, l'échappement à l'apoptose, une capacité réplicative illimitée, l'induction de la néoangiogenèse, l'invasion tissulaire, la formation de métastases, un métabolisme énergétique spécifique et l'échappement à la réponse immunitaire. L'instabilité génétique et l'inflammation tissulaire favorisent l'acquisition progressive de ces propriétés. Les cellules tumorales présentent de multiples anomalies génétiques, dont le type et la fréquence varient selon les cancers. Ces anomalies incluent des mutations, des amplifications ou délétions géniques, des translocations chromosomiques et des modifications épigénétiques telles que la méthylation de l'ADN. Certaines populations cellulaires particulières, comme les cellules souches tumorales ou les cellules engagées dans une transition épithélio-mésenchymateuse, jouent un rôle central dans la résistance aux traitements, les rechutes et la dissémination métastatique.

1.2. Définitions des principaux critères d'évaluation en oncologie

En oncologie, l'évaluation de l'efficacité des traitements et la caractérisation des trajectoires cliniques reposent sur des critères standardisés, définis afin de permettre des comparaisons reproductibles entre études cliniques. Ces critères constituent un cadre indispensable pour interpréter la notion de survie ou de réponse exceptionnelle.

La **survie globale (Overall Survival, OS)** correspond au temps écoulé entre une date de référence, le plus souvent l'initiation du traitement ou le diagnostic, et le décès quelle qu'en soit la cause. Elle est considérée comme le critère de jugement le plus robuste et le plus cliniquement pertinent, car elle reflète directement le bénéfice pour le patient.

La **survie sans progression (Progression-Free Survival, PFS)** désigne le temps écoulé entre le début du traitement et la survenue d'une progression tumorale objectivée ou du décès. Elle permet d'évaluer le contrôle tumoral indépendamment des lignes thérapeutiques ultérieures, mais peut être influencée par la fréquence et les modalités des évaluations radiologiques.

La **réponse tumorale** est classiquement évaluée à l'aide de critères radiologiques standardisés, principalement les critères RECIST (Response Evaluation Criteria In Solid Tumors). Selon ces critères, une réponse complète (Complete Response, CR) correspond à la disparition de toutes les lésions tumorales mesurables, tandis qu'une réponse partielle (Partial Response, PR) est définie par une diminution d'au moins 30 % de la somme des diamètres des lésions cibles. À l'inverse, la progression de la

maladie est caractérisée par une augmentation significative de la charge tumorale ou l'apparition de nouvelles lésions.

Ces critères constituent aujourd'hui la base de l'évaluation clinique en cancérologie, tant dans les essais cliniques que dans la pratique courante. Ils permettent de définir de manière opérationnelle des trajectoires de réponse ou de survie atypiques, et servent de référence pour identifier les patients qualifiés de survivants ou répondeurs exceptionnels.

1.3 Concept des « survivants exceptionnels »

Après avoir rappelé les bases nécessaires à une bonne compréhension du cancer, nous pouvons à présent aborder le concept qui constitue le cœur de ce travail : celui des ***survivants exceptionnels***. Ce terme, parfois désigné comme ***répondeurs exceptionnels*** **ou encore** ***outliers*****,** fait référence à des patients qui présentent une survie ou une réponse thérapeutique très largement supérieure à ce qui est attendu d'un point de vue statistique. De manière opérationnelle, une réponse exceptionnelle est généralement définie comme une réponse complète (CR) ou partielle (PR) observée dans un contexte où moins de 10 % des patients répondent habituellement au même traitement (1), ou encore comme une durée de réponse supérieure à trois fois la médiane classiquement observée. Ces situations cliniques, bien que rares, interrogent directement notre compréhension des mécanismes biologiques de la maladie et de la réponse aux traitements. L'étude de ces cas atypiques nécessite des méthodologies adaptées, en raison des effectifs réduits et de la forte hétérogénéité interindividuelle. Cela implique le recours à des approches spécifiques, qu'il s'agisse d'analyses comparatives de type cas-témoins, de modèles multivariés robustes, de méthodes d'apprentissage automatique adaptées aux petits effectifs, ou encore de l'intégration de données multi-omiques afin de capter des signaux biologiques complexes.

Parmi les rares études ayant spécifiquement analysé des réponses prolongées en contexte métastatique, celle de **Griguolo et al. (4)**constitue un exemple particulièrement illustratif. Cette étude rétrospective visait à caractériser des patientes atteintes de cancer du sein métastatique ayant présenté une réponse exceptionnelle à un traitement systémique. Les auteurs ont identifié **58 patientes** considérées comme *répondeuses exceptionnelles*, définies par une réponse durable largement supérieure aux attentes habituelles pour le traitement administré. Les résultats rapportés montrent une survie globale à cinq ans de 94,6 %, ainsi qu'une survie sans progression à cinq ans de 89,1 %, des valeurs très supérieures à celles classiquement observées dans cette population. L'analyse met en évidence plusieurs facteurs cliniques associés à ces trajectoires de survie prolongée, notamment **le sous-type HER2 positif**, l'obtention d'une réponse complète ou d'un statut sans signe de maladie, ainsi qu'un nombre limité de sites métastatiques. Bien que cette étude repose exclusivement sur des données cliniques, elle souligne clairement l'existence de profils de patients présentant une évolution hors normes. Elle met ainsi en évidence l'intérêt d'une caractérisation biologique plus fine de ces cas afin de mieux comprendre les mécanismes distinguant les patients « outliers » de la population générale, et soutient la pertinence d'approches intégrant l'analyse multi-omique pour identifier des signatures associées à la survie exceptionnelle.

2. Revue des stratégies de caractérisation biologique

2.1. objectif

L'objectif de cette partie est de présenter et d'expliquer les méthodes biologiques et moléculaires permettant de caractériser les survivants exceptionnels du cancer. Elle vise à définir clairement les différentes approches omiques mobilisées dans la recherche actuelle, ainsi qu'à en décrire les intérêts et les limites. Les concepts et terminologies seront introduits progressivement, tout en abordant les enjeux méthodologiques spécifiques liés à l'analyse de petites cohortes. Cette partie pose ainsi les fondations biologiques et technologiques indispensables à l'étude multi-omique des survivants exceptionnels. Nous interrogerons ensuite une experte pour cette partie le Dr Naouel Zerrouk qui donnera son avis sur les meilleurs approches à adopter pour la mise en place d'une étude sur des petites cohortes de survivants exceptionnels.

2.2. Approches génomiques : WES vs WGS

La génomique : analyser la séquence de l'ADN

La génomique constitue la première couche d'exploration des tumeurs et vise à caractériser l'ensemble des altérations présentes dans la séquence de l'ADN. Ces altérations comprennent notamment les mutations somatiques, les délétions ou amplifications de segments chromosomiques, ainsi que les variations du nombre de copies (Copy Number Variations, CNV). Chacune de ces anomalies peut modifier le comportement tumoral, influencer la sensibilité ou la résistance à un traitement, ou encore affecter la capacité de la tumeur à interagir avec le système immunitaire. L'analyse génomique repose aujourd'hui principalement sur deux technologies de séquençage à haut débit. Le **Whole Exome Sequencing (WES)** cible spécifiquement les régions codantes du génome, qui représentent environ 1 à 2 % de l'ADN total mais concentrent la majorité des mutations biologiquement interprétables. À l'inverse, le **Whole Genome Sequencing (WGS)** permet d'explorer l'intégralité du génome, incluant les régions non codantes et régulatrices.

Afin de mieux comprendre la différence entre ces deux techniques, la figure 3 est proposée pour en faciliter la compréhension.

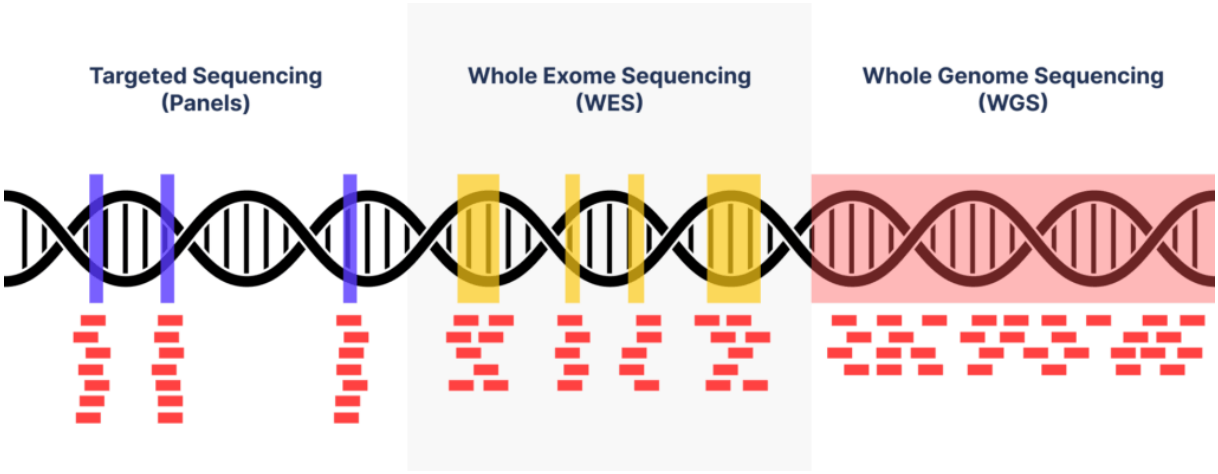

Figure 3 Différences de couverture et de résolution entre les approches de séquençage ciblé, le WES et le WGS (5)

Le choix entre le séquençage de l'exome (WES) et le séquençage du génome complet (WGS) constitue une étape méthodologique déterminante dans l'étude des survivants exceptionnels, car il conditionne directement la nature, la précision et la qualité des signaux génomiques pouvant être détectés. Cette décision influence non seulement la profondeur analytique, mais également la capacité à identifier des altérations rares dans des cohortes de taille limitée.

Dans le contexte des petites cohortes, le WES présente plusieurs avantages majeurs. En ciblant exclusivement les régions codantes du génome, soit environ 1 à 2 % de l'ADN total (6), il concentre l'analyse sur les zones où se trouvent la majorité des mutations ayant un impact fonctionnel ou thérapeutique potentiel. Cette focalisation permet d'obtenir une **profondeur de séquençage élevée**, indispensable pour détecter avec confiance des mutations rares ou présentes à faible fréquence dans la tumeur. Le WES se révèle ainsi particulièrement adapté à l'**identification de mutations actionnables**, dans une perspective de *target discovery*. De plus, le volume de données généré reste relativement limité, rendant les analyses bio-informatiques plus accessibles et facilitant l'interprétation biologique des résultats. Pour ces raisons, il représente souvent la stratégie de choix dans l'étude de cohortes restreintes.

À l'inverse, le WGS offre une **couverture complète du génome**,(7) incluant non seulement les régions codantes mais aussi les régions non codantes, les éléments régulateurs et l'ensemble des structures génomiques complexes telles que les réarrangements, insertions, duplications ou événements de chromothripsis c'est-à-dire un mécanisme d'instabilité chromosomique caractérisé par des réarrangements complexes résultant d'une fragmentation massive et simultanée de segments chromosomiques. Cette exhaustivité constitue un atout pour l'exploration de mécanismes de régulation ou d'altérations structurelles potentiellement impliqués dans la survie prolongée, notamment lorsqu'il s'agit de comprendre des phénomènes rares qui dépassent le cadre strict des gènes codants.

Cependant, cette richesse informationnelle s'accompagne de limites notables dans des contextes où les effectifs sont réduits. Le WGS génère un volume de données considérablement plus élevé, ce qui accroît mécaniquement le bruit technique et biologique. La profondeur de séquençage est généralement moindre que celle obtenue en WES, ce qui peut compromettre la détection fiable d'événements rares ou faiblement représentés. En l'absence d'une cohorte importante ou de jeux de données externes permettant d'ancrer l'interprétation, les signaux issus du WGS peuvent devenir difficiles à distinguer du bruit de fond, rendant certaines conclusions incertaines. De plus, les travaux récents montrent que la mise en œuvre du WGS en clinique nécessite une standardisation poussée des pipelines analytiques afin d'exploiter efficacement la masse d'informations générée. Dans l'étude de Kim et al.(8), même avec un pipeline certifié et des capacités techniques avancées, l'interprétation du génome complet a exigé des étapes dédiées pour corriger les artefacts liés aux échantillons FFPE (échantillons fixés au formol et inclus en paraffine), intégrer les signatures mutationnelles et distinguer précisément les variants somatiques des variants germinaux. Ces exigences méthodologiques soulignent que, malgré son potentiel, le WGS impose une sophistication analytique difficilement transposable à des études de petite taille. Dans le cadre des survivants exceptionnels, où les effectifs sont limités et où la robustesse des signaux prime sur l'exhaustivité, ces contraintes renforcent la pertinence d'approches plus ciblées et plus profondes comme le WES.

Ainsi, dans les études portant sur des survivants exceptionnels caractérisées par des effectifs restreints et la nécessité d'identifier des altérations rares mais robustes le WES apparaît comme l'approche la plus adaptée, en raison de sa profondeur, de sa stabilité analytique et de sa meilleure interprétabilité. En pratique, le choix entre WES et WGS peut également être guidé par des considérations de coût (6)et de capacité analytique, le WES offrant un excellent compromis entre performance et accessibilité. Le WGS conserve néanmoins une place pertinente dans des investigations plus exploratoires, lorsque l'objectif est de cartographier l'ensemble des altérations génomiques ou d'explorer des mécanismes régulateurs au-delà des régions codantes.

2.3. Approches transcriptomiques

L'analyse transcriptomique complète la génomique en offrant une vision dynamique de l'activité cellulaire. En examinant l'expression des ARN messagers, elle permet d'identifier quelles voies biologiques sont activées ou réprimées et d'explorer les mécanismes fonctionnels sous-jacents aux profils de survie prolongée. Trois approches principales peuvent être mobilisées, chacune apportant un niveau de résolution différent et des informations complémentaires.

Le bulk RNA-seq est l'approche la plus classique. En mesurant l'expression moyenne sur un ensemble de milliers de cellules, il fournit une signature globale stable, particulièrement utile pour comparer des groupes de patients dans des cohortes de taille réduite. Bien qu'il ne permette pas de distinguer les différentes populations cellulaires (8)présentes dans la tumeur, il constitue un socle analytique fiable pour identifier des voies transcriptionnelles majoritairement activées ou réprimées. Dans le cadre des survivants exceptionnels, le bulk peut ainsi révéler des patterns globaux distinctifs c'est-à-dire tendances moléculaires générales et cohérentes qui caractérisent un groupe de patients par rapport à un autre ; par exemple une activation immunitaire accrue ou une modulation particulière de voies de stress ou de réparation.

Le **single-cell RNA-seq (scRNA-seq)** permet de dépasser les limites du bulk en analysant le transcriptome cellule par cellule. Cette approche met en évidence l'hétérogénéité intra-tumorale et identifie des sous-populations cellulaires aux profils transcriptionnels distincts, y compris des populations rares susceptibles de jouer un rôle disproportionné dans la réponse thérapeutique ou la survie prolongée. Cette capacité à révéler la diversité interne des tumeurs est largement décrite dans la littérature, notamment dans l'étude publié dans *nature* où les auteurs soulignent que le scRNA-seq *« profiles the dynamic landscapes of cellular and genetic alterations […] and sheds light on intratumoral heterogeneity »* (9)

Le scRNA-seq permet donc de :

- **distinguer différents clones tumoraux** *(groupes de cellules cancéreuses partageant des altérations génétiques ou transcriptomiques communes)***;**
- **identifier des sous-types immunitaires infiltrants** *(différentes populations de cellules du système immunitaire présentes au sein de la tumeur, telles que les lymphocytes T, les macrophages ou les cellules NK) ;*
- **détecter des états cellulaires transitoires ou adaptatifs.** *(changements temporaires de l'état fonctionnel des cellules en réponse au microenvironnement tumoral ou aux traitements)*.

Cette résolution fine est particulièrement utile pour étudier les survivants exceptionnels, chez lesquels des mécanismes cellulaires atypiques minoritaires mais fonctionnellement déterminants peuvent contribuer à une meilleure réponse ou à une résistance naturelle à la progression tumorale.

La transcriptomique spatiale combine l'analyse de l'expression génique avec la localisation des cellules dans le tissu, permettant de visualiser directement comment les cellules tumorales et non tumorales s'organisent et interagissent dans leur microenvironnement. Contrairement au single-cell RNA-seq, qui perd l'information spatiale lors de la dissociation cellulaire, cette approche préserve l'architecture tissulaire et offre un accès in situ aux interactions cellule-cellule. Cette capacité à maintenir la dimension spatiale est essentielle, comme le souligne l'article de **Jiazhou Ye** publié dans *precision oncology* (9)

« provides high-quality transcriptional data for visualization of the spatial dynamic landscape, breaking through the limitation of cell location information lost by scRNA-seq » . (9)

En permettant de localiser précisément les populations cellulaires et d'observer leurs interactions, cette méthode met en lumière la structure fonctionnelle des niches tumorales, qu'il s'agisse de zones d'infiltration immunitaire, de régions stromales actives ou de microdomaines favorisant ou limitant la progression tumorale. Bien que sa robustesse dépende en partie de la qualité des tissus analysés, notamment dans des cancers fibreux comme le pancréas, la transcriptomique spatiale demeure un outil clé pour comprendre comment l'organisation tissulaire influence les trajectoires évolutives et les réponses observées chez certains survivants exceptionnels.

conclusion

Ces trois approches transcriptomiques peuvent être envisagées comme complémentaires :

- **le bulk fournit une base robuste pour comparer des signatures globales,**
- **le single-cell révèle l'hétérogénéité fonctionnelle et les signaux rares,**
- **le spatial replace ces signaux dans l'organisation réelle du tissu.**

Leur intégration est essentielle pour appréhender l'ensemble des mécanismes transcriptionnels impliqués dans la survie exceptionnelle et pour relier la dynamique génique à la structure du microenvironnement tumoral.

2.4. Autres approches omiques : protéomique, immunomique, épigénomique, métabolomique

Au-delà du génome et du transcriptome, d'autres approches omiques permettent d'explorer des niveaux supplémentaires de régulation et de fonction cellulaire, offrant une vision plus complète des mécanismes susceptibles de contribuer à la survie prolongée.

La **protéomique** s'intéresse aux protéines et à leurs modifications post-traductionnelles, qui représentent l'étage fonctionnel directement impliqué dans les processus cellulaires. Elle permet notamment d'identifier des voies de signalisation activées, des altérations de phosphorylation ou des états fonctionnels spécifiques des cellules tumorales et immunitaires. Bien que cette approche soit parfois limitée par la difficulté à quantifier de faibles abondances protéiques, elle reste une source d'information essentielle pour relier les programmes transcriptionnels à leurs conséquences biologiques concrètes.

L'**immunomique** constitue une autre dimension particulièrement pertinente dans l'étude des survivants exceptionnels. Elle permet de caractériser la composition et l'état fonctionnel des cellules immunitaires infiltrant la tumeur, d'analyser le répertoire des récepteurs des lymphocytes T (TCR) ou encore d'explorer l'activation des cellules myéloïdes. Cette approche éclaire les interactions entre la tumeur et le système immunitaire, souvent déterminantes pour comprendre les réponses prolongées observées chez certains patients.

L'**épigénomique** explore les mécanismes de régulation indépendants de la séquence ADN, tels que la méthylation de l'ADN ou les modifications des histones. Ces modifications influencent l'accessibilité de l'ADN et l'expression des gènes, jouant un rôle clé dans les phénomènes d'adaptation tumorale, la plasticité cellulaire ou la transition entre états fonctionnels.

Enfin, la **métabolomique** étudie les métabolites et les voies énergétiques mobilisées par la tumeur. Elle permet de caractériser les dépendances métaboliques spécifiques, les adaptations au stress ou les échanges métaboliques avec le microenvironnement, autant d'éléments pouvant contribuer à expliquer une survie exceptionnelle.

Dans leur ensemble, ces approches omiques représentent des couches d'information complémentaires. Leur intégration ne vise pas à multiplier les données, mais à renforcer la compréhension des mécanismes biologiques en combinant des signaux convergents provenant de différents niveaux d'organisation cellulaire. Plus ces couches sont cohérentes entre elles, plus la confiance dans l'interprétation globale est élevée et donc justifiable.

2.5. Intégration multi-omique et contraintes sur petits effectifs

La valeur ajoutée des études multi-omiques réside dans leur intégration. Deux grandes stratégies émergent :

- **Approches non supervisées** (ex. MOFA) : elles permettent une intégration exploratoire, sans a priori, en identifiant des facteurs latents communs aux différentes couches de données. Elles sont adaptées pour générer des hypothèses nouvelles, mais peuvent être difficiles à interpréter biologiquement.
- **Approches supervisées et biologiquement guidées** (ex. mixOmics, intégration manuelle) : elles reposent sur l'intersection progressive de listes de gènes ou de cibles entre plusieurs "omics" (mutations, expression, localisation spatiale). Cette approche, plus intuitive, facilite l'interprétation et le reverse engineering, entendu ici comme la reconstruction a posteriori des mécanismes biologiques plausibles à partir des signaux observés dans les données, ce qui est particulièrement utile dans des études à effectif limité.

