Université de Lille
Année Universitaire 2020/2021

Faculté de pharmacie Lille

THESE
POUR LE DIPLOME D'ETAT
DE DOCTEUR EN PHARMACIE

Soutenue publiquement le 22 octobre 2021
Par Mr Augustin BOUDRY

THESE EN VUE DU DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

Développement d'un nouvel algorithme pour la détection et la quantification des duplications en tandem de *FLT3* dans les leucémies aiguës myéloïdes

---

Membres du jury :

Président : Dupont Annabelle, PU-PH à la faculté de pharmacie de Lille

Assesseur(s) : Quesnel Bruno, PU-PH à la faculté de médecine de Lille
Duployez Nicolas, MCU-PH à la faculté de médécine de Lille
Figeac Martin, IR à la faculté de médecine de Lille

Directeur de thèse : Preudhomme Claude, PU-PH à la faculté de médecine de Lille

Faculté de Pharmacie
de Lille

Université de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX
☎ 03.20.96.40.40 - 🖷 : 03.20.96.43.64

Université de Lille

Président :	Jean-Christophe CAMART
Premier Vice-président :	Nicolas POSTEL
Vice-présidente formation :	Lynne FRANJIÉ
Vice-président recherche :	Lionel MONTAGNE
Vice-président relations internationales :	François-Olivier SEYS
Vice-président stratégie et prospective	Régis BORDET
Vice-présidente ressources	Georgette DAL
Directeur Général des Services :	Pierre-Marie ROBERT
Directrice Générale des Services Adjointe :	Marie-Dominique SAVINA

Faculté de Pharmacie

Doyen :	Bertrand DÉCAUDIN
Vice-doyen et Assesseur à la recherche :	Patricia MELNYK
Assesseur aux relations internationales :	Philippe CHAVATTE
Assesseur aux relations avec le monde professionnel :	Thomas MORGENROTH
Assesseur à la vie de la Faculté :	Claire PINÇON
Assesseur à la pédagogie :	Benjamin BERTIN
Responsable des Services :	Cyrille PORTA
Représentant étudiant :	Victoire LONG

Liste des Professeurs des Universités - Praticiens Hospitaliers

Civ.	Nom	Prénom	Laboratoire
Mme	ALLORGE	Delphine	Toxicologie et Santé publique
M.	BROUSSEAU	Thierry	Biochimie
M.	DÉCAUDIN	Bertrand	Biopharmacie, Pharmacie Galénique et Hospitalière
M.	DEPREUX	Patrick	Institut de Chimie Pharmaceutique Albert LESPAGNOL
M.	DINE	Thierry	Pharmacologie, Pharmacocinétique et Pharmacie clinique
Mme	DUPONT-PRADO	Annabelle	Hématologie
Mme	GOFFARD	Anne	Bactériologie - Virologie

M.	GRESSIER	Bernard	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	ODOU	Pascal	Biopharmacie, Pharmacie Galénique et Hospitalière
Mme	POULAIN	Stéphanie	Hématologie
M.	SIMON	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	STAELS	Bart	Biologie cellulaire

Liste des Professeurs des Universités

Civ.	Nom	Prénom	Laboratoire
M.	ALIOUAT	El Moukhtar	Parasitologie - Biologie animale
Mme	AZAROUAL	Nathalie	Biophysique et Laboratoire d'application de RMN
M.	CAZIN	Jean-Louis	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	CHAVATTE	Philippe	Institut de Chimie Pharmaceutique Albert LESPAGNOL
M.	COURTECUISSE	Régis	Sciences Végétales et Fongiques
M.	CUNY	Damien	Sciences Végétales et Fongiques
Mme	DELBAERE	Stéphanie	Biophysique et application de RMN
Mme	DEPREZ	Rebecca	Médicaments et molécules pour agir sur les systèmes vivants
M.	DEPREZ	Benoît	Médicaments et molécules pour agir sur les systèmes vivants
M.	DUPONT	Frédéric	Sciences Végétales et Fongiques
M.	DURIEZ	Patrick	Physiologie
M.	FOLIGNÉ	Benoît	Bactériologie - Virologie
M.	GARÇON	Guillaume	Toxicologie et Santé publique
Mme	GAYOT	Anne	Pharmacotechnie industrielle
M.	GOOSSENS	Jean-François	Chimie analytique
M.	HENNEBELLE	Thierry	Pharmacognosie
M.	LEBEGUE	Nicolas	Chimie thérapeutique
M.	LEMDANI	Mohamed	Biomathématiques
Mme	LESTAVEL	Sophie	Biologie cellulaire

Mme	LESTRELIN	Réjane	Biologie cellulaire
Mme	MELNYK	Patricia	Chimie thérapeutique
M.	MILLET	Régis	Institut de Chimie Pharmaceutique Albert LESPAGNOL
Mme	MUHR-TAILLEUX	Anne	Biochimie
Mme	PERROY	Anne-Catherine	Législation et Déontologie pharmaceutique
Mme	ROMOND	Marie-Bénédicte	Bactériologie - Virologie
Mme	SAHPAZ	Sevser	Pharmacognosie
M.	SERGHERAERT	Éric	Législation et Déontologie pharmaceutique
M.	SIEPMANN	Juergen	Pharmacotechnie industrielle
Mme	SIEPMANN	Florence	Pharmacotechnie industrielle
M.	WILLAND	Nicolas	Médicaments et molécules pour agir sur les systèmes vivants

Liste des Maîtres de Conférences - Praticiens Hospitaliers

Civ.	Nom	Prénom	Laboratoire
Mme	BALDUYCK	Malika	Biochimie
Mme	GARAT	Anne	Toxicologie et Santé publique
Mme	GENAY	Stéphanie	Biopharmacie, Pharmacie Galénique et Hospitalière
M.	LANNOY	Damien	Biopharmacie, Pharmacie Galénique et Hospitalière
Mme	ODOU	Marie-Françoise	Bactériologie - Virologie

Liste des Maîtres de Conférences

Civ.	Nom	Prénom	Laboratoire
M.	AGOURIDAS	Laurence	Chimie thérapeutique
Mme	ALIOUAT	Cécile-Marie	Parasitologie - Biologie animale
M.	ANTHÉRIEU	Sébastien	Toxicologie et Santé publique
Mme	AUMERCIER	Pierrette	Biochimie
M.	BANTUBUNGI-BLUM	Kadiombo	Biologie cellulaire

Mme	BARTHELEMY	Christine	Biopharmacie, Pharmacie Galénique et Hospitalière
Mme	BEHRA	Josette	Bactériologie - Virologie
M.	BELARBI	Karim-Ali	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	BERTHET	Jérôme	Biophysique et Laboratoire d'application de RMN
M.	BERTIN	Benjamin	Immunologie
M.	BLANCHEMAIN	Nicolas	Pharmacotechnie industrielle
M.	BORDAGE	Simon	Pharmacognosie
M.	BOSC	Damien	Médicaments et molécules pour agir sur les systèmes vivants
M.	BRIAND	Olivier	Biochimie
M.	CARNOY	Christophe	Immunologie
Mme	CARON-HOUDE	Sandrine	Biologie cellulaire
Mme	CARRIÉ	Hélène	Pharmacologie, Pharmacocinétique et Pharmacie clinique
Mme	CHABÉ	Magali	Parasitologie - Biologie animale
Mme	CHARTON	Julie	Médicaments et molécules pour agir sur les systèmes vivants
M.	CHEVALIER	Dany	Toxicologie et Santé publique
Mme	DANEL	Cécile	Chimie analytique
Mme	DEMANCHE	Christine	Parasitologie - Biologie animale
Mme	DEMARQUILLY	Catherine	Biomathématiques
M.	DHIFLI	Wajdi	Biomathématiques
Mme	DUMONT	Julie	Biologie cellulaire
M.	EL BAKALI	Jamal	Chimie thérapeutique
M.	FARCE	Amaury	Institut de Chimie Pharmaceutique Albert LESPAGNOL
M.	FLIPO	Marion	Médicaments et molécules pour agir sur les systèmes vivants
Mme	FOULON	Catherine	Chimie analytique
M.	FURMAN	Christophe	Institut de Chimie Pharmaceutique Albert LESPAGNOL
M.	GERVOIS	Philippe	Biochimie

Mme	GOOSSENS	Laurence	Institut de Chimie Pharmaceutique Albert LESPAGNOL
Mme	GRAVE	Béatrice	Toxicologie et Santé publique
Mme	GROSS	Barbara	Biochimie
M.	HAMONIER	Julien	Biomathématiques
Mme	HAMOUDI-BEN YELLES	Chérifa-Mounira	Pharmacotechnie industrielle
Mme	HANNOTHIAUX	Marie-Hélène	Toxicologie et Santé publique
Mme	HELLEBOID	Audrey	Physiologie
M.	HERMANN	Emmanuel	Immunologie
M.	KAMBIA KPAKPAGA	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	KARROUT	Younes	Pharmacotechnie industrielle
Mme	LALLOYER	Fanny	Biochimie
Mme	LECOEUR	Marie	Chimie analytique
Mme	LEHMANN	Hélène	Législation et Déontologie pharmaceutique
Mme	LELEU	Natascha	Institut de Chimie Pharmaceutique Albert LESPAGNOL
Mme	LIPKA	Emmanuelle	Chimie analytique
Mme	LOINGEVILLE	Florence	Biomathématiques
Mme	MARTIN	Françoise	Physiologie
M.	MOREAU	Pierre-Arthur	Sciences Végétales et Fongiques
M.	MORGENROTH	Thomas	Législation et Déontologie pharmaceutique
Mme	MUSCHERT	Susanne	Pharmacotechnie industrielle
Mme	NIKASINOVIC	Lydia	Toxicologie et Santé publique
Mme	PINÇON	Claire	Biomathématiques
M.	PIVA	Frank	Biochimie
Mme	PLATEL	Anne	Toxicologie et Santé publique
M.	POURCET	Benoît	Biochimie
M.	RAVAUX	Pierre	Biomathématiques / service innovation pédagogique

Mme	RAVEZ	Séverine	Chimie thérapeutique
Mme	RIVIÈRE	Céline	Pharmacognosie
M.	ROUMY	Vincent	Pharmacognosie
Mme	SEBTI	Yasmine	Biochimie
Mme	SINGER	Elisabeth	Bactériologie - Virologie
Mme	STANDAERT	Annie	Parasitologie - Biologie animale
M.	TAGZIRT	Madjid	Hématologie
M.	VILLEMAGNE	Baptiste	Médicaments et molécules pour agir sur les systèmes vivants
M.	WELTI	Stéphane	Sciences Végétales et Fongiques
M.	YOUS	Saïd	Chimie thérapeutique
M.	ZITOUNI	Djamel	Biomathématiques

Professeurs Certifiés

Civ.	Nom	Prénom	Laboratoire
Mme	FAUQUANT	Soline	Anglais
M.	HUGES	Dominique	Anglais
M.	OSTYN	Gaël	Anglais

Professeur Associé - mi-temps

Civ.	Nom	Prénom	Laboratoire
M.	DAO PHAN	Haï Pascal	Médicaments et molécules pour agir sur les systèmes vivants
M.	DHANANI	Alban	Législation et Déontologie pharmaceutique

Maîtres de Conférences ASSOCIES - mi-temps

Civ.	Nom	Prénom	Laboratoire
Mme	CUCCHI	Malgorzata	Biomathématiques
M.	DUFOSSEZ	François	Biomathématiques
M.	FRIMAT	Bruno	Pharmacologie, Pharmacocinétique et Pharmacie clinique

M.	GILLOT	François	Législation et Déontologie pharmaceutique
M.	MASCAUT	Daniel	Pharmacologie, Pharmacocinétique et Pharmacie clinique
M.	ZANETTI	Sébastien	Biomathématiques

AHU

Civ.	Nom	Prénom	Laboratoire
Mme	CUVELIER	Élodie	Pharmacologie, Pharmacocinétique et Pharmacie clinique
Mme	DEMARET	Julie	Immunologie
M.	GRZYCH	Guillaume	Biochimie
Mme	HENRY	Héloïse	Biopharmacie, Pharmacie Galénique et Hospitalière
Mme	MASSE	Morgane	Biopharmacie, Pharmacie Galénique et Hospitalière

ATER

Civ.	Nom	Prénom	Laboratoire
M.	GHARBI	Zied	Biomathématiques
Mme	FLÉAU	Charlotte	Médicaments et molécules pour agir sur les systèmes vivants
Mme	N'GUESSAN	Cécilia	Parasitologie - Biologie animale
M.	RUEZ	Richard	Hématologie
M.	SAIED	Tarak	Biophysique et Laboratoire d'application de RMN
Mme	VAN MAELE	Laurye	Immunologie

Enseignant contractuel

Civ.	Nom	Prénom	Laboratoire
M.	MARTIN MENA	Anthony	Biopharmacie, Pharmacie Galénique et Hospitalière

Université de Lille

Faculté de Pharmacie de Lille

Faculté de Pharmacie de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX
Tel. : 03.20.96.40.40 - Télécopie : 03.20.96.43.64
http://pharmacie.univ-lille2.fr

L'Université n'entend donner aucune approbation aux opinions émises dans les thèses ; celles-ci sont propres à leurs auteurs.

Remerciements

À ma présidente de thèse :

Madame le Professeur Annabelle DUPONT
Professeur des Universités – Praticien Hospitalier
Institut d'Hématologie & Transfusion
Centre Hospitalier Universitaire de Lille

Vous me faites l'honneur de présider cette thèse. Je vous remercie pour votre disponibilité, de m'avoir donné goût à l'hématologie et de m'avoir suivi tout au long de mon cursus. Soyez assurée de mon profond respect.

À mes juges :

Monsieur le Professeur Bruno QUESNEL
Professeur des Universités – Praticien Hospitalier
Service des Maladies du Sang
Centre Hospitalier Universitaire de Lille

Monsieur le Docteur Nicolas DUPLOYEZ
Maitre de Conférence Universitaire et Praticien Hospitalier
Institut d'Hématologie & Transfusion
Centre Hospitalier Universitaire de Lille

Monsieur Martin FIGEAC
Ingénieur de recherche
Plate-forme de génomique fonctionnelle et structurale Go@L-GFS
UMS 2014 - US 41
Université de Lille
Cellule bio-informatique, plateau commun de biologie moléculaire
Centre Hospitalier Universitaire de Lille

Je vous remercie d'avoir accepté d'évaluer cette thèse et de l'intérêt que vous portez à ce travail. Veuillez trouver ici l'expression de mes sentiments les plus sincères.

À mon directeur de thèse :

Monsieur le Professeur Claude Preudhomme
Professeur des Universités – Praticien Hospitalier
Chef de service du laboratoire d'Hématologie
Centre Hospitalier Universitaire de Lille

Je tiens à vous remercier chaleureusement de m'avoir confié ce travail, de m'avoir accepté dans votre service durant mon année de FST, de m'avoir fait confiance, ainsi que pour votre bienveillance et votre gentillesse.

À mes collègues :
A toutes les équipes de techniciens incroyables que j'ai rencontrées durant mon parcours ! St-Phi, l'immuno, la cyto et surtout l'hémostase je me rends compte de la chance que j'ai eue de travailler avec vous, dans cette complicité et cette bonne humeur. À mon laboratoire actuel l'onco-hématologie, je m'y sens bien grâce à vous tous ! Aux techniciens du lymphoïde Aurore, Camille, Coralie, Danièle, Élise, Lucie, Marie, Mélanie, Paulette et Steeve merci pour vos enseignements et votre aide constante même en pause-café ;). À la team LAM 48h Hélène Olivia Phiphi merci de m'avoir aidé sur ce projet. À ma Karine merci de m'avoir accueilli à bras ouverts dès le premier jour, pour ta gentillesse et pour ta disponibilité. Je n'oublie pas Sophie pour son accueil très très chaleureux et pour m'avoir enseigné l'art du gel en polyacrylamide.

Merci à tous les biologistes qui m'ont encadré et formé pendant mes années d'internat. À Manu pour son énergie et sa joie de vivre qui me donne le sourire tous les jours. À Florent, pour tous ces partages en informatique/cytologie. À Élise et Coralie pour leur enseignement en cytologie. À Alice, pour l'aide que tu m'apportes au quotidien en BM et pour celle que tu m'as apportée sur ce projet. À Olivier, pour ses précieux conseils en biologie et sa pédagogie. À Nathalie, pour ma formation en lymphoïde et les discussions sur la peinture. À Laurène et Nicolas, qui forment le meilleur duo d'encadrants que j'ai pu avoir. J'ai de la chance de vous avoir et je suis très fier de travailler à vos côtés ! Nicolas, merci pour ton humour, merci de toujours trouver réponse à mes questions. Laurène merci de m'avoir appris la rédaction et la rigueur scientifique (il y a encore du boulot), merci pour ton aide indéfectible. Et surtout merci pour votre bienveillance.

À Fanny, pour son efficacité à toute épreuve et pour m'apporter une grande aide quand je suis en train de me noyer.

Merci aux ingénieurs Aurélie, Maxime, Nathalie et Sandrine pour leur disponibilité et leurs précieux conseils, leur aide en manip avec le petit singe.

À Mikaël, sans qui ce projet ne serait rien ! Merci pour tes précieux enseignements, tes longues explications aux tableau et ta patience ; un pédagogue comme on en fait plus !

Merci à la Plateforme de génomique Shéhérazade, Céline, Malo, JP et mon Frédo merci pour vos conseils, votre aide, c'est toujours un plaisir de passer vous voir. À Martin, tu es pour moi ce qui s'approche le plus d'un maitre en bio-informatique, ton aide quotidienne me fait beaucoup grandir. Nos dessins au tableau font partie des moments-clés de mon internat j'espère qu'il y en aura beaucoup d'autres !

À vous les copains, Chadi, Christophe, Émilie, Fabrice, Piétro et Simon : la quantité de connaissances que vous m'avez transmise n'est pas mesurable. Merci pour tous ces singularity foireux que vous avez du builder, les dépannages nextflow d'urgence, les runs sur dragenv3 :p, ton apprentissage de nextflow Chadi, tes métaphores beaucoup trop marquantes Christophe, ton expertise en BM et ton soutien moral sur les nouveaux projets Émilie, tes petits cours plus qu'intéressants Fab, tes exposés sur la physique qui me mindblow Piétro, tes explications magiques en électronique Simon et votre aide indéfectible à tous. Sans vous le petit parasite du cluster ne s'en serait jamais sorti !

À Monsieur le grand professeur très musclé Leo-Paul Bancel, merci d'être différent.

À ma co-interne en or Romane, travailler à tes côtés est un véritable bonheur ! Merci pour toutes ces relectures, tous ces conseils, tous ces rires, tous ces awkward moments. Je pense ne pas me tromper en disant que tu es largement dans le top 5 des personnes les plus bienveillantes que j'ai pu rencontrer. Merci d'exister.

À mes amis :

Merci à mes meilleurs copains JoleJo Rou Meyer pour tous ces bons moments passés ensemble que ça soit dévaler les contrées d'Azeroth, passer des soirées à slacker sur le /b/, nos soirées endiablées dans des caves surtout la mémorable avec Domi et Martine… Vous êtes mes piliers, vous rendez la vie beaucoup plus douce !

Jo, mon ppv pour m'avoir accueilli dès la PACES (m'avoir carry avec manu et tibar), passé le non anniversaire de Simon en regardant des supers films au lit et pour tous nos duos de costumes plus que réussis. Tu es mon roc !

Sim, mon Bébel merSim d'avoir pris de nombreuses photos pour les souvenirs et de relever le niveau enfin pas trop quand même (les fois où tu ne faisais pas de syncope…).

Mey, pour nos soirées philosophique chasse et pêche à déguster des Dwitch et à chercher les chips dans les murs ou appeler l'hôpital pour les copains.

Il n'est pas de mot en elfique, en entique ou en langages des Hommes pour qualifier notre amitié.

Merci à Jirnos aka jeremylink pour sa gentillesse, sa générosité et son calme olympien pour gérer cette bande de cons (range bien ton katana). MINUSCULE METEORE DE MELF naver forgetti.

Merci également aux autres pro-gamers du teamspeak pour les bons moments, tous ces rires et leur patience avec le joueur d'assiettes.

À Marie et Lulu pour m'avoir souvent bichonné en soirée même lorsque je me battais avec des poubelles.

Je n'oublie pas également ~~Alice~~ maman notre britney pref à tous, merci de nous avoir gérés tous, avec amour malgré nos petits défauts <3.

Merci à Botte pour tous ces bons moments, tous ces concerts, toutes ces découvertes. J'ai coché un paquet de truc de la liste « À faire avant 30 ans » avec toi j'espère qu'on aura l'occas d'en cocher d'autres ! Je ne vous oublie pas aussi le reste de l'amical des Kancer bisous de dauphin surtout à TON PÈRE Djamel.

Merci à Josette pour cet éveil à la peinture, à la musique et à la cuisine. Tu es ma petite fenêtre hors du monde de la science. Notre amitié pour moi est juste irésumable.

<u>À ma famille :</u>
Papa, tu as toujours prétendu qu'il fallait imaginer Sisyphe heureux, j'ai mis beaucoup de temps à saisir ta maxime mais aujourd'hui je pense avoir enfin compris. Merci pour tous les enseignements que tu m'as apportés, tout cet amour, je ne serais pas ce que je suis sans toi aujourd'hui.

Maman, ton bâton de vieillesse devient enfin grand ! Merci d'avoir toujours été là, de m'avoir donné tant d'amour, d'avoir tenu bon alors que je t'en faisais voir de toutes les couleurs. Personne ne peut imaginer le roc à toute épreuve que tu as été durant toutes ces années…

Hélène, ma deuxième maman, merci d'avoir toujours été présente pour moi.

Guillaume, frérot tu ne peux pas imaginer comme je suis fier de toi, petit chevalier blanc, le monde irait mieux s'il y avait plus de gens comme toi.

Sopq, soeurette, on a vécu des moments très durs à deux mais ça a forgé une relation immuable. Ohana signifie famille, famille signifie que personne ne doit être abandonné, ni oublié.

Angèle, car elle est tout.

Liste des abréviations

ADN : acide désoxyribonucléique

AMM : autorisation de mise sur le marché

AR : allelic ratio

ARN : acide ribonucléique

AUC : area under the curve

bcl : binary base call

BWA : Burrows-Wheeler Aligner

CBF : core biding factor

CD : cluster de différenciation

CIVD : coagulation intravasculaire disséminée

ddNTP : didésoxynucléotide

ELN : European Leukemia Net

FAB : franco-américano-britannique

FISH : fluorescent in situ hybridization

FLK-2 : fetal liver kinase-2

FLT3 : FMS-like Tyrosine Kinase 3

FLT3-ITD : *FLT3*-internal tandem duplication

FLT3-TKD : *FLT3*-tyrosine kinase domain

FLT3i : inhibiteur de FLT3

FLT3L : FLT3 ligand

GATK : genome analysis toolkit

GO : gemtuzumab ozogamicine

LAL : leucémie aiguë lymphoblastique

LAM : leucémie aiguë myéloïde

LAM-CN : LAM à caryotype normal

LAP : leucémie aiguë promyélocytaire

LMMC : leucémie myélomonocytaire chronique

NGS : séquençage de nouvelle génération

PCR : polymerase chain reaction

RC : rémission complète

Se : sensibilité

SMD : syndrome myélodysplasique

SMP : syndrome myéloprolifératif

Sp : spécificité

TdT : terminal deoxynucleotidyl transferase

VAF : variant allele frequency

VPP : valeur prédictive positive

Table des matières

Index des figures

Index des Tableaux

1 Introduction

1.1 Leucémies aiguës myéloïdes

1.1.1 Définition

Les leucémies aiguës myéloïdes (LAM) forment un groupe de maladies hématologiques, phénotypiquement et génétiquement hétérogènes, caractérisées par une accumulation anormale de cellules myéloïdes blastiques au niveau de la moelle osseuse, du sang périphérique et parfois dans d'autres tissus. Ainsi, l'infiltration de ces blastes dans la moelle s'accompagne, presque invariablement d'une insuffisance médullaire.

1.1.2 Épidémiologie

Les LAM représentent 90% de toutes les leucémies aiguës chez les adultes (1), avec un nombre estimé de 3428 nouveaux cas en 2018 d'après le Réseau des registres du cancer (FRANCIM). L'incidence annuelle est d'environ 4,3 pour 100 000 et augmente avec l'âge (2), avec un risque environ 10 fois plus élevé pour les sujets de plus de 65 ans (20,1 cas pour 100 000) que pour ceux de moins de 65 ans (2 cas pour 100 000) (1). L'âge médian au moment du diagnostic est d'environ 68 ans (2). Chez l'adulte, la survie globale à 5 ans des patients atteints de LAM est de 24%, ce qui la positionne au 5$^{\text{ème}}$ rang des cancers aux pronostics les plus défavorables (1). Chez l'enfant, elle est d'environ 70 % (3). Les hommes sont 1,2 à 1,6 fois plus susceptibles que les femmes de développer une LAM au cours de leur vie (4,5).

1.1.3 Étiologie

Dans la plupart des cas de LAM, aucune étiologie ne peut être identifiée.

1.1.3.1 Facteurs environnementaux

Seuls quatre facteurs environnementaux ont fait l'objet d'études de cohorte et sont donc des agents causaux établis : l'exposition chronique à de fortes doses de benzène (6–10), l'exposition à de fortes doses de rayonnements (11–13), le tabagisme chronique (14,15), et certains antinéoplasiques (agents alkylants et inhibiteurs de topoisomérases II) (16–20).

Récemment, une étude de cohorte incrimine l'exposition longue au glyphosate avec un niveau de preuve très bas (RR 2.04 [1.05 ; 3.97]), d'autres études sont donc nécessaires afin de pouvoir conclure à une causalité (21).

L'obésité constitue un facteur environnemental endogène augmentant de façon significative le risque de développer une LAM. En effet, des études de cohorte et des méta-analyses mettent en évidence un risque accru chez les hommes et les femmes ayant un indice de masse corporelle élevé et ceci est particulièrement notable pour la leucémie aiguë promyélocytaire (LAP) (22).

Des études cas-témoins ont parfois mis en évidence une relation entre les LAM et les solvants organiques (23) et les pesticides (24) mais à des niveaux de preuves inférieurs à ceux cités précédemment et parfois controversés (25).

1.1.3.2 Maladies prédisposantes

Plusieurs maladies peuvent être prédisposantes au développement des LAM :

- Maladies héréditaires : anémie de Fanconi (26) (défauts de réparation de l'ADN), dyskératose congénitale (27) (téloméropathies), thrombopénies constitutionnelles avec mutations de *RUNX1*.
- Syndrome de Down (trisomie 21 constitutionnelle).
- Acutisation d'autres hémopathies : syndrome myélodysplasique (SMD) (28), syndrome myéloprolifératif (SMP) (29), leucémie myélomonocytaire chronique (LMMC) (30,31). L'instabilité génomique qui conduit à de nouvelles mutations pourrait expliquer cette progression clonale. Cependant, l'impact potentiel des différents chimiothérapies et/ou radiothérapies ne doit pas être écarté.

1.2 Présentation clinique

Le tableau clinique est peu spécifique de la maladie et très variable allant de formes pauci à hypersymptomatiques.

Des signes non spécifiques peuvent être présents comme une altération de l'état général, une anorexie, une perte de poids, des suées et une hyperthermie.

Une insuffisance médullaire est quasi systématiquement retrouvée avec (i) des symptômes d'anémie : pâleur, fatigue, palpitations et/ou dyspnée à l'effort (ii) des symptômes de thrombopénie : ecchymoses, pétéchies, épistaxis et des saignements prolongés après une blessure mineure (iii) des infections en cas d'agranulocytose.

Une hyperleucocytose est observée chez environ 5 à 20% des patients (32) et peut induire dans de rares cas une hépatomégalie et/ou splénomégalie. Une leucostase peut survenir en cas d'hyperleucocytose $>$ 100 G/L, ce qui conduit à une augmentation de la viscosité du sang et favorise la formation de thrombi aboutissant à une baisse de la perfusion tissulaire au niveau du système nerveux central, des poumons et d'autres organes.

Certains patients peuvent présenter une coagulation intravasculaire disséminée (CIVD), en particulier ceux atteints de leucémie aiguë promyélocytaire. De ce fait, l'évaluation de la numération plaquettaire, du temps de prothrombine, du temps de céphaline activée et du fibrinogène doit faire partie du bilan biologique au diagnostic.