Un enjeu majeur reste la **pondération des signaux** : accorder moins de poids aux analyses issues de données bruitées (bulk FFPE) et plus aux données résolutives et robustes (single-cell, spatial) afin de limiter l'influence du bruit technique et de renforcer la fiabilité des signatures biologiques identifiées.

2.6. Perspectives et freins actuels

L'ensemble de ces stratégies vise à transformer des signaux statistiques en pistes biologiques exploitables cliniquement. Toutefois, plusieurs freins subsistent :

- **Qualité hétérogène des échantillons** (notamment FFPE).
- **Reproductibilité limitée** des expériences, liée à la dispersion des données dans la communauté scientifique.
- **Manque de mutualisation** : de nombreuses données générées restent inaccessibles ou sous-exploitées.(10)

2.7. Entretien avec l'expert

ENTRETIEN: point de vue de l'experte – Dr. Naouel Zerrouk

Cette section présente et structure l'entretien réalisé avec le Dr. Naouel Zerrouk, computational biologist spécialisée en génomique, transcriptomique et intégration multi-omique. Son expertise couvre à la fois la modélisation biologique, l'analyse de données complexes et le développement de pipelines intégrés en santé.

1. WES vs WGS : pertinence du choix dans les petites cohortes

Le Dr Zerrouk insiste sur le fait que, dans le contexte des survivants exceptionnels, le choix de la technologie génomique n'est pas anodin. Selon elle :

« Dans un contexte aussi complexe, il est primordial d'obtenir une profondeur maximale sur les régions codantes. Le WES profond est plus adapté et plus straightforward. »

Le Dr Zerrouk précise que le WGS, bien que plus exhaustif, introduit davantage de bruit, surtout sur petits effectifs, et reste plus difficile à interpréter en l'absence de cohortes larges ou de validations externes.

2. Intégration des données publiques : une stratégie raisonnée

Le Dr Zerrouk met en garde contre l'usage naïf des bases publiques comme « effectifs supplémentaires » :

« Ces bases sont souvent sparse, peu annotées cliniquement. Il ne faut pas les intégrer comme des échantillons de plus. »

En revanche, elle recommande leur utilisation comme sources d'information externe (essentialité, dépendances, fréquence mutationnelle), en complément des données internes.

3. Transcriptomique : articulation entre bulk, single-cell et spatial

Le Dr Zerrouk considère que les trois approches sont indispensables mais jouent des rôles distincts :

- **Bulk RNA-seq** : base solide, même sur FFPE, utile pour stabiliser les signaux globaux.
- **scRNA-seq** : approche prioritaire, essentielle pour capturer l'hétérogénéité et les populations rares.

- **Spatial transcriptomics** : technologie apportant une preuve spatiale des interactions, supérieure à la seule corrélation transcriptomique.

Elle note :

« Le spatial permet de montrer la co-localisation ligand–récepteur. Cela augmente énormément la confiance. »

4. Limites techniques des approches omiques

Le Dr Zerrouk décrit plusieurs biais à anticiper :

- **Spatial** : qualité hétérogène sur tissus fibreux (pancréas notamment).
- **Single-cell** : zéros techniques dus à une profondeur insuffisante.
- **Bulk** : dépendance à la qualité du FFPE.

Ces limites doivent guider la pondération des signaux lors de l'intégration.

5. Autres omiques : intérêt et réalisme méthodologique

Le Dr Zerrouk souligne que l'ajout de couches omiques est toujours souhaitable, mais rarement réaliste :

- **Protéomique** : riche mais difficile à analyser (beaucoup de zéros).
- **Immunomique** : très pertinente pour comprendre les réponses prolongées.
- **Métabolomique / épigénomique** : utiles mais rarement prioritaires.

Selon elle :

« Le compromis réaliste, c'est génomique + transcriptomique + immunologique/protéomique. Le reste, si possible, mais pas indispensable au départ. »

6. Intégration multi-omique : combinaison des approches supervisées et non supervisées

Le Dr Zerrouk distingue deux logiques complémentaires :

- **Approches non supervisées (MOFA, etc.)** : utiles pour détecter des facteurs latents transversaux.
- **Approches supervisées couche-par-couche** : plus interprétables, permettant de valider un signal dans plusieurs omiques.

Elle ajoute :

« On pondère les signaux selon la fiabilité de la technique : moins de poids au bulk FFPE, plus au single-cell. »

7. Vision translationnelle et principales limitations actuelles

Le Dr Zerrouk insiste sur la nécessité d'une vision systémique :

« Impossible de réduire à une seule omique. Tout interagit. »

Elle identifie trois freins majeurs :

- manque de reproductibilité,
- absence de mutualisation,
- faible réexploitation des données existantes.

Selon elle :

« Ce n'est pas le manque de données qui bloque, mais leur fragmentation. »

Synthèse

L'étude des survivants exceptionnels nécessite non seulement une méthodologie adaptée, mais aussi un cadre de mutualisation et de validation externe pour dépasser les limitations actuelles de la recherche.

Conclusion générale de l'entretien

Les recommandations formulées par le Dr Naouel Zerrouk convergent vers une stratégie méthodologique, particulièrement adaptée à l'analyse de petites cohortes et à l'étude de phénotypes rares tels que celui des survivants exceptionnels du cancer. Elles soulignent l'importance d'une approche progressive et hiérarchisée, capable de tirer parti de données complexes tout en limitant les biais inhérents à ce type de population. Ces recommandations s'inscrivent dans la continuité des choix méthodologiques présentés dans ce travail. L'utilisation prioritaire du séquençage de l'exome (WES) apparaît comme un socle robuste pour l'identification d'altérations génétiques pertinentes, tandis que la combinaison raisonnée d'analyses bulk, single-cell et de transcriptomique spatiale permet de capturer à la fois les signaux globaux et l'hétérogénéité fine des tissus tumoraux et de leur microenvironnement. L'intégration de données publiques est envisagée avec prudence, non comme un substitut aux cohortes internes, mais comme un outil complémentaire de validation et de contextualisation des signaux observés. Par ailleurs, l'accent mis sur la hiérarchisation des résultats en fonction de la robustesse des

techniques utilisées, la prise en compte stricte des biais techniques, ainsi que le recours à une approche multi-omique hybride combinant analyses non supervisées et supervisées, s'inscrit pleinement dans les bonnes pratiques actuelles. Enfin, l'approche défendue par le Dr Naouel Zerrouk, intégrant de manière cohérente les différentes échelles d'analyse, depuis les altérations moléculaires jusqu'à leurs implications cliniques potentielles, renforce l'ambition de ce travail. Celui-ci ne se limite pas à l'identification de signaux statistiques, mais vise à produire des résultats biologiquement interprétables, susceptibles d'être mobilisés à terme dans une démarche de médecine de précision.

3. Exploration des outils biostatistiques et méthodologiques

3.1 Objectif

L'objectif de cette partie est d'identifier et de justifier les méthodes statistiques et bio-informatiques les plus adaptées à l'étude de patients dits « survivants exceptionnels », c'est-à-dire des individus atteints de cancers réputés incurables mais ayant présenté une survie beaucoup plus longue que prévue. L'étude de ces profils repose sur des effectifs restreints et sur des données complexes issues de plusieurs couches biologiques (génomique, transcriptomique, épigénomique, protéomique, immunologique). Il est donc essentiel de sélectionner des approches analytiques robustes, adaptées aux petits effectifs et capables d'extraire des signaux fiables tout en limitant les biais statistiques et le surapprentissage. Plus précisément, cette partie a pour but de définir les modèles statistiques pertinents selon la nature des données et la question étudiée, de présenter les outils de référence utilisés pour l'analyse des données omiques pour pouvoir décrire les approches d'intégration multi-omique permettant d'analyser conjointement différentes couches biologiques, de mettre en évidence les stratégies pour éviter le surapprentissage et garantir la robustesse des résultats, enfin de clarifier les critères de qualité d'un modèle dans le contexte spécifique des survivants exceptionnels :

Afin de mieux comprendre ces enjeux méthodologiques et de consolider les choix présentés, deux entretiens ont été réalisés avec des experts reconnus dans leur domaine :

- **Dr Rémy Nicolle**, responsable de l'équipe *Génomique translationnelle de l'hétérogénéité des néoplasies pancréatiques (GeNeHetX)* à l'INSERM, dont les travaux portent sur l'inflammation oncogénique et le développement tumoral.
- **M. Samuel Blanck**, ingénieur d'études rattaché à l'équipe d'accueil EA2694, spécialisé dans les méthodes biostatistiques appliquées aux données omiques.

Leurs contributions ont permis d'éclairer les aspects techniques, d'affiner le choix des outils statistiques, et de consolider la cohérence entre la stratégie d'analyse et les contraintes propres aux petites cohortes de survivants exceptionnels.

3.2. Choix des methode statistiques

L'**analyse de survie** regroupe l'ensemble des méthodes permettant d'étudier le délai écoulé avant la survenue d'un événement (décès, rechute, progression…). Parmi ces outils, la **courbe de Kaplan-Meier** constitue une représentation graphique de la probabilité de survie au cours du temps, en tenant compte des observations censurées, c'est-à-dire des patients pour lesquels l'événement d'intérêt n'a pas été observé au cours de la période de suivi, qu'il s'agisse de patients perdus de vue ou encore de patients toujours vivants ou sans événement à la fin du suivi. Le **modèle de Cox proportionnel**, quant à lui, permet d'estimer l'effet de variables explicatives (telles que des gènes, mutations ou facteurs cliniques) sur le risque de décès de progression ou de rechute. Deux approches principales peuvent donc être envisagées selon la manière dont on définit l'**endpoint** (le critère étudié).

3.2.1. Analyse de survie (Kaplan-Meier, modèles de Cox)

Les modèles de survie sont les plus utilisés en oncologie lorsqu'il s'agit d'évaluer le délai jusqu'à un événement (décès, rechute, progression). Le Kaplan-Meier permet une visualisation descriptive, tandis que le modèle de Cox proportionnel estime l'effet de covariables sur le risque d'événement. Cependant, ces méthodes nécessitent un effectif suffisamment important et surtout un nombre élevé d'événements pour être robustes. Dans le cas des **survivants exceptionnels**, qui sont par définition peu nombreux, cette approche perd rapidement en pertinence. Elle peut néanmoins fournir une illustration utile, comme le montre **l'étude de Kamath et al. sur l'adénocarcinome pancréatique (11)**, où une survie significativement prolongée était observée chez les patients présentant moins de trois mutations non synonymes *(Figure 3)*.

FIGURE 3. **Outcomes for Patients With High Vs Low Number of Genomic Mutations**

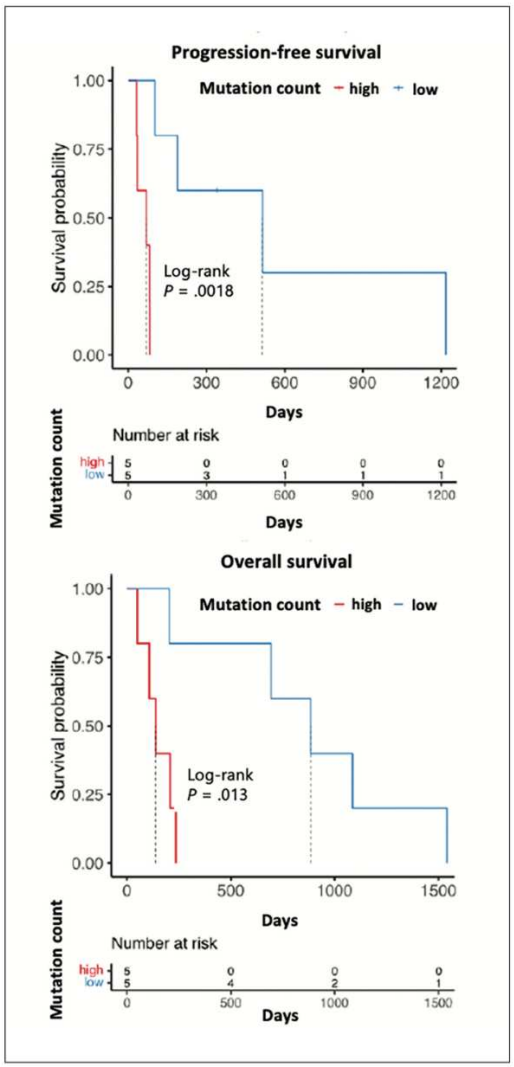

Figure 4 Résultats de survie chez les patients présentant un nombre élevé versus faible de mutations génomiques (11)

3.2.2. Approche binaire (case–control) et outils d'analyse associés

Dans les cohortes réduites, l'approche la plus pertinente consiste à comparer directement deux groupes : les **survivants exceptionnels** d'un côté et les **autres patients** de l'autre. Cette stratégie, dite **case–control**, ne prend pas en compte la dimension temporelle, mais s'attache à identifier des différences de caractéristiques biologiques (expression génique, altérations moléculaires, profils immunitaires) entre deux populations distinctes. Elle constitue l'alternative la plus robuste aux modèles de survie classiques lorsque le nombre d'événements est trop limité pour fournir une puissance statistique suffisante.

3.2.3. Analyse différentielle de l'expression génique : DESeq2

Pour les données transcriptomiques (ARN-seq), la méthode de référence est **DESeq2**, un algorithme développé pour identifier les gènes **exprimés différemment** entre deux

groupes (par exemple, survivants exceptionnels vs non survivants). DESeq2 repose sur un **modèle de comptage négatif binomial** qui corrige les biais liés à la profondeur de séquençage et à la variance inter-échantillons, rendant les comparaisons fiables même avec un effectif modéré.

En pratique, DESeq2 génère trois types de graphiques standards :

- le **MA Plot**, qui visualise les variations d'expression en fonction de l'abondance moyenne des gènes,
- le **Volcano Plot**, qui combine le niveau de changement d'expression et sa significativité statistique,
- et la **Heatmap de clustering**, permettant d'observer les regroupements de profils transcriptomiques similaires.

Pour illustrer son fonctionnement, un **Volcano Plot fictif** a été généré à partir d'un jeu de données simulé dans la figure 5

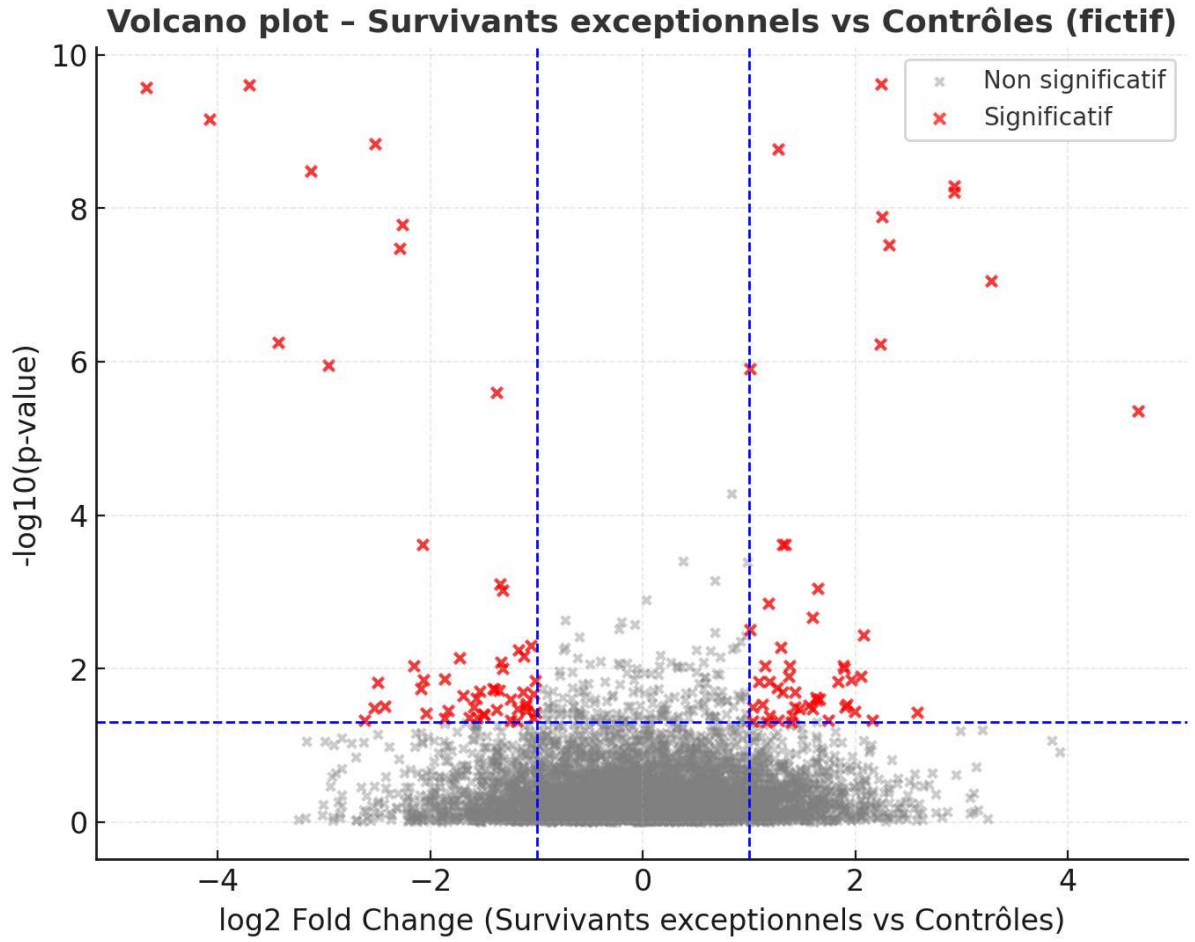

Figure 5 Volcano plot illustrant l'analyse différentielle de l'expression génique entre survivants exceptionnels et contrôles (données simulées)

Dans ce type de représentation :

- chaque **point** correspond à un gène,
- **l'axe X** représente la différence d'expression (log2 Fold Change) entre survivants exceptionnels et contrôles,
- **l'axe Y** indique la significativité statistique (–log10 p-value),
- les **points rouges** traduisent les gènes significativement différentiels (*à gauche* : sous-exprimés chez les survivants exceptionnels, *à droite* : surexprimés),
- les **points gris** représentent les gènes non significatifs,
- les **lignes bleues** matérialisent les seuils classiques de significativité (|log2FC| > 1 et $p < 0,05$).

Cette approche permet d'identifier visuellement les gènes les plus discriminants et d'évaluer la robustesse statistique de leur expression différentielle. Lorsqu'elle est appliquée à des données single-cell ou spatiales, il est recommandé d'agréger les comptages par type cellulaire pour générer des profils "pseudo-bulk" (par exemple, cellules tumorales, macrophages, fibroblastes). Cette stratégie combine la richesse du single-cell avec la robustesse statistique du bulk RNA-seq, évitant la dispersion des signaux sur de trop petits sous-groupes cellulaires.

3.2.4. Analyse des données discrètes : test exact de Fisher

Pour les données discrètes (mutations présentes/absentes, altérations génétiques, amplifications, etc.), le test exact de Fisher s'impose comme la méthode la plus adaptée. Ce test compare la fréquence d'une mutation ou d'un événement binaire entre deux groupes de petite taille, sans hypothèse d'échantillonnage normal. Cependant, lorsqu'on analyse plusieurs centaines ou milliers de gènes ou de mutations en parallèle, il existe un risque que certaines différences apparaissent statistiquement significatives par pur hasard. C'est ce que l'on appelle le problème des tests multiples. Pour corriger ce biais, on applique une correction du risque de faux positifs, le plus souvent à l'aide de la méthode de Benjamini–Hochberg (BH). La méthode de Benjamini–Hochberg repose sur le concept de **False Discovery Rate (FDR)**, ou taux de fausses découvertes. Le FDR correspond à la proportion attendue de résultats faux positifs parmi l'ensemble des résultats déclarés significatifs. Concrètement, un FDR fixé à 5 % signifie que, parmi 100 gènes considérés comme significatifs, environ 5 peuvent en réalité correspondre à des associations dues au hasard. Cette approche permet d'accepter un niveau contrôlé d'erreurs afin de préserver la capacité à détecter des signaux biologiques pertinents. À titre de comparaison, la correction de **Bonferroni** adopte une stratégie beaucoup plus stricte. Elle consiste à ajuster le seuil de significativité en le divisant par le nombre total de tests effectués. Si cette méthode réduit fortement le risque de faux positifs, elle entraîne également une perte importante de sensibilité. Dans le contexte des études omiques, où plusieurs centaines ou milliers de tests sont réalisés, la correction de Bonferroni conduit souvent à éliminer la quasi-totalité des signaux, y compris des associations potentiellement pertinentes sur le plan biologique, en particulier lorsque les effectifs sont limités. Dans ce contexte, la méthode de Benjamini–Hochberg préserve un meilleur équilibre entre rigueur statistique et sensibilité biologique. Elle permet donc de détecter les signaux réels tout en gardant un contrôle raisonnable sur le nombre de faux positifs, ce qui la rend particulièrement adaptée aux études omiques à petit effectif, où chaque signal potentiel doit être exploité avec précaution mais sans excès de conservatisme.