1.3 Présentation biologique

1.3.1 Hémogramme

A l'instar du tableau clinique, il existe une grande variabilité au niveau des anomalies de l'hémogramme qui peuvent être retrouvées au diagnostic.

La plupart des patients atteints de LAM présentent une anémie, le plus souvent normochrome normocytaire et arégénérative liée à l'insuffisance médullaire.

Une thrombopénie est présente dans la majorité des situations, parfois majeure (< 10 G/L). Elle s'explique par une insuffisance de production et parfois par une consommation excessive (CIVD).

Le statut leucocytaire est le paramètre le plus variable d'un patient à l'autre, pouvant aller d'une leucopénie franche (< 1G/L) sans blastes circulants, jusqu'à l'hyperleucocytose majeure (100-500G/L) essentiellement blastique. La plupart des patients sont neutropéniques et des anomalies morphologiques (hyposegmentation, hypogranulation …) peuvent être observées dans les neutrophiles résiduels.

1.3.2 Myélogramme

Dans les années 1970, le système de classification franco-américano-britannique (FAB) a établi que le diagnostic de LA pouvait être posé dès lors que plus de 30% de blastes étaient présents dans la MO. Une distinction faite selon des critères morphologiques et cytochimiques ainsi que selon le degré de différenciation des blastes et le type de lignée engagée a permis d'établir huit sous-types de LAM (M0 à M7) (33).

Cette classification a été modifiée afin d'améliorer la concordance entre les différents observateurs et d'intégrer les nouvelles découvertes des études immunologiques et cytogénétiques (34).

Actuellement, le diagnostic de LAM est posé lorsque les blastes myéloïdes constituent 20 % ou plus des cellules de la moelle osseuse (35). Pour les LA à composante monocytaire, les monoblastes et promonocytes sont inclus dans ce pourcentage.

- LAM 0 : Leucémie aiguë myéloblastique sans maturation

Les blastes sont sans granulation, sans corps d'Auer et la cytochimie de la myéloperoxydase est négative. Le diagnostic de certitude pour ce sous-type ne peut se faire que par immunophénotypage.

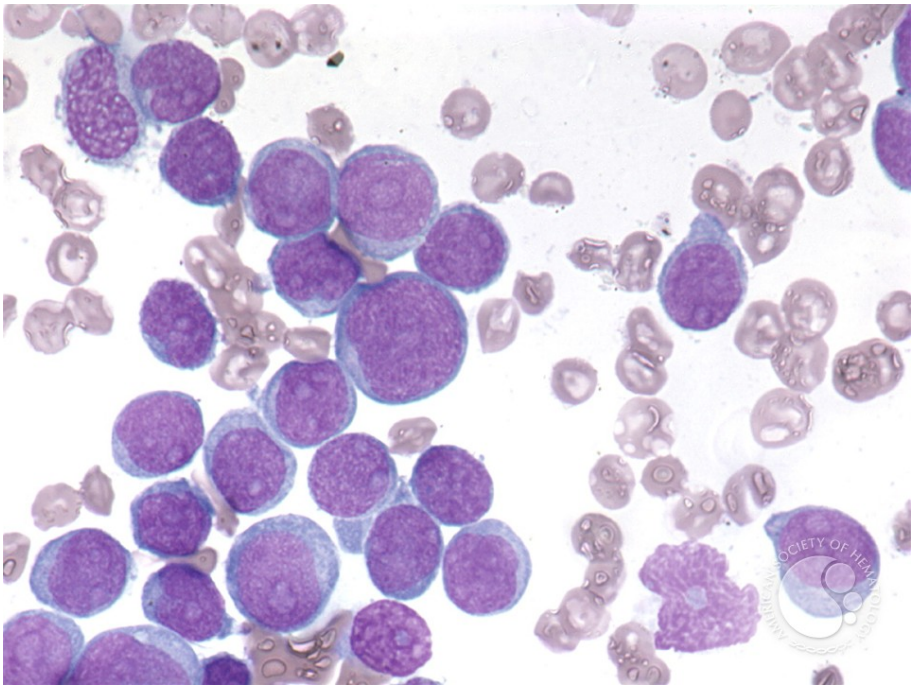

Figure 1 : Leucémie aiguë myéloblastique sans maturation

- LAM1 : Leucémie aiguë myéloblastique peu différenciée

Les blastes sont >90% avec une chromatine fine (1 à 3 nucléoles), ont des noyaux ronds avec des quantités modérées de cytoplasme (N/C = 0,8 - 0,9) qui peut être légèrement granulé et qui peut contenir des bâtonnets d'Auer. Il y a peu de signes de maturation myéloïde, avec < 10% de cellules au-delà du stade du promyélocyte. La peroxydase est positive, focale dans plus de 3 % des blastes. Il arrive dans certaines LAM sous-type FAB 1 que les blastes prennent la forme de "miroir à mains" mais cette particularité peut être également retrouvée dans certaines leucémies aiguës lymphoblastiques (LAL).

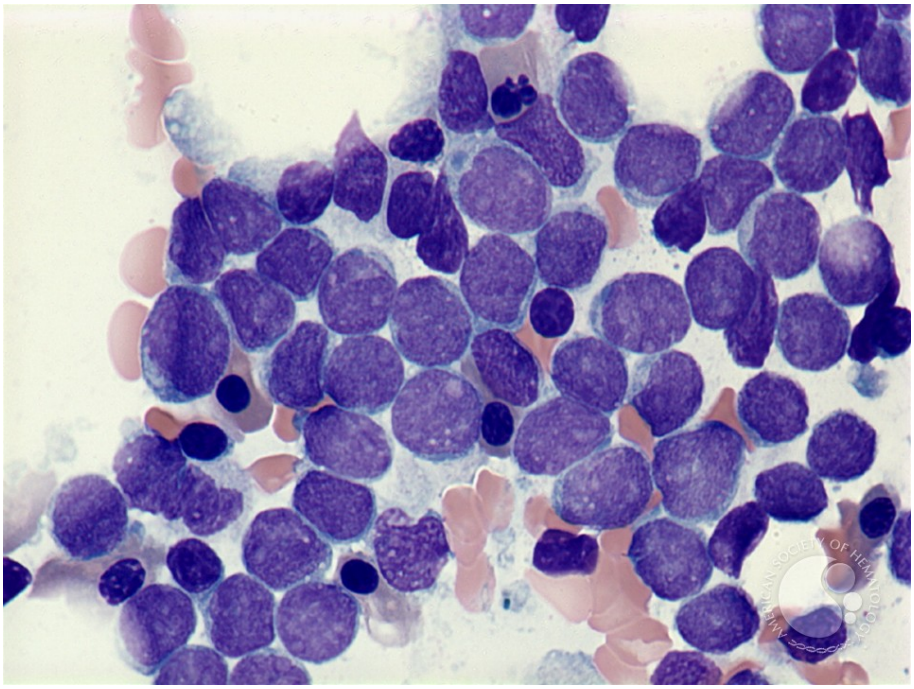

Figure 2 : Leucémie aiguë myéloblastique peu différenciée

- LAM2 : Leucémie aiguë myéloblastique avec maturation

La blastose est comprise entre 30 et 90%. La lignée myéloïde poursuit sa maturation avec la présence de promyélocytes, de myélocytes et souvent d'éléments myéloïdes plus matures > 10%. La lignée monocytaire représente <20% des cellules. Les blastes présentent des granulations franches et/ou des corps d'Auer. Environ 20% des patients atteints de LAM M2 FAB présentent une translocation caractéristique entre les chromosomes 8 et 21 t(8;21)(q22;q22)(36,37) se traduisant par la présence de corps d'Auer fins en boussole piquant le noyau.

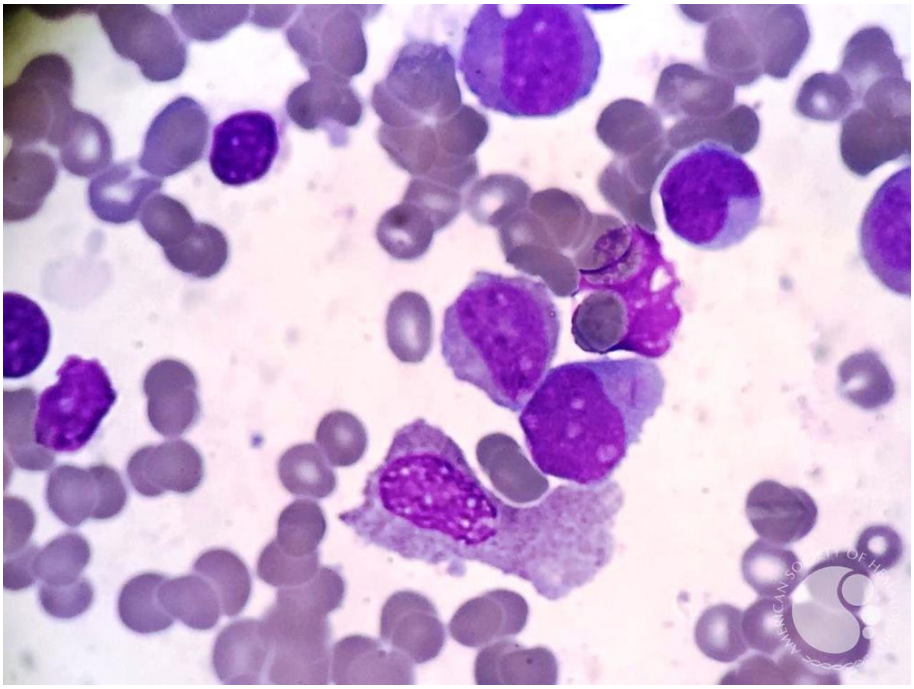

Figure 3 : Leucémie aiguë myéloblastique avec maturation

- LAM3 : Leucémie aiguë promyélocytaire

La LAP est l'un des sous-types les plus caractéristiques de LAM concernant les caractéristiques morphologiques, cliniques et cytogénétiques. Chez la plupart des patients, les blastes ressemblent à des promyélocytes fortement granulés. Les noyaux sont ronds ou bilobés, avec des nucléoles évidents, et le cytoplasme est rempli de multiples granules azurophiles, de grande taille. On observe généralement des bâtonnets d'Auer superposés dits en "fagot". La majorité des patients atteints de LAP présentent une translocation caractéristique impliquant les chromosomes 15 et 17 t(15;17)(q24;q21).

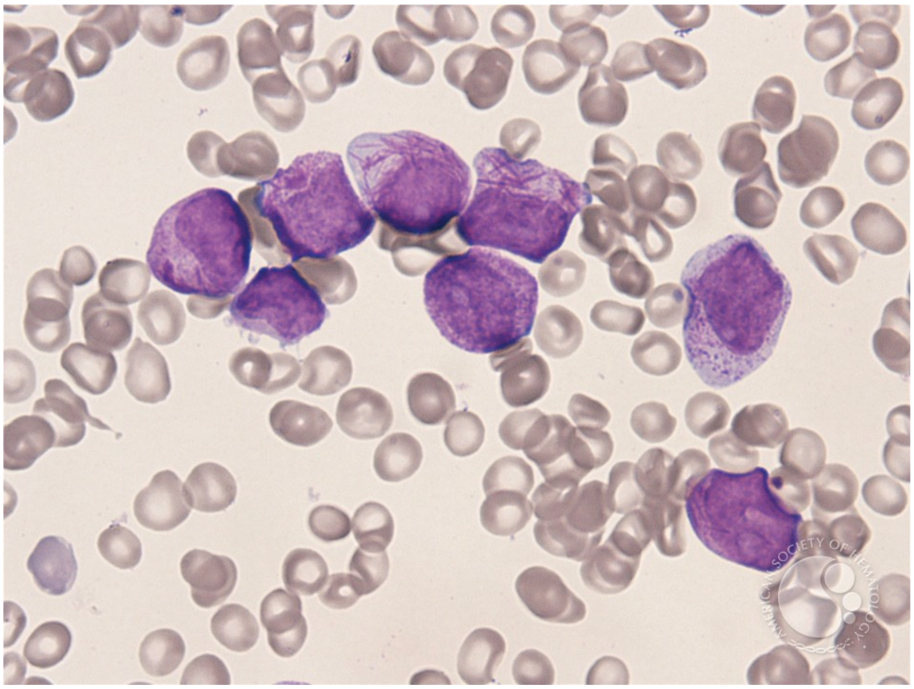

Figure 4 : Leucémie aiguë promyélocytaire

Il existe une forme variante minoritaire dans laquelle les blastes paraissent hypogranuleux (variante microgranulaire) et sont souvent des cellules à noyau bilobé ou lobulé dit en « ailes de papillon ». Contrairement à la présentation leucopénique typique de la LAP, les patients atteints de la variante microgranulaire ont tendance à présenter une hyperleucocytose.

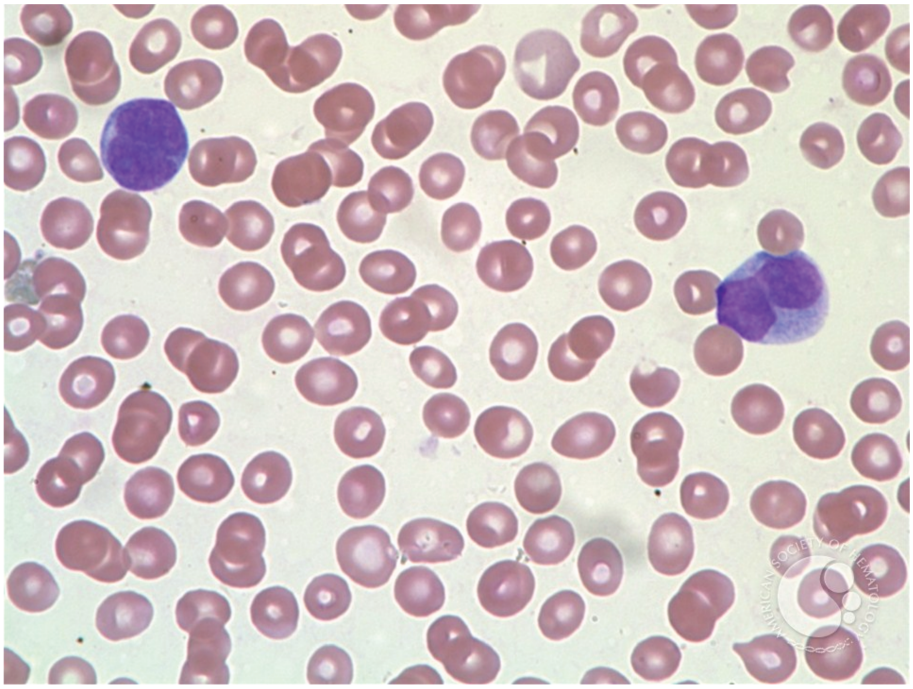

Figure 5 : Leucémie aiguë promyélocytaire forme variante

- LAM4 : Leucémie aigue myélomonocytaire

La leucémie aiguë myélo-monocytaire se caractérise morphologiquement par un mélange d'éléments myéloïdes $\geq$ 20% et monocytaires (promonocyte/monocyte) $\geq$ 20% avec une monocytose sanguine $>$ 5G/L.

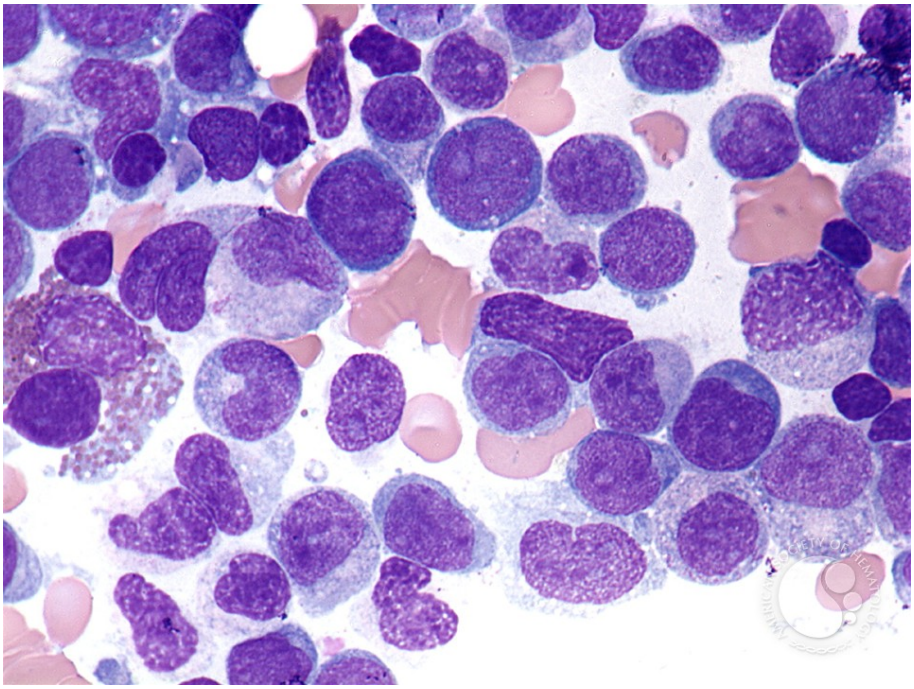

Figure 6 : Leucémie aiguë myélomonocytaire

Une forme variante de la leucémie aiguë myélomonocytaire existe, se caractérisant par la présence d'un contingent d'éosinophiles médullaires dysplasiques à différents stades de maturation (LAM4Eo). Ces cellules contiennent généralement de gros granules basophiles en plus des granules éosinophiles typiques. Pour la grande majorité des patients atteints de cette forme, on retrouve au caryotype une anomalie du chromosome 16 (inv(16)(p13q22) ou t(16;16)(p13;q22)).

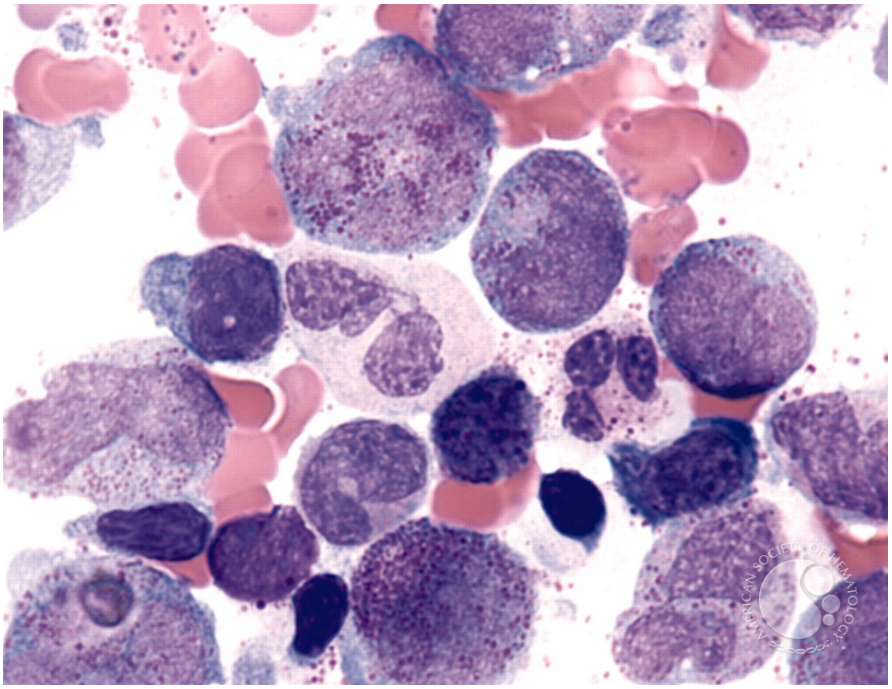

Figure 7 : Leucémie aiguë myélomonocytaire avec éosinophiles dysplasiques

- LAM5 : Leucémie aiguë monoblastique

Deux variantes de la leucémie aiguë monoblastique ont été décrites ; dans les deux cas, > 80% des blastes sont de lignée monocytaire. Dans la LAM5a, les blastes monocytaires ont des noyaux ronds et de petites quantités de cytoplasme fortement basophile sans preuve de différenciation morphologique. Dans la LAM5b, la différenciation est plus nette : au moins 20 % des blastes ont l'aspect de promonocytes à noyaux repliés et à cytoplasme abondant, légèrement granulé, sans bâtonnets d'Auer. La peroxydase est légèrement positive et dispersée. Le Naphtol-ASD-acétate (NASDA) est positive et inhibée en présence de fluorure de sodium (NaF).

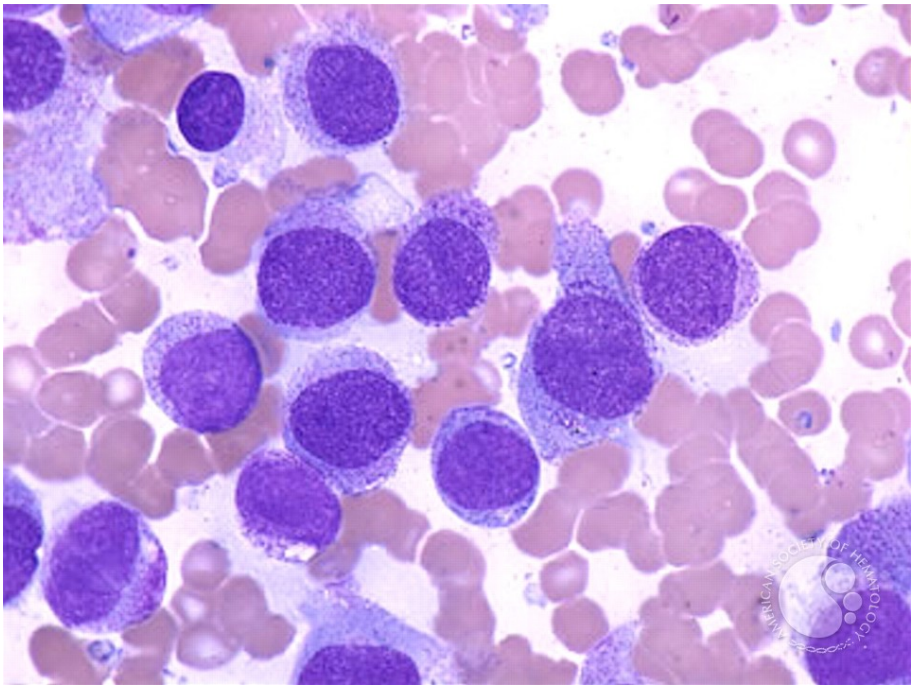

Figure 8 : Leucémie aiguë monoblastique

- LAM6 : Erythroleucémie

L'érythroleucémie est une variante de LAM dans laquelle les anomalies morphologiques de l'érythropoïèse sont les plus importantes. Des modifications dysplasiques marquées sur les trois lignées hématopoïétiques sont observées. Parallèlement à l'augmentation des blastes d'apparence myéloïde, on observe la persistance d'anomalies morphologiques dans la série érythroïde avec mégaloblastose profonde, multinucléarité, caryorrhexie, augmentation du nombre de mitoses et sidéroblastes annelés.

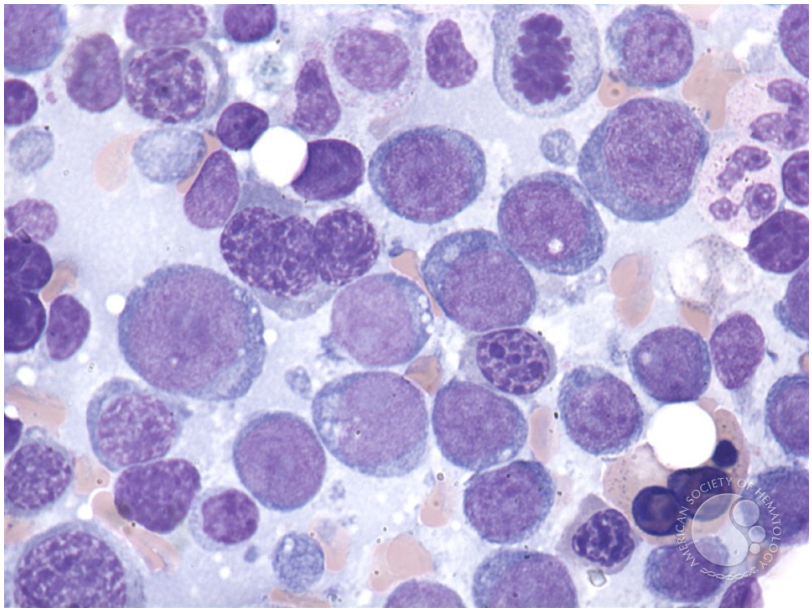

Figure 9 : Erythroleucémie

- LAM7 : Leucémie aiguë mégacaryoblastique

Les anomalies morphologiques de la mégacaryopoïèse, généralement caractérisées par la présence de micromégacaryocytes, sont particulièrement importantes chez les patients atteints de LAM7. La majorité des blastes observés est d'origine mégacaryocytaire. La myéloperoxydase est négative et le diagnostic s'effectue par immunophénotypage.

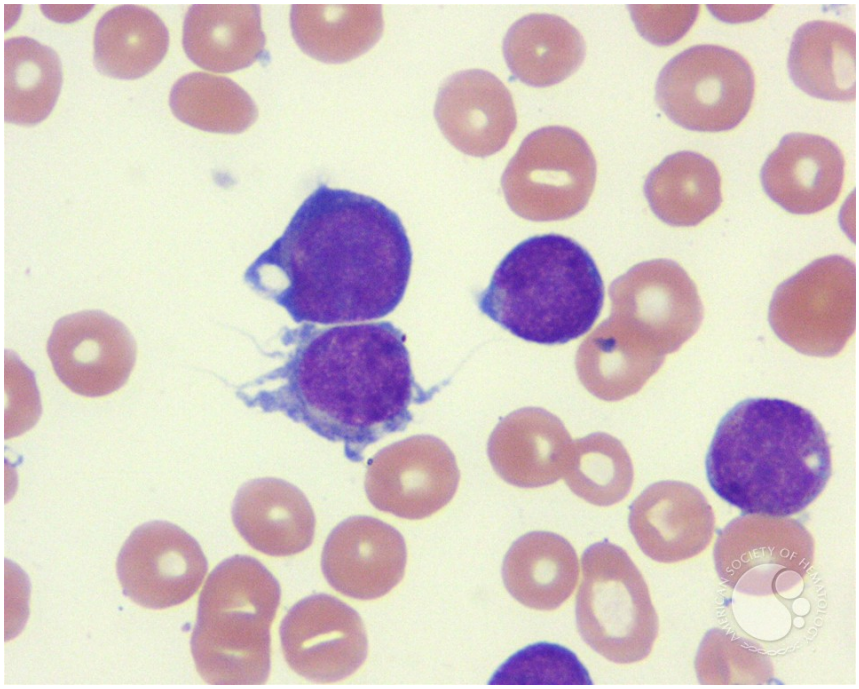

Figure 10 : Leucémie aiguë mégacaryoblastique

1.3.3 Cytométrie en flux

L'immunophénotypage par cytométrie de flux est réalisé à l'aide d'instruments multicanaux qui permettent d'évaluer simultanément la taille de la cellule (faisceau forward scatter FSC), le contenu de la cellule (faisceau side scatter SCC) et l'expression d'antigènes membranaires, cytoplasmiques et nucléaires (faisceau side fluorescence light SFL).

Dans le cas des leucémies aigües myéloïdes, elle permet (i) d'identifier la lignée cellulaire en cause (indispensable au diagnostic de LAM0 et plus rarement à celui de la LAM7), (ii) d'identifier le stade de maturation des blastes, (iii) d'effectuer un suivi en maladie résiduelle, (iv) d'identifier les rares leucémies de phénotype mixte ou ambigu.

1.3.4 Cytogénétique

L'analyse cytogénétique permet, à l'aide d'une coloration au Giemsa, de constituer le caryotype et d'étudier les bandes chromosomiques de cellules en métaphase. Il peut fournir des preuves de clonalité et ainsi confirmer qu'une maladie est néoplasique lorsque d'autres

preuves sont absentes. Elle permet en outre, dans certains cas, de catégoriser la maladie concernée et de fournir des informations sur le pronostic. Néanmoins, le caryotype présente plusieurs limites (i) impossibilité de détection des remaniements complexes et cryptiques, (ii) métaphases de mauvaise qualité empêchant l'interprétation du caryotype, (iii) impossibilité de mettre en évidence des anomalies spécifiques.

La technique FISH (fluorescent in situ hybridization) permet de rattraper ce manque de spécificité en utilisant des sondes oligonucléotidiques marquées qui se lient à des séquences précises d'ADN. Il peut s'agir de sondes spécifiques d'un locus, de sondes centromériques, de sondes télomériques et de peintures de chromosomes entiers. Cette technique peut être utilisée sur des noyaux en interphase ainsi qu'en métaphase. Elle peut être utilisée pour détecter les gains et les pertes de matériel génétique ainsi que les translocations.

1.3.5 Biologie moléculaire

Il est également possible de caractériser les LAM en recherchant l'éventuelle présence de mutations génétiques acquises (38–40). La biologie moléculaire, en mettant en évidence ces altérations permet de prédire l'évolution de la maladie (stratification pronostique) et d'orienter la stratégie thérapeutique et/ou d'identifier une cible thérapeutique (impact théranostique).

Bien que ces mutations puissent se retrouver dans n'importe quel groupe cytogénétique, leur découverte a eu l'impact le plus important sur le groupe des patients atteints de LAM cytogénétiquement normales (LAM-CN). Les LAM-CN représentent à elles seules environ 40 à 50% des LAM totales. La biologie moléculaire a permis pour ces patients autrefois considérés comme un groupe homogène au pronostic équivalent (41), de les stratifier en fonction de la présence ou non de certaines mutations aux conséquences cliniques hétérogènes (38).

Actuellement, la biologie moléculaire prend part intégrante dans le diagnostic et le suivi des LAM, avec une grande diversité de techniques disponibles et ne cessant pas d'évoluer.

Le séquençage de nouvelle génération (NGS) ou séquençage à haut débit a tout particulièrement révolutionné ce domaine en permettant un séquençage rapide et relativement peu coûteux de l'ADN et de l'ARN. Cela a entraîné une généralisation de son utilisation résultant en une augmentation conséquente du volume de données générées par les nouvelles technologies de séquençage. Des compétences de bio-informatique sont aujourd'hui primordiales pour assurer le traitement, le stockage et l'analyse de ces données. Le NGS a permis la découverte de nouvelles mutations récurrentes au niveau moléculaire et a ainsi permis une meilleure compréhension de la génomique complexe des LAM.

Cependant, seul un petit sous-ensemble de celles-ci s'est avéré contribuer à la leucémogenèse avec un impact évident sur le pronostic (38). Ce sous-ensemble de mutations génétiques récurrentes qui ont été impliquées dans la leucémogenèse est connu sous le nom de "driver mutations"(42). Pour celles qui sont retrouvées dans les LAM mais n'étant pas directement impliquées dans la leucémogenèse, on parle de "passenger mutations". Celles-ci n'ont pas d'impact direct sur l'évolution de la maladie (43). L'hypothèse est qu'une ou plusieurs mutations driver déclenchent le processus de leucémogènèse, tandis que d'autres agissent de manière coopérative, accélérant ou encourageant le processus leucémogène et, dans certains cas, contribuant à la résistance à la chimiothérapie et/ou au risque de rechute. Diverses formes de coopération (*NPM1*/*FLT3*-ITD) ainsi que l'exclusivité mutuelle (*RUNX1-RUNX1T1*/*CBFB-MYH11*) entre différents gènes ont également été découvertes (44).

Ces marqueurs moléculaires désormais bien connus sont intégrés en routine dans le suivi et le diagnostic biologiques des LAM. C'est notamment le cas de la duplication en tandem du gène *FLT3* (*FLT3*-ITD), bien identifiée dans le processus de leucémogénèse, considérée comme une mutation driver et possédant donc un rôle clef dans la stratification pronostique ainsi qu'en théranostique.

1.4 Classification OMS

La première version de cette classification a été créée en 2001 (45) reprenant de nombreux critères de la classification FAB et en y intégrant les données de cytogénétique. Elle a été révisée en 2008 (40) et en 2016 (35) pour y intégrer progressivement les anomalies détectées en biologie moléculaire de façon récurrente dans les LAM (Tableau 1). Une approche pluridisciplinaire est donc nécessaire afin de collecter l'ensemble de ces données cytologiques, cytogénétiques et moléculaires et d'ainsi classer précisément les LAM.

Tableau 1 : LAM classification OMS 2016.

LAM avec anomalies cytogénétiques récurrentes (30%)
LAM avec t(8;21)(q22;q22.1);*RUNX1-RUNX1T1*
LAM avec inv(16)(p13.1q22) ou t(16;16)(p13.1;q22);*CBFB-MYH11*
LAM avec *PML-RARA*
LAM avec t(9;11)(p21.3;q23.3);*MLLT3-KMT2A*
LAM avec t(6;9)(p23;q34.1);*DEK-NUP214*
LAM avec inv(3)(q21.3q26.2) ou t(3;3)(q21.3;q26.2); *GATA2*, *MECOM*
LAM (mégacaryoblastique) avec t(1;22)(p13.3;q13.3);*RBM15-MKL1*
LAM avec *NPM1* muté
LAM avec mutations *CEBPA* bi-alléliques
Entités provisoires : - LAM avec *BCR-ABL1* - LAM avec *RUNX1* muté
LAM avec dysplasie multilignée (10-15%) : dysmyélopoïèse ou antécédent de SMD ou anomalie cytogénétique de type SMD
LAM secondaires à une chimiothérapie (10-15%)
Autres (40-50%) : classées selon la classification FAB
Sarcome granulocytaire

1.5 Classification de l'European LeukemiaNet

La classification de l'European Leukemia Net (ELN) a été publiée initialement en 2010 et a été mise à jour en 2017 (38). Cette classification, qui a été largement adoptée, intègre des données cytogénétiques et moléculaires pour définir 3 catégories de pronostic : favorable, intermédiaire et défavorable (Tableau 2).

Tableau 2 : Classification ELN 2017.

Catégories	Anomalies génétiques
Favorable	t(8;21)(q22;q22.1); *RUNX1-RUNX1T1*
	inv(16)(p13.1q22) ou t(16;16)(p13.1;q22);*CBFB-MYH11*
	NPM1 muté sans *FLT3*-ITD ou avec *FLT3*-ITD faible[1]
	Mutations bi-alléliques de *CEBPA*
Intermédiaire	*NPM1* muté et *FLT3*-ITD élevé
	NPM1 sauvage sans *FLT3*-ITD ou avec *FLT3*-ITD faible[1] (sans anomalie génétique défavorable)
	t(9;11)(p21.3;q23.3); *MLLT3-KMT2A*
	Anomalies cytogénétiques non classées comme favorables ou défavorables
Défavorable	t(6;9)(p23;q34.1); *DEK-NUP214*
	t(v;11q23.3); *KMT2A* réarrangé
	t(9;22)(q34.1;q11.2); *BCR-ABL1*
	inv(3)(q2l.3q26.2) ou t(3;3)(q21.3;q26.2); *GATA2, MECOM(EVI1)*
	−5 ou del(5q); −7; −17/abn(17p)
	Caryotype complexe ou caryotype monosomal
	NPM1 sauvage et *FLT3*-ITD élevé[2]
	RUNX1 muté
	ASXL1 muté
	TP53 muté

[1] ratio allélique < 0.5

[2] ratio allélique ≥ 0.5

L'intégration de ces marqueurs pronostiques est particulièrement importante afin de stratifier les patients atteints de LAM pour lesquels les issues cliniques sont hétérogènes. Cette classification permet d'établir des diagnostics qui définissent plus clairement les groupes de maladies, avec une évolution et une réponse différente aux traitements.

1.6 Duplications en tandem de *FLT3* (*FLT3*-ITD)

Les duplications internes en tandem dans le gène *FLT3* constituent un marqueur biologique de choix. En effet, outre leur intérêt pronostique, elles jouent un rôle théranostique puisque leur présence justifie l'introduction d'un traitement par inhibiteur de FLT3 (FLT3i) (46). Elles figurent parmi les mutations les plus fréquemment retrouvées dans les LAM (47–49) (20 à 30% des cas).

Ainsi, elles doivent être recherchées et quantifiées systématiquement en cas de suspicion de LAM.

1.6.1 Le récepteur FLT3

Le gène *FLT3*, également connu sous le nom de FLK-2 (fetal liver kinase-2) et STK-1 (human stem cell kinase-1) est situé sur le chromosome 13q12 et code pour un récepteur tyrosine kinase de classe III (50). *FLT3* est physiologiquement exprimé par les progéniteurs de la lignée myéloïde et de la lignée lymphoïde B ainsi que par les blastes (51,52). Il y joue un rôle clef dans la myélopoïèse, la lymphopoïèse mais également dans la régulation du système immunitaire (53). Son activation est dépendante du ligand FLT3 (FLT3L).

FLT3L est une protéine transmembranaire de type 1, qui contient un peptide de signalisation amino-terminale, quatre domaines hélicoïdaux extracellulaires, un domaine transmembranaire et un domaine cytoplasmique (54). Le gène codant pour cette protéine est situé sur le chromosome 19q13. Le ligand peut également être libéré sous forme de protéine homodimérique soluble (55). La forme soluble et la forme transmembranaire sont toutes deux capables d'activer le récepteur FLT3 (56). Il est exprimé essentiellement par les cellules endothéliales mais aussi par de nombreux types de cellules et stimule les cellules hématopoïétiques (57,58). En effet, à lui seul il ne peut induire la prolifération cellulaire des progéniteurs hématopoïétiques : il agit en synergie avec d'autres facteurs de croissance (59). Il se lie spécifiquement au récepteur FLT3 et entraine son activation.

Le récepteur FLT3 est monomérique et se compose des domaines suivants : domaine extracellulaire de liaison au ligand, domaine transmembranaire, domaine juxtamembranaire (JMD) et domaine tyrosine kinase (TKD). Dans un état non muté, le récepteur FLT3 se dimérise en se liant au ligand FLT3, entrainant un changement conformationnel découvrant les sites TKD pouvant ainsi se lier à l'ATP. Cette dimérisation provoque une autophosphorylation et une activation de l'activité tyrosine kinase, ce qui active les voies de signalisation en aval. Les voies de signalisation impliquées sont RAS/MAP kinase, la PI3 kinase/AKT et JAK/STAT. Elles permettent la prolifération, la différentiation ainsi que la survie cellulaire (49,60) (Figure 11).

Il existe deux grands types de mutations *FLT3* : les mutations situées sur le domaine tyrosine kinase (*FLT3*-TKD) et les duplications dans le JMD et ou TKD1 (JMD/TKD1 ; *FLT3*-ITD). Elles déclenchent toutes deux l'activité de la kinase FLT3. Ces deux types de mutation sont responsables d'une prolifération et d'une survie cellulaire augmentées mais seule *FLT3*-ITD en touchant le JMD/TKD1 a un impact clinique clairement établi (47). En effet, le JMD du FLT3, comme de nombreux autres récepteurs, exerce une influence régulatrice négative sur l'activité tyrosine kinase (61,62) (Figure 11). Des mutations dans ce domaine juxtamembranaire peuvent perturber ses fonctions régulatrices négatives. Les mutations *FLT3*-ITD touchant le TKD1 confèrent quant à elle une activation constitutive du récepteur (63).

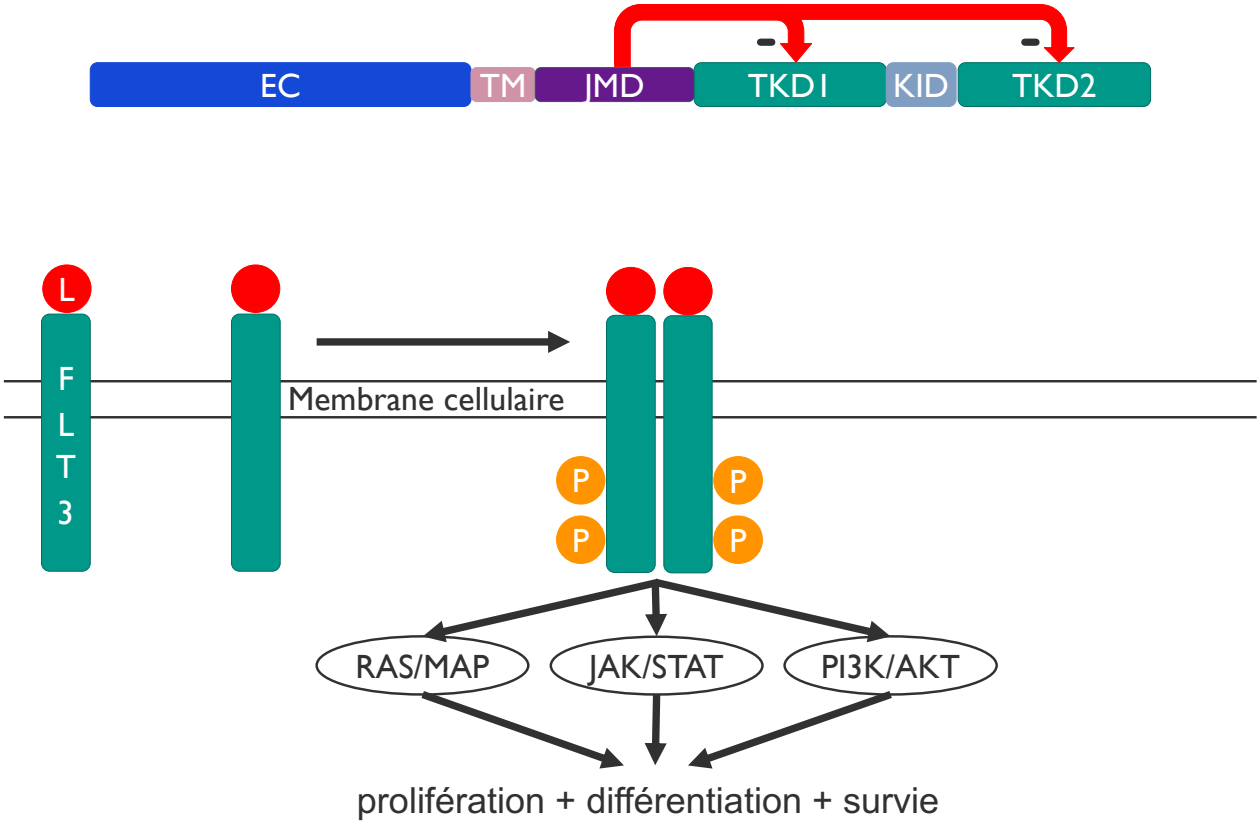

Figure 11 : Structure et activation du récepteur FLT3

EC : Extracellular domain
TM : Transmembrane domain
JMD : Juxtamembrane domain
TKD1/ TKD2 : Tyrosine kinase domain
KID : Kinase insert domain

1.6.2 Mécanismes moléculaires

Les mutations *FLT3*-ITD ont été mises en évidence pour la première fois en 1996 (64). Les ITDs se traduisent par des duplications de taille allant de 3 à 400 paires de bases sans

décalage du cadre de lecture dans le JMD/TKD1 et perturbent son activité (49). Leur site d'insertion se situe sur l'exon 14 et/ou 15 (transcrit NM_004119.2). Le mécanisme moléculaire qui sous-tend leur formation reste mal connu. Récemment *J. Borrow et al.,* en étudiant les séquences au niveau du point de cassure de la duplication incriminent l'enzyme lymphoïde TdT dans l'amorçage des *FLT3*-ITD (65).

L'ITD de ce mutant *FLT3* produit un récepteur tyrosine kinase activé de manière constitutive, subissant une autophosphorylation en l'absence de ligand, ce qui entraîne l'activation de voies de signalisation en aval, responsable du syndrome myéloprolifératif (66) (Figure 12). En effet, les mutations *FLT3*-ITD ne suffisent pas à elles seules à provoquer une LAM. En pratique, chez le modèle murin, seul l'état myéloprolifératif causé par l'induction d'une mutation *FLT3*-ITD a pu être mis évidence (66). Ainsi, le mécanisme exact par lequel la perte de répression et l'activation constitutive des domaines tyrosine kinase FLT3 aboutissent finalement à la leucémogenèse reste mal compris.

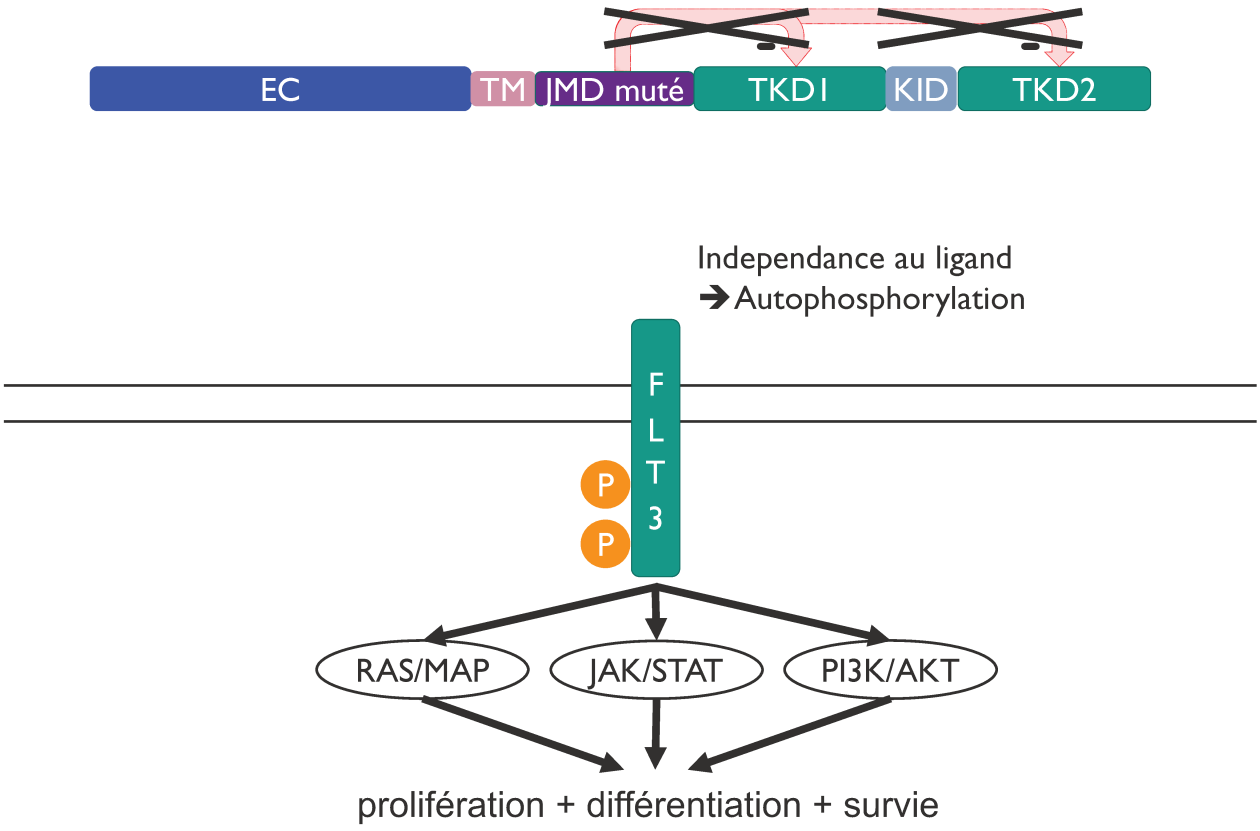

Figure 12 : Structure et activation du récepteur FLT3 muté ITD

EC : Extracellular domain
TM : Transmembrane domain
JM : Juxtamembrane domain
TKD1/TKD2 : Tyrosine kinase domain
KID : Kinase insert domain

1.6.3 Pronostic

Au diagnostic, les patients porteurs de *FLT3*-ITD présentent un taux de leucocytes dans le sang périphérique et une quantité de blastes dans la moelle osseuse significativement plus élevés que les patients qui n'en sont pas porteurs (67). De plus, ils présentent un risque de rechute plus élevé, une survie sans événement et une survie globale diminuées (68). L'importance du ratio allélique (AR), évalué en divisant la quantité d'allèle muté par la quantité d'allèle sauvage, a été mise en évidence dans plusieurs études (69,70). Un AR élevé serait associé à une diminution de la survie globale et de la survie sans rechute. Il est important de noter que pour ces publications les patients n'ont pas été mis sous inhibiteurs de FLT3. Actuellement, l'ELN a fixé une valeur seuil de l'AR à 0.5, permettant ainsi de former 2 groupes distincts : AR faible < 0.5 et AR élevé ≥ 0.5. D'autres paramètres liés à cette duplication pourraient avoir une signification pronostique comme sa taille (71), son site d'insertion (72) ainsi que sa structure (73,74). L'impact de la taille est toujours discuté (67,75).

Plusieurs études suggèrent que la présence de la mutation *FLT3*-ITD peut être associée à un pronostic péjoratif, non seulement au moment du diagnostic mais aussi lors de la première rechute. Les patients présentant des mutations *FLT3*-ITD à la première rechute ont une probabilité réduite d'obtenir une seconde rémission complète (RC) et une survie globale diminuée (76–78).

FLT3-ITD est également utilisée comme outil de stratification : les patients dont l'AR est élevé doivent si possible bénéficier d'une greffe de cellules souches hématopoïétiques allogénique (allo-HSCT) (38,70,79). Seule la présence conjointe de la mutation *NPM1* chez les patients présentant un faible AR *FLT3*-ITD permet de moduler le recours à cette dernière (80,81). Toutefois, plusieurs études ont démontré que le pronostic était défavorable dans les LAM mutées *NPM1* avec *FLT3*-ITD à faible AR lorsque l'allo-HSCT n'était pas réalisée en première rémission complète (82,83).

La fréquence élevée et l'impact pronostic négatif de la mutation *FLT3*-ITD en font un candidat idéal pour une thérapie ciblée. Plusieurs inhibiteurs ont été développés et commercialisés ou font encore actuellement l'objet d'essais cliniques.

1.6.4 Inhibiteurs de tyrosine kinase

Le traitement général des LAM s'effectue toujours en deux temps : phase d'induction et phase de consolidation. L'objectif de la chimiothérapie d'induction est d'obtenir une rémission afin de maîtriser la maladie et de rétablir une hématopoïèse normale. La phase de consolidation post-rémission est également nécessaire. En effet, les patients qui ne reçoivent pas de thérapie post-rémission peuvent rechuter dans les 6 à 9 mois (84). Le choix de la

thérapie de consolidation dépend de la stratification du risque du patient (38). Les thérapies ciblées comme les inhibiteurs de tyrosine kinase (ITKs) sont venues compléter ce schéma classique du traitement des LAM. L'utilisation de la midostaurine en association avec les thérapies usuelles a été approuvée en 2017 pour le traitement des LAM *FLT3* mutées nouvellement diagnostiquées, durant les phases d'induction et de consolidation (46). En 2018, le giltéritinib a obtenu une autorisation de mise sur le marché (AMM) pour les LAM en rechute ou réfractaires avec *FLT3* muté (85,86). Des données récentes issues d'essais randomisés ont apporté la preuve que le traitement d'entretien ciblé post-HSCT à base de sorafénib pourrait devenir un nouveau paradigme thérapeutique dans la LAM mutée *FLT3*-ITD (87). Après des décennies de stagnation thérapeutique, l'arrivée de ces molécules a fait entrer les LAM dans une ère de thérapie moléculaire ciblée pour un vaste sous-ensemble de patients.

Les mécanismes d'action des ITKs diffèrent en fonction de leur type. Les ITKs de type I se lient au récepteur FLT3 dans la conformation active et sont actifs contre les mutations ITDs et TKDs. Les ITKs de type II se lient au récepteur FLT3 dans la conformation inactive au niveau d'une région adjacente au TKD. En raison de cette affinité de liaison, les inhibiteurs FLT3 de type II empêchent l'activité des mutations ITDs mais ne ciblent pas les mutations TKD (47) (Tableau 3).

Le spectre d'action varie selon la génération de l'ITKs. La première génération serait assez peu spécifique et pourrait toucher d'autres récepteurs comme KIT, VEGFR et PDGFR ou des protéines de signalisation comme JAK2, NRAS et RAF. La deuxième génération serait plus spécifique du récepteur FLT3 (Tableau 3).

Tableau 3 : Inhibiteurs FLT3 commercialisés et/ou en développement

1ère Génération	2ème Génération
Type 1	
midostaurine sunitinib lestaurtinib	**giltéritinib** crénolanib
Type 2	
sorafénib	quizartinib
Non connu	
tandutinib	

Bien que les ITKs aient amélioré les résultats historiquement médiocres du traitement des LAM mutées *FLT3*, la réponse clinique à un inhibiteur de FLT3 peut être de courte durée dans les cas de rechute ou de réponse réfractaire. En général, la résistance aux FLT3i peut être classée soit comme une résistance primaire du clone, soit comme le développement d'une résistance secondaire suite à une première réponse (88). La résistance primaire est

principalement due à des mutations de *FLT3* rendant la protéine insensible aux ITKs spécifiques, ou à l'activation de voies de survie alternatives. De nombreuses causes peuvent induire une résistance secondaire : (i) mutations de résistance au niveau du site de liaison à l'ATP, (ii) augmentation du FLT3L, (iii) activation de voies alternatives telles que les mutations du gène *NRAS* (89).

Les thérapies ciblant le récepteur FLT3 ne sont pas les seules à démontrer leur efficacité chez les patients porteurs de *FLT3*-ITD : les anti-CD33, dont fait partie le MYLOTARG®, montrent des résultats probants et peuvent être indiqués chez les patients *FLT3*-ITD (90).

1.6.5 Gemtuzumab ozogamicine

Le gemtuzumab ozogamicine (GO) est un conjugué anticorps anti-CD33-calichéamicine sélectif. Une fois lié au cluster de différenciation 33 (CD33), le complexe CD33/GO s'internalise puis il est acheminé vers des lysosomes où est libérée la calichéamicine. Elle exerce son effet cytotoxique en se liant au petit sillon de l'ADN et induit une cassure double brin. Ces dommages à l'ADN provoquent un arrêt du cycle cellulaire et l'induction de l'apoptose (91–93).

Chez les individus sains, le CD33 se trouve essentiellement sur les précurseurs myéloïdes mais est absent de la surface des cellules souches hématopoïétiques (94). On le retrouve également sur les blastes myéloïdes dans 85-90% des cas de LAM (95,96). Des essais de phase III ont montré que l'utilisation du GO en association avec les thérapies usuelles augmente la survie globale (97–99) ainsi que la survie sans événement (100) pour les LAM CD33+. La FDA a donc approuvé le 1er septembre 2017 l'utilisation de GO en association pour les LAM à CD33 positif nouvellement diagnostiquées chez les adultes et pour les LAM à CD33 positif en rechute ou réfractaires chez les patients âgés de 2 ans et plus.

L'ajout de GO a démontré également une amélioration de la survie globale, de la survie sans rechute et de la survie sans évènement chez les patients adultes atteints de LAM présentant des mutations *FLT3*-ITD (100,101). Cette efficacité pourrait s'expliquer par la surexpression du CD33 chez les patients présentant des duplications internes en tandem de *FLT3* (102). En France, la GO a donc reçu en 2018 une AMM pour le traitement des patients adultes atteints de LAM *de novo*, ayant une cytogénétique favorable ou intermédiaire ou une mutation *FLT3*-ITD, en association durant la phase de d'induction.

Plusieurs mécanismes de résistance contre la GO sont décrits dans la littérature : (i) l'ATP-Binding Cassette Subfamily B Member 1 (ABCB1) et la Multidrug Resistance-Associated Protein 1 (MRP1) peuvent provoquer un efflux de la calichéamicine libre hors de la cellule compromettant l'efficacité de la GO (103,104) (ii) le statut de méthylation de la Suppressor Of Cytokine Signaling 3 (SOCS3), puisque l'hyperméthylation des îlots CpG augmenterait

la survie globale chez les patients sous GO (105) (iii) la modification de l'épitope du CD33 pouvant altérér la liaison avec la GO (106).

1.6.6 Détection et quantification

D'après les recommandations de l'ELN émises en 2017, l'analyse de fragments est la méthode de référence pour détecter et quantifier *FLT3*-ITD (38). Elle comprend une étape d'amplification par PCR (polymerase chain reaction), suivie d'une étape de séparation des amplicons par électrophorèse capillaire (107,108). Un ou plusieurs pics sont obtenus en fonction de la présence ou non de(s) duplication(s). La taille de l'ITD est estimée en soustrayant la taille du fragment sauvage à celle du fragment muté. L'AR sera quant à lui évalué en divisant l'aire sous la courbe (AUC) du fragment muté par l'AUC du fragment sauvage (Figure 13).