3.2.5. Application aux données single-cell et spatiales

Les technologies de transcriptomique unicellulaire (single-cell RNA-seq) et de transcriptomique spatiale permettent une caractérisation fine des populations cellulaires et de leur organisation dans le microenvironnement tumoral. Elles offrent une richesse d'information considérable, en révélant des sous-populations rares, par exemple, des clones tumoraux particuliers ou des cellules immunitaires activées susceptibles de jouer un rôle clé dans la survie prolongée observée chez les patients exceptionnels. Cependant, cette granularité accrue s'accompagne d'une limite statistique importante : plus les données sont résolues cellule par cellule, plus la taille de chaque sous-groupe est réduite, entraînant une diminution de la puissance statistique et une augmentation du bruit technique. Pour pallier cette contrainte, une approche fréquemment utilisée consiste à regrouper les cellules par type cellulaire(cellules tumorales, macrophages, fibroblastes, lymphocytes, etc.) afin de créer des profils agrégés, appelés "pseudo-bulk".

Définition des termes "bulk" et "pseudo-bulk"

- Le terme bulk RNA-seq désigne les analyses classiques de séquençage d'ARN effectuées sur un mélange de milliers de cellules. Cette approche donne une mesure moyenne de l'expression génique pour l'ensemble du tissu ou de l'échantillon analysé. Elle présente l'avantage d'être statistiquement robuste mais masque l'hétérogénéité cellulaire, car toutes les cellules sont regroupées dans une seule mesure globale.
- À l'inverse, le single-cell RNA-seq permet de mesurer l'expression génique cellule par cellule, révélant les différences entre sous-populations cellulaires. Cependant, le volume de données et la variabilité technique rendent les analyses statistiques complexes, en particulier dans les petits effectifs.
- Le pseudo-bulk est une approche hybride qui combine les avantages des deux méthodes : on regroupe les données du single-cell par type cellulaire ou par patient en additionnant ou moyennant les comptages de gènes pour chaque groupe homogène. Cela permet de réduire le bruit, d'augmenter la stabilité des estimations, tout en préservant la spécificité biologique des populations d'intérêt.

Cette transformation rend possible l'utilisation d'outils statistiques robustes développés pour les données bulk, notamment DESeq2, qui reste l'étalon d'or pour l'analyse différentielle transcriptomique, même lorsqu'il est appliqué à des jeux de données issus du single-cell agrégé. Pour les données discrètes, telles que la présence ou l'absence d'altérations génétiques, le test exact de Fisher, associé à la correction de Benjamini–Hochberg pour contrôler le False Discovery Rate, demeure la méthode la plus pertinente. Ces stratégies permettent ainsi de tirer parti de la richesse biologique du single-cell et du spatial tout en maintenant une rigueur statistique indispensable dans les études à

effectif restreint. Ces stratégies permettent d'exploiter la résolution biologique apportée par les approches single-cell et spatiales tout en s'appuyant sur des cadres statistiques adaptés aux petits effectifs, limitant ainsi l'impact du bruit technique sur l'interprétation des résultats.

Applications concrètes

Ces approches sont illustrées dans l'étude pancréatique de Kamath et al.,(11) qui a montré que les survivants exceptionnels présentaient un nombre réduit de mutations non synonymes *(Figure 5)* et moins d'altérations biologiquement actionnables *(Figure 6)* par rapport aux patients contrôles. Ces observations démontrent qu'une analyse adaptée à petit effectif peut révéler des différences biologiques pertinentes, invisibles dans des modèles de survie standard.

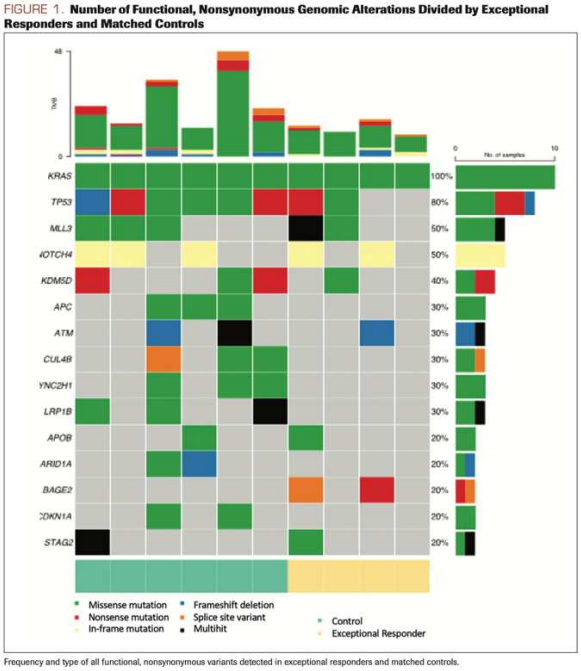

Figure 6 Nombre d'altérations génomiques fonctionnelles non synonymes réparties entre les survivants exceptionnels et les patients contrôles appariés(11)

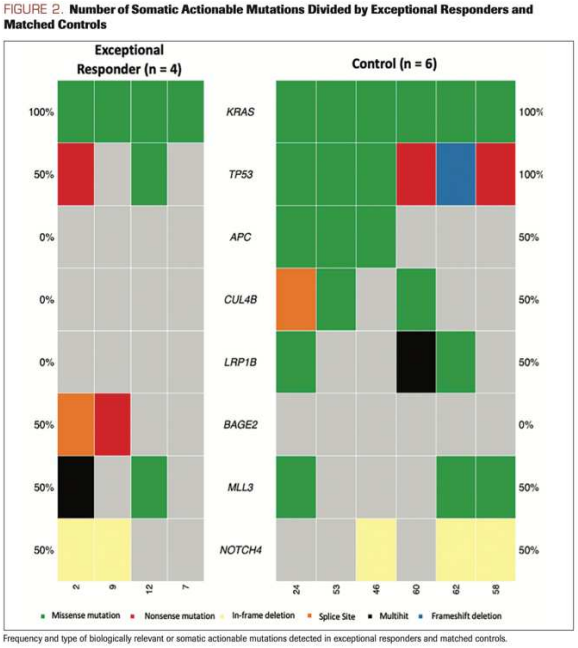

Figure 7 Nombre de mutations somatiques actionnables divisé entre les survivants exceptionnels et les patients contrôles appariés (11)

On peut donc convenir que, dans l'étude des survivants exceptionnels, les approches binaires (case–control) et l'utilisation combinée d'outils comme DESeq2 pour l'expression génique et Fisher pour les données discrètes apparaissent mieux

adaptées que les modèles classiques de survie, dont la puissance statistique est limitée par le faible nombre d'événements observés.

3.3 Intégration des données multi-omiques

L'un des enjeux majeurs dans l'étude des survivants exceptionnels réside dans la capacité à intégrer des données issues de différentes couches biologiques, génomiques, transcriptomiques, épigénétiques, protéomiques et immunologiques. Cette approche vise à dépasser les analyses gène par gène pour identifier des signaux intégrés reflétant la complexité biologique de la tumeur et de son microenvironnement. Les approches multi-omiques reposent de plus en plus sur des modèles statistiques probabilistes capables d'extraire de l'information pertinente à partir d'un grand nombre de variables pour un faible nombre d'échantillons. Parmi ces modèles, le cadre bayésien occupe une place centrale. Il repose sur un principe simple : combiner les données observées avec des connaissances a priori. Autrement dit, plutôt que de considérer les estimations comme fixes, le modèle bayésien met à jour en continu la probabilité qu'une hypothèse soit vraie à mesure que de nouvelles données sont intégrées. Ce raisonnement permet d'intégrer naturellement l'incertitude et d'obtenir des estimations plus stables, ce qui est particulièrement utile dans le cadre de cohortes à petit effectif, comme celles des survivants exceptionnels. Dans ce contexte, le cadre bayésien contribue à pondérer les signaux rares tout en évitant que les résultats soient dominés par le bruit aléatoire inhérent aux petits échantillons.(12)

3.3.1 Réduction de dimension : de la PCA aux approches non linéaires

Les données omiques sont dites à grande dimension, car elles comportent un nombre de variables, souvent plusieurs milliers de gènes ou de protéines, très supérieur au nombre de patients analysés. Une première étape indispensable consiste donc à réduire cette dimensionnalité afin de synthétiser les grandes tendances présentes dans les données et de mettre en évidence les principales sources de variation biologique. L'analyse en composantes principales (PCA) est classiquement utilisée comme première étape de réduction de dimensionnalité en analyse de données omiques (Jolliffe et Cadima, 2016).(13) Elle résume les données en un petit nombre d'axes indépendants, appelés composantes principales, chacun expliquant une part de la variance totale observée dans le jeu de données. Chaque point du nuage de projection correspond à un patient, positionné selon ses coordonnées sur les deux premiers axes (PC1 et PC2). Dans un exemple fictif, les patients contrôles (en rouge) et les survivants exceptionnels (en bleu) se répartissent distinctement dans cet espace bidimensionnel. Cette séparation visuelle suggère que les survivants exceptionnels présentent un profil moléculaire globalement différent de celui des patients témoins.

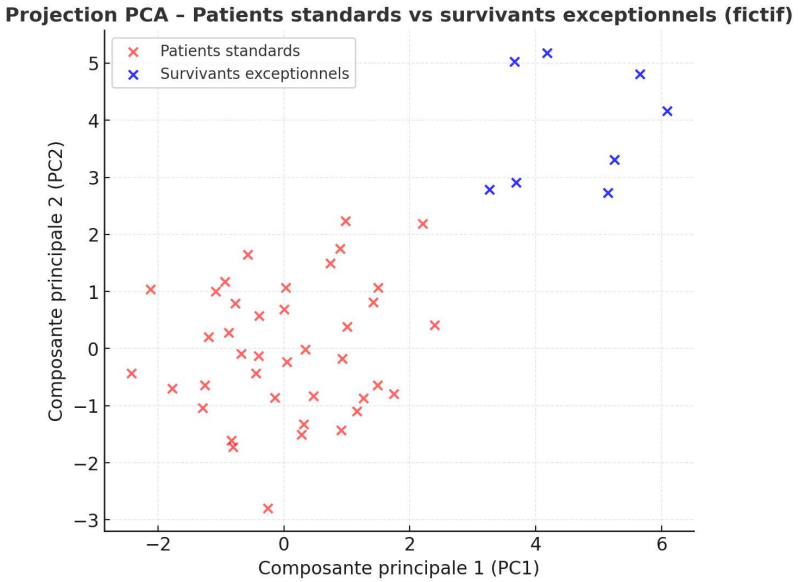

Figure 8 Projection en analyse en composantes principales (PCA) des patients contrôles et des survivants exceptionnels (données simulées)

D'autres méthodes, telles que t-SNE (t-Distributed Stochastic Neighbor Embedding) ou UMAP (Uniform Manifold Approximation and Projection), offrent une projection non linéaire des données multidimensionnelles en deux dimensions. Elles permettent de révéler d'éventuelles structures latentes, sous-groupes ou relations complexes entre échantillons. Ces approches, dites non supervisées, ne prennent pas en compte les étiquettes de groupe (cas ou contrôle). Leur objectif est purement exploratoire : identifier les structures globales présentes dans les données sans a priori, afin d'émettre des hypothèses sur les différences biologiques potentielles entre patients.

3.3.2 Intégration multi-omique : identification

L'un des principaux défis des analyses multi-omiques est d'identifier des signaux cohérents entre différentes couches biologiques. Des outils tels que MOFA (Multi-Omics Factor Analysis), mixOmics ou iClusterBayes permettent de relier plusieurs types de données par exemple le transcriptome, le génome et l'immunome, en extrayant des facteurs latents communs.

Ces méthodes s'appuient sur des modèles factorielles latents, qui considèrent chaque jeu de données comme la combinaison de quelques composantes cachées expliquant les covariations entre variables. Leur avantage est double :

- elles réduisent la dimensionnalité tout en préservant les structures biologiques pertinentes ;
- elles limitent l'inflation des faux positifs liée à la multiplicité des tests statistiques.

Un exemple fictif permet d'en illustrer le principe :

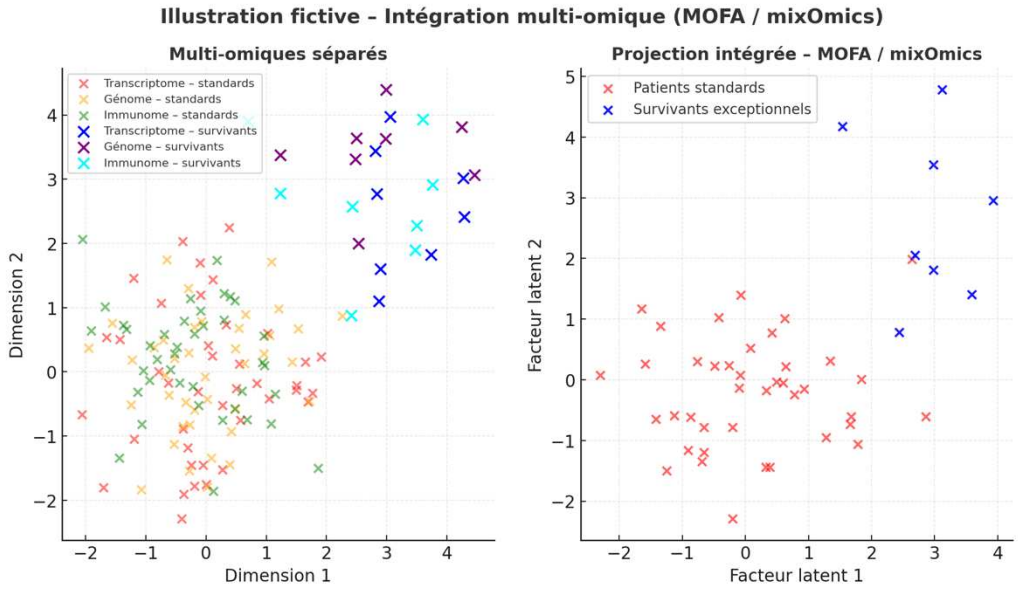

Figure 9 Illustration fictive de l'intégration multi-omique par facteurs latents (MOFA / mixOmics)

- à gauche, les différentes couches omiques analysées séparément (transcriptomique, génomique, immunologique) génèrent des signaux dispersés, difficiles à interpréter de manière intégrée;
- à droite, l'intégration par MOFA ou mixOmics permet de regrouper les survivants exceptionnels (en bleu) dans un espace latent distinct de celui des patients standards (en rouge), traduisant une cohérence multi-omique de leur profil.

Plus récemment, des approches bayésiennes supervisées, comme SPEAR (14)ont été proposées pour analyser les données multi-omics et realiser prédiction de phénotype

dans un même cadre probabiliste. Concrètement, SPEAR vise à combiner plusieurs couches de données biologiques (par exemple génomiques et transcriptomiques) afin de prédire un phénotype d'intérêt, comme la réponse thérapeutique ou la survie, tout en tenant compte de l'incertitude liée au faible nombre d'échantillons. Ces méthodes illustrent une évolution conceptuelle vers des approches plus directement orientées vers la prédiction clinique. Toutefois, leur application reste encore largement exploratoire dans des contextes de petits effectifs et n'a pas été évoquée par les chercheurs interrogés dans le cadre de ce travail.

3.3.3. Méthodes supervisées et non supervisées

Les approches non supervisées (PCA, MOFA, mixOmics, iCluster) explorent les données à détecter des structures globales. Les approches supervisées, quant à elles, utilisent les étiquettes connues (survivant exceptionnel vs contrôle) pour identifier les variables les plus discriminantes. Parmi celles-ci, les méthodes LASSO (Least Absolute Shrinkage and Selection Operator) et Partial Least Squares (sPLS) sont particulièrement adaptées aux données de haute dimension.

Dans le contexte des petits effectifs, comme celui des survivants exceptionnels, la stratégie la plus adaptée repose sur une approche méthodologique en deux temps. Dans un premier temps, une analyse exploratoire non supervisée (telle que MOFA, mixOmics ou iCluster) permet d'identifier les grands axes de variation biologique partagés entre les différentes couches omiques, sans a priori sur le phénotype étudié. Cette étape vise à faire émerger des structures globales et des signaux convergents au sein des données.

Dans un second temps, une analyse supervisée et parcimonieuse (par exemple LASSO ou sPLS) est utilisée pour isoler un nombre restreint de variables réellement discriminantes entre groupes, tout en conservant une interprétabilité biologique. Cette combinaison d'approches permet de réduire la complexité des données, de limiter le risque de faux positifs et de prioriser des candidats biologiques robustes en vue de validations ultérieures. Elle constitue ainsi un cadre méthodologique pertinent pour hiérarchiser les signaux et dégager des pistes mécanistiques exploitables, tout en respectant les contraintes statistiques imposées par la rareté des survivants exceptionnels.(12)

3.3.4. Limiter le sur-apprentissage

L'un des principaux écueils de l'analyse de petites cohortes, comme celles des survivants exceptionnels, réside dans le risque de surapprentissage (overfitting). Ce phénomène correspond à la construction de modèles qui décrivent très précisément les données observées, mais dont les performances se dégradent fortement lorsqu'ils sont appliqués à d'autres ensembles d'échantillons. Dans ce cas, le modèle capture davantage le bruit aléatoire que les signaux biologiques réellement informatifs.

Pour limiter ce biais, une vigilance méthodologique accrue est indispensable. La réduction de dimension constitue une première étape clé, en permettant de concentrer l'analyse sur les structures biologiques les plus stables et reproductibles, plutôt que sur des variations locales ou spécifiques à un échantillon. Les approches factorielles non supervisées, telles que la PCA, MOFA, mixOmics ou iCluster, s'inscrivent dans cette logique en résumant la variabilité biologique en un nombre limité d'axes dominants partagés entre les différentes couches omiques. Par ailleurs, la correction des tests multiples doit être systématiquement appliquée afin de limiter la détection de faux signaux statistiques. Dans ce contexte, la méthode de Benjamini–Hochberg, fondée sur le contrôle du taux de fausses découvertes (False Discovery Rate, FDR), représente un compromis adapté entre rigueur statistique et sensibilité biologique. À l'inverse, des corrections plus strictes comme celle de Bonferroni tendent à éliminer un grand nombre de signaux potentiellement pertinents lorsqu'elles sont appliquées à des analyses de grande dimensionnalité.

Enfin, la validation des résultats constitue l'étape la plus déterminante pour évaluer la robustesse d'un modèle. La validation croisée permet d'en tester la stabilité interne, à condition de séparer strictement les phases de sélection des variables et d'évaluation des performances. Lorsque cela est possible, la validation externe apporte un niveau de confiance supplémentaire, en confrontant les résultats à des bases de données indépendantes telles que The Cancer Genome Atlas (TCGA) ou l'International Cancer Genome Consortium (ICGC). TCGA est un programme international de référence ayant généré des données multi-omiques (génomiques, transcriptomiques, épigénomiques et cliniques) sur plusieurs milliers de tumeurs issues de nombreux types de cancer. ICGC est un consortium international visant à cartographier de manière systématique les altérations génomiques des cancers à l'échelle mondiale. Bien que ces ressources ne comprennent pas spécifiquement de survivants exceptionnels, elles offrent des référentiels précieux pour apprécier la plausibilité biologique et la généralisabilité des signatures identifiées.

3.4. Entretien avec les experts

Cette section synthétise et confronte les positions de deux experts interrogés dans le cadre de ce travail : **Rémy Nicolle**, spécialiste en génomique translationnelle et analyses multi-omiques appliquées à l'oncologie, UMRS1149 CRI , et **Samuel Blanck**, biostatisticien à l'EA2694, expert des méthodes statistiques pour petites cohortes. Leurs propos permettent de consolider les décisions méthodologiques adoptées pour l'étude des survivants exceptionnels.

1. Pertinence des analyses de survie dans le contexte des survivants exceptionnels

Le Dr Rémy Nicolle rejette d'emblée l'idée d'utiliser un modèle de Cox ou une analyse de survie classique dans ce contexte :

« *Kaplan–Meier, c'est un plot, pas un test. Pour analyser une survie, il faut un modèle de Cox, mais dans notre cas ça n'a aucun sens : on a deux groupes, survivants vs contrôles.* »

Il insiste sur l'inadéquation d'un modèle basé sur le temps jusqu'à événement lorsque l'étude repose sur une comparaison binaire.

Samuel Blanck confirme la non-pertinence de ces modèles dès lors qu'il existe deux groupes clairement définis.

2. Méthodes d'analyse différentielle : usage de DESeq2 sur petits effectifs

Le Dr Rémy Nicolle met fortement en avant DESeq2 :

« *DESeq2, c'est vraiment ce qu'on a de plus robuste.* »

Selon lui, l'intérêt de DESeq2 s'accentue lorsqu'on applique une stratégie de pseudo-bulk, permettant d'utiliser des outils éprouvés même à partir de données single-cell.

Même constat de monsieur Blanck :

« *Pour les petites cohortes, DESeq2 et les tests modérés sont les plus efficaces. Ils stabilisent la variance et limitent les faux positifs.* »

Il rappelle le principe méthodologique : la variance est modélisée en partageant l'information entre gènes similaires, ce qui améliore la fiabilité statistique lorsque n est faible.

3. Données single-cell et transformation pseudo-bulk

Le Dr Rémy Nicolle décrit une stratégie éprouvée et déjà utilisée dans son équipe :

« *Quand tu fais du single-cell, tu transformes en bulk : tu prends les cellules tumorales, les macrophages, les fibroblastes… et tu appliques DESeq2. C'est le plus robuste.* »

Cette démarche permet de contourner l'insuffisance du n effectif par sous-population cellulaire.

Pour monsieur Blanck, bien qu'il n'ait pas pratiqué directement le single-cell, il valide totalement la logique :

« *L'important est d'avoir un modèle statistique stable. Si les données sont comparables, DESeq2 peut être appliqué.* »

4. Analyses sur données discrètes : test exact de Fisher et correction FDR

Le Dr Rémy Nicolle est très clair :

« *Pour les données discrètes, tu fais un test de Fisher, puis une correction. Benjamini–Hochberg évidemment.*»

Même position pour Monsieur Blanck:

« *Pour mutations et altérations, Fisher + FDR est indispensable, vu le nombre de tests.* »

5. Correction des tests multiples : choix incontournable du FDR

Le Dr Rémy Nicolle écarte explicitement Bonferroni :

« *Bonferroni est trop violent. On utilise tous Benjamini–Hochberg.* »

Monsieur Blanck renforce exactement la même idée :

« *La correction FDR est essentielle pour conserver du pouvoir statistique.* »

6. Réduction de dimension et intégration multi-omique

Le Dr Rémy Nicolle précise que toutes les méthodes modernes d'intégration reposent, explicitement ou non, sur des variantes de PCA :

« *Ces outils, MOFA, mixOmics, iCluster… c'est toujours une PCA plus ou moins supervisée. On cherche des signaux covariants.* »

Pour lui, l'objectif n'est pas de trouver « un gène isolé » mais des **signaux globaux robustes**, moins sensibles au bruit statistique.

Monsieur Blanck adopte une position plus prudente :

« *Je connais mixOmics, mais je n'ai pas de retour pratique. Sur petit effectif, mieux vaut rester simple et robuste.* »

7. Risque de surapprentissage et stratégies de limitation

Le Dr Rémy Nicolle insiste sur le danger principal :

« *Quand tu fais 3 vs 3, tu peux dire n'importe quoi. Il faut chercher un signal global, pas un gène isolé.* »

Monsieur Blanck formule la même préoccupation en termes méthodologiques :

« *Toujours séparer apprentissage et test. Validation croisée répétée. Et idéalement validation externe.* »

8. Validation externe et usage des bases publiques

Le Dr Rémy Nicolle est catégorique :

« *La vraie solution contre le surapprentissage, ce sont les cohortes externes.* »

Il recommande explicitement l'usage de :

« *TCGA/ICGC comme contrôles.* »

Même position pour monsieur Blanck:

« *L'idéal est de tester sur une cohorte indépendante. TCGA ou ICGC sont de bons référentiels de contrôles.* »

9. Taille d'échantillon et implications méthodologiques

Le Dr Rémy Nicolle rappelle que même des études multi-omiques avec 30 patients peuvent être exploitables, à condition d'utiliser la bonne méthodologie :

« *La réduction de dimension, c'est ce qui marche le mieux.* »

Monsieur Blanck considère que la cohorte envisagée (~200 vs 200) est « déjà très confortable », mais souligne :

« *La correction FDR et la validation restent indispensables.* »

Conclusion générale de la discussion

L’analyse croisée des entretiens montre une convergence marquée entre les deux experts. Malgré des profils différents, leurs recommandations dessinent un cadre méthodologique clair, cohérent et adapté aux spécificités de l’étude des survivants exceptionnels.

4. Données et validation biologique

4.1. Objectif :

L'étude des survivants exceptionnels s'appuie sur l'identification de signatures multi-omiques, Toutefois, pour qu'une telle signature puisse prétendre à une valeur biologique ou clinique, elle doit être validée à plusieurs niveaux c'est-à-dire au niveau analytique, fonctionnel, biologique, et enfin translationnel. Cette démarche, qui vise à transformer une observation statistique en preuve biologique, est au cœur de la médecine de précision moderne (10)(Hasin et al., *Genome Biology*, 2017). Dans ce contexte, l'expertise du Dr Paloma Cejas, chercheuse translationnelle spécialisée dans les approches multi-omiques appliquées à des cohortes cliniques complexes, a constitué un appui précieux pour nourrir la réflexion méthodologique présentée dans cette partie. Son regard critique et son expérience ont permis de mettre en perspective les choix méthodologiques discutés, ainsi que d'en aborder les limites et les enjeux translationnels, sans prétendre à une validation exhaustive.

4.2. Découverte de la signature multi-omique

L'approche multi-omique combine plusieurs couches de données comme on l'a vu précédemment par exemple, mutations génétiques (WES/WGS), profils d'expression génique (RNA-seq), et niveaux protéiques afin de capturer la complexité du phénotype tumoral. Cette intégration permet d'identifier des signatures moléculaires spécifiques associées à la résistance, la sensibilité ou la survie exceptionnelle.

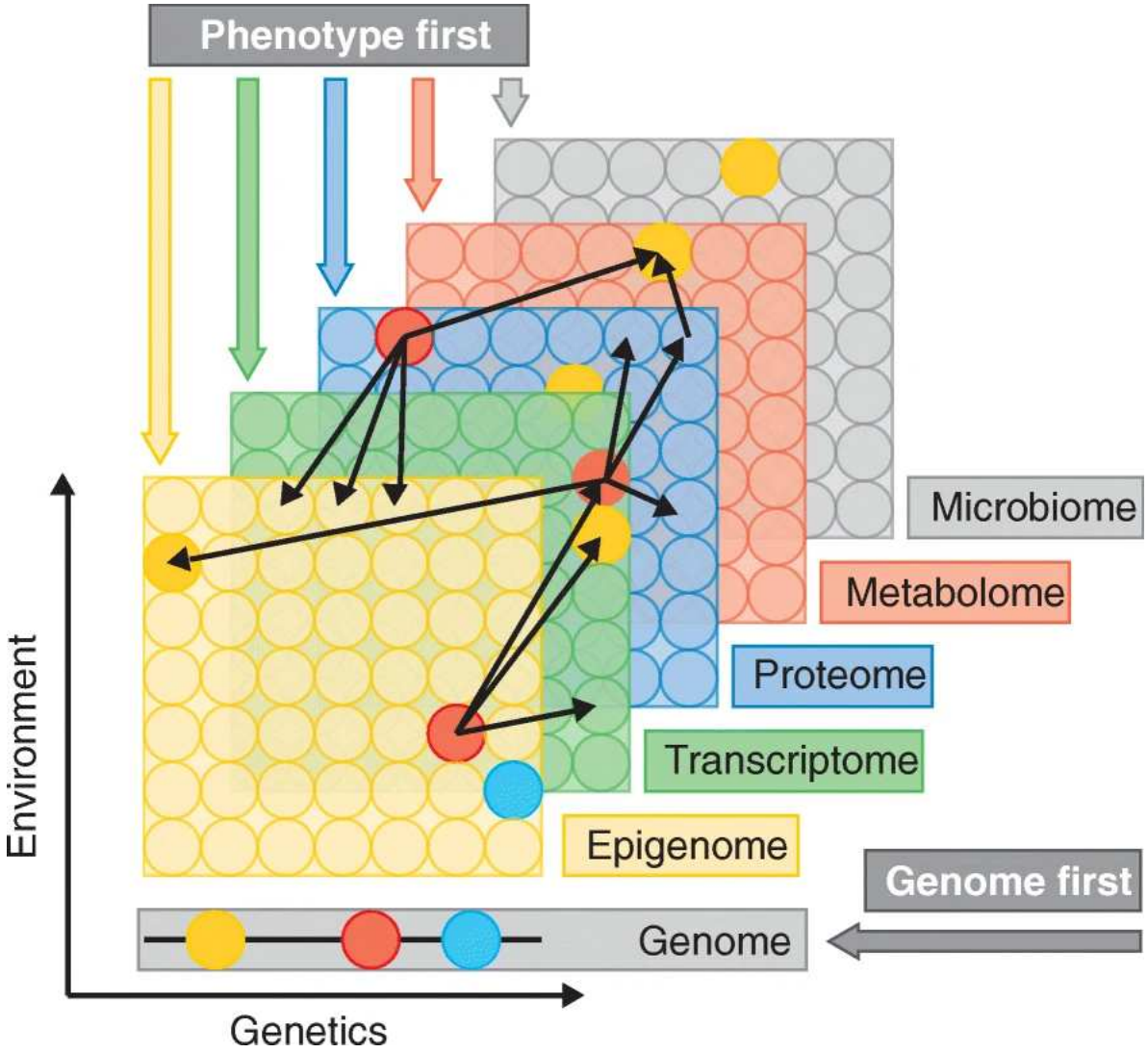

Figure 10 Représentation schématique de l'intégration des différentes couches multi-omiques dans l'étude des phénotypes complexes (10)

Hasin et al. (10)(ont démontré que l'intégration de données multi-omiques améliore la robustesse des découvertes en cancérologie, en reliant des altérations génétiques à des phénotypes observables. Plus récemment, (*Jennifer J. Knox, Gun Ho Jang, 2025)* (15)ont utilisé une approche similaire pour décrire les mécanismes de longévité dans le cancer du pancréas, révélant que la convergence d'altérations dans les voies de réparation de l'ADN et la réponse immunitaire adaptative pouvait expliquer certaines survies prolongées. Dans la littérature récente, plusieurs exemples concrets illustrent cette

valeur ajoutée des approches multi-omiques. Dans une première étude (16) un score élevé de cellules immunitaires cytotoxiques était associé à une survie plus longue, tandis qu'un score important de fibroblastes associés au cancer (CAF) corrélait avec une progression tumorale plus rapide. Ce type de combinaison de marqueurs immunitaires et stromaux montre qu'il est possible d'identifier des signatures pronostiques multi-omiques mesurables, parfois même à partir d'échantillons sanguins.

Dans l'étude *"Integrative multi-omics analysis reveals the role of toll-like receptor signaling in pancreatic cancer"* (Scientific Reports, 2025) (17) On s'est intéressée au rôle des Toll-like receptors (TLR). Les TLR sont des récepteurs de l'immunité innée capables de détecter des signaux de danger et d'activer des voies inflammatoires. En cancérologie, leur effet est ambivalent c'est-à-dire qu' une activation contrôlée peut soutenir l'immunité antitumorale, tandis qu'une activation excessive peut favoriser la croissance tumorale. Dans l'étude citée, certains TLR étaient fortement exprimés dans des cellules immunitaires et endothéliales, et leur activation modérée semblait associée à une meilleure évolution clinique. La signature identifiée incluant NT5E, TGFBI, ANLN et FAM83A regroupe des gènes impliqués respectivement dans la modulation de l'immunité et du microenvironnement tumoral (NT5E, TGFBI), la dynamique du cytosquelette et la prolifération cellulaire (ANLN), ainsi que dans des processus de signalisation oncogénique et de plasticité tumorale (FAM83A). Cette combinaison illustre comment des signaux immunitaires et transcriptionnels combinés peuvent caractériser un sous-groupe de patients à survie exceptionnelle.

Ces deux exemples, bien qu'issus de contextes différents, convergent vers la même conclusion, les signatures multi-omiques validées, qu'elles proviennent du plasma ou du tissu tumoral, permettent de relier des mécanismes biologiques mesurables (immunité, inflammation, signalisation cellulaire) à des phénotypes cliniques tels que la survie prolongée.

4.3. Validation analytique : robustesse et fiabilité du signal

Une fois la signature identifiée, il est essentiel de vérifier qu'elle résulte d'un phénomène biologique réel et non d'un artefact expérimental. Cette étape de validation analytique vise à évaluer la répétabilité, la reproductibilité inter-plateformes et la robustesse statistique du signal. En effet, les biais techniques liés à la préparation des échantillons, aux effets de lot ("batch effects") ou aux différences de plateformes de séquençage peuvent fausser les résultats et conduire à de fausses interprétations biologiques. Pour corriger ces biais et harmoniser les données entre laboratoires, des outils comme ComBat sont couramment utilisés. Cet algorithme ajuste mathématiquement les écarts dus aux conditions expérimentales (différences de machine, de date ou de site d'analyse), tout en préservant les différences biologiques réelles entre les groupes étudiés, par exemple entre survivants exceptionnels et témoins.
Son utilisation permet d'obtenir des jeux de données plus homogènes, comparables et fiables, condition indispensable avant toute validation biologique ou intégration multi-omique. À titre d'illustration, une étude récente sur le cancer gastrique associé à *Helicobacter pylori* publié par le Oxford university press (18)a utilisé l'algorithme ComBat pour intégrer des données issues de trois jeux de microarrays (GSE27411, GSE60427, GSE233973) et d'un jeu RNA-seq (TCGA-STAD).

- **Les microarrays** sont des lames contenant des milliers de sondes d'ADN. On y dépose l'ARN extrait d'un échantillon, et plus un gène est exprimé, plus le signal fluorescent associé à sa sonde sera intense. C'est une technologie ancienne, fiable mais limitée : elle ne mesure que ce qui est déjà connu et dépend beaucoup de la qualité des sondes.
- **Le RNA-seq**, au contraire, séquence directement toutes les molécules d'ARN présentes dans l'échantillon. C'est une méthode plus récente, plus sensible, qui permet de détecter aussi bien des gènes très exprimés que très faiblement exprimés, et même des transcrits inconnus.

Bien que le RNA-seq soit désormais la technologie de référence, les microarrays restent utilisés en raison de la disponibilité de vastes jeux de données historiques et de leur intérêt dans des analyses intégratives ou de validation, sous réserve d'une correction rigoureuse des biais techniques.

Ces deux technologies produisent donc des données qui ne sont pas directement comparables, car le type de signal, son échelle, sa sensibilité et ses sources de bruit sont très différents. Sans correction statistique, ces différences techniques peuvent masquer les véritables variations biologiques liées à la maladie. C'est précisément pour cela que des algorithmes comme ComBat sont utilisés. ComBat ajuste les données pour corriger ces différences techniques appelées *effets de lot* ou plus couramment batch effect, tout en conservant les variations biologiques réelles. Autrement dit, il rend les mesures issues

de microarrays et de RNA-seq comparables sur une même échelle, sans effacer les différences dues à l'évolution de la maladie.

Dans l'étude prise ici comme exemple, cette harmonisation a permis d'observer des profils d'expression cohérents pour des gènes tels que TPX2, MKI67, EXO1 et CTHRC1, dont l'expression augmentait progressivement au fil des stades de la maladie (infection → gastrite → atrophie → cancer). Cela montre que ComBat n'a pas altéré le signal biologique d'intérêt mais au contraire, il a rendu la comparaison entre plateformes plus fiable et a renforcé la robustesse statistique de l'analyse.

D'autre part la validité analytique d'une signature repose également sur sa capacité à être reproduite sur des jeux de données indépendants. De plus au-delà des aspects biologiques, la robustesse d'une signature multi-omique dépend également des outils statistiques utilisés pour analyser les données. Parmi eux, l'un des outils les plus largement adoptés en génomique est le **package limma**, développé d'abord pour les microarrays puis étendu au RNA-seq. Limma repose sur des modèles linéaires associés à une approche dite *empirical Bayes*, particulièrement utile lorsque les cohortes sont petites, une situation fréquente dans les études portant sur des survivants exceptionnels. L'approche *empirical Bayes* consiste à « emprunter de l'information » à l'ensemble des gènes pour stabiliser les estimations obtenues pour chaque gène pris individuellement. Dans les petites cohortes, les variances calculées gène par gène sont souvent très instables ; l'empirical Bayes ajuste ces variances en les ramenant vers une valeur moyenne estimée sur l'ensemble du transcriptome. Cela réduit le bruit, augmente la puissance statistique et permet d'identifier des signaux fiables même lorsque **n** est faible. Autrement dit, cette approche diminue le risque de faux positifs et de faux négatifs liés à la faible taille d'échantillon. Dans leur publication de référence (*Nucleic Acids Research*, 2015), Ritchie et al.(19) précisent que limma a été conçu pour *« promouvoir une recherche reproductible en génomique »*

Cette ambition se traduit concrètement par plusieurs caractéristiques :

- la possibilité d'appliquer la même stratégie d'analyse à différents jeux de données (microarrays, RNA-seq via voom).
- la stabilisation des variances dans les petits effectifs grâce à l'empirical Bayes,
- une intégration dans des pipelines standardisés permettant de réanalyser des jeux externes sans modifier la méthodologie.

Ces propriétés font de limma un outil central pour tester la robustesse d'une signature. Ainsi, une signature identifiée dans une cohorte interne de survivants exceptionnels par exemple une signature liée à la réparation de l'ADN doit être réévaluée dans des bases indépendantes telles que **TCGA** ou **ICGC**. La capacité à retrouver le même signal dans

des données externes constitue une étape essentielle pour s'assurer que la signature ne reflète pas un artefact propre à la cohorte d'origine, mais bien un mécanisme biologique généralisable.

4.4. Interprétation fonctionnelle : du signal au sens biologique

Une signature multi-omique n'acquiert sa valeur scientifique que si elle peut être reliée à un mécanisme biologique identifiable. Cette interprétation biologique repose classiquement sur des analyses d'enrichissement fonctionnel et de pathways, ces derniers correspondant à des ensembles structurés de gènes ou de protéines impliqués dans une même voie biologique, métabolique ou de signalisation. Ces analyses permettent de relier les listes de gènes identifiées à des voies moléculaires et des processus biologiques connus (Hasin et al., 2017)(10). Dans ce cadre, les outils largement utilisées incluent :

- **Gene Ontology (GO)** pour l'annotation des processus biologiques, fonctions moléculaires et composants cellulaires (Gene Ontology Consortium, *Nature Genetics*, 2021) ;
- **Reactome**, base de données de voies biochimiques et de signalisation (Fabregat et al., *Nucleic Acids Research*, 2018) ;
- **KEGG**, qui relie les gènes à des voies métaboliques ou pathologiques ;
- **MSigDB**, notamment ses "Hallmark gene sets" (Liberzon et al., *Cell Systems*, 2015) ;
- **GSEA**, méthode statistique d'enrichissement fonctionnel (Subramanian et al., *PNAS*, 2005).

KEGG (Kanehisa & Goto, 2000) est une base de connaissances permettant de relier un ensemble de gènes à des voies représentées sous forme de réseaux métaboliques et de cascades de signalisation. L'ambition de KEGG est de *« lier l'information génomique à l'information fonctionnelle de plus haut niveau »(20)* Pour mieux comprendre, voici un schéma (*Figure 10)* tiré de l'article *KEGG: (Kyoto Encyclopedia of Genes and Genomes)*

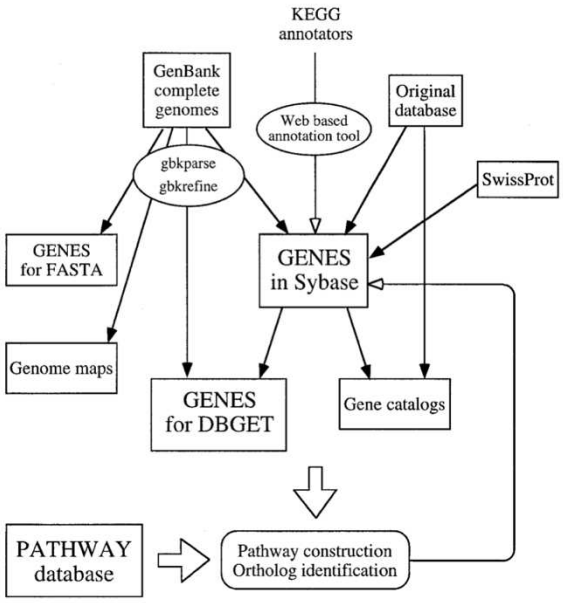

Figure 1. Procedures used to organize and annotate the GENES database.