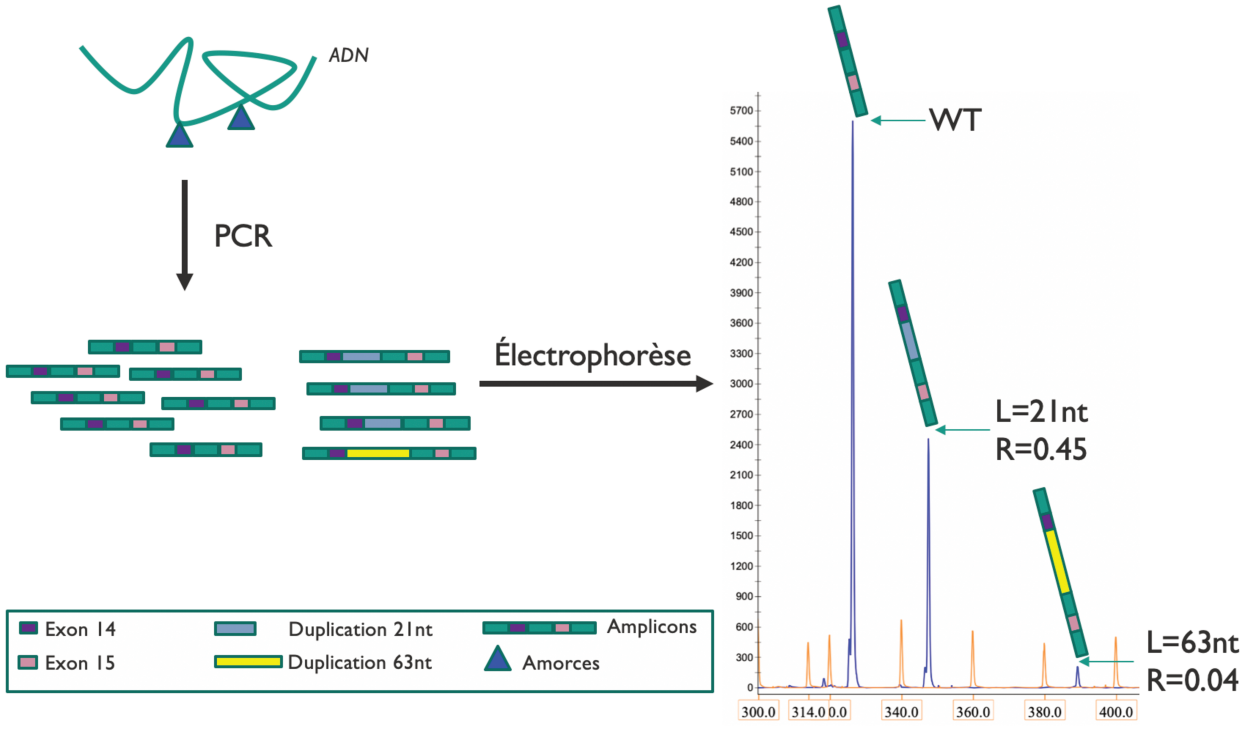

Figure 13 : Vue d'ensemble de l'analyse de fragments

Cette technique bien que robuste présente plusieurs inconvénients : (i) une limite de détection élevée (AR 1-3% (109,110)) pouvant entrainer une interprétation délicate en cas d'AR faible (ii) un manque de résolution pour déterminer la taille de l'ITD (précision à 1pb) (iii) elle ne fournit pas la séquence nucléotidique de l'ITD ni son site d'insertion (iv) cette technique n'est pas multiplexable.

Le NGS pourrait, à terme, remplacer cette technique et présenterait de nombreux avantages. La détection des ITDs ne se ferait plus par l'interprétation de pic(s) mais par le

dénombrement de reads, permettant ainsi une estimation plus sensible et plus objective. La séquence nucléotidique exacte de la duplication serait déterminée, ce qui limiterait les approximations concernant sa longueur. L'avantage majeur en effectuant du NGS serait la possibilité de rechercher simultanément plusieurs mutations d'intérêt. Malheureusement, de par son hétérogénéité en termes de taille et de site d'insertion (111), *FLT3*-ITD est difficile à mettre en évidence avec les outils d'alignement classiques. Pour mettre en évidence les mutations, il est nécessaire d'aligner l'ensemble des reads sur un génome de référence. Selon la longueur de l'ITD, l'alignement des reads peut se faire de trois façons (i) soit l'ITD est courte et lors de l'alignement des reads elle sera détectée comme une insertion (ii) soit elle est de longueur moyenne, l'alignement s'effectuera alors sur l'ITD ou sur la portion restante du read et la duplication ne sera pas détectée (iii) soit elle est de grande taille et dépasse la longueur du read et elle sera alors alignée sur le génome sans être détectée (Figure 14).

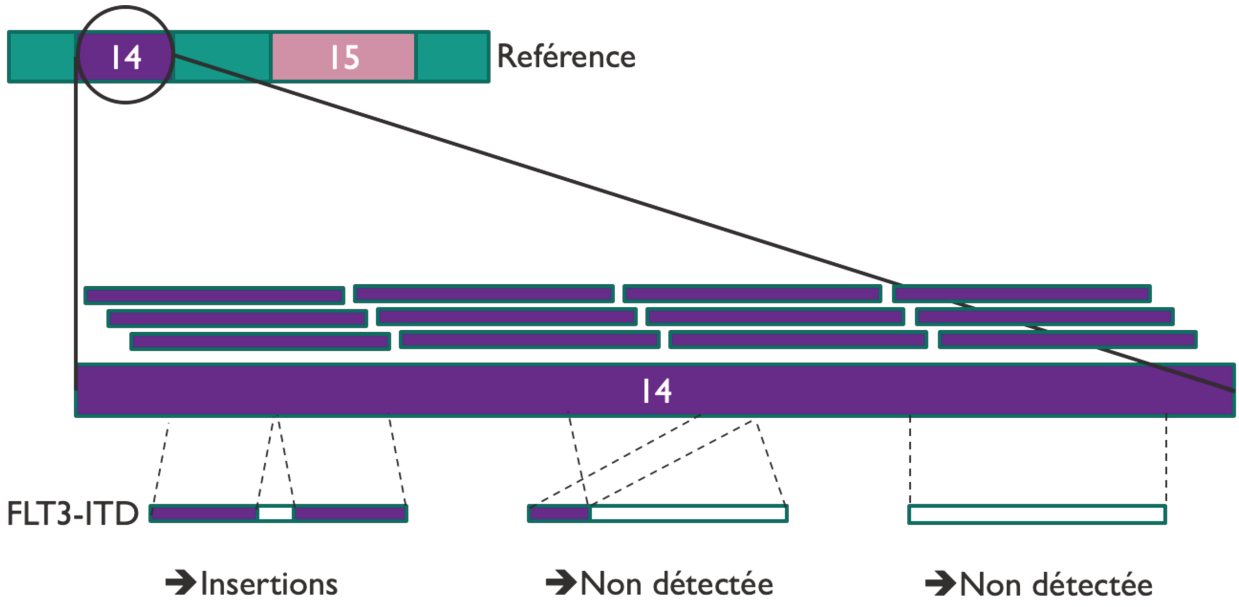

Figure 14 : Problématique d'alignement de FLT3-ITD

1.7 Objectifs

Actuellement, les processus de séquençage de nouvelle génération pour le dépistage des *FLT3*-ITD sont bien documentés. Cependant, la plupart sont incapables de détecter et d'annoter toutes les ITDs testées (112,113). En effet, les ITDs sont hétérogènes en termes de taille et/ou de site d'insertion et sont issues de la duplication parfaite ou presque parfaite de la séquence de type sauvage, soulevant des problèmes pour le traitement et l'analyse de l'information en NGS (112). Plusieurs algorithmes ont été créés afin de détecter et de quantifier cette anomalie. Outre la détection de *FLT3*-ITD, la détermination de l'AR avec précision est indispensable.

Le but de ce travail était (i) d'étudier et de comparer les algorithmes existants pour la détection et la quantification de *FLT3*-ITD puis (ii) de participer à la mise au point d'un algorithme appelé FiLT3r permettant de détecter et de quantifier les mutations *FLT3*-ITD (iii) tout en évaluant celui-ci en s'assurant qu'il soit bien compatible avec la technique de séquençage haut-débit avec enrichissement par capture ainsi qu'avec le panel de gènes utilisé en routine au laboratoire analysé conjointement à *FLT3*-ITD.

2 Matériels et méthodes

2.1 Échantillons

Afin d'éviter un phénomène de sur-ajustement de FiLT3r par rapport aux autres algorithmes testés, un jeu de données "d'entrainement" et un jeu de données "test" ont été formés. L'optimisation du paramétrage de FiLT3r sera effectuée sur le jeu de données "d'entrainement" puis l'ensemble des algorithmes sera comparé sur un jeu de données indépendant, le jeu de données "test".

2.1.1 Jeu de données "d'entrainement" : BIG1

L'optimisation des différents paramètres de FiLT3r a été effectuée sur un jeu de données extrait de la cohorte BIG1. Cette cohorte, toujours en cours actuellement, inclut des patients atteint de LAM de novo ou secondaire (CBF et LAP exclus), âgés de 18 à 59 ans. La base testée comprend 114 échantillons *FLT3*-ITD+ (81 échantillons de moelle osseuse et 21 de sang périphérique) ainsi que 71 échantillons *FLT3*-ITD- (57 échantillons de moelle osseuse et 14 de sang périphérique). Le statut des échantillons vis à vis de la mutation *FLT3*-ITD a été déterminé à l'aide de l'analyse par fragments.

2.1.2 Jeu de données "test": ALFA-0701, ALFA-0702 et ALFA 1200

Les algorithmes ont été évalués à l'aide d'un jeu de données rassemblant plusieurs cohortes : ALFA-0701, ALFA-0702, ALFA-1200 (Tableau 4).

Tableau 4 : Tableau récapitulatif des cohortes utilisées pour le jeu de données "test"

Acronyme	**Période de Recrutement**	**Population cible**	**n théorique**	**n réel**
ALFA-0701	2008-2010	Patients atteints de LAM de novo (LAP exclus), âgés de 50 à 70 ans.	271	196
ALFA-0702	2009-2013	Patients atteints de LAM de novo (LAP exclus), âgés de 18 à 59 ans.	713	349
ALFA 1200	2012-2016	Patients atteints de LAM de novo (LAP exclus), âgés de 60 ans et plus.	509	443

n théorique : correspond au nombre de patients publiés pour la cohorte.

n réel : correspond au nombre de patients pour lesquels des échantillons étaient encore disponibles.

Au total le jeu de données test comprend 222 échantillons *FLT3*-ITD+ ainsi que 766 échantillons *FLT3*-ITD-. Le statut des échantillons vis-à-vis de la mutation *FLT3*-ITD a été déterminé à l'aide de l'analyse par fragments.

2.2 Extraction d'ADN

Dans un premier temps les cellules mononuclées ont été isolées du prélèvement par technique FICOLL. Elle permet, après centrifugation, de séparer les cellules mononuclées plus légères, des cellules plus denses. Les cellules mononuclées sont réparties en culot sec de 5 à 10 millions de cellules.

L'ADN de ces cellules a été ensuite extrait sur colonnes. Cette extraction s'effectue sur l'automate Hamilton NIMBUS® en utilisant le kit Nucléospin 8 Blood® pour le sang et le kit Nucléospin 8 Tissus® pour la moelle conformément aux recommandations du fabricant.

2.3 Analyse de fragments

FLT3 a été amplifié par PCR avec des amorces dont l'une est marquée par fluorescence (amorces sens marquées par fluorescence : 5'FAM_GCA-ATT-TAG-GTA-TGA-AAG-CCA-GC_3' ; amorce antisens : 5'_CTT-TCA-GCA-TTT-TGA-CGG-CAA-CC_3') flanquant les exons 14 et 15. La PCR a été réalisée dans un volume final de 25 µL contenant

de l'ADN génomique (25 ng/µL) avec le kit HotstarTaq® (QIAGEN). Les conditions d'amplification étaient les suivantes : dénaturation à 95° C pendant 15 min, suivie de 30 cycles de 94°C pendant 30 sec, 56°C pendant 30 secondes, et 72°C pendant 45 secondes, avec une étape finale à 72°C pendant 10 min. Les produits PCR ont été dilués au 1:3 et séparés par électrophorèse capillaire avec l'analyseur d'ADN ThermoFisher® 3130xl. Les données ont été analysées à l'aide du logiciel ABI GeneMapper® version 4.1 avec une validation biologique en aval.

2.4 NGS

Le protocole utilisé en routine pour réaliser les NGS des patients au diagnostic a été appliqué aux échantillons . Il s'agit d'un séquençage en paired-end, couvrant un panel de gènes impliqué dans les hémopathies myéloïdes. Ce panel évolue en fonction des données de la littérature, et comprend actuellement 67 gènes (*ANKRD26*, *ASXL1*, *ASXL2*, *ATRX*, *BCOR*, *BCORL1*, *BRAF*, *CALR*, *CBL*, *CEBPA*, *COG1*, *CRLF2*, *CSF3R*, *CUX1*, *DDX41*, *DNMT3A*, *ETNK1*, *ETV6*, *EZH2*, *FLT3*, *GATA1*, *GATA2*, *GNAS*, *GNB1*, *HRAS*, *IDH1*, *IDH2*, *IKZF1*, *IL2RG*, *IL7R*, *JAK1*, *JAK2*, *JAK3*, *KDM6A*, *KIT*, *KRAS*, *MPL*, *NF1*, *NFE2*, *NPM1*, *NRAS*, *PAX5*, *PHF6*, *PPM1D*, *PTPN11*, *RAD21*, *RIT1*, *RUNX1*, *SAMD9*, *SAMD9L*, *SETBP1*, *SF3B1*, *SH2B3*, *SMC1A*, *SMC3*, *SRP72*, *SRSF2*, *STAG2*, *STAT3*, *STAT5B*, *TERC*, *TERT*, *TET2*, *TP53*, *U2AF1*, *WT1* et *ZRSR2*.). Le NGS comprend 4 grandes étapes (i) étape de préparation de librairie (ii) étape d'amplification clonale (iii) étape de séquençage (iv) et l'étape de traitement bio-informatique (Figure 15).

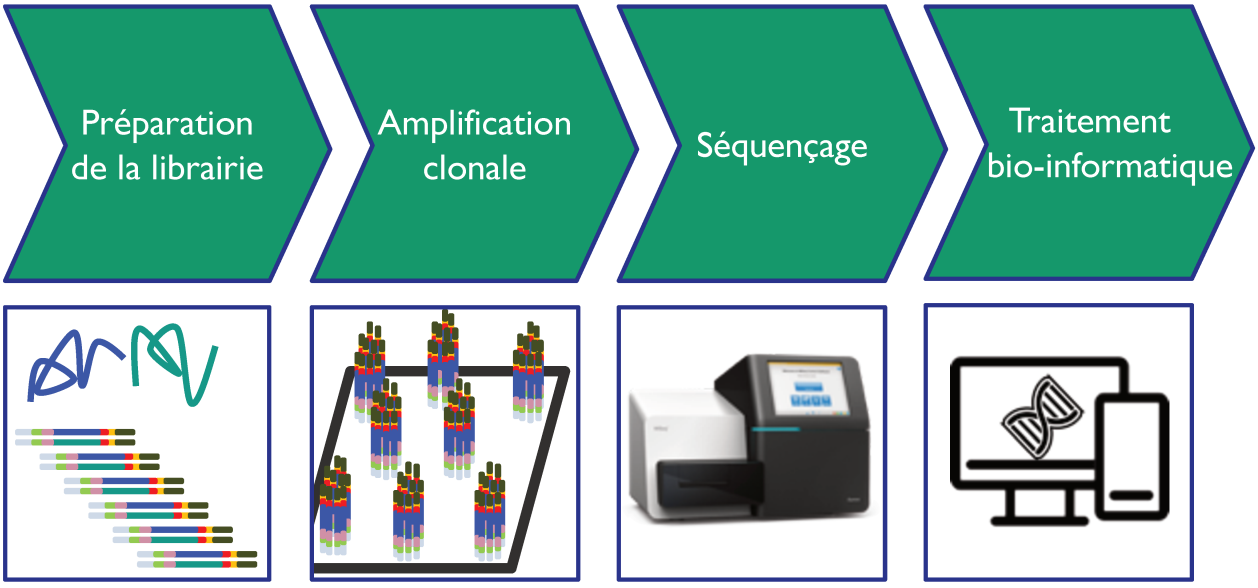

Figure 15 : Vue d'ensemble du NGS

2.4.1 Préparation de la librairie

Pour l'ensemble des échantillons, la préparation de la librairie est réalisée à l'aide de l'enrichissement par capture. En premier lieu, l'ADN est fragmenté de façon aléatoire par une fragmentase. Les produits de fragmentation sont purifiés à l'aide de billes magnétiques, afin de ne conserver que les fragments d'au moins 150 paires de bases. Les amorces (SP1, SP2) nécessaires au séquençage sont alors liées aux fragments d'ADN par PCR. Puis, les régions d'intérêt sont capturées à l'aide de sondes complémentaires liées à des billes de streptavidine ; cette étape est suivie d'un lavage permettant de retirer les régions non capturées. Le produit de capture est amplifié par PCR afin d'enrichir en régions d'intérêts l'échantillon et de lier l'index permettant l'identification de l'échantillon (index 5, index 7), ainsi que les adaptateurs permettant l'hybridation a la Flow Cell (P5, P7). Enfin, le produit d'amplification des différents échantillons est dosé, normalisé et poolé avec l'ensemble des échantillons du run (Figure 16).

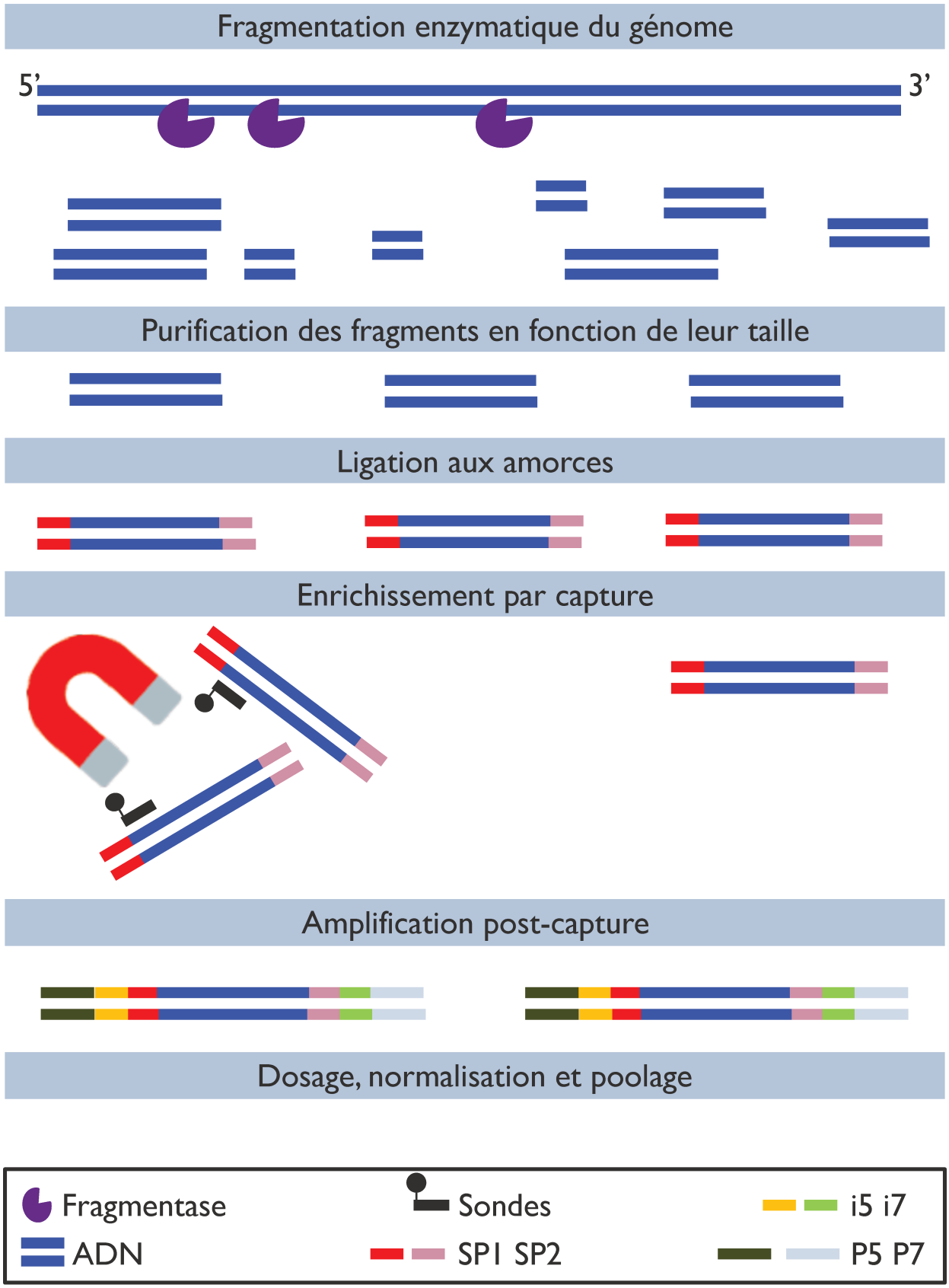

Figure 16 : Préparation de la librairie par capture

La librairie provenant du jeu de données "d'entrainement" a été élaborée à l'aide du kit SureSelect-QXT AGILENT® conformément aux recommandations du fabricant.

La librairie provenant du jeu de données "test" a été élaborée à l'aide du kit Twist BIOSCIENCE® conformément aux recommandations du fabricant.

Préalablement dénaturées, les librairies ont été chargées dans une cassette de séquençage ILLUMINA®, contenant les réactifs nécessaires à la suite du protocole. Le séquençage a été réalisé à l'aide du séquenceur MiSeq® pour le jeu de données "d'entrainement", et du séquenceur NovaSeq® pour le jeu de données "test".

2.4.2 Amplification clonale par clusterisation de la Flow Cell

Des centaines de millions d'ADN monobrins sont hybridés à une couche d'adaptateurs oligonucléotididiques (P5 et P7) immobilisés à la surface de la Flow Cell. Ensuite, durant l'étape d'élongation, ces ADN monobrins sont copiés à partir des adaptateurs jouant le rôle d'amorce. Enfin, les brins d'ADN originaux sont dénaturés, tandis que les brins néoformés restent immobilisés sur la Flow Cell (Figure 17 ***A***).

Les brins néoformés s'hybrident aux adaptateurs adjacents, formant ainsi des ponts qui permettent l'élongation des brins à partir des adaptateurs (bridges d'amplification). Les ponts d'ADN double brins sont ensuite dénaturés pour former deux brins d'ADN monobrins. Ces étapes d'hybridation, élongation, dénaturation sont répétées pour créer des millions de groupes de copies nommés clusters (Figure 17 ***B***).

Les brins antisens sont ensuite éliminés par clivage de P5. Les extrémités 3' des adaptateurs P5 ont été préalablement bloquées pour éviter toute interférence avec la réaction de séquençage (Figure 17 ***C***).

A la fin de l'amplification clonale, la Flow Cell contient plusieurs centaines de millions de clusters avec ~1 000 molécules d'ADN sens par cluster et est prête à être séquencée (Figure 17 ***D***).

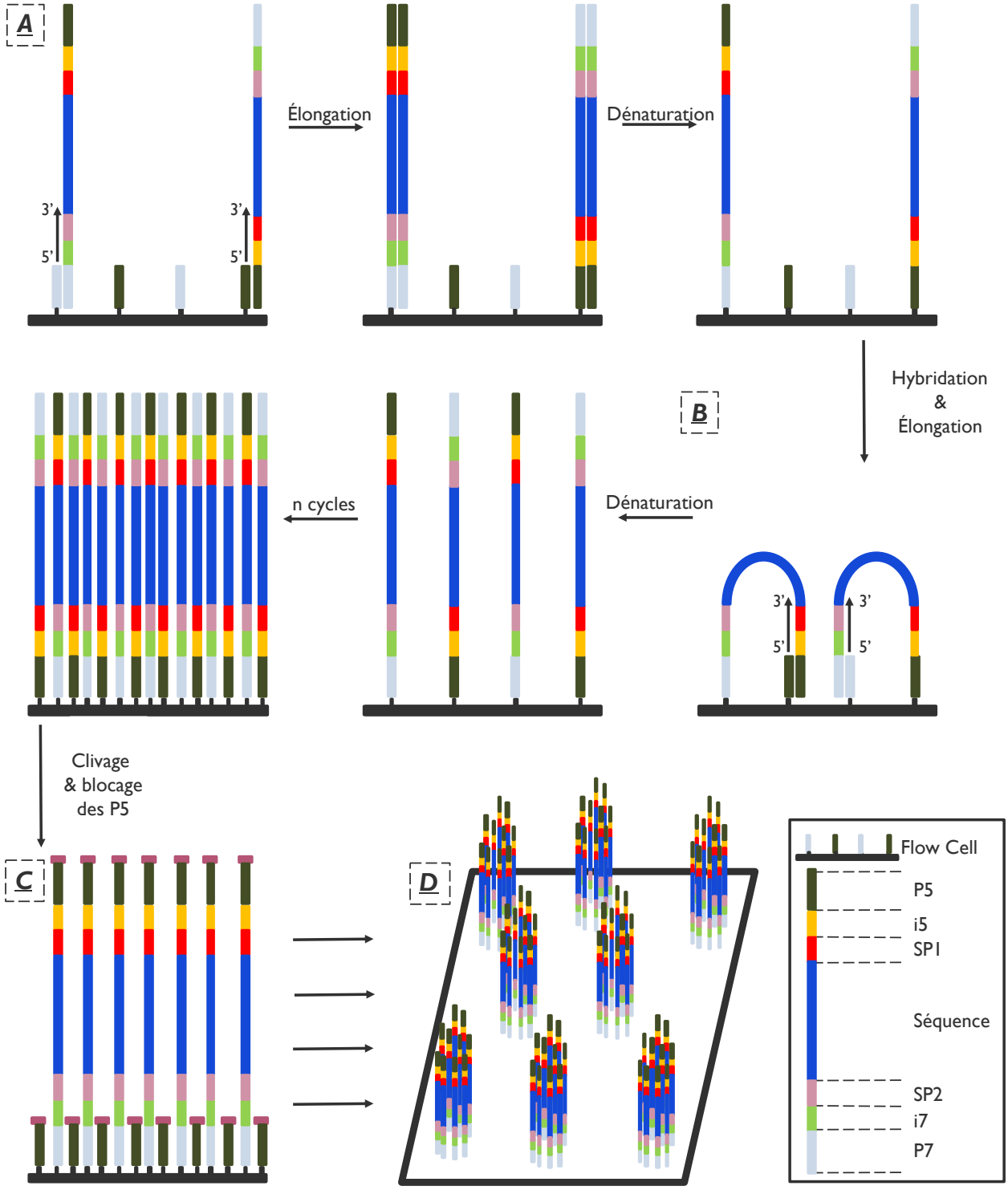

Figure 17 : Clusterisation de la Flow Cell

2.4.3 Séquençage

Chaque cycle de séquençage comprend (i) l'incorporation d'un type de nucléotide terminateur réversible (didésoxynucléotide ddNTP) A, T, C ou G marqué par un fluorophore caractéristique du nucléotide (ii) un lavage des ddNTPs non incorporés (iii) la prise d'image de la fluorescence par une caméra haute résolution (iv) le clivage de la partie terminatrice/fluorophore (v) un lavage de cette partie.

La première étape correspond au séquençage sens du brin (R1). Elle débute par l'hybridation d'une amorce complémentaire au SP1 permettant l'initiation du séquençage. Dans un second temps, s'il existe une complémentarité au brin, un ddNTP est incorporé. Après l'incorporation, les ddNTP non incorporés restants sont ôtés par lavage. Une prise d'image est ensuite effectuée pour déterminer l'identité du nucléotide incorporé. L'étape de clivage permet par la suite de casser la liaison entre le nucléotide et son groupe terminateur/fluorophore, ce dernier est alors éliminé par une étape de lavage avant de recommencer le cycle n fois. Lorsque l'intégralité du brin sens est séquencé, le brin néoformé ayant servi pour le séquençage est dénaturé (Figure 18 ***A***).

Dans un second temps, l'index i7 est séquencé à l'aide d'une amorce complémentaire à SP2 suivant le même cycle que pour le brin sens. Le brin néo formé est dénaturé et le 3'OH du P5 est débloqué (Figure 18 ***B***).

Une fois le brin sens et l'index i7 séquencés, l'étape de préparation du paired-end peut être initiée. Cette étape comprend une succession d'hybridations, d'élongations et de dénaturations afin de former le brin antisens. Le brin sens est éliminé par clivage de P7. Les extrémités 3' des adaptateurs P7 sont bloquées (Figure 18 ***C***).

L'index i5 est séquencé à l'aide d'une amorce complémentaire à SP1, précédant une étape de dénaturation (Figure 18 ***D***).

En dernier lieu, le brin antisens est séquencé (R2) (Figure 18 ***E***).

Les images des différents cycles générées par la caméra haute résolution sont stockées dans des fichiers .bcl (Binary Base Call) contenant l'ensemble des échantillons du run.

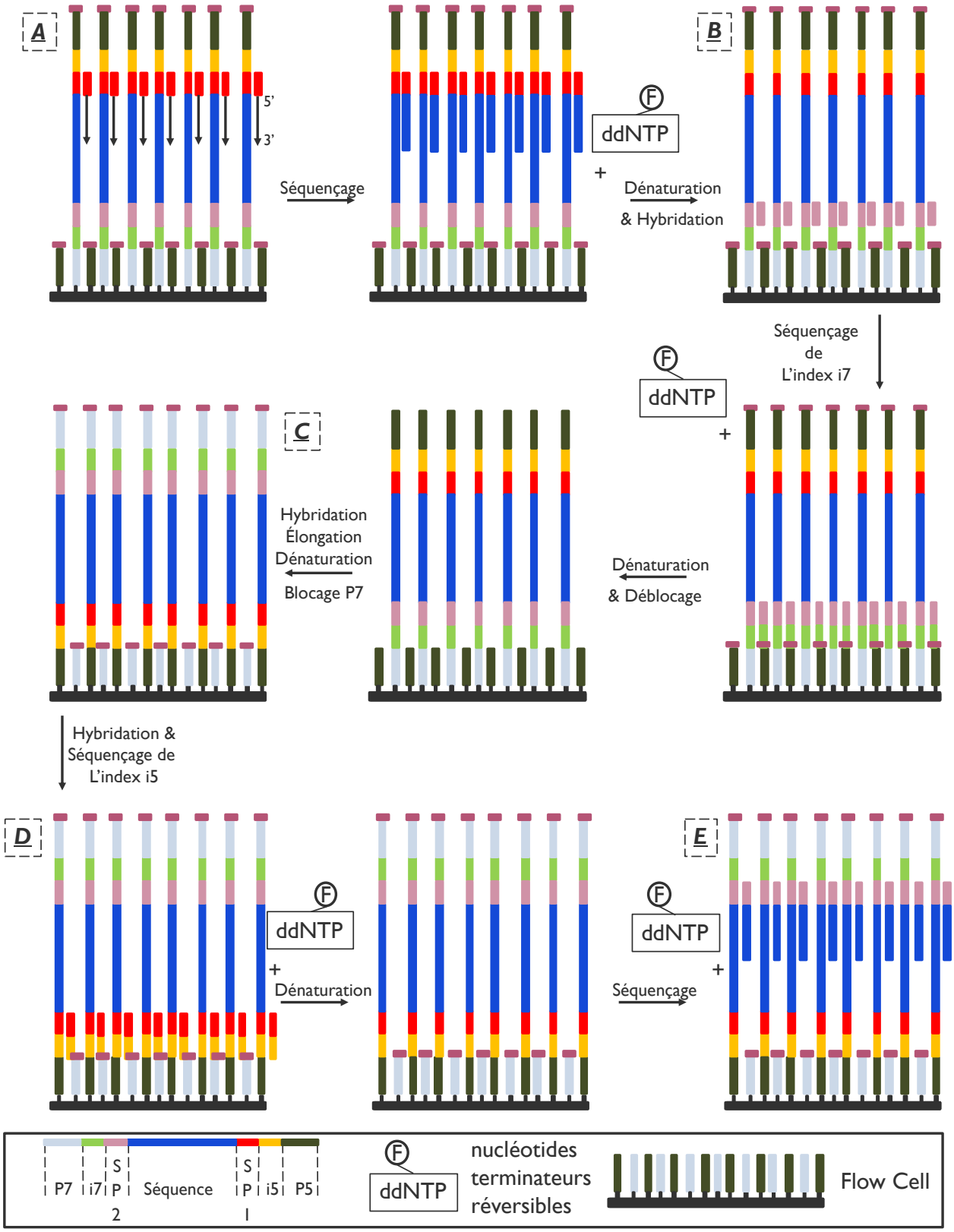

Figure 18 : Séquençage paired-end

2.4.4 Traitement bio-informatique

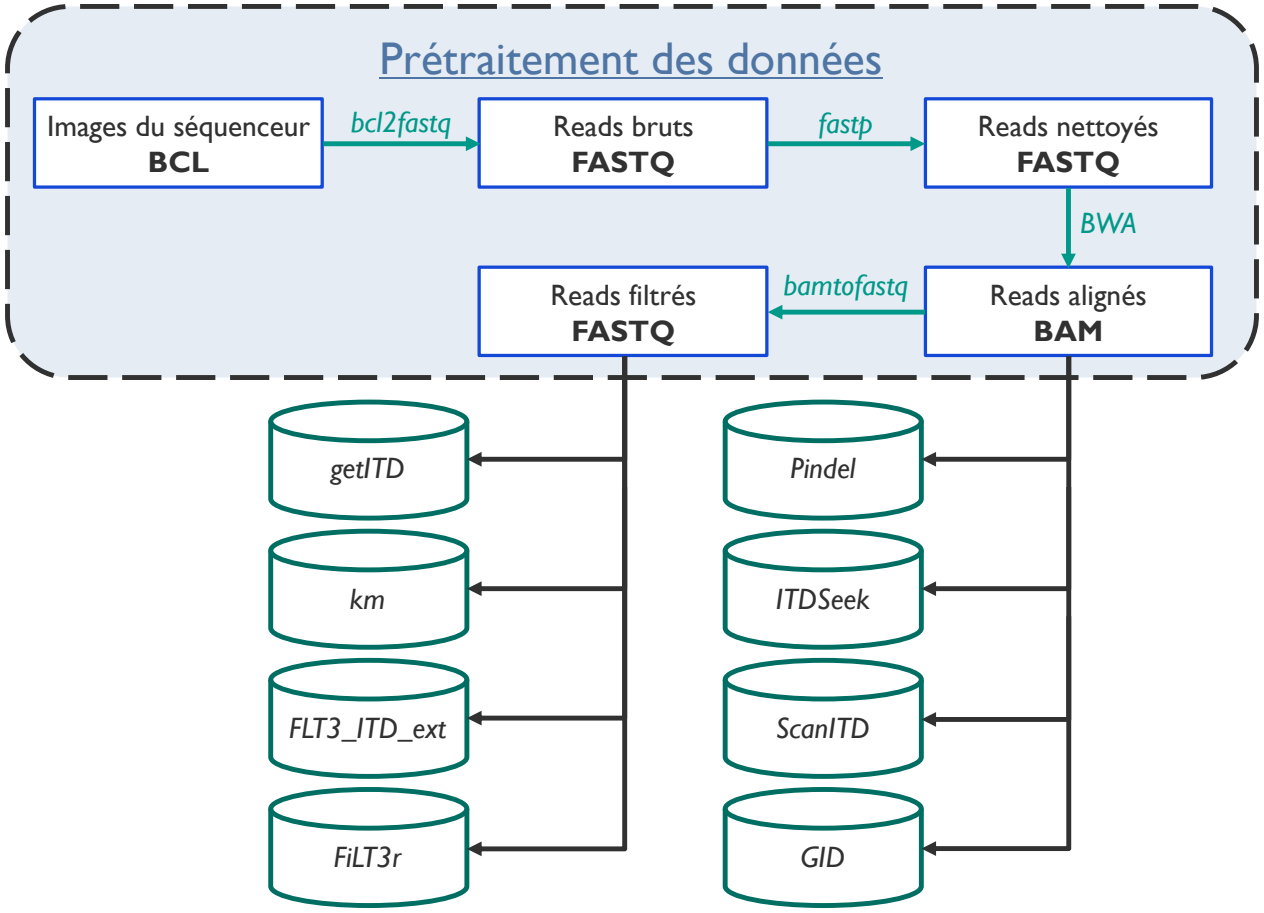

Figure 19 : Vue d'ensemble du pipeline

2.4.4.1 Prétraitement des données

L'ensemble des processus est chainé à l'aide d'un gestionnaire de pipelines *Nextflow*.

Les fichiers BCL du Run sont démultiplexés à l'aide d'un fichier tabulé comprenant les index (i5 et i7) propres à l'échantillon. L'outil *bcl2fastq* prend en entrée ces fichiers et crée un fichier FASTQ par échantillon.

Chaque fichier FASTQ est composé d'enregistrements qui sont en blocs de quatre lignes correspondant à un read. La première ligne, qui commence par le symbole @, correspond à l'identifiant du read (Figure 20 ***A***). La deuxième ligne contient la séquence composée des nucléotides G, A, T, C, la présence d'un nucléotide non déterminé étant signalée par la lettre N (Figure 20 ***B***). La troisième ligne commence par le symbole + et contient généralement ce seul caractère (Figure 20 ***C***). La quatrième ligne comprend les scores de qualité correspondant à chaque base (stockés en tant que caractères ASCII) (Figure 20 ***D***). L'outil *fastp* utilise ce score de qualité pour effectuer un nettoyage des reads associés à un score

insuffisant et permet également de rogner les index (i5 i7) ayant servi au démultiplexage étant désormais superflus.

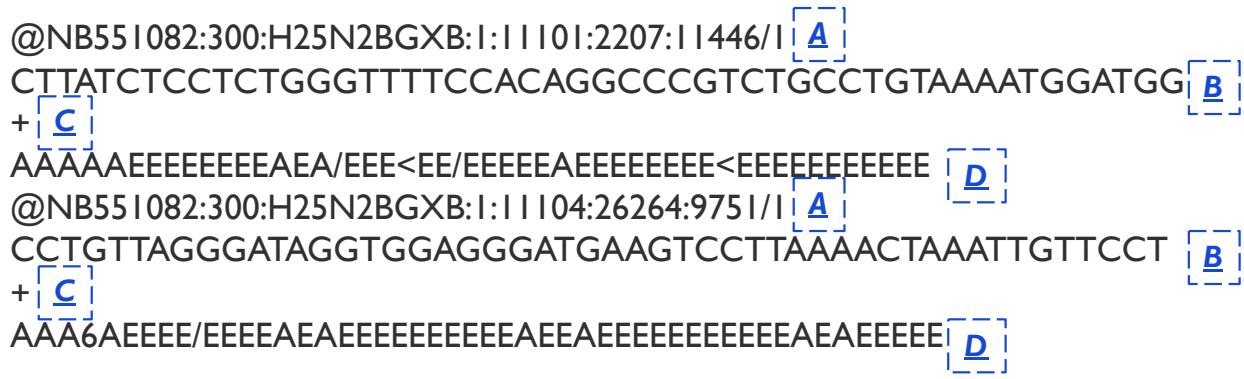

Figure 20 : Exemple de deux reads extraits d'un FASTQ

L'étape d'alignement utilise la suite d'outils GATK (Genome Analysis Toolkit) reposant sur le principe de l'algorithme de Burrows-Wheeler Aligner (BWA), prenant en entrée des fichiers FASTQ et un génome de référence indexé au format FASTA (hg19 UCSC) et donnant en sortie un fichier binary alignment/map (BAM). Le format BAM comprend une section d'en-tête (dont les lignes commencent par le caractère @) et une section d'alignement. Cette section d'alignement est délimitée par des tabulations composées le plus souvent de 11 champs (Tableau 5).

Tableau 5 : Section d'alignement BAM

QNAME	NOM de la séquence requête
FLAG	bitwise FLAG contenant les informations concernant l'alignement
RNAME	NOM de la séquence de référence
POS	POSition 5' de l'alignement
MAPQ	Qualité MAPping
CIGAR	Compact Idiosyncratic Gapped Alignment Record Chaîne de caractères CIGAR décrivant l'alignement
RNEXT	Nom du Read apparié
PNEXT	Position du Read apparié
TLEN	Longueur de séquence alignée
SEQ	SEQuence
QUAL	Qualités de base ASCII selon l'échelle Phred QUALity

En raison de la taille conséquente du panel, une étape de pré-filtrage est effectuée pour les algorithmes ne nécessitant pas d'alignement afin de diminuer le temps d'exécution. Les FASTQs sont extraits des BAMs à l'aide l'outils *bamtofastq* (provenant de la suite d'outils *bedtools*), le BAM étant de toute façon réalisé car nécessaire à l'appel de variants classiques.

L'extraction est effectuée sur les reads couvrant la région *FLT3* (chr13:28577411-28674713) ainsi que les reads non alignés.

De nombreux algorithmes ont été testés parmi lesquels : *pindel* (112), *ITDseek* (114), *ScanITD* (115), *getITD* (116), *FLT3_ITD_ext* (117), *km* (118), *Genomon ITDetector (GID)* (119), *ITD-Assembler* (120) ainsi que l'algorithme *FiLT3r* développé au laboratoire en partenariat avec le CNRS. L'algorithme *ITD-Assembler* (120) *a* été exclu parce qu'il ne fonctionnait pas dans notre environnement de calcul. Cette difficulté d'installation a été également rapportée par d'autres équipes (115,121).

2.4.4.2 Algorithmes testés

Tableau 6 : Résumé des différents algorithmes testés

Outils	Basé sur l'alignement	Versions	Variants détectés	Année de publication
pindel	OUI	v0.2.5b9	Indels et SV	2009
GID	NON	Ø[1]	ITDs	2014
ITDSeek	OUI	v1.2	ITDs	2016
getITD	OUI	v1.5.10	ITDs	2019
km	NON	v2.0.1	SNP, indels et fusions	2019
ScanITD	OUI	Ø[1]	ITDs	2020
FLT3_ITD_ext	OUI	Ø[1]	ITDs	2020
FiLT3r	NON	Ø[1]	Indels	2021

[1]*Ø : Pas de version*

2.4.4.2.1 Algorithmes basés sur l'alignement

2.4.4.2.1.1 *pindel*

pindel comprend une stratégie split-read afin de détecter des variants structuraux (SV) ainsi que les ITDs. Il aligne d'abord toutes les lectures sur le génome de référence, puis sélectionne les paires de reads dont une seule extrémité a été alignée avec succès et divise l'extrémité non alignée en deux parties, afin d'aligner ce qui peut être aligné (Figure 21). Enfin, si la partie restante non alignée sur la position théorique (en jaune sur la Figure 21) s'aligne à une position en amont ou en aval il s'agit alors d'une duplication.

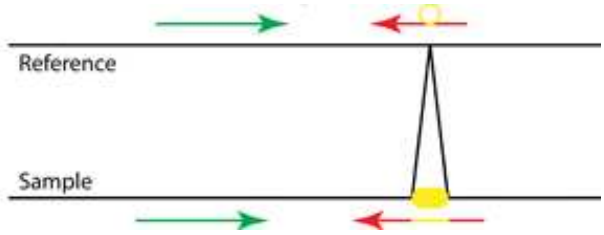

Figure 21 : Stratégie split-read par pindel

Les paramètres d'usage ont été employés conformément aux recommandations provenant du manuel utilisateur de l'équipe l'ayant mis au point (122).

2.4.4.2.1.2 *ITDseek*

ITDseek prend en entrée les alignements SAM/BAM et crée une sortie VCF contenant uniquement les mutations ITDs. Son principe repose sur l'alignement secondaire des nucléotides "soft-clipped" (lettre "S" dans le code CIGAR) déterminés par BWA MEM. Les nucléotides soft-clipped sont définis comme les parties 5' ou 3' des reads qui ne s'alignent pas sur la séquence de référence durant l'alignement à la suite d'un point de cassure (Figure 22). Le réalignement des nucléotides soft-clipped est un moyen d'identifier les éventuelles duplications. La longueur de l'ITD est extrapolée par la distance entre le point soft-clipped et le début du réalignement (114).

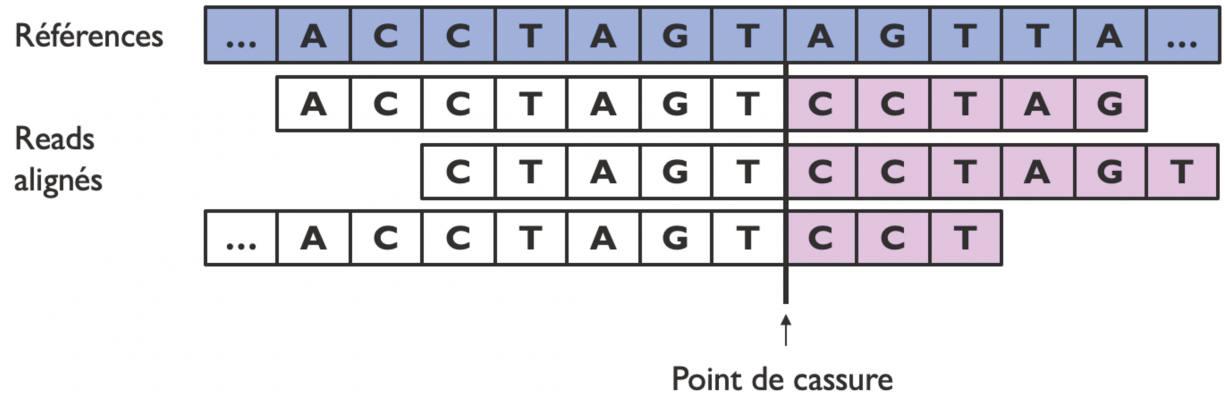

Figure 22 : Reads soft-clippés

La séquence de référence est indiquée en bleu.
Les nucléotides en blanc correspondent à des nucléotides correctement alignés.
Les nucléotides en rose correspondent à des nucléotides soft-clipped.

Les paramètres d'utilisation ont été utilisés conformément aux recommandations provenant du git de l'équipe l'ayant mis au point (123).

2.4.4.2.1.3 ScanITD

A l'instar d'*ITDseek, ScanITD* utilise l'alignement secondaire et nécessite donc l'utilisation de BWA MEM pour l'alignement. Afin de détecter les ITDs, *ScanITD* analyse le fichier BAM en suivant deux étapes.

Dans la première étape, *ScanITD* reconstruit les ITDs en recombinant les lectures chimériques par les procédures suivantes : (i) identification des reads soft-clipped à l'aide du CIGAR (ii) clusterisation des alignements primaires et secondaires (l'alignement secondaire correspond à l'alignement des nucléotides soft-clipped (Figure 23 ***B***) alignés sur le même chromosome et le même brin (iii) estimation de la position génomique et de la taille de l'ITD

à partir du décalage de distance entre les alignements primaire et secondaire. Selon la taille de l'ITD, deux situations peuvent être observées. Dans la première situation, la taille de l'ITD est inférieure à la longueur du read et *ScanITD* reconstruit les ITDs en modifiant le CIGAR par l'ajout de la position de départ de la lecture chimérique. Il ajoutera (n)I dans la chaîne CIGAR redéfinie, où (n) est la taille de l'ITD et "I" indique une insertion (Figure 23 ***B***). Dans la deuxième situation, la taille de l'ITD est supérieure à la longueur du read et l'algorithme ajoutera une nouvelle balise SV dans les reads chimériques (créés à partir des reads soft-clipped) au lieu de modifier leurs chaînes CIGAR. Le format de la balise SV est de type (TDUP, POS, SIZE), où TDUP indique qu'il s'agit d'un événement ITD, et la position et la taille de l'ITD sont déduites comme illustré dans la (Figure 23 ***B***).

La deuxième étape consiste à calculer la fréquence allélique (VAF) en recherchant l'ITD dans les reads chimériques. La VAF est calculée par AO/DP, où AO (alternate allele observation count, nombre d'observations d'allèles alternatifs) est le nombre de lectures soutenant l'ITD et DP (depth) la profondeur totale de lecture.

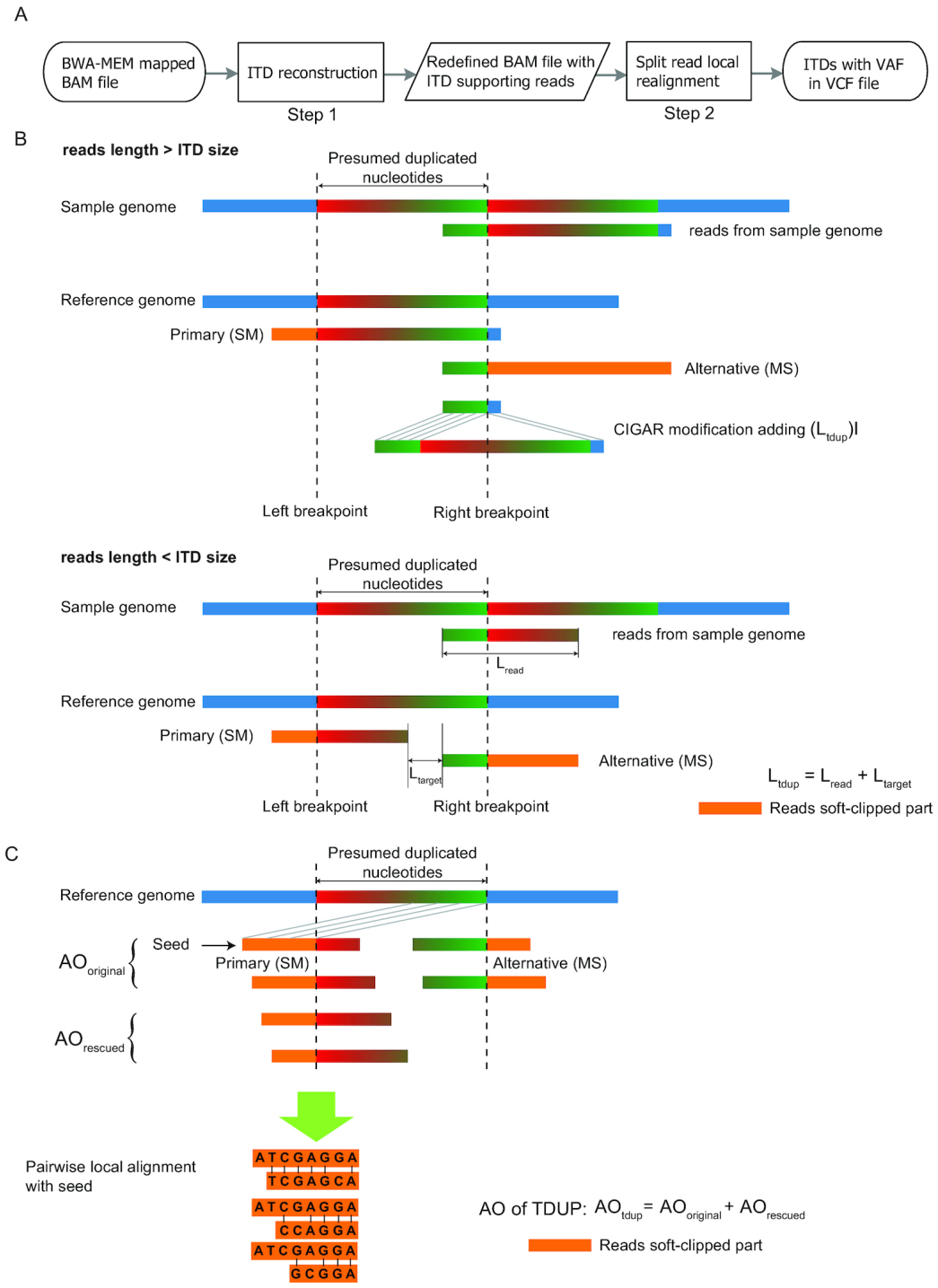

Figure 23 : Vue d'ensemble de ScanITD (115)

A) Vue d'ensemble schématique de ScanITD

B) Les ITDs sont récupérées à partir des alignements primaire et secondaire (soft-clipped)

C) Réalignement local des lectures fractionnées pour calculer le nombre réel de lectures porteuses de l'ITD (appelé AO)

ScanITD a été utilisé conformément aux recommandations provenant du git de l'équipe l'ayant mis au point (123).

2.4.4.2.1.4 *getITD*

getITD prend en entrée les FASTQ. Son principe repose sur un alignement de Needlemann-Wunsch. *getITD* sélectionne les reads de haute qualité (base quality score > 30) alignés sur le génome de référence du gène *FLT3* pour identifier les insertions, puis détermine si les insertions sont des ITDs. L' ensemble des ITDs est filtré selon deux critères : VAF <0.006% et <2 reads mutés.

getITD a été utilisé conformément aux recommandations provenant d'échanges épistolaires avec l'équipe l'ayant mis au point.

2.4.4.2.1.5 *FLT3_ITD_ext*

FLT3_ITD_ext est basé sur deux principes, l'utilisation du soft-clipping et la clusterisation à l'aide de sumaclust. Les différentes étapes sont décrites ci-dessous (Figure 24)

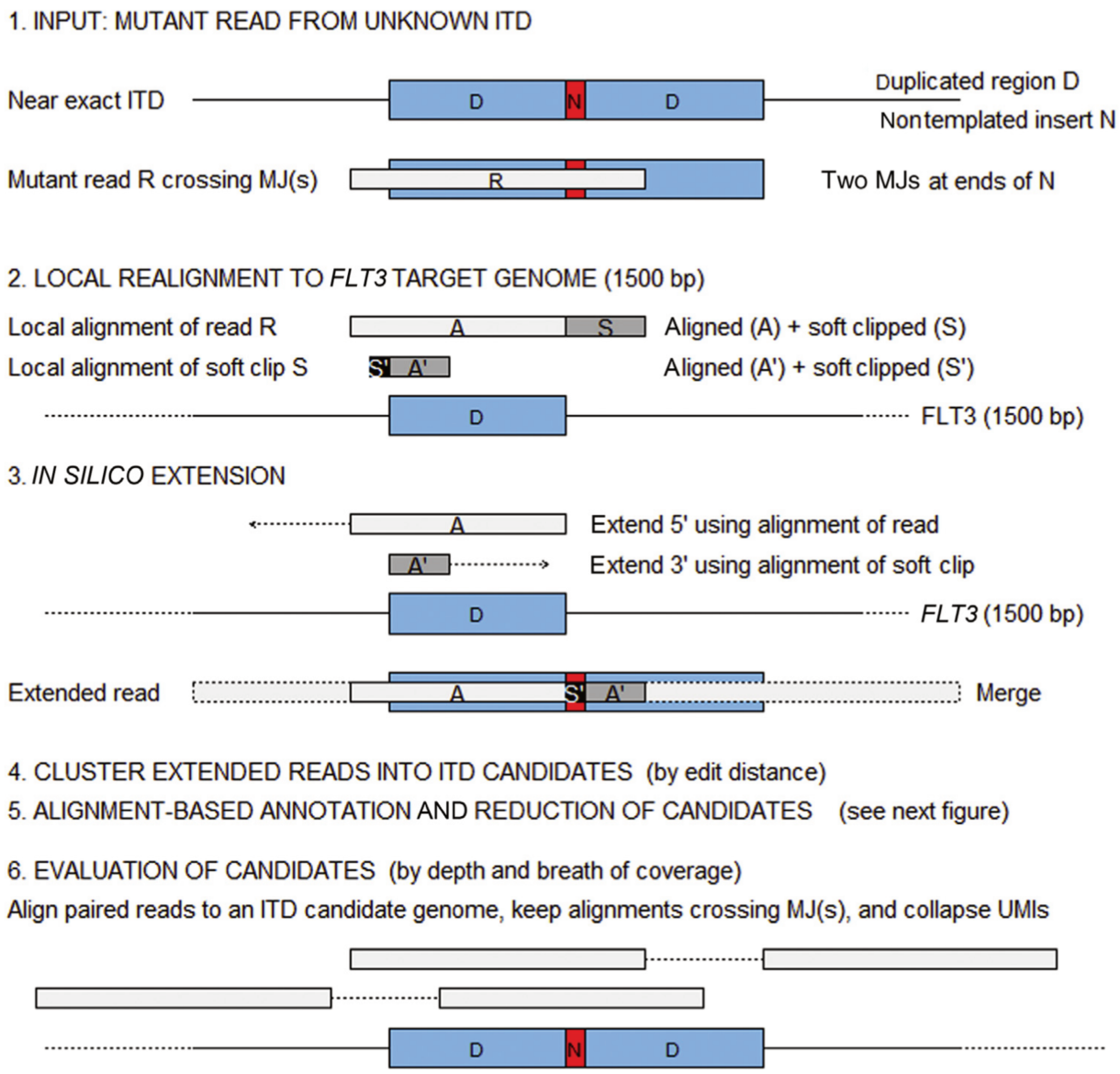

Figure 24 : Vue d'ensemble de FLT3_ITD_ext (117)

1 : En entrée un fichier BAM ou FASTQ contenant des reads (R ; zone encadrée gris clair) chevauchants la jonction mutante (MJ) d'une ITD avec un segment dupliqué (D ; zone encadrée bleu clair) et une insertion (N ; zone encadrée rouge).