Figure 11 Procédures utilisées pour organiser et annoter la base de données GENES de KEGG (20)

Celui-ci illustre comment KEGG construit ses voies à partir de sources génomiques fiables.

Tout part des génomes complets déposés dans GenBank, qui contiennent la liste exhaustive des gènes d'un organisme. Ces informations sont extraites et standardisées, puis regroupées dans une base centrale appelée GENES, qui est le cœur de KEGG. À cette étape, des annotateurs experts, par annotateurs experts, on entend des biologistes et bio-informaticiens spécialisés dans la curation de données, chargés de valider et d'enrichir manuellement les annotations fonctionnelles des gènes à partir de données expérimentales publiées et de ressources de référence, enrichissent la base à l'aide d'informations issues d'autres ressources comme SwissProt ou des catalogues de gènes déjà connus. Une fois ces gènes correctement annotés, KEGG les utilise pour reconstruire les voies biologiques : il identifie quels gènes sont orthologues entre différentes espèces, quelles enzymes ils codent, et comment ces enzymes interagissent dans des réactions biochimiques. C'est cette étape représentée en bas du schéma ce qui permet de générer les pathway maps, c'est-à-dire les diagrammes de voies de signalisation et de métabolisme qui font la spécificité de KEGG.

MSigDB (21)regroupe des milliers de jeux de gènes classés en ensembles cohérents appelés *gene sets*. Ces ensembles peuvent représenter : des voies canoniques, des signatures publiées ou des processus biologiques hautement conservés, comme

les *Hallmarks* (inflammation, interféron, hypoxie, réparation de l'ADN…). L'intérêt de MSigDB est de fournir des catégories fonctionnelles standardisées pour interpréter biologiquement une signature multi-omique.

GSEA (22) est l'outil statistique utilisé pour tester si un *gene set* est globalement activé ou réprimé dans une condition donnée. Contrairement aux approches gène-par-gène, GSEA évalue si les gènes d'un ensemble se concentrent parmi les plus exprimés (activation) ou les moins exprimés (répression). Cette approche est particulièrement utile dans les données bruitées ou de petite taille, car elle détecte des signaux faibles mais cohérents entre gènes.

Ainsi, pour mieux comprendre l'utilisation concrète de ces outils, l'étude *Integrative multi-omics analysis reveals the role of Toll-like receptor signaling in pancreatic cancer* (Nature Communications, 2025) (17)constitue une illustration parfaite de l'utilisation qui peut être fait de ces outils. Dans ce travail multi-omique incluant transcriptomique, single-cell et spatial transcriptomics, les auteurs analysent une signature centrée sur la voie des Toll-like receptors (TLR). Pour interpréter cette signature :

- **GO** est utilisé pour identifier les processus biologiques associés, révélant une activation marquée des réponses inflammatoires et de l'immunité innée ;
- **KEGG** permet de relier les gènes identifiés à des voies immunitaires et métaboliques, notamment celles du récepteur des cellules T et B, de l'apoptose et du métabolisme énergétique ;
- **MSigDB** met en évidence l'activation coordonnée de signatures *Hallmark* telles que *Inflammatory Response* et *Interferon Signaling*, confirmant un programme inflammatoire amplifié dans certains sous-types de patients.

.

4.5. Validation biologique expérimentale

Une fois qu'une signature multi-omique a été interprétée biologiquement, elle doit être validée expérimentalement pour confirmer qu'elle correspond bien à un mécanisme fonctionnel et non à une simple corrélation statistique. Cette validation repose généralement sur trois niveaux complémentaires.

Validation sur d'autres cohortes de patients.

Cette étape permet de vérifier que la signature apparaît également dans des jeux de données indépendants (par exemple TCGA ou ICGC). Si le même signal est retrouvé dans plusieurs cohortes, la signature gagne en solidité et en généralisabilité.

Validation in vitro.

Les lignées cellulaires, comme celles référencées dans le Cancer Cell Line Encyclopedia (CCLE), servent à tester directement l'effet des gènes ou des voies identifiés. Il est par exemple possible d'inhiber un gène clé à l'aide de technologies d'interférence génique (siRNA) ou d'édition génomique de type CRISPR-Cas9, une méthode permettant d'inactiver de manière ciblée un gène en coupant son ADN. Les conséquences de cette inhibition peuvent alors être évaluées sur la survie, la prolifération ou la sensibilité des cellules tumorales aux traitements.

Validation in vivo.

Les modèles animaux, tels que les xénogreffes dérivées de patients (PDX), permettent d'évaluer si la signature influence réellement la croissance tumorale, la réponse immunitaire ou la sensibilité thérapeutique dans un environnement biologique complet. Dans l'étude multi-omique consacrée à la voie TLR dans le cancer du pancréas (15), cette stratégie a été illustrée concrètement : après avoir identifié transcriptionnellement une activation de la voie TLR, les auteurs ont confirmé son rôle fonctionnel via des analyses cellulaires et tissulaires montrant que cette activation était associée à une dérégulation immunitaire mesurable au niveau des cellules tumorales et de leur microenvironnement. Ce type d'approche, dont le cœur repose sur des modèles in vivo tels que les xénogreffes dérivées de patients, s'appuie également sur des validations complémentaires in vitro et sur l'analyse de données humaines. Cette combinaison permet de passer d'une association statistique à une démonstration mécanistique cohérente. Ainsi, la validation expérimentale, en particulier lorsqu'elle inclut des modèles in vivo, constitue une étape clé pour transformer une signature multi-omique en hypothèse biologique testable.

4.6 Entretien avec l'experte

Présentation de l'experte

Un entretien a pu être réalisé avec le Dr Paloma Cejas, chercheuse translationnelle, formée en Espagne (PhD soutenue en 2003), puis recrutée au Dana-Farber Cancer Institute, centre de référence affilié à la Harvard Medical School, où elle a développé une expertise centrée sur l'analyse moléculaire d'échantillons cliniques afin de répondre à des questions directement médicales. Aujourd'hui membre de la faculté du Dana-Farber en tant que Head of Innovation, elle dirige un laboratoire utilisant des approches multi-omiques à visée translationnelle, avec un intérêt particulier pour l'étude de tissus cliniques de type FFPE (formalin-fixed paraffin-embedded). Elle souligne que ces échantillons, historiquement considérés comme difficiles à exploiter en raison de leur dégradation, deviennent de plus en plus informatifs grâce aux méthodologies récentes, notamment le single-cell et la transcriptomique spatiale, qui permettent d'analyser finement les cellules tumorales et leur microenvironnement.

L'entretien ayant été mené en anglais, une retranscription aussi fidèle que possible du point de vue de l'experte a été réalisée, en intégrant ses citations directes lorsque cela était pertinent.

Quels critères sont essentiels pour distinguer une signature biologiquement pertinente d'un signal purement statistique ou d'un artefact ?

Paloma Cejas insiste d'abord sur la nécessité d'une posture critique face aux résultats, rappelant qu'il est relativement facile de produire des listes de gènes différentiellement exprimés et d'en tirer une interprétation a posteriori. Elle souligne que la première barrière contre les artefacts réside dans la qualité des échantillons et des données, en particulier dans les approches single-cell et spatiales. Comme elle l'exprime clairement :

"The first thing is to make sure that the quality of the sample is good enough to put it in the analysis, because otherwise all the data are biased."

Elle rappelle que des pipelines d'analyse existent, mais que leur validité dépend fortement du respect de contrôles qualité stricts. Travailler avec des échantillons de mauvaise qualité revient, selon elle, à accepter des biais structurels qui compromettent toute interprétation ultérieure.

Au-delà du contrôle qualité, elle souligne qu'une signature biologiquement crédible repose sur une convergence d'indices. Elle privilégie les résultats qui se répètent entre

patients, plutôt que ceux observés dans un cas isolé, car l'objectif est d'identifier des mécanismes partageables. Elle insiste également sur l'importance de croiser plusieurs méthodes analytiques et de vérifier que le signal est retrouvé par différents angles :

"You need to see that it is consistently observed, not only with one methodology, but from different angles."

Enfin, elle accorde une grande importance à la cohérence avec des observations tissulaires directes. Certains résultats moléculaires peuvent être corroborés par l'histologie classique, par exemple une infiltration lymphocytaire ou une abondance de fibroblastes visibles sur les lames, ce qui renforce fortement la plausibilité biologique :

"If you see more lymphocytes or fibroblasts, you can even see it histologically. That is a very big proof."

Pour elle, la robustesse d'une signature repose donc sur un ensemble cohérent combinant qualité des données, répétition inter-échantillons, validation croisée par plusieurs approches, et ancrage solide dans la biologie et la littérature existante.

Validation dans des cohortes indépendantes et bases publiques

La validation dans des cohortes indépendantes et des bases publiques (TCGA, ICGC) est-elle un prérequis avant l'interprétation approfondie ?

Selon Paloma Cejas, la validation dans des cohortes indépendantes est toujours souhaitable, notamment pour tester la stabilité d'un signal au-delà d'un contexte local. Elle souligne l'avantage des cohortes internationales, qui permettent de vérifier si un résultat persiste dans des arrière-plans cliniques et génétiques variés :

"It's very important to validate in independent cohorts, especially because patients come from different backgrounds."

Concernant les grandes bases publiques comme TCGA, elle en souligne les limites méthodologiques, en particulier l'absence de résolution unicellulaire, qui restreint leur capacité à valider directement des signaux issus du single-cell ou de la transcriptomique spatiale. Néanmoins, elle explique les utiliser très régulièrement comme point de comparaison, notamment pour explorer des associations avec la survie ou vérifier si un gène ou une signature observée se retrouve dans une population plus large :

"Whenever I look at a result, I am always comparing it to TCGA."

Elle insiste toutefois sur un principe méthodologique clé : lorsque l'objectif est de valider un signal issu d'une technologie spécifique, la comparaison la plus pertinente consiste à utiliser des jeux de données indépendants générés avec une méthodologie similaire, souvent disponibles via les dépôts associés aux publications scientifiques. Autrement

dit, TCGA est utile comme référence générale, mais la validation la plus stricte repose sur des datasets externes comparables en termes de résolution et de design expérimental.

Une fois une signature multi-omique identifiée, quelles approches sont les plus pertinentes pour en comprendre le sens biologique ?

Paloma Cejas décrit l'analyse d'enrichissement fonctionnel comme une étape essentielle pour passer d'une liste de gènes à une hypothèse mécanistique cohérente. L'objectif est d'identifier si les gènes différentiellement exprimés reflètent un nombre limité de programmes biologiques dominants, tels que la prolifération, la réparation de l'ADN, l'inflammation ou la réponse immunitaire :

"You want to know what are the pathways they are representing, because they may all be representing the same pathway."

Ce type d'analyse permet de proposer une lecture mécanistique globale, par exemple des contrôles dominés par des voies de prolifération, tandis que les survivants exceptionnels pourraient présenter une activation de programmes immunitaires contribuant au contrôle tumoral.

Elle précise toutefois que l'interprétation ne se limite pas à l'application d'un pipeline standardisé. Une fois les voies suggérées, l'étape déterminante reste la mise en contexte biologique : lecture approfondie de la littérature, compréhension des mécanismes connus, et jugement de biologiste du cancer. Elle évoque l'apport potentiel des outils d'IA pour orienter l'analyse, tout en soulignant leurs limites :

"AI tools can help, but they don't substitute knowledge. You need criteria to understand what is written there."

Quand plusieurs voies sont enrichies simultanément, comment décider lesquelles explorer en priorité ?

Sur ce point, Paloma Cejas souligne que la priorisation n'est pas uniquement une décision biologique. Elle dépend de contraintes multiples et d'arbitrages, et se construit souvent collectivement. Elle rappelle que l'analyse scientifique doit s'appuyer sur la répétition des signaux et leur cohérence biologique, mais que d'autres considérations peuvent intervenir :

"It's not only a biological answer."

Ici, l'élément central reste que la priorisation doit être fondée sur la robustesse du signal et sa convergence entre patients et méthodes. Elle reconnaît toutefois que ces décisions

s’inscrivent dans un contexte plus large, où les ressources disponibles et la stratégie globale de recherche influencent inévitablement les choix effectués.

Les signatures basées sur des voies biologiques sont-elles plus robustes que celles centrées sur un gène unique ?

Paloma Cejas insiste sur une idée fondamentale : une stratégie pertinente doit reposer sur des éléments robustes et reproductibles entre patients. Elle rappelle que l’objectif n’est pas de développer des approches adaptées à un seul individu, mais d’identifier des mécanismes partageables :

“You are not doing drugs for one person. You want something that can be used.”

Elle observe que, dans la pratique, les gènes qui émergent individuellement s’inscrivent souvent dans un même programme biologique, ce qui rend l’interprétation par voie plus naturelle et plus robuste. Elle propose ainsi une logique en deux temps : identifier la voie convergente dominante, puis déterminer à quel niveau de cette voie une intervention pourrait être envisagée.

Quel niveau d’évidence est nécessaire pour considérer une signature comme fonctionnelle et crédible ?

Pour Paloma Cejas, une signature gagne en crédibilité fonctionnelle lorsqu’elle est observée dans plusieurs patients, qu’elle s’inscrit dans une biologie cohérente, et que les données sous-jacentes sont solides. Elle souligne que le niveau de validation requis dépend fortement du contexte :

“If it’s something completely new, then you need to show it. Otherwise you don’t go far.”

À l’inverse, si un mécanisme est déjà bien documenté dans la littérature, l’objectif n’est pas de tout redémontrer, mais de montrer sa cohérence et sa spécificité dans le contexte particulier des survivants exceptionnels. Cette approche graduée permet de distinguer une simple corrélation statistique d’une hypothèse biologique crédible et exploitable.

.

5. Apports cliniques et translationnels

5.1 Objectif :

Dans cette partie, nous allons d'abord nous intéresser à l'impact clinique direct que peut avoir l'identification de signatures exploitables, c'est-à-dire comment celles-ci peuvent nous aider dans la prédiction de la réponse à un traitement, la stratification des patients et le repositionnement thérapeutique. Dans un second temps, nous aborderons le passage à la translation, c'est-à-dire comment, à partir de signaux statistiques issus de données multi-omiques, nous pouvons aboutir à l'identification d'une piste thérapeutique.

5.2 L'Impact Clinique

L'impact clinique se divise en trois volets, pour lesquels nous allons tout d'abord définir la nature dans leur sens global, puis examiner les moyens par lesquels ils peuvent être développés à partir des découvertes issues de l'étude des données outliers.

5.2.1 Prédiction de la réponse à un traitement

La première utilité clinique d'une signature multi-omique réside dans sa capacité à prédire la réponse d'un patient à un traitement donné. Un biomarqueur prédictif permet d'anticiper l'efficacité d'une thérapie, de sélectionner les patients susceptibles d'en bénéficier et d'éviter l'exposition inutile à des traitements toxiques ou inefficaces. Cette dimension est centrale dans l'oncologie de précision et constitue l'un des principaux objectifs de l'étude des survivants exceptionnels. L'histoire récente de la cancérologie fournit plusieurs exemples où l'identification d'un profil moléculaire spécifique a transformé la prise en charge thérapeutique. L'un des cas les plus emblématiques est celui des mutations du gène **EGFR** dans le cancer du poumon non à petites cellules.

En effet un exemple récent illustre bien comment une altération génomique, même rare, peut modifier radicalement la réponse à un traitement. Dans une étude de 2024(23) portant sur 57 patients atteints de cancer pulmonaire à petites cellules, les auteurs ont montré que les tumeurs porteuses d'une mutation **EGFR** présentaient :

- **un taux de réponse plus élevé** aux inhibiteurs de tyrosine kinase (80 % vs 57 %),
- **une survie sans progression significativement prolongée** (8,2 mois vs 3,0 mois),
- **une survie globale doublée** (16,7 mois vs 7,4 mois).

Ces résultats démontrent que, même dans un cancer réputé homogène et sans cible thérapeutique comme le SCLC (small cell lung cancer) , l'identification d'un biomarqueur peut révéler un sous-groupe biologiquement distinct et thérapeutiquement exploitable. De la même manière, l'identification du statut **MSI-H** (instabilité micro-satellitaire élevée) dans les tumeurs digestives a profondément modifié le paradigme thérapeutique : elle a prédit une probabilité élevée de réponse aux inhibiteurs de checkpoints immunitaires, conduisant à l'autorisation du pembrolizumab dans une indication "tissue-agnostic"

Au-delà de ces exemples établis, des biomarqueurs plus rares ont également démontré leur pouvoir prédictif lorsqu'ils étaient correctement identifiés et validés. Les **fusions NTRK**, longtemps considérées comme anecdotiques car dispersées dans de petites sous-populations tumorales, ont révélé des réponses spectaculaires aux inhibiteurs de TRK lorsqu'elles ont été regroupées et étudiées de manière spécifique. Leur impact clinique a été tel qu'elles ont débouché sur une des premières autorisations FDA basées uniquement sur un biomarqueur moléculaire, indépendamment du type tumoral. Cela illustre directement la valeur translationnelle d'une altération moléculaire : une information génomique permet d'anticiper la réponse à un traitement ciblé et de proposer une prise en charge spécifique.

Dans le contexte des survivants exceptionnels, l'identification de signatures multi-omiques pourrait permettre de révéler des mécanismes de sensibilité thérapeutique aujourd'hui sous-estimés.

5.2.2 Valeur prédictive du pronostique

Au-delà de leur rôle prédictif, certaines signatures multi-omiques peuvent également offrir une valeur pronostique, c'est-à-dire informer sur l'évolution naturelle de la maladie indépendamment du traitement administré. Un biomarqueur pronostique permet ainsi d'identifier des patients ayant un risque plus élevé ou plus faible de progression ou de décès, ce qui constitue un élément essentiel pour ajuster la surveillance, orienter le choix thérapeutique ou décider de l'intensité du traitement. Classiquement, plusieurs altérations moléculaires se sont imposées comme des facteurs pronostiques majeurs dans différents cancers. Dans le cancer du poumon non à petites cellules, par exemple, les mutations du gène TP53 ou les co-altérations EGFR/TP53 sont associées à une survie globale réduite,(2) indépendamment de la ligne thérapeutique. Dans les cancers digestifs, l'expression élevée de signatures inflammatoires telles que IL6, STAT3 ou CXCL8 corrèle avec un phénotype tumoral plus agressif, un microenvironnement immunosuppresseur et un risque accru de rechute. De même, dans les tumeurs cérébrales, le statut mutationnel IDH1 et la méthylation du promoteur MGMT constituent des facteurs pronostiques robustes permettant d'anticiper la survie à long terme.

Dans cette continuité, les signatures dérivées de cohortes de survivants exceptionnels apportent un éclairage inédit sur les mécanismes associés à une évolution clinique particulièrement favorable. Une étude récente menée chez des patients atteints de cancer du pancréas avancé en fournit une illustration en adéquation avec notre sujet actuel.

Dans leur étude *Genomic Predictors Associated With Exceptional Response to Systemic Therapy in Advanced Pancreatic Cancer*, Kamath et al. (11)ont montré que les survivants exceptionnels présentaient une charge mutationnelle non synonyme nettement plus faible que les patients au pronostic habituel (2,25 contre 5,17 mutations en médiane) La « charge mutationnelle non synonyme » correspond au nombre de mutations entraînant un changement d'acide aminé dans la protéine codée ; elle reflète donc l'impact fonctionnel potentiel des altérations génétiques sur la biologie tumorale. Un seuil inférieur à trois mutations non synonymes était fortement associé à une survie globale prolongée (29,4 mois versus 4,6 mois) ainsi qu'à une survie sans progression significativement supérieure (17,2 mois versus 2,3 mois)[1]. Ces résultats suggèrent qu'une moindre accumulation d'altérations fonctionnelles pourrait caractériser une tumeur biologiquement moins agressive et contribuer à la survie exceptionnellement longue observée chez ces patients. Cependant, il convient de souligner que cette étude repose sur une cohorte de **21 patients** seulement (8 survivants exceptionnels et 13 contrôles). Une telle taille d'échantillon, bien qu'inévitable dans le contexte des survivants exceptionnels, limite la puissance statistique et rend difficile d'exclure la possibilité que les observations reflètent un artefact expérimental plutôt qu'un phénomène biologique robuste. Cet exemple illustre précisément les défis méthodologiques posés par des

effectifs extrêmement réduits, pour lesquels même des approches statistiques adaptées aux petites cohortes atteignent leurs limites, rendant les résultats difficilement comparables et nécessitant une interprétation prudente.

Les signatures transcriptomiques ou épigénétiques découvertes dans d'autres cohortes de survivants exceptionnels pourraient s'inscrire dans cette même logique. Par exemple, une signature enrichie en gènes impliqués dans la réparation de l'ADN, la stabilité génomique ou la réponse immunitaire adaptative pourrait traduire une capacité intrinsèque de contrôle tumoral, se manifestant par une évolution plus lente de la maladie et une survie prolongée. Enfin, la valeur pronostique est indispensable pour comprendre la physiopathologie des survivants exceptionnels eux-mêmes. La survie prolongée observée dans ces patients peut résulter d'une biologie tumorale intrinsèquement moins agressive ou d'un microenvironnement immunitaire plus efficace. Identifier et caractériser ces signatures pronostiques permet non seulement de mieux comprendre les trajectoires atypiques de survie, mais aussi de transposer ces mécanismes à des populations plus larges.

5.2.3 Entretien avec l'expert

Dr Yaovi Eric Amela

Chef du pôle d'oncologie médicale – Comité de cancérologie urologique
Format : entretien (visio)

1. Définition clinique du survivant exceptionnel : une notion intrinsèquement contextuelle

Pour le Dr Amela, **la définition d'un survivant exceptionnel dépend entièrement du type de cancer**.

Il précise :

« Dans le pancréas métastatique, une survie de deux ans est déjà exceptionnelle. On ne peut pas appliquer les mêmes seuils que dans d'autres cancers. »

Il rappelle que l'identification des survivants exceptionnels est techniquement faisable via les parcours hospitaliers, mais **la constitution d'un bras contrôle strictement comparable reste un défi majeur**.

2. Valeur clinique actuelle des biomarqueurs : entre rares certitudes et nombreuses zones grises

Selon lui, seuls **deux biomarqueurs ont une utilité clinique démontrée** dans le pancréas :

- **BRCA1/2** → réponse aux inhibiteurs de PARP (essai POLO).(24)
- **MSI-H** → réponse très probable à l'immunothérapie.

Il souligne que d'autres marqueurs – notamment **KRAS** – font l'objet de recherches actives, mais **sans impact thérapeutique immédiat**.

Il insiste également sur les limites de la médecine de précision actuelle : « Même avec un biomarqueur valide, certains patients ne répondent pas. EGFR, PD-L1… ce n'est jamais absolu. »

3. De la statistique à la clinique : une transition exigeante et sous contraintes

Le Dr Amela rappelle que transformer un signal statistique en recommandation thérapeutique nécessite :

- des **preuves répétées**,

- des **cohortes suffisamment grandes**,
- et souvent des essais stratifiés dès le **design initial**.

Il cite l'exemple du PD-L1, prédictif dans certaines indications mais **non pertinent dans le rein**, illustrant la **forte dépendance contextuelle des biomarqueurs**.

Il apporte également un exemple clé :

deux études comportant chacune ~40 patients ne montraient aucun signal isolément, mais combinées, elles révélaient un effet significatif.

4. Survivants exceptionnels et repositionnement thérapeutique

Le Dr Amela confirme que l'analyse de patients "hors normes" a déjà transformé la pratique Il cite notamment les mutations de BRCA comme ouvrant l'accès aux inhibiteurs de PARP, ainsi que le statut MSI comme prédictif d'une réponse à l'immunothérapie, et précise que ces exemples montrent que l'étude d'outliers peut conduire à l'émergence de nouveaux standards thérapeutiques.

5. Freins et conditions d'adoption : robustesse, simplicité, bénéfice clinique

Le Dr Amela identifie trois critères indispensables pour qu'une signature soit intégrée en pratique :

1. **Méthode simple et acceptable pour le patient** (ex : biopsie liquide).
2. **Reproductibilité statistique démontrée**.
3. **Bénéfice clinique clair**, justifiant une modification du parcours de soins.

Il précise :

« Il n'y a pas de frein éthique. Si c'est robuste et utile pour le patient, tout le monde est preneur. »

6. Vision prospective : du biomarqueur isolé aux signatures multi-omiques intégrées

Pour lui, l'avenir repose sur :

- des **signatures multi-couches** (génomique + transcriptomique + microenvironnement),
- une **sélection intelligente des patients**,

- et l'usage d'**IA** permettant d'extraire des signaux robustes dans des ensembles complexes.

Conclusion générale

La perspective du Dr Amela soutient la pertinence clinique de l'étude des survivants exceptionnels, tout en rappelant les exigences nécessaires pour transformer une signature statistique en outil thérapeutique. Son point de vue renforce l'importance :

- d'une définition contextuelle et rigoureuse,
- d'une stratification précise,
- de signatures robustes et reproductibles,
- et d'une intégration progressive dans des essais cliniques.

5.3 Impact translationnel

5.3.1 De la signature au mécanisme : établir un lien fonctionnel

Lors de la partie 5 nous avons pu détailler comment des outils d'enrichissement fonctionnel tels que GO, KEGG, MSigDB ou GSEA peuvent permettre de relier une liste de gènes à des voies biologiques cohérentes et interprétables. Grâce à ces approches, une signature n'est plus une simple collection de gènes différentiellement exprimés, mais devient l'expression d'un programme fonctionnel. Dans une perspective translationnelle l'objectif est d'identifier quelle vulnérabilité thérapeutique découle de ce mécanisme. Un exemple dans un article publié dans *nature scientific* report par Kamii et al.en 2025(25) illustre parfaitement ce processus.

Dans cette publication, les auteurs analysent le transcriptome de cellules cancéreuses HR-proficientes c'est-à-dire présentant une réparation de l'ADN par recombinaison homologue fonctionnelle, traitées par un inhibiteur de PARP. Leur analyse différentielle identifie 866 gènes modulés, dont une proportion importante appartient à des voies inflammatoires. Les analyses KEGG et GSEA montrent une activation coordonnée de la signalisation TNF-α/NFκB, de la réponse interféron et des voies JAK-STAT, indiquant que le traitement déclenche non seulement une réponse au stress, mais un véritable programme inflammatoire structuré, part là, on entend l'activation inflammatoire n'est pas limitée à quelques gènes isolés, mais correspond à l'activation coordonnée et cohérente de plusieurs voies biologiques interconnectées. Cette signature transcriptionnelle est ensuite reliée à un mécanisme biologique précis : la mise en place d'un état de sénescence. Les auteurs confirment expérimentalement l'apparition d'un phénotype sénescent (morphologie cytomégalique, activité SA-β-Gal, augmentation de la granularité cytoplasmique). Ils montrent par ailleurs que cette sénescence est médiée par l'activation de la voie cGAS–STING, une voie de l'immunité innée impliquée dans la détection de l'ADN anormalement localisé dans le cytoplasme. Le traitement par inhibiteur de PARP entraîne la formation de micronoyaux, structures contenant de l'ADN endommagé, qui sont reconnus par le capteur cGAS et déclenchent une réponse inflammatoire. L'implication de cette voie est confirmée par la disparition du signal inflammatoire dans des cellules génétiquement invalidées pour STING (STING-KO). Enfin, cette compréhension mécanistique conduit à l'identification d'une vulnérabilité thérapeutique. Les auteurs démontrent que les cellules sénescentes libèrent un ensemble de facteurs SASP capables d'attirer des cellules immunitaires, suggérant que les inhibiteurs de PARP pourraient, même en absence d'altération BRCA, favoriser une activation immunitaire du microenvironnement tumoral. Cette observation ouvre la voie à des combinaisons rationnelles entre PARP inhibiteurs et immunothérapies, transformant une simple signature statistique en une hypothèse thérapeutique testable

Cet exemple montre comment une signature multi-omique, d'abord identifiée par des méthodes statistiques, peut être progressivement interprétée, validée et transformée en

piste thérapeutique. C'est précisément cette logique qui doit guider l'étude des survivants exceptionnels : partir d'un signal rare, en établir la cohérence biologique, puis identifier la vulnérabilité thérapeutique susceptible d'expliquer et potentiellement de reproduire leur survie atypique et extraordinaire

5.3.2 Priorisation thérapeutique : transformer le mécanisme en cible

Une fois le mécanisme biologique identifié, l'étape suivante consiste à déterminer s'il peut être exploité thérapeutiquement. Ce processus est appelé **priorisation de cible**, il repose en grande partie sur le concept de **druggabilité**, c'est-à-dire la capacité d'une protéine à être modulée de manière efficace et sûre par un médicament. Cette notion a été formalisée par Hopkins et Groom dans un article fondateur publié dans *Nature Reviews Drug Discovery* en 2002(26), où ils montrent que seule une fraction limitée du génome humain code des protéines réellement compatibles avec une modulation pharmacologique. Selon leur analyse, environ **10 à 14 %** des gènes humains appartiennent à ce qu'ils appellent le **"druggable genome"**, c'est-à-dire des protéines possédant des sites de liaison structurés, susceptibles d'interagir avec des petites molécules thérapeutiques. Ils soulignent également que parmi ces protéines théoriquement « druggables », seule une portion plus restreinte représente des **cibles thérapeutiques pertinentes**, c'est-à-dire des gènes à la fois impliqués dans une pathologie et modulables par un médicament. Ce rappel est essentiel dans le contexte translationnel car un mécanisme peut être biologiquement passionnant et émettre beaucoup d'espoir sans pour autant constituer une cible exploitable.

Ensuite pour évaluer si un mécanisme biologique peut être exploité thérapeutiquement, les chercheurs s'appuient aujourd'hui sur plusieurs outils complémentaires qui permettent de relier gènes, médicaments et dépendances cellulaires.
Les bases de données pharmacologiques, telles que **DrugBank** ou **DGIdb(27)**, recensent les médicaments existants et les protéines qu'ils ciblent. Elles permettent ainsi de vérifier rapidement si un gène ou une voie identifiée a déjà été modulé pharmacologiquement, et si un composé est disponible pour envisager soit un repositionnement thérapeutique, soit une extension d'indication. Autrement dit, ces ressources servent à établir un premier lien entre une signature moléculaire et des options thérapeutiques concrètes.

Les projets de **cartographie de dépendances tumorales,** comme **DepMap**,(28) reposent sur des technologies de criblage génétique à grande échelle. À l'aide de systèmes **CRISPR-Cas9,** chaque gène est inactivé de manière ciblée dans des centaines de lignées cellulaires cancéreuses, permettant d'identifier ceux dont la perte compromet fortement la survie ou la prolifération cellulaire. Un gène qualifié d'essentiel dans ce contexte représente une vulnérabilité potentielle, car son inhibition pharmacologique pourrait reproduire cet effet délétère sur la cellule tumorale. En complément, les bases de données de pharmacogénomique comme le **Genomics of Drug Sensitivity in Cancer**

(GDSC) associent les profils moléculaires des lignées tumorales à leur réponse à une large gamme de composés anticancéreux. En comparant la sensibilité ou la résistance de ces cellules à différents médicaments, il devient possible de relier une signature génétique ou transcriptomique donnée à une réponse pharmacologique spécifique. Ces analyses aident ainsi à prédire quels patients pourraient bénéficier d'un traitement ciblé, et à identifier des interactions gène–médicament pertinentes pour une stratégie translationnelle. On peut donc constater que la priorisation thérapeutique consiste à transformer une observation mécanistique en une hypothèse d'intervention, à l'aide de critères pharmacologiques, d'outils bioinformatiques et de données expérimentales, ouvrant la voie à la validation préclinique puis clinique.

5.3.3 Validation préclinique et passage vers le développement clinique

Jusqu'ici, les étapes que je vous ai présenté étaient spécifiques à l'analyse multi-omique et à l'identification de signaux rares propres aux survivants exceptionnels ; la suite du processus s'inscrit dans un cadre plus général. En effet il s'agit de transformer une hypothèse mécanistique en une intervention thérapeutique potentiellement utile en clinique. Les étapes décrites ci-dessous relèvent ainsi du développement préclinique et clinique traditionnel, mais demeurent toutefois indispensables pour évaluer la valeur translationnelle d'une cible issue de l'analyse des survivants exceptionnels.

Validation préclinique :

L'objectif de cette étape est de vérifier que la modulation de la cible identifiée reproduit bien l'effet attendu comme par exemple l'inhibition de la croissance cellulaire, l'induction d'apoptose, la sensibilisation à un traitement, ou la modification du microenvironnement tumoral.

Les approches utilisées reposent sur trois niveaux :

- **Validation in vitro :**

  Utilisation de siRNA, CRISPR ou inhibiteurs pharmacologiques pour inhiber la cible dans des lignées cellulaires tumorales. Ces modèles permettent de tester rapidement si la perturbation du gène ou de la voie identifiée a un impact fonctionnel cohérent avec le mécanisme observé.

- **Validation in vivo** :

  Modèles dérivés de patients (PDX) ou organoïdes tumoraux permettent d'évaluer l'effet de l'inhibition dans un contexte biologique complet.
  L'objectif est de déterminer si la modulation de la cible entraîne un ralentissement tumoral ou une modification du microenvironnement immunitaire.

- **Corrélations pharmacogénomiques** :

  Grâce aux données issues de plateformes telles que **DepMap** ou **GDSC**, il est possible d'évaluer si l'expression ou l'altération de la cible est corrélée à la sensibilité à certains médicaments dans de vastes ensembles de modèles cellulaires.
  Ces analyses complètent les validations expérimentales en apportant une vision populationnelle.

Ainsi, la validation préclinique permet de passer d'une relation statistique à une démonstration fonctionnelle, établissant que la cible identifiée est effectivement pertinente biologiquement.

Passage au développement clinique :

Le développement clinique repose classiquement sur une succession d'essais de phases croissantes, chacun répondant à un objectif spécifique. Les essais de phase I visent avant tout à évaluer la tolérance et la sécurité de la molécule, à déterminer la dose maximale tolérée et à caractériser les principaux effets indésirables pouvant survenir. Ils peuvent également fournir des premières données pharmacocinétiques et pharmacodynamiques. Les essais de phase II ont pour objectif principal d'explorer l'activité antitumorale du traitement dans une population ciblée, souvent enrichie sur la base du biomarqueur ou de la signature identifiée. Cette phase permet de confirmer l'intérêt biologique et clinique de la cible, ainsi que la pertinence de la stratégie de stratification des patients. Enfin, les essais de phase III visent à démontrer le bénéfice clinique du traitement par comparaison au standard de soin (first-in-class ou best-in-class) dans des essais randomisés. C'est à ce stade que sont évalués les critères cliniques majeurs, tels que la survie globale, la survie sans progression ou la qualité de vie, conditionnant l'autorisation de mise sur le marché. Ainsi, les signatures issues de l'étude des survivants exceptionnels trouvent leur pleine valeur lorsqu'elles permettent d'orienter le développement clinique vers des essais mieux stratifiés et plus rationnels. Si leur rôle est central pour générer des hypothèses mécanistiques innovantes, leur validation finale repose nécessairement sur les standards classiques de l'évaluation clinique, garantissant la transposabilité et la robustesse des bénéfices observés.

5.2.4. Entretien avec l'expert

Dans cette partie, l'analyse est éclairée par l'expertise du Dr Paloma Cejas.

Impact clinique des signatures issues des survivants exceptionnels

Le Dr Paloma Cejas souligne d'emblée que la notion de survie exceptionnelle doit être interprétée dans le contexte propre à chaque pathologie. Dans certains cancers de très mauvais pronostic, une survie relativement modeste en valeur absolue peut déjà représenter un phénotype atypique. Comme elle le formule explicitement :

"In some cancers like pancreatic cancer, where survival is almost nonexistent, surviving 18 or 24 months can already be considered exceptional."

Cette observation implique que les signatures cliniques ne doivent pas être évaluées selon des seuils universels, mais à l'aune du pronostic habituel de la maladie étudiée. Dans ce cadre, l'étude des survivants exceptionnels permet d'identifier des signaux cliniquement pertinents, même lorsqu'ils concernent un nombre limité de patients.

Elle insiste également sur l'intérêt des cohortes internationales et hétérogènes, qui offrent une opportunité unique de tester la robustesse clinique des signatures. Selon elle, un signal clinique crédible doit émerger de tendances cohérentes observées dans différents contextes, et non d'un effet isolé propre à une population restreinte.

Du mécanisme à la piste thérapeutique

Priorisation d'une voie ou d'une cible

Pour le Dr Paloma Cejas, la transition entre mécanisme biologique et cible thérapeutique repose sur une série d'arbitrages qui dépassent la seule biologie. Elle rappelle que la priorisation dépend à la fois de la robustesse scientifique, de la faisabilité technique et du paysage concurrentiel. Elle souligne ainsi que :

"It's not only a biological answer. It may be a commercial answer also."

Une voie très bien caractérisée biologiquement peut s'avérer peu attractive si elle est déjà saturée de développements concurrents. À l'inverse, une cible trop nouvelle peut être difficile à défendre cliniquement si elle apparaît trop spéculative. La décision repose donc sur un équilibre entre innovation, crédibilité biologique et potentiel de développement.

Le Dr Cejas insiste également sur l'importance de raisonner au niveau de la voie, puis de sélectionner le meilleur point d'intervention. Comme elle l'exprime :

"You need to see which is the pathway that everybody's telling you is important, and then which level and which is the best target to target with your methodology."

Druggabilité et choix de la modalité thérapeutique

Le Dr Paloma Cejas rappelle que toutes les cibles biologiques ne sont pas équivalentes sur le plan pharmacologique. Le choix d'une cible dépend étroitement de la modalité thérapeutique envisagée. Pour les petites molécules, la présence d'un site de liaison exploitable est déterminante, ce qui exclut de facto certains gènes pourtant biologiquement centraux.

Elle souligne également l'intérêt croissant de certaines approches, notamment les anticorps conjugués, qu'elle perçoit comme plus rapides à initier dans certains contextes translationnels :

"Antibody-based strategies can be faster to develop, and they are becoming very specific, reducing toxicity."

Elle rappelle néanmoins que des contraintes biologiques, comme la barrière hémato-encéphalique dans les tumeurs cérébrales, peuvent limiter l'usage de certaines modalités, et doivent être intégrées très tôt dans la réflexion translationnelle.

Repositionnement thérapeutique

Le repositionnement est, selon le Dr Paloma Cejas, une stratégie particulièrement attractive lorsqu'elle est possible, car elle permet d'accélérer considérablement le passage vers la clinique. Elle le résume simplement :

"Anything that can make you go faster, I would take advantage of it."

Cependant, elle souligne une limite structurelle importante : les molécules déjà développées sont rarement facilement accessibles, car les entreprises conservent leurs brevets et leurs actifs, même lorsqu'ils ne sont pas exploités immédiatement. Ainsi, si le repositionnement est scientifiquement séduisant, sa mise en œuvre dépend fortement de contraintes industrielles et de propriété intellectuelle.

Validation préclinique adaptée à la nature de la cible

Le Dr Paloma Cejas insiste sur le fait qu'il n'existe pas de modèle expérimental universel. Le choix du modèle dépend directement du mécanisme étudié. Pour les mécanismes strictement tumoraux, les approches de type CRISPR, y compris les criblages génome-entier, constituent selon elle un outil très puissant :

"CRISPR is the most efficient way to knock out something when you want to test cell-autonomous mechanisms."

En revanche, lorsqu'un mécanisme implique des interactions entre la tumeur et le microenvironnement, en particulier le système immunitaire, des modèles plus complexes sont nécessaires. Elle précise clairement que :

"If the mechanism involves immune communication, you need immune cells there. You cannot rely only on organoids."

Le Dr Paloma Cejas rappelle également les limites des modèles PDX classiques, généralement dépourvus de système immunitaire fonctionnel, et souligne que le niveau de validation requis dépend du degré de nouveauté de la cible. Plus une piste est éloignée de la littérature existante, plus une démonstration fonctionnelle devient nécessaire.

Passage au clinique et contraintes de temporalité

Le Dr Paloma Cejas met en évidence que la recherche académique et le développement translationnel répondent à des logiques différentes. Alors que la recherche académique vise prioritairement l'exploration et la production de connaissances, les approches translationnelles sont davantage contraintes par des impératifs de priorisation et de temporalité, liés à l'objectif d'une application clinique.

"We are not an academic lab. We need to focus and move things to the clinic."

Elle insiste sur la nécessité de limiter la dispersion des efforts, car multiplier des validations très complexes peut ralentir l'émergence d'actifs cliniquement testables. Le passage à l'essai clinique exige donc un compromis entre profondeur scientifique et efficacité opérationnelle.

Limites actuelles des approches multi-omiques

Enfin, le Dr Paloma Cejas identifie une limite majeure des approches multi-omiques actuelles : leur complexité analytique. Ces technologies génèrent des volumes de données considérables, alors même que les outils computationnels et les cadres analytiques sont encore en évolution. Elle souligne que :

"These methodologies are very new, and even in academia people don't fully know how to analyze them yet."

Elle conclut toutefois sur une note résolument constructive : cette complexité représente autant une difficulté qu'une opportunité, car la capacité à extraire des mécanismes actionnables à partir de ces données pourrait ouvrir de nouvelles voies thérapeutiques, notamment dans des cancers longtemps restés sans avancées majeures.