2 : Ces reads sont alignés (alignement local) sur le locus cible FLT3 : le point d'ancrage (A ; zone encadrée en gris clair), la partie soft-clipped (S ; zone encadrée en gris foncé) ainsi que la partie MJ [S' (zone encadrée en noir) ; A' (zone encadrée en gris)].

3 : A partir des alignements (A ; A') une extension en 5' et 3' respectivement est effectuée in silico.

4 : Des clusters sont formés à partir de ces séquences étendues en formant ainsi des ITDs candidates.

5 : Agrégation des ITDs candidates et annotation basée sur l'alignement

6 : L'étendue et la profondeur de la couverture sont évaluées à l'aide des reads paired puis l'AR est estimé.

FLT3_ITD_ext a été utilisé conformément aux recommandations provenant du git de l'équipe l'ayant mis au point (124).

2.4.4.2.2 *Algorithmes sans alignement*

2.4.4.2.2.1 *km*

L'algorithme *km* effectue une analyse approfondie d'une seule séquence définie par l'utilisateur, appelée séquence cible qui correspond à la référence et contient la région d'intérêt. Cette séquence cible est décomposée en k-mers (sous-séquences de longueur k) qui se chevauchent de k-1, k étant suffisamment grand pour produire un graphe orienté.

En parallèle, un tableau de comptage des k-mers est préparé, indiquant l'occurrence de chaque k-mer des reads de l'échantillon.

Les mutations sont mises en évidence par l'apparition d'un chemin alternatif reliant les k-mers de départ et d'arrivée (Figure 25).

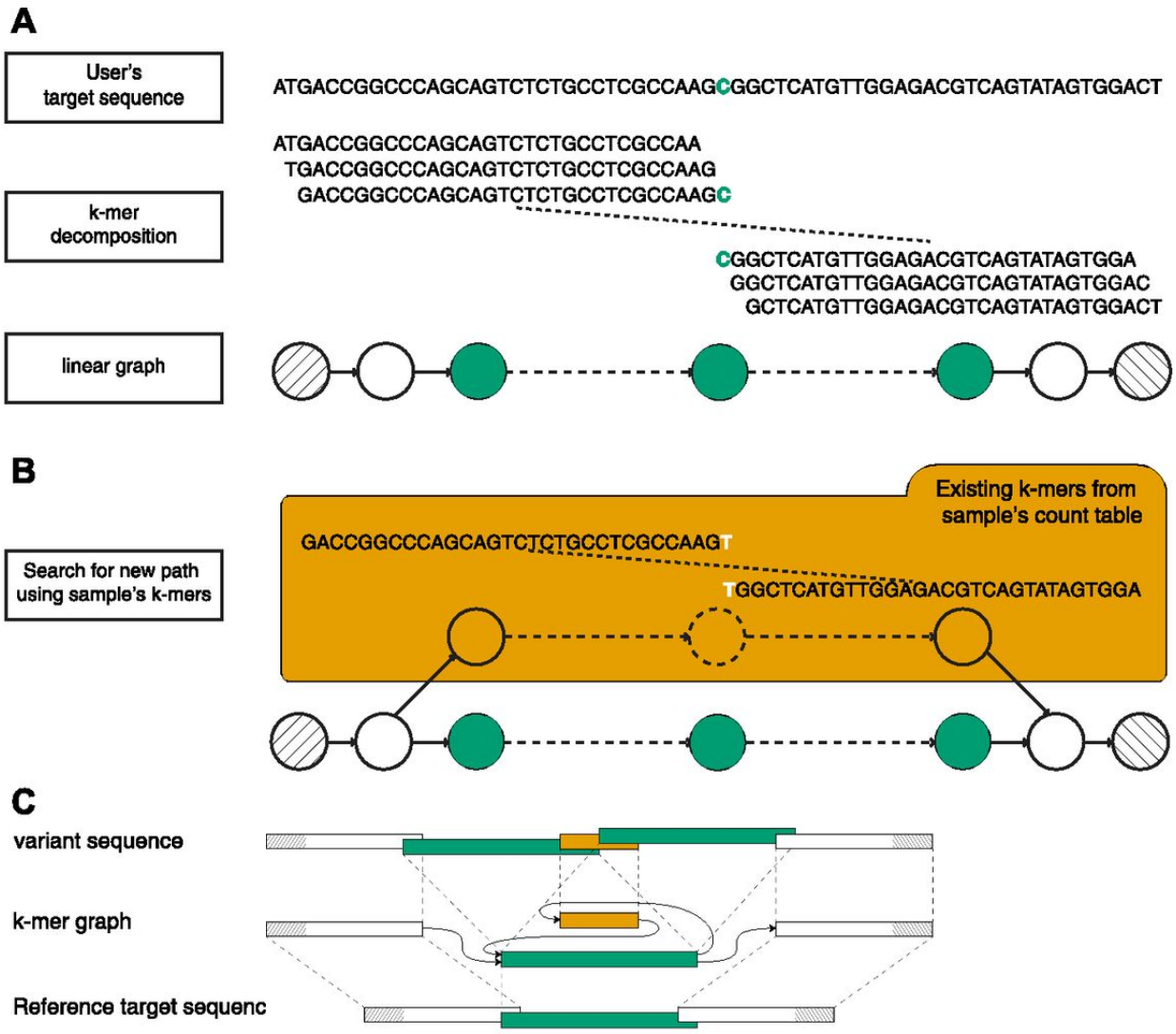

Figure 25 : Vue d'ensemble de km (118)

(A) La séquence de référence est segmentée en k-mers pour créer un graphe orienté, qui représente l'espace de recherche délimité par les k-mers de début et de fin (hachurés).

(B) Une variante sera représentée par un nouveau chemin entre les deux extrémités. Ce chemin est découvert en se déplaçant le long du graphe orienté et en suivant les nouveaux k-mers qui se chevauchent (non visibles dans la séquence cible), interrogés à partir du tableau de comptage d'un échantillon.

(C) Représentation schématique d'un graphe pour une ITD, les chemins indiquant la séquence variante et la séquence cible sont représentés

L'algorithme *km* a été utilisé conformément aux recommandations provenant de la publication princeps de l'équipe l'ayant mis au point (125).

2.4.4.2.2.2 Genomon ITDetector

Genomon ITDetector (GID) se base sur une stratégie d'assemblage et sur l'utilisation des reads soft-clipped. GID collecte les reads soft-clipped d'une longueur de plus de 20pb. Ces reads sont réalignés sur le génome de référence humain afin d'identifier les paires de breakpoints des ITDs (ITD-BPP), (dont les séquences soft-clipped correspondant aux breakpoints gauche et droit s'alignant sur la position des autres breakpoints). Les reads porteurs des ITD-BPPs sont assemblés pour générer une séquence consensus (Figure 26).

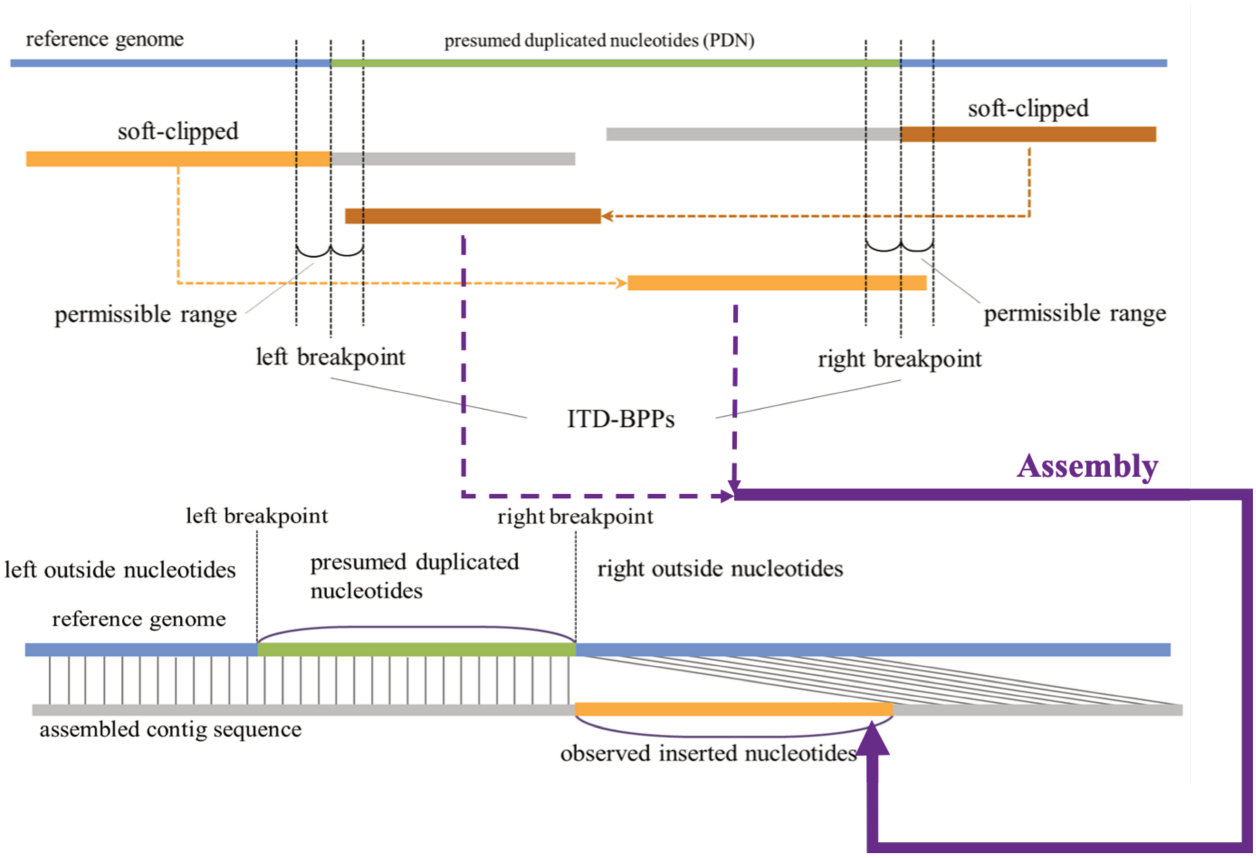

Figure 26 : Vue d'ensemble de Genomon ITDetector

L'algorithme *GID* a été utilisé conformément aux recommandations de l'équipe l'ayant mis au point (126).

2.4.4.2.2.3 FiLT3r

FiLT3r se base également sur la notion de k-mers. Il a pour principe de simplifier la lecture des séquences nucléotidiques : au lieu de les analyser paire de bases par paire de bases, l'algorithme analyse k-mer par k-mer consécutifs, de façon à avoir une vision plus globale. Chaque read est alors associé à une suite d'index correspondant aux indices des k-mers retrouvés dans la référence (ou un 'N' en cas de non-correspondance). Ainsi, une duplication en tandem se caractérise par une suite naturelle (correspondant à la succession des k-mers dans la séquence originelle), séparée par une coupure (nommée point d'arrêt) d'une seconde

suite naturelle (commençant après le début de la première et avant sa fin). La détection ne porte plus sur les séquences nucléotidiques mais sur les suites naturelles. Les différentes étapes de fonctionnement de *FiLT3r* sont décrites ci-dessous (Figure 27).

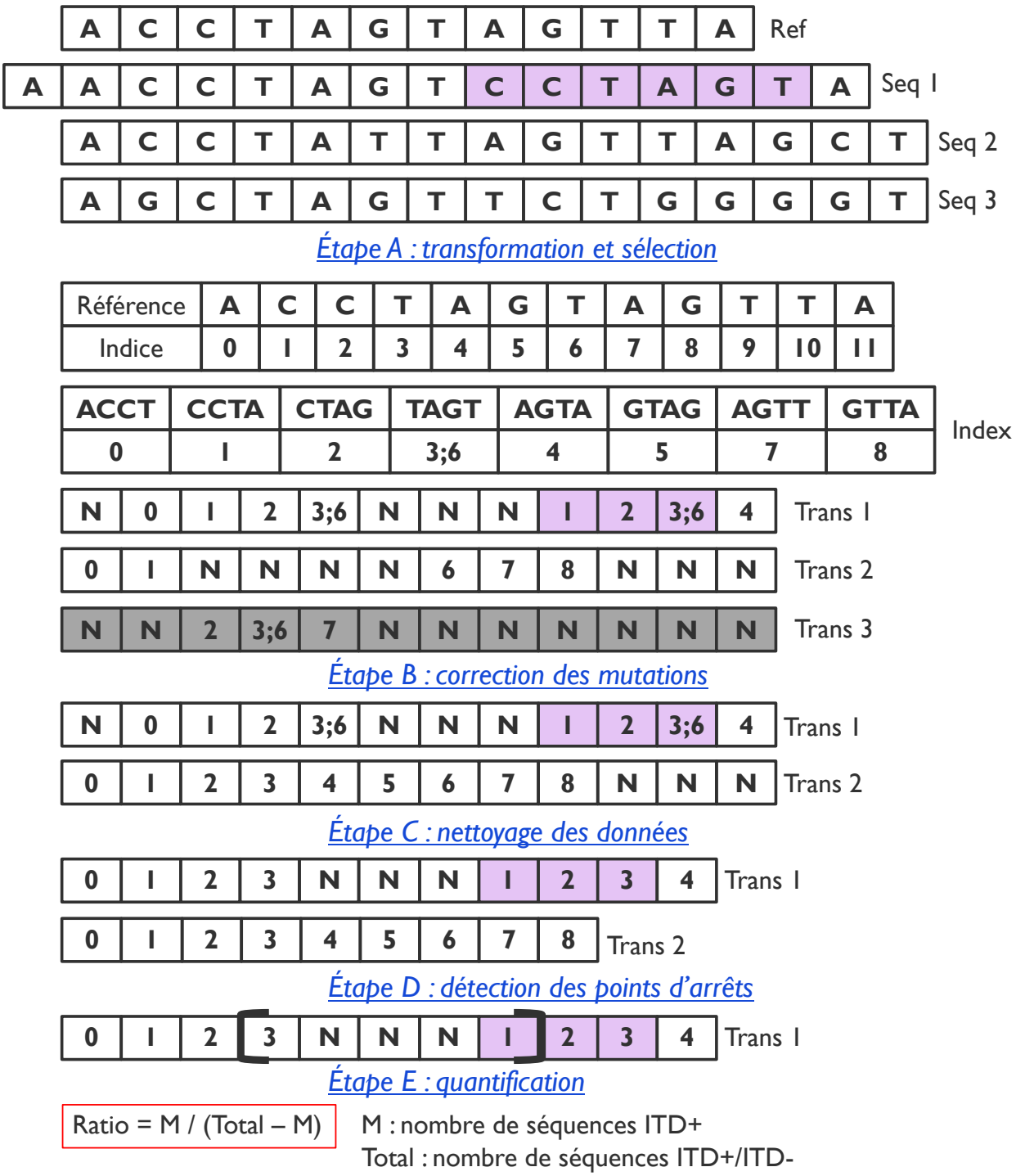

Figure 27 : Vue d'ensemble de FiLT3r

On pose : $k = 4$ la longueur du k-mer, Ref la référence, Seq 1 séquence nucléotidique porteuse d'une duplication, Seq 2 séquence nucléotidique porteuse d'une mutation et Seq 3 séquence non retrouvée dans la référence.

Étape A : création d'un index permettant la transformation de chaque séquence (Seq i) en transformé (Trans i) et sélection des transformés à l'aide d'un seuil (t) de correspondance

à la référence (30 % dans notre cas). Le Trans 3 grisé possède 25% d'homologie avec les k-mers de la référence et a donc été éliminé après la sélection.

Étape B : Cette étape a pour but d'éliminer les N non utiles à la détection des points d'arrêt. Le postulat repose sur le fait que si les N sont inclus dans une suite naturelle et que leur nombre correspond à la longueur k (0 1 N N N N 6 7 8), il s'agit d'une substitution. Ainsi, les N du transformé sont remplacés par la suite naturelle manquante sans altérer la séquence.

Étape C : Cette étape consiste à rendre les transformés facilement analysables :

- *En supprimant les N en 5' et en 3' non utiles à la détection des points d'arrêt (Trans RC 2)*
- *En sélectionnant l'indice du k-mer le plus adapté lorsqu'il correspond à plusieurs indices (l'indice retenu possède la distance la plus courte avec l'indice précédent).(Trans 1)*

Étape D : Les transformés sont parcourus, lorsque des N sont détectés un test est effectué traduisant un point d'arrêt caractéristique d'une ITD si "borne gauche" + "nombre de N" + 1 ≥ "borne droite".
Un point d'arrêt est observé pour le Trans 1 3+3+1 ≥ 1.

Étape E : Quantification de l'AR à l'aide des séquences porteuses du point d'arrêt détecté précédemment pour déterminer l'allèle muté (M). L'allèle sauvage sera estimé en soustrayant le total au muté.

2.4.5 Évaluation

L'analyse comparative des algorithmes a porté principalement sur quatre éléments. (i) La détection : comparaison de la capacité à distinguer correctement les ITDs dont les principaux indicateurs de mesure sont la sensibilité (Se), la spécificité (Sp), la valeur prédictive positive (VPP), et le score F_1. (ii) La quantification : comparaison de la capacité à quantifier correctement l'AR par rapport à la méthode de référence évaluée à l'aide du coefficient de corrélation linéaire et des AR dits aberrants (écart de plus d'un log). (iii) Pouvoir discriminant : comparaison de la capacité à distinguer les patients en regard du seuil de 0.5 d'AR défini par l'ELN.

3 Résultats

3.1 Jeu de données "d'entrainement"

3.1.1 Caractéristiques des ITDs

3.1.1.1 Tailles

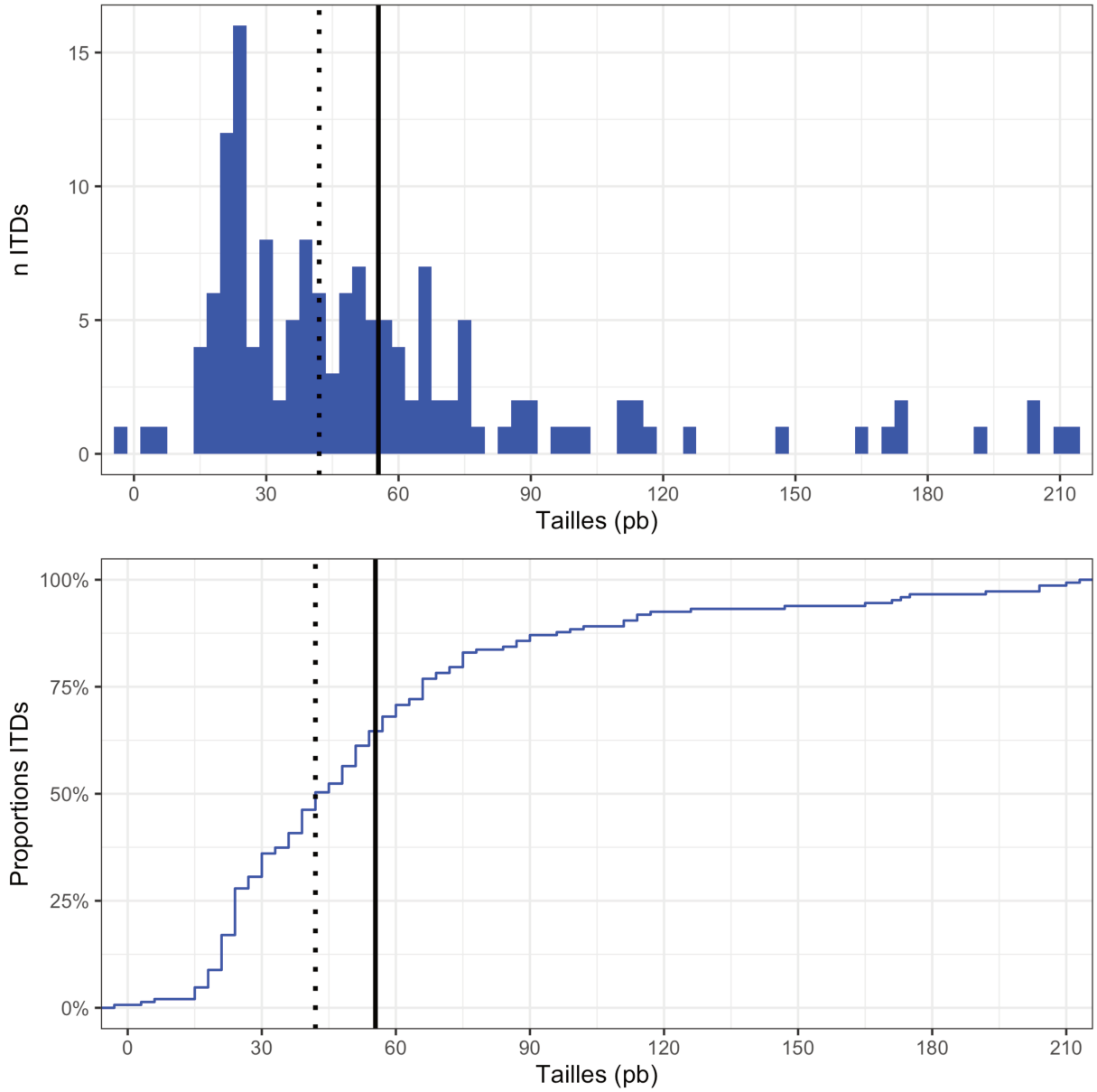

Figure 28 : Distribution des différentes tailles d'ITDs obtenues par analyse de fragments formant le jeu de données d'entrainement

La droite en pointillé correspond à la médiane et la droite pleine correspond à la moyenne.

3.1.1.2 Ratio allélique

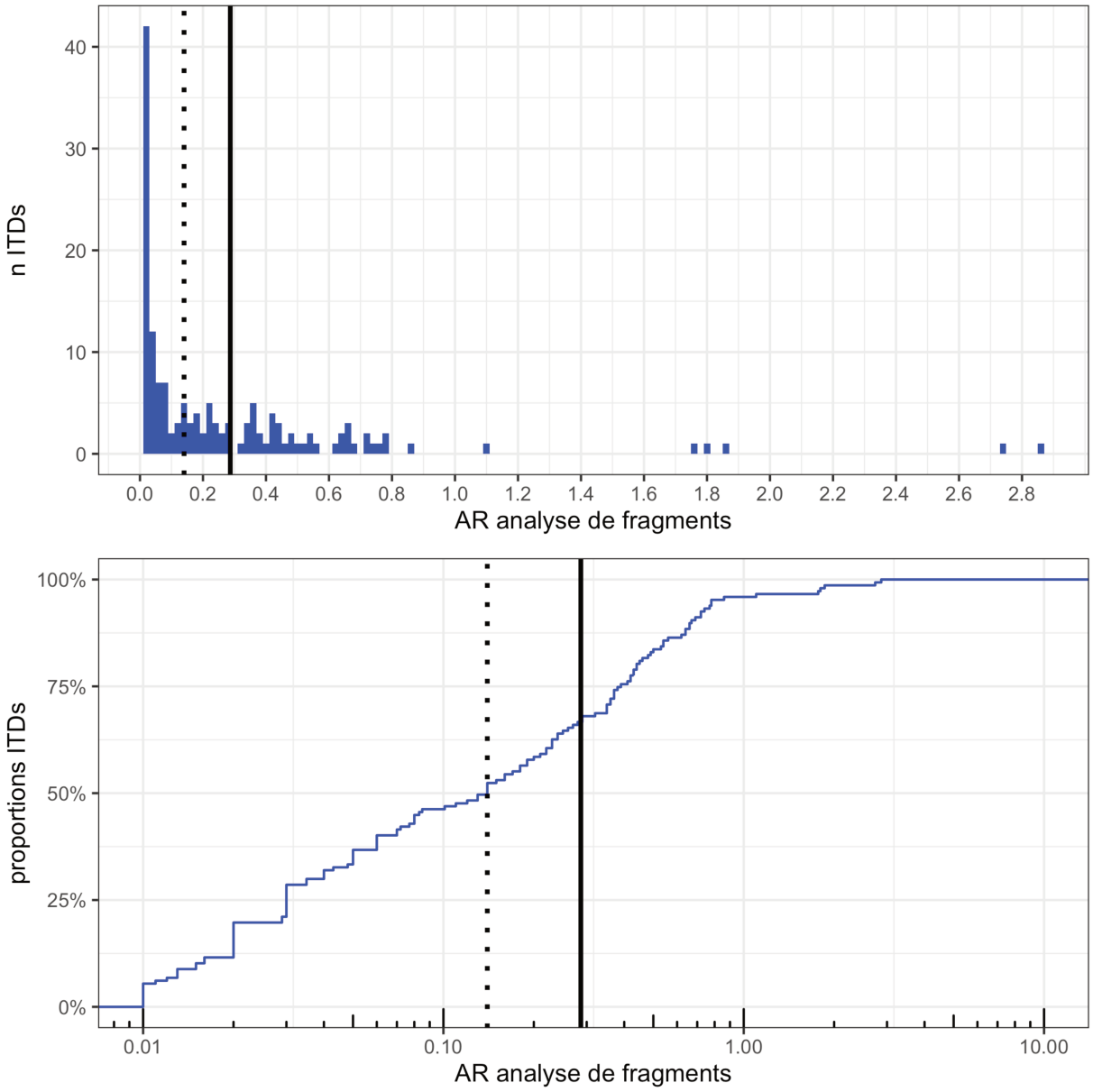

Figure 29 : Distribution des AR obtenus par analyse de fragments formant le jeu de données d'entrainement

La droite en pointillé correspond à la médiane et la droite pleine correspond à la moyenne.

3.1.2 Optimisation du paramétrage

L'optimisation du paramétrage de l'algorithme a été effectuée en modifiant les variables : k la longueur du k-mer et t le seuil de sélection. Ainsi, 98 couples k/t ont été formés à partir des longueurs de k-mer (7 à 20) et des seuils (0.1 à 0.7). Cette optimisation a été évaluée à l'aide des indicateurs clés : la sensibilité et la spécificité.

La sensibilité est équivalente pour la majeure partie des couples de paramètres k/threshold. On observe une légère diminution de cette dernière pour des k-mers de longueur k 15-20 à un threshold de 0.7 (Figure 30).

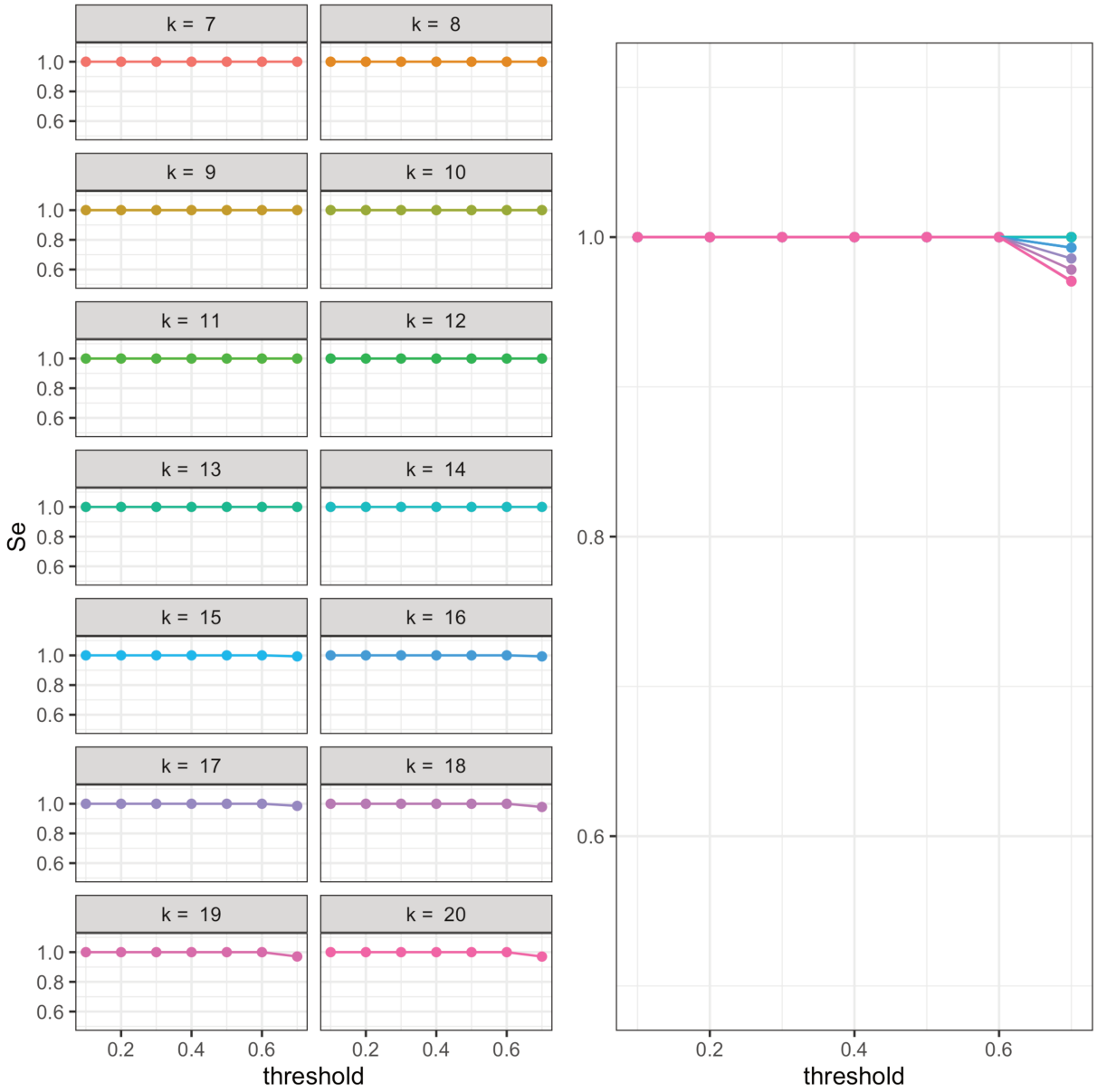

Figure 30 : Sensibilité en fonction de la longueur du kmer (k) et du seuil (threshold)

L'algorithme ayant une limite de quantification plus basse que la méthode de référence, la spécificité est difficile à évaluer pour l'ensemble des couples. Les faux positifs sont donc comptabilisés uniquement s'ils sont retrouvés à un AR supérieur à la limite de quantification de la méthode de référence (2%).

La spécificité reste constante peu importe le seuil et augmente avec la longueur du k-mer jusqu'à atteindre un plateau pour une longueur à 12 (Figure 31).

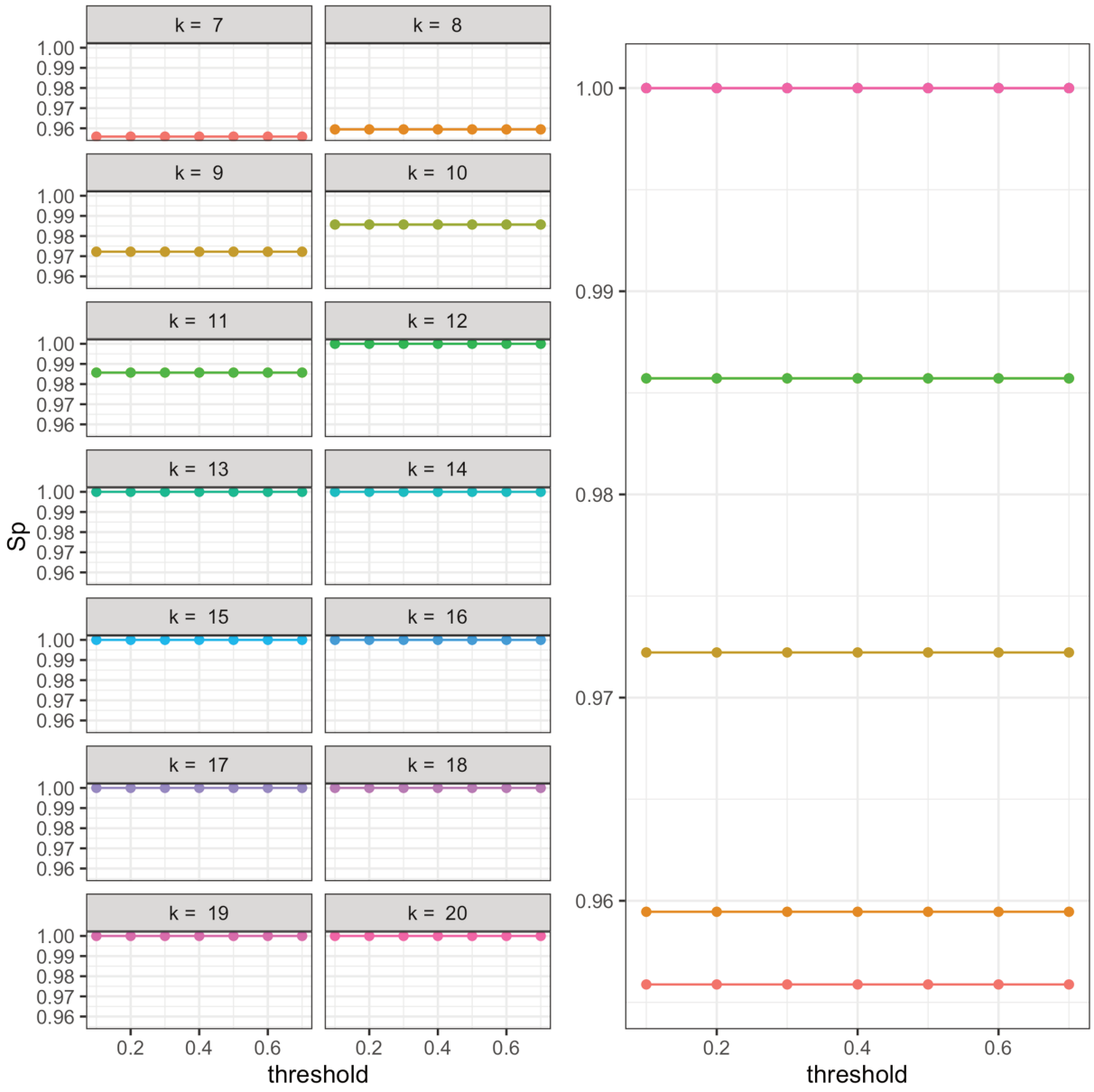

Figure 31 : Spécificité en fonction de la longueur du kmer (k) et du seuil (threshold)

En se basant sur les résultats de sensibilité et de spécificité, le couple k = 12 / t = 0.3 a été choisi pour la comparaison des algorithmes sur le jeu de données "test". La longueur k = 12 a été choisie car elle correspond à la longueur la plus courte ayant la spécificité la plus élevée. L'utilisation d'une longueur courte d'un k-mer permet notamment de mieux couvrir les extrémités des reads, d'être moins impacté par les substitutions et par conséquent plus sensible. Le seuil de 0.3 a été choisi pour filtrer au maximum les reads durant l'étape de transformation et de sélection, permettant ainsi un temps d'exécution plus court.

3.2 Jeu de données "test"

3.2.1 Caractéristiques des ITDs

3.2.1.1 Tailles

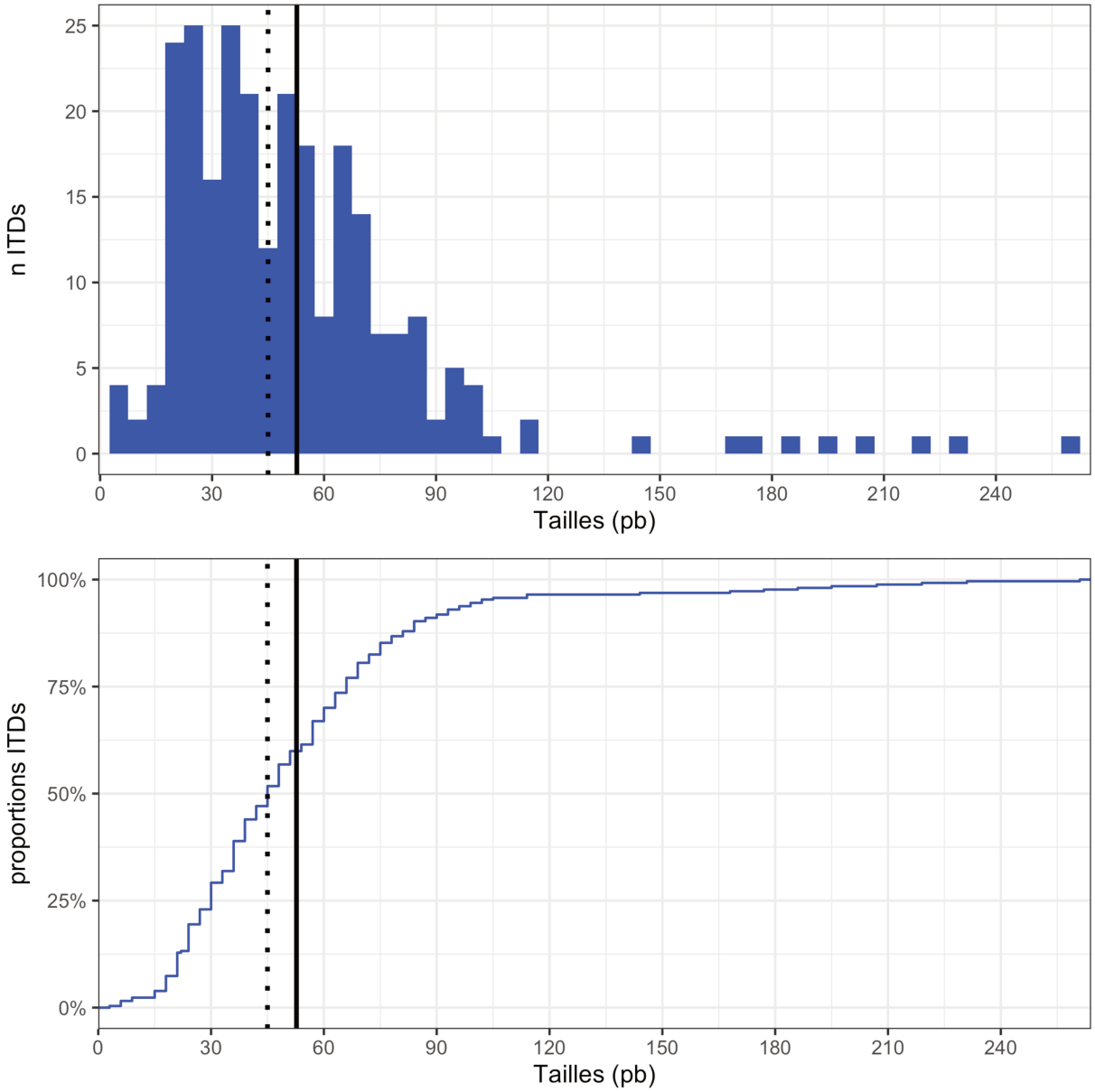

Figure 32 : Distribution des différentes tailles d'ITDs obtenues par analyse de fragments formant le jeu de données test

La droite en pointillé correspond à la médiane et la droite pleine correspond à la moyenne.

3.2.1.2 Ratio allélique

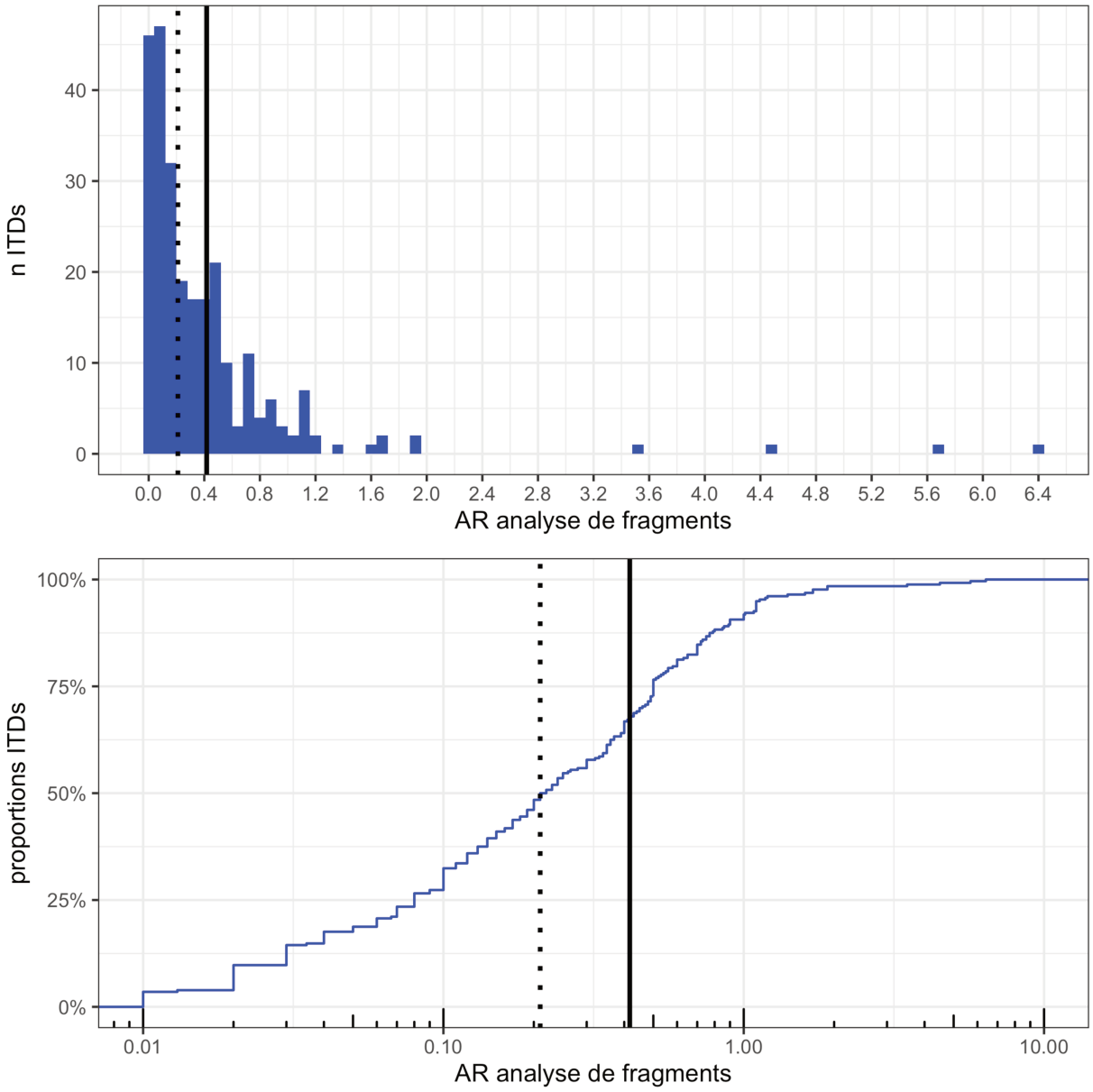

Figure 33 : Distribution des AR obtenus par analyse de fragments formant le jeu de données test

La droite en pointillé correspond à la médiane et la droite pleine correspond à la moyenne.

3.2.2 Évaluations

3.2.2.1 Détection

Les résultats émanant des différents algorithmes ont été comparés à ceux provenant de l'analyse de fragments. Toute duplication détectée par un algorithme, quel que soit son ratio et également identifiée par la méthode de référence a été considérée comme un vrai positif. Toute duplication détectée par un algorithme avec un AR supérieur à 2% et qui n'a pas été identifiée par la méthode de référence a été considérée comme un faux positif.

Les indicateurs clés utilisés comprenaient la sensibilité (Se), se référant à la proportion d'ITD positives détectées par l'outil parmi les ITDs déterminées par l'analyse de fragments ; la spécificité (Sp), se référant à la proportion d'ITDs négatives détectées par l'outil parmi les échantillons jugés négatifs par l'analyse de fragments ; la valeur prédictive positive (VPP), qui se réfère à la proportion d'ITDs correctement détectées au regard de l'analyse de fragments parmi l'ensemble des ITDs détectées par l'outil ; et le score F_1, défini comme la moyenne harmonique de la Se et de la VPP.

Les performances de *pindel*, *FLT3_ITD_ext* et *FiLT3r* ont été nettement supérieures à celles des autres algorithmes, avec un score F_1 et une spécificité élevés. *FiLT3r* a présenté la sensibilité la plus élevée (Se = 1) ; comparable aux sensibilités des outils *pindel* et *FLT3_ITD_ext* (0.98 et 0.99 respectivement). L'ensemble des algorithmes à l'exception de ITDSeek ont révélé une spécificité équivalente.

Globalement, si l'on considère les indicateurs d'évaluation de la détection, *pindel*, *FLT3_ITD_ext* et *FiLT3r* ont obtenu de meilleurs résultats (Tableau 7)

Tableau 7 : Tableau agrégé des résultats de l'ensemble des algorithmes testés sur le jeu de données test

	VP	FN	VN	FP	Se	Sp	VPP	F_1
pindel	251	6	766	0	0.98	1	1	0.99
ITDSeek	182	75	686	80	0.71	0.9	0.69	0.7
ScanITD	244	13	765	1	0.95	1	1	0.97
getITD	238	19	766	0	0.93	1	1	0.96
FLT3_ITD_ext	254	3	766	0	0.99	1	1	0.99
km	243	14	753	13	0.96	0.98	0.95	0.95
GID	231	26	760	6	0.90	0.99	0.97	0.93
FiLT3r	257	0	766	0	1	1	1	1

$$Se = \frac{VP}{VP+FN} \; Sp = \frac{VN}{VN+FP} \; VPP = \frac{VP}{VP+FP} \; F_1 = 2 \cdot \frac{VPP \cdot Se}{VPP+Se}$$

3.2.2.2 Quantification

La quantification fait référence à l'estimation de l'AR de *FLT3*-ITD avec l'algorithme testé par rapport à celle de l'analyse de fragments. L'AR quantifié par *pindel*, *FLT3_ITD_ext* et *FiLT3r* présentait une forte corrélation avec l'estimation de l'analyse de fragments, contrairement aux autres algorithmes (Tableau 8, Figure 34).

Les AR sont considérés comme aberrants lorsque leur écart avec l'AR estimé à l'aide de la méthode de référence excède un log. A nouveau, *pindel*, *FLT3_ITD_ext* et *FiLT3r* ont obtenu les meilleurs résultats, présentant respectivement 3, 2 et 1 AR aberrants. Il faut également noter que la seule ITD largement sous-estimée par FiLT3r est également sous-estimée par l'ensemble des algorithmes.

Tableau 8 : Tableau agrégé de la quantification par les différents algorithmes

	r	Δ 1log
pindel	0.86	3
ITDSeek	0.36	82
ScanITD	0.65	53
getITD	0.74	10
FLT3_ITD_ext	0.86	2
km	0.65	34
GID	0.73	13
FiLT3r	0.87	1

r correspond au coefficient de corrélation de Pearson des AR log-transformés entre l'analyse de fragments et l'algorithme testé

Δ 1log correspond au nombre de patients pour lesquels on observe un écart d'un log entre les deux techniques

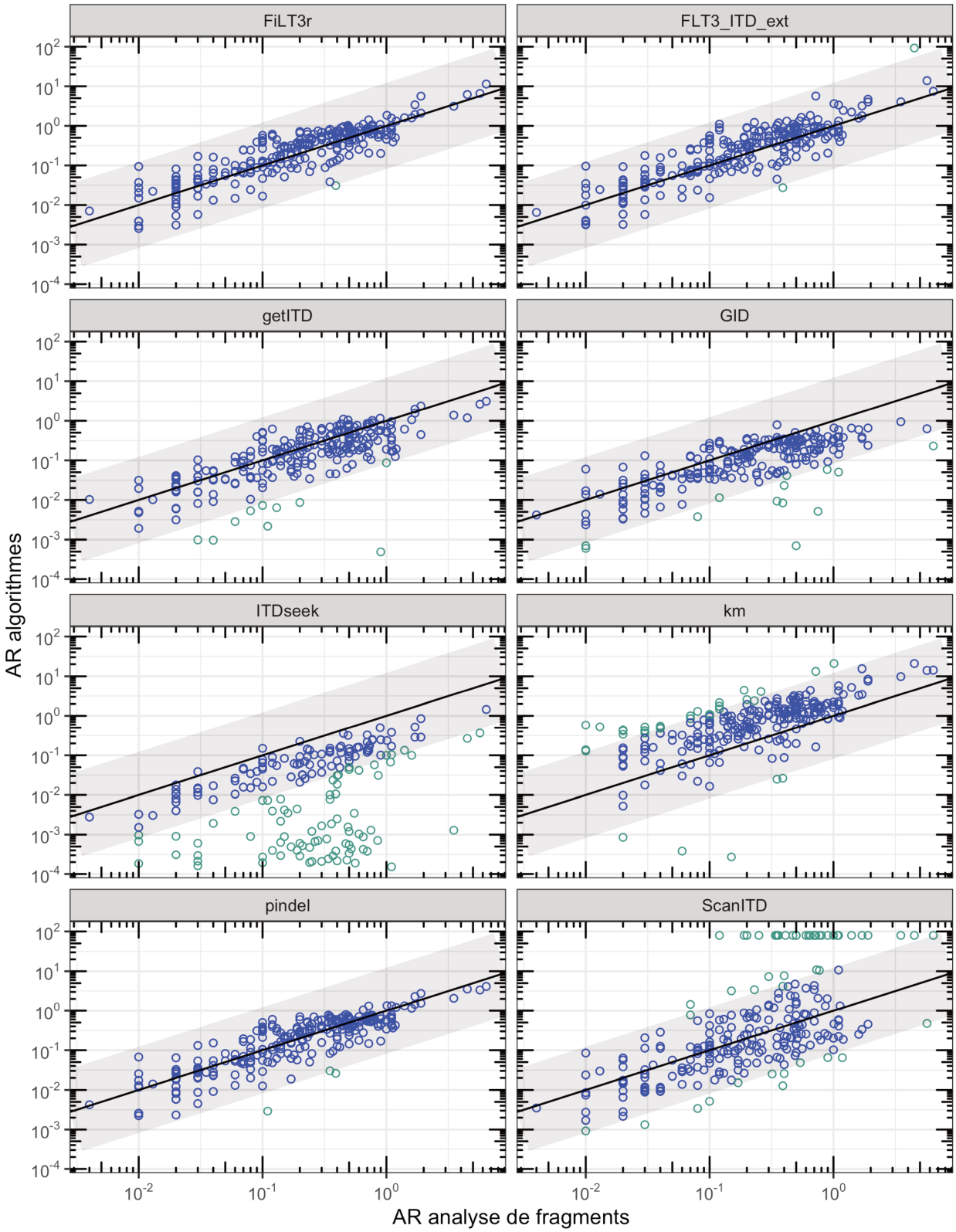

Figure 34 : AR log-transformé des duplications détectées par l'ensemble des algorithmes par rapport à l'analyse de fragments

La ligne droite grise correspond à $y = x$ et idéalement les points devraient être alignés le long de cette ligne. La zone grise correspond à un écart de moins d'un log par rapport à la ligne $y = x$, les points verts correspondent aux ITDs dépassant cet écart.

3.2.2.3 Pouvoir discriminant

Le pouvoir discriminant correspond à la proportion de patients pour lesquels il existe une adéquation entre la méthode de référence et l'algorithme au regard du seuil de 0.5 d'AR défini par l'ELN (38). Toutes les ITDs détectées par un algorithme sont ramenées à l'échelle du patient, et en cas d'ITDs multiples la somme des AR est utilisée. Un patient est considéré comme bien classé lorsque les AR calculés avec la méthode de référence ainsi qu'avec l'algorithme testé sont < 0.5 ou ≥ 0.5. L'incertitude de mesure pour l'analyse de fragments ainsi que pour la technique TWIST® a été prise en compte ($U_{FA} = 0.17$ et $U_{TWIST} = 0.11$).

pindel, *FLT3_ITD_ext* et *FiLT3r* possédaient le meilleur pouvoir discriminant : respectivement de 95%, 93% et 96% (Figure 35).

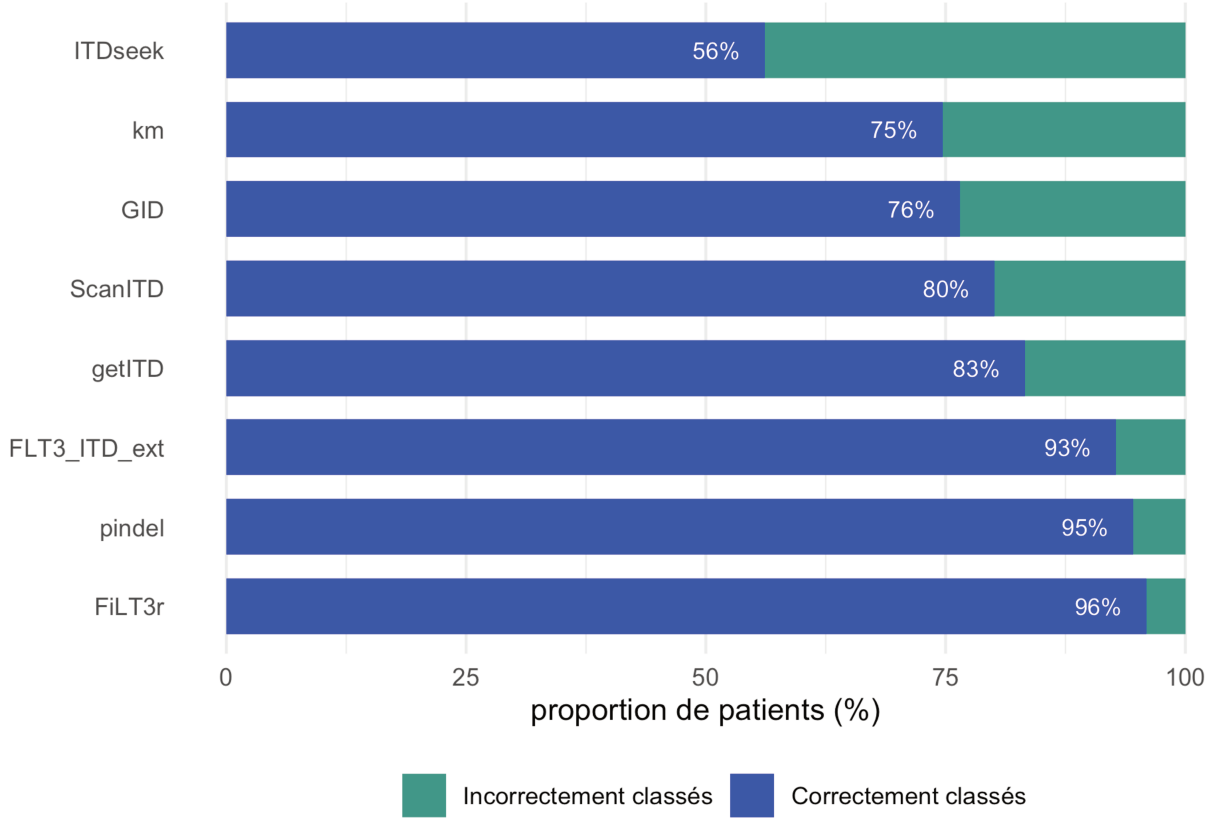

Figure 35 : Pouvoir discriminant des différents algorithmes par rapport à la méthode de référence (seuil de l'ELN < 0.5)

4 Discussion

Avec le développement du séquençage haut débit, un certain nombre d'outils bio-informatiques ont été mis au point pour détecter *FLT3*-ITD, permettant de surmonter efficacement les limites de l'analyse par fragments (112,113,121,127–129). Dans l'ensemble, des progrès significatifs ont été réalisés dans la détection et la quantification de *FLT3*-ITD par NGS (121). En raison de leur impact majeur sur le pronostic (68–70,76,78) ainsi que sur la prise en charge thérapeutique du patient (39,46,85,87,87,90,91,93,130), la détection et la quantification de *FLT3*-ITD doivent être effectuées par une technique aux performances analytiques parfaitement maîtrisées.