Conclusion

Cette analyse met en évidence que la notion de survie exceptionnelle doit être définie de manière relative au pronostic propre à chaque type tumoral. Elle montre également que des signaux issus de cohortes de taille limitée peuvent conserver une pertinence clinique lorsqu'ils sont biologiquement cohérents, reproductibles et interprétés dans un cadre méthodologique rigoureux. La priorisation thérapeutique repose ainsi sur l'intégration raisonnée des données biologiques, pharmacologiques et expérimentales, en privilégiant une approche par voies biologiques plutôt que par gènes isolés.

L'échange souligne enfin que la validation préclinique et la translation vers la clinique ne suivent pas un schéma unique, mais doivent être adaptées à la nature du mécanisme étudié et au contexte clinique considéré. Dans ce cadre, l'étude des survivants exceptionnels constitue un modèle pertinent pour explorer des mécanismes biologiques rares, à condition d'en maîtriser strictement les limites méthodologiques.

6. RESULTAT

6.1. Synthèse

Cette thèse a été structurée de manière progressive afin que chaque partie réponde à une question précise, tout en contribuant à un cadre méthodologique et conceptuel cohérent pour l'étude des survivants exceptionnels du cancer.

La **Partie 1**, consacrée au contexte clinique et scientifique, a permis de poser les bases du sujet en mettant en évidence la singularité des survivants exceptionnels. Elle a montré que ces patients constituent un sous-groupe atypique, souvent invisible dans les analyses classiques centrées sur la moyenne des cohortes. Cette partie a également souligné les limites des approches traditionnelles en oncologie, qui peinent à expliquer des trajectoires cliniques extrêmes, et a justifié l'intérêt d'une démarche spécifiquement orientée vers l'étude des outliers.

La **Partie 2** a ensuite abordé les outils biologiques et omiques mobilisables pour caractériser ce phénotype rare. Elle a montré que la compréhension des survivants exceptionnels ne peut reposer sur une seule couche de données, mais nécessite une approche multi-échelle intégrant la génomique, la transcriptomique, ainsi que les technologies unicellulaires et spatiales. Cette partie a également insisté sur la nécessité d'une hiérarchisation raisonnée des technologies, en tenant compte des contraintes liées à la qualité des échantillons, aux effectifs disponibles et aux objectifs biologiques poursuivis.

La **Partie 3** s'est focalisée sur les méthodes statistiques adaptées aux petits effectifs, une contrainte centrale dans l'étude des survivants exceptionnels. Elle a démontré que les modèles statistiques standards, conçus pour de larges cohortes, sont souvent inadaptés dans ce contexte. Cette partie a ainsi justifié l'utilisation de stratégies analytiques hybrides, combinant des approches non supervisées pour explorer les structures globales des données et des méthodes supervisées, parcimonieuses et robustes, pour identifier des signaux biologiques discriminants tout en limitant le risque de surapprentissage.

La **Partie 4** a porté sur l'interprétation biologique des signatures multi-omiques et leur validation. Elle a établi que la valeur d'une signature ne réside pas uniquement dans sa significativité statistique, mais dans sa cohérence mécanistique, sa reproductibilité et sa capacité à être validée par des approches croisées, qu'elles soient analytiques, biologiques ou issues de cohortes indépendantes. Cette partie a mis en évidence l'importance de relier les signaux omiques à des voies biologiques identifiables et à des mécanismes fonctionnels plausibles.

Enfin, la **Partie 5** a exploré les apports cliniques et translationnels de ces signatures. Elle a montré comment des signaux issus de l'étude des survivants exceptionnels peuvent

devenir actionnables, à condition d'intégrer dès l'amont les contraintes de druggabilité, de priorisation des cibles et de développement clinique. Cette partie a souligné que la translation ne constitue pas une simple étape finale, mais un processus structurant qui doit orienter les choix méthodologiques tout au long de l'analyse.

6.2. Limites et points de vigilance

Comme toute démarche scientifique exploratoire appliquée à des phénotypes rares, l'étude des survivants exceptionnels présente un certain nombre de limites et de points de vigilance, qui ont été régulièrement soulignés par les experts interrogés au cours de ce travail. L'identification et la prise en compte explicite de ces limites constituent un élément central de la robustesse méthodologique de l'approche proposée.

La première limite concerne la **qualité et l'hétérogénéité des échantillons biologiques**. Les survivants exceptionnels sont, par définition, des patients identifiés a posteriori, avec des échantillons collectés dans des contextes cliniques variés, souvent sur de longues périodes. Comme l'ont rappelé les experts, notamment dans le cadre des analyses single-cell et spatiales, la qualité des tissus conditionne directement la fiabilité des résultats. Des échantillons dégradés ou insuffisamment contrôlés peuvent introduire des biais structurels qui se propagent tout au long de l'analyse. Cette contrainte impose un contrôle qualité rigoureux en amont et limite parfois l'exploitabilité de certaines technologies pourtant très informatives sur le plan biologique.

Une deuxième limite majeure réside dans **l'hétérogénéité intrinsèque des cohortes**. Les survivants exceptionnels proviennent de contextes cliniques, génétiques et thérapeutiques très divers, ce qui complique l'interprétation des signaux observés. Les experts ont souligné que cette hétérogénéité constitue à la fois une richesse et une difficulté : elle permet de tester la robustesse des signatures dans des contextes variés, mais elle complique l'identification de mécanismes universels. Cette réalité impose de privilégier des signaux convergents et reproductibles plutôt que des associations observées dans des sous-groupes très restreints.

La **complexité analytique des données multi-omiques** représente également une limite importante. Les technologies mobilisées génèrent des volumes de données considérables, dont l'intégration requiert des compétences computationnelles avancées et des cadres méthodologiques encore en évolution. Comme l'ont souligné plusieurs experts, ces approches sont relativement récentes, et les pipelines analytiques ne sont pas encore totalement standardisés. Cette situation augmente le risque d'interprétations divergentes et impose une grande prudence dans la construction des modèles et dans la hiérarchisation des résultats.

Par ailleurs, l'**accès à des validations expérimentales complètes** constitue un frein fréquent, en particulier dans un cadre translationnel. Les experts ont rappelé que le

niveau de validation requis dépend fortement du degré de nouveauté des mécanismes identifiés. Lorsque les signaux observés s'inscrivent dans une littérature existante, des validations lourdes ne sont pas toujours indispensables. En revanche, pour des pistes très innovantes, l'absence de validation fonctionnelle approfondie peut limiter la portée des conclusions. Cette contrainte est accentuée par des considérations pratiques, telles que le coût, la durée et la disponibilité des modèles expérimentaux pertinents.

Enfin, plusieurs experts ont insisté sur le risque d'une **surinterprétation des signaux statistiques**, particulièrement dans des contextes de petits effectifs. Même lorsque des outils adaptés sont utilisés, le danger du surapprentissage demeure. Cette limite impose une approche parcimonieuse, fondée sur la convergence des résultats, la validation croisée et la prudence dans la formulation des hypothèses biologiques.

Dans leur ensemble, ces limites n'invalident pas l'approche développée dans cette thèse. Elles en définissent au contraire le **cadre d'application**, en soulignant que l'étude des survivants exceptionnels requiert une méthodologie rigoureuse, critique et contextualisée. En intégrant ces points de vigilance dès la conception des analyses, cette thèse s'inscrit dans une démarche de synthèse critique des approches existantes et vise à en clarifier les conditions d'application dans le cadre de l'étude des survivants exceptionnels.

7. CONCLUSION

Les survivants exceptionnels du cancer constituent un paradoxe clinique, caractérisé par des trajectoires de survie inattendues dans des contextes de pronostic très défavorable. L'objectif de cette thèse n'était pas d'identifier un biomarqueur unique ou une signature définitive, mais de proposer une démarche méthodologique structurée permettant d'aborder ces phénotypes rares de manière rigoureuse, reproductible et biologiquement interprétable.

En s'appuyant sur une approche multi-omique intégrée, ce travail met en évidence que l'étude des survivants exceptionnels nécessite une lecture multi-échelle du cancer, attentive à la fois aux altérations tumorales et à leurs interactions avec le microenvironnement. Il souligne également l'inadéquation fréquente des modèles analytiques classiques pour de petits effectifs, justifiant le recours à des stratégies hybrides combinant exploration, parcimonie statistique et validation croisée.

Les limites méthodologiques associées à ces approches, largement discutées dans ce travail, définissent avant tout un cadre d'application exigeant, fondé sur la prudence interprétative et la hiérarchisation des signaux biologiques.

Enfin, cette thèse défend l'idée que l'exception, longtemps considérée comme un bruit statistique, peut constituer une source de connaissance biologiquement informative lorsqu'elle est étudiée à l'aide d'outils et de méthodologies adaptés. En structurant une approche reproductible et transférable, ce travail contribue à repositionner les survivants exceptionnels comme un modèle pertinent pour l'exploration des mécanismes de résistance, de réponse durable et de vulnérabilités thérapeutiques en oncologie.

Au-delà du cancer, ce cadre méthodologique pourrait également trouver un écho dans l'étude d'autres pathologies rares ou hétérogènes, caractérisées par des trajectoires cliniques extrêmes, telles que certaines maladies auto-immunes, génétiques, neurodégénératives ou infectieuses. Dans le VIH, l'étude de patients dits *elite controllers*, capables de maintenir une charge virale indétectable sans traitement antirétroviral, illustre comment l'analyse de phénotypes exceptionnels pourrait permettre d'identifier des mécanismes de contrôle durable de la maladie (Gebara et al., 2019)(29). Dans ces contextes, l'analyse approfondie de phénotypes atypiques pourrait, de la même manière, révéler des mécanismes de résilience ou de protection jusqu'alors sous-estimés.

Ce travail s'inscrit ainsi dans une perspective élargie de médecine de précision, où l'étude rigoureuse des trajectoires cliniques hors normes ne vise pas seulement à expliquer l'exception, mais à en extraire des principes biologiques susceptibles d'éclairer la prise en charge de populations plus larges.

Annexe I Compte rendu d'entretien Remy Nicolles, PhD

Entretien avec Rémy Nicolle

Expertise : Translational genomics of pancreatic cancer heterogeneity
Objectif : Explorer les outils statistiques et méthodologiques adaptés à l'étude de cohortes multi-omiques modérées sur des survivants exceptionnels.

Question 2 : Choix de modèles statistiques

- **Kaplan-Meier** : « C'est un plot, pas un test statistique en soi. Pour analyser des survies, il faut un modèle de Cox, mais cela nécessite un nombre suffisant de patients et surtout d'événements. Dans le cadre des survivants exceptionnels, où l'on compare souvent deux groupes (survivants vs non-survivants, outliers vs contrôles), le Cox n'a pas beaucoup de sens. »
- **Approche binaire (case-control)** :

  Ici, la comparaison se fait le plus souvent entre deux groupes. Dans ce cadre, des outils comme **DESeq2** sont privilégiés pour l'analyse différentielle, particulièrement robuste pour les données transcriptomiques.
- **Analyse single-cell / spatial transcriptomics** :

  Pour pallier les limites statistiques, la pratique courante est de « transformer » le single-cell en bulk après sélection (par exemple : ne garder que les cellules tumorales, ou seulement les macrophages, fibroblastes, etc.). Cela permet d'appliquer des méthodes plus robustes comme DESeq2 sur des données agrégées, tout en bénéficiant de l'information fine du single-cell.
- **Données discrètes (altérations génétiques, mutations rares)** : Dans ces cas, un **test exact de Fisher** est recommandé, avec correction pour tests multiples (Benjamini-Hochberg pour contrôler le False Discovery Rate).

Question 3 : Réduction de dimension & intégration multi-omique

- Il n'existe pas encore de standard pour l'intégration multi-omique sur petits effectifs.
- Les approches comme **PCA, t-SNE, UMAP** peuvent être utilisées, mais elles posent des problèmes de reproductibilité et d'interprétation.

- Les outils comme **MOFA, mixOmics, iCluster** reposent souvent sur des variantes de PCA supervisée ou non supervisée (par ex. sparse PLS). Leur intérêt principal est de rechercher des signaux covariants entre plusieurs omics, plutôt que des différences isolées case-control.
- Ces méthodes sont jugées « assez robustes » car elles réduisent les risques liés aux multiples tests en cherchant des **patterns globaux** plutôt que des gènes isolés.

Question 4 : Fiabilité sur petits effectifs

- **Réduction de dimension supervisée** = la meilleure stratégie pour limiter le sur-apprentissage.
- Cependant, « la vraie solution reste l'utilisation de cohortes externes » pour validation.
- Les données de contrôle peuvent être issues de bases publiques (TCGA, ICGC), ce qui permet de limiter le coût et de maximiser la puissance statistique.

Question 5 : Validation croisée et robustesse

- La validation croisée reste utile, mais l'essentiel est d'avoir des **cohortes externes indépendantes**.
- Lorsque ce n'est pas possible, l'utilisation de bases publiques est une alternative robuste.

Question 6 : Retour d'expérience

- Plus petite étude menée : **30 patients** (multi-omique : exome, RNA-seq, méthylation, métabolomique).
- Enseignement principal : la **réduction de dimension** est ce qui marche le mieux, mais certains omics peuvent dominer l'analyse et masquer les signaux attendus des autres.
- Conclusion : se concentrer sur les omics les plus robustes/stables, et hiérarchiser leur importance.

Question 7 : Outils / Packages recommandés

- **DESeq2** : référence pour analyse différentielle en transcriptomique.
- **Benjamini-Hochberg** : correction multiple adaptée (préférée à Bonferroni trop conservateur).
- Approches bulk sur données single-cell : considérées comme les plus robustes actuellement.

Question 8 : Conseils et contacts

- Importance de disposer de suffisamment de cas rares pour détecter des signaux spécifiques à des sous-groupes (~10 % des patients).
- Contacts recommandés : **Davy Vernerey** (Besançon, unité méthodologique de statistiques de l'hôpital, statistique des essais cliniques GERCOR/PRODIGE).

ANNEXE II Compte rendu d'entretien – Dr. Naouel Zerrouk, PhD

Computational Biologist – Paris-Saclay

1. Parcours et expérience

« J'ai commencé par une école d'ingénieur en biotechnologie à Angers. C'était une formation assez généraliste et multidisciplinaire : biologie moléculaire et cellulaire, manipulations expérimentales au laboratoire, mais aussi sciences de l'ingénieur (mathématiques, physique, électronique).

En troisième année, j'ai choisi de me réorienter vers un **Master en bio-informatique à Paris-Saclay**, avec une spécialisation en génomique et transcriptomique. L'idée était d'apprendre à analyser des données complexes : mutations, variations du nombre de copies (CNV), expression génique (bulk, single-cell), etc.

Pendant mes stages, je me suis spécialisée en **informatique de la santé**. J'ai travaillé sur la **polyarthrite rhumatoïde**, en développant des approches de **modélisation mathématique de systèmes biologiques complexes**. Ces modèles servaient à simuler l'effet de l'activation ou de l'inhibition de certains gènes, pour voir si cela inversait un phénotype pathologique (ex. réduire l'inflammation).

Ma thèse, en collaboration avec Sanofi, a porté sur les mêmes thématiques. L'idée était d'exploiter ces modèles pour explorer de nouvelles cibles thérapeutiques dans la polyarthrite. Mais les données internes étaient difficiles d'accès ; nous avons beaucoup travaillé avec des **bases publiques**, ce qui posait problème : données incomplètes, annotation clinique très limitée, beaucoup de bruit. C'est une difficulté récurrente dans la recherche. »

2. Approches génomiques (WES vs WGS)

« Dans un projet comme celui des survivants exceptionnels, la question du **WES (Whole Exome Sequencing) vs WGS (Whole Genome Sequencing)** est centrale.

Nous avons fait un pilote en comparant les deux. Finalement, nous avons choisi le **WES**. Pourquoi ? Parce que dans un contexte aussi complexe, **il est primordial d'obtenir une profondeur maximale sur les régions codantes**, là où se situent la plupart des mutations directement *druggable*.

Le **WGS**, de son côté, capture aussi des régions non codantes et des éléments régulateurs, mais cela génère beaucoup plus de bruit et reste plus difficile à interpréter sur de petits effectifs. Pour une mission claire de **target discovery**, le **WES profond est plus adapté et plus *straightforward***.

Le WGS n'est pas inutile – il permettrait d'explorer des mécanismes régulateurs à distance – mais pour les survivants exceptionnels, je recommanderais de privilégier le WES. »

3. Intégration de données publiques

« Même si l'on a des données internes, il est **fondamental d'intégrer des bases publiques** (TCGA, ICGC, DepMap, etc.).

Attention cependant : ces bases sont souvent **sparse** et peu annotées cliniquement, donc elles introduisent du bruit. Ce que je recommande, c'est de **ne pas les utiliser comme des échantillons supplémentaires**, mais plutôt comme une **source d'information externe**.

Par exemple : j'analyse ma cohorte interne (bulk RNA-seq, exome, etc.), j'obtiens une liste de gènes candidats. Ensuite, je vais dans TCGA ou d'autres bases pour récupérer des informations de type **gènes essentiels**, scores d'essentialité ou profils de dépendance, que je peux **merger** avec ma liste. Cela renforce l'analyse sans ajouter du bruit inutile. »

4. Approches transcriptomiques (bulk, single-cell, spatial)

« Pour moi, ces approches sont **complémentaires**.

- **Bulk RNA-seq** : il reste pertinent, même si les échantillons FFPE sont souvent de mauvaise qualité. Le bulk apporte une vision globale et peut servir de base solide, à condition de pondérer les résultats selon la qualité du séquençage.
- **Single-cell RNA-seq (scRNA-seq)** : je le prioriserais, car il permet de capturer l'hétérogénéité cellulaire et d'identifier des populations rares potentiellement critiques.
- **Transcriptomique spatiale** : c'est un apport unique. Là où le single-cell ne donne que des corrélations entre ligands et récepteurs, le spatial permet de montrer leur **co-localisation dans le tissu**. Cela augmente énormément la confiance dans l'interprétation des interactions.

Si je devais hiérarchiser, je dirais **single-cell en premier**, complété par le spatial, mais je maintiendrais quand même du **bulk** pour consolider et comparer les signaux. »

5. Biais techniques à prendre en compte

« Chaque technologie a ses **points faibles** :

- **Spatial transcriptomics (Visium par ex.)** : dans certains tissus comme le pancréas, très fibreux, la qualité du séquençage est hétérogène, parfois médiocre. On peut perdre des signaux essentiels.
- **Single-cell RNA-seq** : la **profondeur de séquençage** peut être limitée, ce qui conduit à des zéros techniques (absence apparente d'expression pour certains gènes/cellules, alors qu'ils sont présents).
- **Bulk RNA-seq** : la qualité dépend fortement de la préparation des échantillons, notamment sur FFPE, souvent fragmentés. »

6. Autres omiques (protéomique, immunomique, métabolomique)

« Plus on a de couches de données, mieux c'est, mais il faut être réaliste.

- **Protéomique** : utile mais très difficile à analyser. On se retrouve vite avec beaucoup de zéros (protéines non détectées).
- **Immunomique** : très pertinente, car elle donne une idée du rôle du microenvironnement immunitaire.
- **Métabolomique / épigénomique** : intéressantes mais rarement prioritaires faute de moyens.

Le compromis réaliste est souvent : **génomique + transcriptomique + immunologique/protéomique**. Le reste, si possible, mais pas indispensable au départ. »

7. Intégration multi-omique

« Il n'existe pas de méthode unique, et dans la pratique on combine plusieurs approches.

- **Non supervisé (ex. MOFA)** : on laisse le modèle intégrer les données sans a priori. Cela produit des poids par omique et permet de détecter des signaux transversaux.
- **Approche supervisée "couche par couche"** : on part d'une liste de cibles dans une omique (ex. single-cell), puis on vérifie leur présence dans d'autres couches (bulk, WES, spatial). Cela permet une intégration **biologiquement compréhensible**, plus facile à expliquer.

En pratique, on pondère les signaux selon la fiabilité de la technique : par exemple, **moins de poids au bulk FFPE**, **plus au single-cell**. On peut même entraîner un modèle avec des données externes (Open Targets, etc.) pour calibrer ces poids et hiérarchiser les cibles. »

8. Vision translationnelle et freins actuels

« À mon avis, il faut absolument voir le tout comme un **système complexe**, où toutes les couches interagissent. Impossible de réduire à une seule omique.

Les freins actuels sont clairs :

- **Reproductibilité insuffisante** : énormément de données générées, mais rarement ré-analysées ou reproduites.
- **Manque de mutualisation** : beaucoup de données dorment dans des laboratoires ou dans des articles jamais exploités.
- **Accès restreint aux données** : un effort global d'**open data** et de **méta-analyses** permettrait d'accélérer considérablement la recherche.

Aujourd'hui, ce n'est pas tant le manque de données qui bloque, mais leur fragmentation et leur mauvaise exploitation. »

ANNEXE III Compte rendu de l'entretien avec le Dr Yaovi Eric Amela

Chef du pôle d'oncologie médicale – Comité de cancérologie urologique
Format : Visio

1. Définition et perception des survivants exceptionnels

Le Dr Amela souligne que définir un "survivant exceptionnel" est une étape complexe : tout dépend du type de cancer et de la durée de survie retenue comme seuil. Pour le **cancer du pancréas métastatique**, où la survie à 5 ans est pratiquement inexistante, un tel patient est déjà un cas spectaculaire. Mais doit-on retenir 18 mois, 2 ans, ou 5 ans comme cut-off ? Ce choix conditionne toute l'analyse.

Il précise qu'avec les **parcours de soins hospitaliers**, identifier ces patients est faisable rapidement : les logiciels permettent de retracer tous ceux inclus dans un parcours "cancer pancréatique" et de voir ceux qui n'en sont pas sortis. En revanche, la vraie difficulté reste le **bras contrôle** : il faut trouver des patients strictement comparables (âge, contexte, traitement reçu) mais décédés selon la médiane habituelle, ce qui est beaucoup moins marquant et donc plus difficile à extraire dans la pratique courante.

2. Valeur clinique des signatures biologiques

Aujourd'hui, il n'existe que deux biomarqueurs réellement utiles en pratique courante pour le pancréas :

- **BRCA1/2**, mutations présentes dans 5 à 10 % des cas, ouvrant la voie aux inhibiteurs de PARP (essai *POLO*).
- **MSI-H**, rare mais hautement prédictif d'une réponse à l'immunothérapie, avec l'exemple du pembrolizumab approuvé en tissue-agnostic par la FDA.

En dehors de ces deux, d'autres marqueurs comme **KRAS** sont explorés, mais n'ont pas encore modifié les pratiques quotidiennes.

Le Dr Amela rappelle aussi que des **biomarqueurs validés peuvent échouer** : certains patients EGFR+ ou PD-L1+ n'y répondent pas, ce qui souligne la complexité et les limites de la médecine de précision actuelle.

3. Du signal statistique à l'essai clinique

Le passage du signal statistique à l'essai clinique demande des preuves solides.

- Dans le pancréas, cela reste compliqué : il faut des essais stratifiés dès le design initial, ce qui implique de longs délais et de grandes cohortes.
- Il rappelle que selon le type tumoral, un biomarqueur peut être prédictif dans une indication mais pas dans une autre : par exemple, **PD-L1 n'est pas prédictif dans le rein**, alors qu'il l'est dans le **cancer de la vessie**.

Il donne un exemple marquant de **petits sous-groupes** :

- Dans le cancer de la prostate, deux grandes études avaient testé un même biomarqueur sur ~40 patients chacune. Pris séparément, ces sous-groupes dilués dans la population générale n'avaient pas d'impact. Mais en regroupant ces 80 patients, l'analyse devient significative et suggère un effet réel.
  → Pour lui, cela illustre parfaitement que même des **petits effectifs**, si bien exploités et stratifiés, peuvent changer l'interprétation et orienter des essais cliniques.

4. Repositionnement thérapeutique

Il confirme que des altérations issues de survivants exceptionnels ou de sous-groupes rares peuvent révéler des opportunités de repositionnement.

- Exemple concret : **BRCA → inhibiteurs de PARP**, **MSI → immunothérapie**.
- Pour lui, ces cas montrent que l'étude d'outliers peut déboucher sur de nouveaux standards thérapeutiques.

5. Freins et conditions d'intégration

Les principaux freins identifiés sont :

- Le **temps** nécessaire à la validation clinique.
- La **reproductibilité et robustesse** statistique des signaux.

En revanche, il ne voit pas de freins éthiques ou académiques : "si un biomarqueur robuste apporte un bénéfice au patient, tout le monde est preneur".

Conditions minimales pour l'adoption en pratique :

1. Une méthode simple, tolérable pour le patient (ex. biopsie liquide).
2. Une robustesse et reproductibilité démontrées.
3. Un bénéfice clinique concret justifiant un changement de prise en charge.

6. Vision prospective

Pour l'avenir, il insiste sur deux points :

- L'avenir ne sera pas limité aux mutations uniques mais reposera sur des **signatures multi-omiques complexes**, intégrant plusieurs couches de données.
- L'**IA et le big data** joueront un rôle majeur dans la sélection des patients, la hiérarchisation des signaux et la conception des essais cliniques.

ANNEXE IV – Interview Summary with Paloma Cejas, PhD

Translational Researcher – Dana-Farber Cancer Institute

Format: Videoconference

1. Background and Translational Expertise

Paloma Cejas presents herself as a translational cancer researcher with extensive experience in molecular oncology. She obtained her PhD in 2003 and subsequently developed a research program focused on the molecular analysis of clinical samples to address clinically driven questions. Her training includes genetics and epigenetics, notably through work conducted at Dana-Farber Cancer Institute.

She currently leads a laboratory dedicated to multi-omic analyses with a strong translational orientation. Her expertise includes the analysis of FFPE samples, which are frequently encountered in retrospective clinical cohorts and are central to the work conducted at Q51. She also indicates that she is a co-founder of Q51, a company dedicated to the systematic study of exceptional cancer survivors.

2. Translational Context and Relevance of the Q51 Cohort

Paloma Cejas emphasizes that her primary scientific objective is the identification of biomarkers for patient stratification and the discovery of novel therapeutic targets. She highlights the particular relevance of the Q51 cohort, which brings together exceptional survivors across multiple tumor types, including glioblastoma, pancreatic ductal adenocarcinoma, and small-cell lung cancer.

She states that understanding the biological mechanisms underlying these exceptional outcomes requires a structured framework capable of systematically analyzing outliers, something that is difficult to achieve in a purely academic setting. According to her, the scale, consistency, and organization of the Q51 cohort make it particularly suited to this type of investigation.

3. Identification and Robustness of Biological Signatures

When discussing how to distinguish biologically meaningful signatures from statistical artifacts, Paloma Cejas stresses the importance of critical self-assessment throughout the analytical process. She explains that robustness relies on several key elements: the

use of both established and emerging computational pipelines, strict quality control to ensure sample reliability, and a strong grounding in cancer biology.

She underlines that biologically relevant signatures should be observed consistently across samples and methodologies. As an example, she refers to pancreatic ductal adenocarcinoma, where convergent signals can be identified using molecular analyses as well as histological approaches such as H&E staining. She concludes that consistency with existing literature and reproducibility across approaches are essential to constructing a credible biological narrative.

4. Validation Using Independent Cohorts and Public Resources

Paloma Cejas confirms the importance of validating findings in independent cohorts. She notes that the diversity of patients included in the Q51 cohort represents a major strength in this regard. While public resources such as TCGA can be used for contextual validation, she explains that these datasets are not fully suited to their objectives, particularly due to the lack of single-cell resolution.

Instead, she describes the use of publicly available single-cell datasets generated with comparable methodologies, allowing meaningful cross-study comparisons. She also mentions the use of functional resources such as the Cancer Dependency Map from the Broad Institute to explore cellular sensitivities and gene essentiality.

5. Interpretation of Omic Signals and Mechanistic Understanding

For the interpretation of multi-omic signatures, Paloma Cejas explains that gene set–based approaches, such as gene set enrichment analysis, are central to her workflow. These methods allow differentially expressed genes to be interpreted at the pathway level, facilitating the understanding of biological processes such as proliferation, immune activation, or resistance mechanisms.

She also mentions the exploratory use of emerging AI-based tools, such as Perplexity, to assist in hypothesis refinement. However, she strongly emphasizes that such tools cannot replace expert biological interpretation and must remain complementary to domain knowledge and critical reasoning.

6. Prioritization of Biological Pathways and Targets

When multiple pathways emerge from enrichment analyses, Paloma Cejas explains that prioritization cannot rely solely on biological relevance. Additional considerations include the competitive landscape, feasibility of drug development, and strategic positioning. She

notes that well-established targets may be commercially saturated, whereas entirely novel targets may be too speculative for near-term clinical development.

She further explains that prioritization is performed at the level of pathways rather than individual genes, as therapeutic development is not aimed at single patients. Decisions are informed by multidisciplinary discussions involving experts familiar with pharmaceutical development and drug modalities.

7. Experimental Validation and Functional Credibility

According to Paloma Cejas, a biological signature gains functional credibility when it is supported by coherent biological rationale, observed across multiple patients, and aligned with a clear development strategy. She indicates that extensive validation using complex models such as organoids or xenografts is not always required.

She emphasizes that the choice of experimental model must be adapted to the biological question. In particular, immune-related mechanisms cannot be adequately addressed using models lacking immune components, such as standard organoids or PDXs, and instead require co-culture or immune-competent systems.

8. Screening Approaches and Development Constraints

Paloma Cejas notes that high-throughput approaches such as genome-wide CRISPR screens are highly effective for identifying cell-autonomous cancer dependencies but are less informative for mechanisms involving the tumor microenvironment or immune interactions.

She stresses that, in an industrial context, prioritization and focus are essential. As Q51 is not an academic laboratory, resources must be directed toward assets with a realistic path to the clinic. This may require accepting that not every biological aspect can be exhaustively validated prior to clinical translation.

9. Therapeutic Development Strategies

Regarding therapeutic development, Paloma Cejas highlights the importance of selecting targets supported by robust signals across patients and compatible with rapid development strategies. She expresses a preference for antibody-based approaches, including antibody–drug conjugates, due to their specificity, speed of development, and favorable toxicity profiles, citing successful examples in bladder cancer.

She also considers drug repurposing as an attractive strategy when feasible, while acknowledging practical challenges related to intellectual property and patent ownership.

10. Limitations and Future Perspectives

Finally, Paloma Cejas identifies the relative novelty of multi-omic approaches and the ongoing evolution of analytical tools as a key limitation in translational oncology. She emphasizes that the complexity of these datasets necessitates multidisciplinary teams capable of integrating computational, biological, and clinical expertise.

She concludes that, despite these challenges, multi-omics represents a major opportunity to uncover meaningful biological insights, provided that analyses are conducted with rigor, prioritization, and a clear translational objective.

BIBLIOGRAPHIE

1. National Cancer Institute. Exceptional Responders Initiative: Questions and Answers [Internet]. 2018 [cited 2026 Feb 25]. Available from: https://www.cancer.gov/about-cancer/treatment/research/exceptional-responders-initiative-qa

2. Giraud P, Jean SL. Collège de cancérologie. 4th ed. Paris: MED-LINE; 2024.

3. Institut national du cancer. L'Institut national du cancer publie les dernières données en cancérologie dans son Panorama édition 2024 [Internet]. 2024 [cited 2026 Feb 25]. Available from: https://www.cancer.fr/presse/l-institut-national-du-cancer-publie-les-dernieres-donnees-en-cancerologie-dans-son-panorama-edition-2024

4. Griguolo G, Bottosso M, Crema A, Giarratano T, Miglietta F, Bonomi G, Mioranza E, Napetti D, Massa D, Faggioni G, Dieci MV, Guarneri V. Exceptional responses to systemic treatment in metastatic breast cancer: clinical features and long-term outcomes. Eur J Cancer. 2025 Mar 26;219:115321. doi: 10.1016/j.ejca.2025.115321. Epub 2025 Feb 20. PMID: 39987798.

5. Gunukula SR. Which is the best NGS approach for rare disease diagnosis: panels, WES, or WGS? [Internet]. 2018 [cited 2026 Feb 25]. Available from: https://3billion.io/blog/which-is-the-best-ngs-approach-for-rare-disease-diagnosis-panels-wes-or-wgs

6. Zhang Q, Qin Z, Yi S, Wei H, Zhou XZ, Su J. Clinical application of whole-exome sequencing: A retrospective, single-center study. Exp Ther Med. 2021 Jul;22(1):753. doi: 10.3892/etm.2021.10185. Epub 2021 May 12. PMID: 34035850; PMCID: PMC8135134.

7. Kerle, I.A., Gross, T., Kögler, A. *et al*. Translational and clinical comparison of whole genome and transcriptome to panel sequencing in precision oncology. *npj Precis. Onc.* **9**, 9 (2025). https://doi.org/10.1038/s41698-024-00788-3

8. Kim, R., Kim, S., Oh, B.BL. *et al*. Clinical application of whole-genome sequencing of solid tumors for precision oncology. *Exp Mol Med* **56**, 1856–1868 (2024). https://doi.org/10.1038/s12276-024-01288-x

9. Ye J, Lin Y, Liao Z, Gao X, Lu C, Lu L, Huang J, Huang X, Huang S, Yu H, Bai T, Chen J, Wang X, Xie M, Luo M, Zhang J, Wu F, Wu G, Ma L, Xiang B, Li L, Li Y, Luo X, Liang R. Single cell-spatial transcriptomics and bulk multi-omics analysis of heterogeneity and ecosystems in hepatocellular carcinoma. NPJ Precis Oncol. 2024 Nov 15;8(1):262. doi: 10.1038/s41698-024-00752-1. PMID: 39548284; PMCID: PMC11568154.

10. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 May 5;18(1):83. doi: 10.1186/s13059-017-1215-1. PMID: 28476144; PMCID: PMC5418815.

11. Kamath S, Roopkumar J, Ni Y, Shen M, Bejarano P, Allende D, Nagarajan A, Nguyen T, Dergham B, Shepard D, Shapiro MA, McNamara MJ, Estfan BN, Nair KG, Khorana AA. Genomic Predictors Associated With Exceptional Response to Systemic Therapy in Advanced Pancreatic Cancer. Oncology (Williston Park). 2023 Dec 14;37(12):488-495. doi: 10.46883/2023.25921008. PMID: 38133563..

12. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, Buettner F, Huber W, Stegle O. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018 Jun 20;14(6):e8124. doi: 10.15252/msb.20178124. PMID: 29925568; PMCID: PMC6010767.

13. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016 Apr 13;374(2065):20150202. doi: 10.1098/rsta.2015.0202. PMID: 26953178; PMCID: PMC4792409.

14. Gygi JP, Konstorum A, Pawar S, Aron E, Kleinstein SH, Guan L. A supervised Bayesian factor model for the identification of multi-omics signatures. Bioinformatics. 2024 May 2;40(5):btae202. doi: 10.1093/bioinformatics/btae202. PMID: 38603606; PMCID: PMC11078774.

15. Knox, J.J., Jang, G.H., Grant, R.C. *et al.* Whole genome and transcriptome profiling in advanced pancreatic cancer patients on the COMPASS trial. *Nat Commun* **16**, 5919 (2025). https://doi.org/10.1038/s41467-025-60808-z

16. Tao Y, Xing S, Zuo S, Bao P, Jin Y, Li Y, Li M, Wu Y, Chen S, Wang X, Zhu Y, Feng Y, Zhang X, Wang X, Xi Q, Lu Q, Wang P, Lu ZJ. Cell-free multi-omics analysis reveals potential biomarkers in gastrointestinal cancer patients' blood. Cell Rep Med. 2023 Nov 21;4(11):101281. doi: 10.1016/j.xcrm.2023.101281. PMID: 37992683; PMCID: PMC10694666.

17. Peng, J., Sun, J., Yu, Y. *et al.* Integrative multi-omics analysis reveals the role of toll-like receptor signaling in pancreatic cancer. *Sci Rep* **15**, 52 (2025). https://doi.org/10.1038/s41598-024-84062-3

18. Mohamed SH, Hamed M, Alamoudi HA, Jastaniah Z, Alakwaa FM, Reda A. Multi-omics analysis of Helicobacter pylori-associated gastric cancer identifies hub genes as a novel therapeutic biomarker. Brief Bioinform. 2025 May 1;26(3):bbaf241. doi: 10.1093/bib/bbaf241. PMID: 40445003; PMCID: PMC12123523.

19. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47. doi: 10.1093/nar/gkv007. Epub 2015 Jan 20. PMID: 25605792; PMCID: PMC4402510.

20 . Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000 Jan 1;28(1):27-30. doi: 10.1093/nar/28.1.27. PMID: 10592173; PMCID: PMC102409.

21. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011 Jun 15;27(12):1739-40. doi: 10.1093/bioinformatics/btr260. Epub 2011 May 5. PMID: 21546393; PMCID: PMC3106198.

22. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330-D338. doi: 10.1093/nar/gky1055. PMID: 30395331; PMCID: PMC6323945.

23. Xie X, Qiu G, Chen Z, Liu T, Yang Y, You Z, Zeng C, Lin X, Xie Z, Qin Y, Wang Y, Ma X, Zhou C, Liu M. Characteristics and prognosis of EGFR mutations in small cell lung cancer patients in the NGS era. Clin Transl Oncol. 2024 Feb;26(2):434-445. doi: 10.1007/s12094-023-03263-w. Epub 2023 Jul 12. PMID: 37436674; PMCID: PMC10811109.

24. Kindler HL, Hammel P, Reni M, Van Cutsem E, Macarulla T, Hall MJ, Park JO, Hochhauser D, Arnold D, Oh DY, Reinacher-Schick A, Tortora G, Algül H, O'Reilly EM, Bordia S, McGuinness D, Cui K, Locker GY, Golan T. Overall Survival Results From the POLO Trial: A Phase III Study of Active Maintenance Olaparib Versus Placebo for Germline BRCA-Mutated Metastatic Pancreatic Cancer. J Clin Oncol. 2022 Dec 1;40(34):3929-3939. doi: 10.1200/JCO.21.01604. Epub 2022 Jul 14. Erratum in: J Clin Oncol. 2024 Jun 10;42(17):2112. doi: 10.1200/JCO.24.00821. PMID: 35834777; PMCID: PMC10476841.

25. Kamii, M., Kamata, R., Saito, H. *et al.* PARP inhibitors elicit a cellular senescence mediated inflammatory response in homologous recombination proficient cancer cells. *Sci Rep* **15**, 15458 (2025). https://doi.org/10.1038/s41598-025-00336-4

26Hopkins, A., Groom, C. The druggable genome. *Nat Rev Drug Discov* **1**, 727–730 (2002). https://doi.org/10.1038/nrd892

27. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, Griffith M, Griffith OL, Wagner AH. Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res. 2021 Jan 8;49(D1):D1144-D1151. doi: 10.1093/nar/gkaa1084. PMID: 33237278; PMCID: PMC7778926.

28. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, Gill S, Harrington WF, Pantel S, Krill-Burger JM, Meyers RM, Ali L, Goodale A, Lee Y, Jiang G, Hsiao J, Gerath WFJ, Howell S, Merkel E, Ghandi M, Garraway LA, Root DE, Golub TR, Boehm JS, Hahn WC. Defining a Cancer Dependency Map. Cell. 2017 Jul 27;170(3):564-576.e16. doi: 10.1016/j.cell.2017.06.010. PMID: 28753430; PMCID: PMC5667678.

29. Gebara NY, El Kamari V, Rizk N. HIV-1 elite controllers: an immunovirological review and clinical perspectives. J Virus Erad. 2019 Sep 18;5(3):163-166. doi: 10.1016/S2055-6640(20)30046-7. PMID: 31700663; PMCID: PMC6816117.

Université de Lille

UFR3S-Pharmacie

DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

Année Universitaire 2025/2026

Nom : RAMDANI
Prénom : Kamil

Étude des survivants exceptionnels du cancer : comment l'analyse approfondie de leurs données cliniques, génomiques et biologiques peut-elle orienter la découverte de nouveaux traitements anticancéreux ?

Mots-clés : Survivants exceptionnels; Oncologie; Multi-omique; Signatures Transcriptomique; Génomique; Méthodologie analytique; Recherche translationnelle

---

Résumé :

Cette thèse s'intéresse aux survivants exceptionnels du cancer, définis comme des patients atteints de cancers de mauvais pronostic mais présentant une survie prolongée inattendue. L'objectif principal n'est pas d'identifier des biomarqueurs spécifiques, mais de présenter une démarche méthodologique rigoureuse permettant d'exploiter ces situations cliniques atypiques.

À travers l'analyse critique d'approches multi-omiques intégrant génomique, transcriptomique, single-cell et transcriptomique spatiale, ce travail explore les protocoles analytiques nécessaires pour identifier des signatures biologiques robustes dans des cohortes rares. Une attention particulière est portée aux enjeux de qualité des données, de reproductibilité, d'interprétation fonctionnelle et de validation.

La thèse discute enfin les implications cliniques et translationnelles de ces signatures, en mettant l'accent sur les conditions nécessaires pour transformer un signal multi-omique en hypothèse thérapeutique exploitable en oncologie.

Membres du jury :

Président : Cazin, Jean Louis, PU Professeur de Pharmacologie et Pharmacie Clinique à UFR3 Pharmacie, Université de lille / chef d'unité au centre Oscar Lambret

Directeur, conseiller de thèse : Pinçon, Claire, Maîtresse de conférences à UFR3, Université de Lille

Assesseur : BENHALIMA, Ilyès, Assistant Hospitalo-Universitaire CHU lyon)