À notre connaissance, aucune comparaison et évaluation systématique de ces algorithmes n'a encore été rapportée sur une large cohorte de patients. En effet, la plupart des articles portant sur le développement de ces outils utilisent des données simulées et un nombre très restreint de données de séquençage réelles (114–116,121).

Ce travail a donc permis une étude comparative des différents algorithmes disponibles actuellement sur des données réelles.

pindel fut l'un des premiers algorithmes utilisés pour la détection de *FLT3*-ITD (112). Bien que ses résultats soient acceptables ($F_1 = 0.99$ et $r = 0.86$), il totalise 6 ITDs non retrouvées rendant son utilisation limitée en routine. Cinq de ces ITDs étaient caractérisées par une taille supérieure à 2 écarts-type de la moyenne. La dernière correspondait à une duplication hors modèle, cette catégorie comprenant les duplications portant des indels, et les duplications dispersées. *pindel* ne sous-estime que très légèrement l'AR avec une pente de régression linéaire estimée à 0.95. Ces résultats sont cohérents avec ce qui est rapporté dans la littérature, à savoir une sous-estimation quasi systématique de l'AR (127,128) et quelques faux négatifs (114,115,121).

Dans l'article traitant de l'utilisation de *ITDseek* (114), *Chun Hang Au et al.* ont principalement effectué des tests sur des données simulées (longueur de l'ITD : 1–201 pb ; VAF 50 % ; read théorique 2 × 275 bp), ne prenant ainsi pas en compte les duplications hors modèle. *ITDSeek* a été également testé sur une cohorte de 50 patients dont 11 étaient mutés *FLT3*-ITD. Cet ensemble n'est donc pas entièrement représentatif du large éventail d'ITDs que l'on peut rencontrer et pourrait expliquer ses résultats ($F_1 = 0.7$, $r = 0.36$) sur nos données. L'absence de similitude entre les faux négatifs (position génomique, longueur, séquence, AR) et le manque d'informations complémentaires malgré les sollicitations auprès des auteurs ne nous ont pas permis d'approfondir les résultats. Les données de la littérature font également part de résultats similaires pour *ITDSeek* (115,121). Contrairement à ce qui a été réalisé dans notre étude, *ITDSeek* a été testé sur des données de séquençage obtenues après enrichissement par PCR. Cette variation de technique pourrait également expliquer

ces résultats. Il s'agit d'une des principales limites de notre travail, puisqu'aucun test n'a été effectué sur des données issues d'un enrichissement par PCR.

ScanITD repose sur le même principe que *ITDseek* avec quelques améliorations permettant ainsi l'obtention de meilleurs résultats que son homologue (F_1 = 0.97 et r = 0.65). Les 13 ITDs non retrouvées ont toutes une longueur inférieure à 50pb. Ces résultats concordent bien avec la publication princeps de *ScanITD* (115), qui est décrit comme plus performant dans la détection des duplications de moyenne et de grande taille (50-300 pb). En sachant que plus de la moitié des ITDs ont une taille inférieure à 50pb (Figure 28, Figure 32), il n'est donc pas un outil de choix dans la détection et la quantification de *FLT3*-ITD.

GID totalise 26 ITDs non retrouvées, il n'est donc pas compatible avec une utilisation en routine. Sept d'entre elles correspondent à des duplications hors modèles, les 19 restantes ne possédaient pas de similitude pouvant expliquer leur non détection. De façon similaire à *ITDSeek*, ces résultats peuvent s'expliquer car il a été principalement évalué sur des données simulées ainsi que sur 20 patients issus de la cohorte TCGA possédant des ITDs courtes (< 51 pb), plus aisées à détecter car incluses intégralement dans le read.

getITD obtient un coefficient de corrélation satisfaisant (r = 0.74), mais il ne semble pas adapté à notre pratique pour plusieurs raisons. Il totalise 18 faux négatifs ; 10 s'expliquant par son incapacité à détecter des duplications inférieures à 6pb ou d'une taille supérieure à la longueur du read moins 6bp (144bp avec la technique TWIST®) (131) ; les 8 ITDs restantes ne possèdent pas de similitude entre elles pouvant expliquer l'absence de détection. *getITD* a été conçu pour la quantification MRD (minimal residual disease) par séquençage ciblé uniquement sur le gène *FLT3* et par conséquent une très grande profondeur de séquençage d'environ 2 millions de séquences est nécessaire (116). Dans notre étude, *FLT3* est séquencé avec un panel de gènes, ce qui entraîne *per se* une diminution de la profondeur avec en moyenne 2000X sur les exons 14-15 de *FLT3*. La variation de technique d'amplification ainsi que la profondeur faible pourraient altérer le fonctionnement optimal de *getITD*.

Les résultats de *km* (118) ne sont pas satisfaisants concernant la détection de *FLT3*-ITD (F_1 = 0.95 et r = 0.65). Les ITDs non retrouvées ne présentent pas de similitude pouvant expliquer l'absence de détection par *km*. L'ensemble des 15 faux positifs ont été également retrouvés par *FLT3_ITD_ext* et *FiLT3r* mais à des ratios inférieurs à 2%. Ce manque de spécificité est dû à un problème de quantification de l'allèle sauvage, le comptage de l'allèle muté étant proche de *FLT3_ITD_ext* et *FiLT3r* pout ces ITDs.

En avril 2021, Yuan et al. (121) ont réalisé une étude comparative des différents outils disponibles pour détecter et quantifier *FLT3*-ITD par NGS. Dans cet article, les auteurs ont qualifié *FLT3_ITD_ext* (117) comme le meilleur outil disponible actuellement. Dans notre travail, il obtient également des résultats très satisfaisants compatibles avec la détection de

FLT3-ITD en routine ($F_1 = 0.99$ et $r = 0.86$). Deux parmi ces 3 faux négatifs correspondent à des duplications hors modèle, les ITDs étant associées à des délétions. *FLT3_ITD_ext* ne prend pas en compte les délétions, ce qui constitue sa principale limitation. Dans le script, un filtre est appliqué pour toute taille < 0 pb, conduisant à l'estimation d'une taille erronée par rapport à l'analyse de fragments, en compilant la délétion et la duplication.

Ce travail a permis la mise au point de *FiLT3r* un nouvel algorithme pour la détection et la quantification de *FLT3*-ITD. En comparaison à l'ensemble des algorithmes testés, *FiLT3r* obtient les meilleurs résultats ($F_1 = 1$ et $r = 0.87$). Ces résultats supérieurs ne sont pas dûs à un effet d'over-fitting, l'utilisation d'un jeu de données indépendant pour le développement et pour le paramétrage ayant permis de supprimer cet effet. Il obtient également le meilleur pouvoir discriminant avec 96% (Figure 35) des patients correctement classés. FiLT3r apparaît donc comme l'algorithme de choix pour la détection de *FLT3*-ITD en routine et pourrait à terme remplacer la méthode par analyse de fragments et ainsi pallier à ses limites.

Le manque de résolution de l'analyse de fragments (précision à 1pb) peut induire des erreurs au niveau de l'estimation de la taille de l'ITD. Dans le jeu de données test, l'ensemble des algorithmes de détection à l'exception de *FLT3*-ITD-ext ont identifié deux patients présentant des ITDs non multiples de 3 (173 et 175 pb). Ces 2 ITDs avaient été détectées à une taille de 174 pb (multiple de 3). Cette discordance pourrait s'expliquer par l'approximation de la longueur des ITDs avec l'analyse de fragments. Un séquençage Sanger sur ADN a donc été effectué confirmant leur taille non multiple de trois. Pourtant un décalage du cadre de lecture entraine généralement la synthèse d'une protéine non fonctionnelle. Ceci n'a pas encore été rapporté dans la littérature concernant *FLT3*-ITD, qui est constamment associée à un gain de fonction. Néanmoins, après une analyse approfondie, ces duplications se situent dans l'intron séparant l'exon 14 et l'exon 15 et pourraient avoir un effet sur l'épissage. Afin de prédire cet effet, l'algorithme Neural Network Splice (NNSplice (132,133)) a été utilisé. Son principe repose sur une fenêtre glissante parcourant la séquence afin de détecter les sites donneur et accepteur d'épissage. L'algorithme prédit pour les 2 patients que l'ITD crée un site donneur et un site accepteur et forme ainsi un potentiel exon additionnel de 81 pb entre l'exon 14 et 15 (gain de fonction potentiel). A l'instar de nombreux outils de bio-informatique, NNSplice fournit seulement une prédiction. Une confirmation par séquençage sanger sur ARN permettrait de prouver cette hypothèse. Toutefois cet exon de taille 81 bp est concordant avec une ITD respectant le cadre de lecture.

L'analyse de fragments ne permet pas de déterminer la position exacte du site d'insertion ni la séquence de l'ITD. Récemment, certains auteurs ont pourtant rapporté que ces deux paramètres pourraient avoir un impact majeur sur le pronostic du patient.

Schwartz et al. ont travaillé sur les caractéristiques structurelles des séquences d'ITDs et leur impact sur la réponse au traitement chez les patients atteints de LAM *FLT3*-ITD mutées. Les auteurs ont catégorisé les ITDs en classe "typique" ou "atypique" en fonction de

leur composition nucléotidique. La séquence d'une ITD typique correspond à une duplication parfaite de l'originale, tandis qu'une ITD atypique correspond à une séquence différente (duplication hors modèle). Leur analyse révèle que la réponse au traitement par FLT3i est corrélée avec la classification. Chez les patients traités par FLT3i être porteur d'une ITD typique réduisait le risque de 66,3 % par rapport à aux porteur d'ITD atypique(HR 0,337 ; CI 0,165-0,689), ce qui suggère que les patients typiques avaient une meilleure survie globale. Néanmoins l'étude présentait plusieurs biais (i) l'effectif réduit de la cohorte (n = 68) ; (ii) l'absence d'étude *in vivo* concernant la conséquence fonctionnelle d'une ITD atypique ; (iii) l'AR et le site d'insertion n'étaient pas pris en compte dans leur modèle multivarié. Aux dates d'inclusion des différentes cohortes utilisées dans notre travail (2008-2019), trop peu de patients étaient sous FLT3i ce qui ne nous a pas permis pas d'inférer sur cette théorie.

En avril 2021, *Rücker et al.* ont démontré l'impact pronostique et prédictif du site d'insertion de l'ITD chez 452 patients mutés *FLT3*-ITD randomisés dans le cadre de l'essai RATIFY (NCT00651261). Les patients étaient répartis dans 3 catégories selon la position des sites d'insertion de leurs ITDs : le groupe JMD si les sites étaient uniquement situés dans le domaine du gène codant pour la partie juxtamembranaire du récepteur, le groupe TKD1 si les sites étaient uniquement situés dans le domaine du gène codant pour le domaine tyrosine kinase du récepteur, et le groupe JMD/TKD1 si les sites étaient situés dans ces deux domaines. La survie globale différait significativement, avec une probabilité de survie globale à 4 ans estimée à 0,44, 0,50 et 0,30 pour JMD, JMD/TKD1 et TKD1, respectivement. Ces résultats confirment l'hétérogénéité moléculaire de *FLT3*-ITD et l'impact pronostic négatif lorsqu'elles concernent uniquement le domaine TKD1 dans la LAM (134). Cette découverte pourrait faire également l'objet d'études concernant le site d'insertion pour les futurs patients du BIG afin de corroborer les résultats de la cohorte RATIFY.

Outre l'intérêt pronostique de ces deux paramètres, leur utilisation conjointe pourrait permettre d'évaluer l'hétérogénéité clonale de *FLT3*-ITD. En effet, un pic en analyse de fragments pourrait correspondre en réalité à des clones différents. Dans notre travail, nous avons fait le choix d'agréger les ITDs de même longueur afin d'avoir un AR le plus représentatif de l'analyse de fragments. Mais il pourrait être intéressant de suivre ces clones individuellement afin d'apprécier l'évolution clonale. *K. Schranz, et al.* ont mis en évidence que les patients présentant une hétérogénéité clonale de *FLT3*-ITD avaient une survie globale et une survie sans rechute significativement diminuées (128).

L'utilisation de *FiLT3r* pour la détection de *FLT3*-ITD par NGS rendrait possible le multiplexage avec d'autres gènes impliquées dans les LAM. En raison des implications pronostiques et thérapeutiques de certaines altérations génétiques, il est impératif de procéder, au moment du diagnostic, à une analyse moléculaire complète des mutations ayant un impact pronostique et thérapeutique. Outre la caractérisation initiale de la maladie, le multiplexage pourrait également inclure une exploration des mutations de résistance aux FLT3i. En effet, *Schmalbrock et al.* ont exploré les mécanismes de résistance à la

midostaurine en étudiant 54 couples diagnostiques rechutes/réfractaires dans le cadre de l'essai RATIFY. Les auteurs ont mis en évidence chez les patients réfractaires, un nombre plus élevé de mutations persistantes et ont acquis moins de nouvelles mutations par rapport aux patients en rechute. Cette observation est conforme à l'hypothèse selon laquelle les mutations de résistance chez les patients réfractaires sont déjà présentes au moment du diagnostic, alors que chez les patients en rechute, l'échec du traitement est dû à la sélection et à la croissance des clones résistants et à l'acquisition de nouveaux clones (135). Cet article souligne l'importance d'un screening large dès le diagnostic et sa répétition chez les patients réfractaire et en rechute afin d'évaluer le changement du paysage mutationnel. Ainsi, la disponibilité d'une plateforme de séquençage rapide pour la recherche de l'ensemble des mutations impliquées dans les LAM, la caractérisation de *FLT3*-ITD (séquence et site d'insertion) et la recherche d'éventuelles mutations de résistance aux FLT3i présenterait un avantage significatif dans la prise en charge du patient.

L'analyse de fragments a une limite de détection élevée, les données de la littérature l'estimant entre 1 et 3% (109,110). Lorsque le ratio est faible, (AR $< 3\%$) il est parfois difficile de distinguer les pics "vrais" des artefacts. L'analyse des données NGS à l'aide de *FiLT3r* est basée sur le dénombrement de reads, et non sur l'interprétation de pics, permettant ainsi une détection plus basse et plus précise. Dans notre travail, nous avons appliqué le même seuil que pour l'analyse de fragments. Cependant, *FiLT3r* était capable de détecter des ITDs $< 1\%$ chez des patients considérés comme négatifs par la méthode de référence. Cette problématique pourrait faire l'objet d'un travail futur dont le propos serait de déterminer la pertinence de traiter les patients présentant une duplication *FLT3*-ITD à très faible ratio. L'hypothèse principale serait qu'il existe un risque accru de rechute chez ces patients non traités par FLT3i, imputable à la persistance du clone minoritaire *FLT3*-ITD muté sous traitement par chimiothérapie conventionnelle.

La détection de *FLT3*-ITD par *FiLT3r* pourrait également ouvrir la voie à la MRD. Les taux de réponse initiale à une chimiothérapie intensive sont assez élevés chez les patients atteints de LAM à *FLT3*-ITD (136). Néanmoins, de nombreux patients finissent par rechuter même après la greffe (79). La disponibilité d'un test sensible et spécifique permettant d'évaluer la MRD pour *FLT3*-ITD, à utiliser chez les patients en rémission clinique, présenterait un avantage significatif pour les cliniciens. Ce test pourrait guider les transplantations et le traitement d'entretien par inhibiteurs de FLT3 après une chimiothérapie intensive ou une transplantation (87,137,138). Le traitement d'entretien par inhibiteurs n'est pas encore clairement positionné dans la LAM *FLT3*-ITD positive après une allo-HSCT (137). En outre, la capacité à déterminer de manière sensible et précise la charge d'allèles *FLT3* mutés pourrait être utilisée afin de mieux caractériser les réponses aux inhibiteurs de *FLT3* et participer à définir des recommandations. Toutefois, un travail important reste à effectuer concernant la détermination de la limite de détection et de quantification dans le contexte de MRD.

Nous avons procédé à un examen complet des principes, des fonctionnalités et des limites des algorithmes existants pour détecter *FLT3*-ITD par NGS. Dans ce travail nous avons mis au point l'algorithme *FiLT3r* sur un jeu de données d'entrainement. Nous l'avons comparé ainsi que l'ensemble des algorithmes disponibles à la méthode de référence (analyse de fragments), sur plusieurs cohortes de patients indépendantes. L'utilisation de FiLT3r représente une approche prometteuse pour la détection sensible, robuste et informative des mutations *FLT3*-ITD par NGS.

5 Annexes

Manuscrit soumis au journal *Bioinformatics* le 20/09/2021.

Frugal alignment-free identification of *FLT3*-internal tandem duplications with FiLT3r

Augustin Boudry[†3], Sasha Darmon[†1,2], Nicolas Duployez[3,4], Martin Figeac[5], Sandrine Geffroy[3], Maxime Bucci[3], Karine Celli-Lebras[6], Matthieu Duchmann[7], Romane Jourdinaud[3], Laurène Fenwarth[3,4], Olivier Nibourel[3], Laure Goursaud[3], Raphael Itzykson[7,6], Hervé Dombret[6], Mathilde Hunault[8], Claude Preudhomme[4,3], and Mikaël Salson[*1]

[1]CRIStAL, Université de Lille, CNRS, Lille F-59000 France
[2]ENS Lyon, France
[3]Laboratory of Hematology, CHU LILLE, Lille, France
[4]U1277, Inserm, Lille, France
[5]Univ. Lille, CNRS, Inserm, CHU Lille, Institut Pasteur de Lille, US 41 - UMS 2014 - PLBS, F-59000 Lille, France
[6]Department of Hematology, Saint Louis Hospital, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France
[7]INSERM/CNRS UMR 944/7212, Saint-Louis Research Institute, Paris Diderot University, Paris, France
[8]Department of Hematology, CHU Angers, Angers, France

[†] These authors contributed equally to this work.
[*] Correspondence: Mikaël Salson <mikael.salson@univ-lille.fr>

Abstract

Internal tandem duplications in the *FLT3* gene, termed *FLT3*-ITDs, are useful molecular markers in acute myeloid leukemia for patient risk stratification and follow-up. *FLT3*-ITDs are increasingly screened through high-throughput sequencing raising the need for robust and efficient algorithms.

We developed a new alignment-free algorithm that aimed to identify and quantify *FLT3*-ITDs in high-throughput sequencing data. The algorithm focuses on the k-mers from the reads that also belong to the exons 14-15 of the *FLT3*-ITD gene. We show that those k-mers bring enough information to accurately identify the *FLT3*-ITD length and to quantify those duplications.

The algorithm is implemented in a free software called FiLT3r. We compare FiLT3r with state-of-the-art alternatives to the fragment analysis gold standard. We show on a reproducible benchmark from a cohort of 185 patients sequenced on Illumina that FiLT3r is more precise (no false positive nor false negative) and is both quicker and more memory-efficient.

Introduction

Monitoring the disease progression in blood cancers requires the identification of pathology-specific molecular markers. In acute myeloid leukemia (AML), one of the markers used is an internal tandem duplication in the *FLT3* gene (*FLT3*-ITD), which occurs in 20-30% of cases (Cao *et al.* (2019); Pratz and Levis (2010); Small (2006); Thiede *et al.* (2002)). *FLT3*-ITD are in-frame duplications of highly variable size, ranging from 3 to more than 400 nucleotides, mostly located within the receptor's autoinhibitory juxatamembrane domain. This mutation represents a strong prognostic biomarker since patients with *FLT3*-ITD are at higher risk of relapse and have decreased event-free and overall survival (Kottaridis *et al.* (2001)). The importance of allelic ratio (AR), as assessed by the ratio between the mutated allele and the wild-type allele, has been demonstrated in several studies (Schlenk *et al.* (2014); Thiede *et al.* (2002)). Besides its prognostic value, *FLT3*-ITD has also a major therapeutic impact as its presence may lead to allogeneic hematopoietic stem cell transplantation (allo-HSCT) (Döhner *et al.* (2017); Hunter and Chen (2020); Schlenk *et al.* (2014)). *FLT3*-ITD is also a therapeutic target with the emergence of *FLT3* inhibitors in the induction and consolidation therapy (Daver *et al.* (2019); Stone *et al.* (2017)) and more recently as maintenance therapy (Xuan *et al.* (2020)).

Historically, according to European LeukemiaNet (ELN) guidelines, the identification and quantification of *FLT3*-ITD was performed with fragment analysis (Döhner *et al.* (2017)). DNA fragments were fluorescently labeled, then separated by capillary electrophoresis (Lyu *et al.* (2020); Murphy *et al.* (2003)). One or more peaks were obtained depending on the presence or absence of ITD(s). The size of the ITD was determined by subtracting the size of the wild type fragment from that of the mutated fragment, using a scale. The AR was evaluated by dividing the area under the curve of the mutated fragment's peak by that of the wild-type fragment's peak. Although robust, this technique has several limitations: (i) the lower limit of quantification of the AR is high, generally

at least 1% (Kim *et al.* (2015); Sakaguchi *et al.* (2020)), (ii) the determination of the size of the ITD using the scale is approximate, (iii) the exact position of insertion and complete sequence of the ITD are not available (Schranz *et al.* (2018), Rücker *et al.* (2021)) and (iv) sample multiplexing cannot be performed. Development of high-throughput sequencing enabled the detection of several markers of interest in a single sequencing run and has the potential to overcome several of those limitations. Initially the data produced were analysed using pindel, a general-purpose algorithm that detects and quantifies indels and structural variants (Ye *et al.* (2009)). Although using a general-purpose software can be appealing, it may be under-optimised to address the specific problem of identifying and quantifying *FLT3*-ITD. Thus many methods were specifically developed to detect *FLT3*-ITD in high-throughput sequencing data (Yuan *et al.* (2021)). Yuan and colleagues distinguished the assembly-based methods from the alignment-based ones, and demonstrated better accuracy of alignment-based methods.

However alignment-free methods have gained importance in bioinformatics (Zielezinski *et al.* (2017)) as they usually have the advantage of using only a fraction of the resources required by alignment methods while providing similar results. km is an example of an alignment-free strategy for *FLT3*-ITD detection (Audemard *et al.* (2019)). It is based on Jellyfish (Marçais and Kingsford (2011)), a k-mer counting algorithm, to efficiently count k-mers in the reads. Then km first builds a linear k-mer graph of the reference seuqence and traverses it using their count table. Any divergent path that can be identified with the count table is a potential duplication. We introduce another alignment-free approach based on k-mers, but contrarily to km our approach doesn't require to count the k-mers from the reads. This allows it to be quicker. For the detection of duplications we were inspired by the methodology used in the work of Philippe *et al.* (2013) by analysing the occurrences of the k-mers in the reads.

Methods

We introduce our heuristic aimed at finding duplications compared to a reference sequence though it can more generally detect insertions and deletions. In what follows what is described for duplication also holds for indels without loss of generality.

To determine if a read contains a duplication we first need to consider all k-mers from the read and identify their positions in the reference sequence. The main principle consists of identifying any read in which k-mer positions on the reference sequence would suddenly go back. Such an event would correspond to a duplication with respect to the reference sequence. To further explain this principle, here is a concrete example of its application. Let T be the reference sequence and R a read. Figure 1 illustrates an example where AGAT is duplicated in R. When querying at which positions the k-mers of R occur in T, the positions of the k-mers gradually increase from 1 to 3 and then go back to 2 at the start of the duplication. This phenomenon is the signature of a duplication. We will use this observation to detect duplications in an efficient way. As will be shown later, this information is also sufficient to determine the length of the duplication.

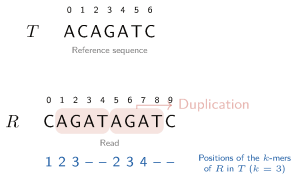

Fig. 1. Identifying duplications from a reference sequence T in a read R with k-mers ($k = 3$). The bottom line represents the positions of the k-mers from R in T. For instance the first element (1) corresponds to the k-mer CAG which appears in position 1 in T. The dashes (–) correspond to k-mers from R that do not occur in T (eg. ATA or TAG).

The FiLT3r algorithm

Our algorithm can be summarised in three main steps:

1. Indexing k-mers of the reference sequence with their original positions.
2. Traversing all the reads: keep the ones with enough k-mers from the reference sequence and determine the position of the read's k-mers in the reference sequence.
3. Detect a duplication event in reads using the k-mers positions in the reference sequence.

The indexing step is trivial as the reference sequence (basically exons 14 to 15 in the *FLT3* gene) is very short. With short values of k, there may exists several occurrences of the same k-mer in the reference sequence (see A in figure 2). Thus, all positions are stored to prevent loss of information. Ambiguities in the positions will be resolved afterwards.

The second step only consists of reading the reads one by one and then querying the index (a hash table) with the k-mers of each read (see C in figure 2). The main difficulty comes from the third step. Figure 1 introduces a simplified version of the problem but other cases may arise that should be handled correctly to prevent false negatives or false positives as will be shown later.

Duplications are only searched in reads whose number of k-mers from the reference sequence is above a given threshold to prevent spending time on reads not coming from the gene of interest. For each of these reads we will now focus on the k-mers occurrences in the reference sequence. We call *trace of R in T*, denoted by t_R, the list of k-mers positions in T from a read R. Then $t_R[i]$ is a list of positions where the k-mer $R[i \dots i+k-1]$ occurs in T (see *traces of kept reads* in figure 2 for examples). This list can obviously be empty whenever the k-mer doesn't occur in T. We call a break (as in Philippe *et al.* (2013)) the maximal span of positions where $t_R[i]$ is empty (denoted by $t_R[i] = ()$). Thus, a break is identified by its starting position and its ending position. For a read R if we have a break (j_1, j_2), then $p_R[j] = ()$, for each $j_1 \leq j \leq j_2$, and either $j_1 = 0$ or $t_R[j_1 - 1] \neq ()$, and either $j_2 = |R| - 1$ or $t_R[j_2 + 1] \neq ()$.

Whenever a duplication occurs in a read R, this will create a break in the trace. Note that the converse is not true: many other events can also create breaks. Actually, any difference

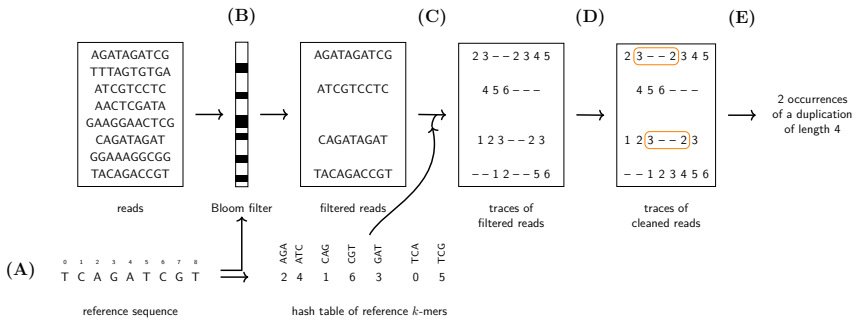

Fig. 2. How FiLT3r processes the reads. (A) The k-mers of the reference sequence are indexed in a hash table and a Bloom filter. The keys of the hash table are k-mers and the values are a list of positions where the key occurs in the reference sequence. (B) The reads matching a sufficient number of k-mers with the reference are kept. (C) The k-mer positions in the reference are considered. (D) Substitutions are removed. (E) Duplications, and more generally indels, are called only using the positions of the k-mers.

between the read and the reference sequence will lead to a break. Also, in some rare cases the duplication will not create a break. However the probability of this phenomenon decreases exponentially with k.

Using the trace of the reads we start by identifying all the breaks in a trace and by determining if the positions before and after the break are compatible with a duplication. Thus, assuming for now that $t_R[j_1-1]$ and $t_R[j_2+1]$ both correspond to a single position in the reference sequence, we check whether $t_R[j_1-1]+g-1 > t_R[j_2+1]$, where g is the size of the break. In such a case, a duplication has been detected (see E in figure 2).

The length of the duplication can be deduced from the k-mer positions. Let (j_1, j_2) be a break corresponding to a duplication. $t_R[j_1-1]+k-1$ is the last nucleotide position before the duplication while $t_R[j_2+1]$ is the first nucleotide in the duplication. Thus the length of the duplication is $t_R[j_1-1]+k-1-t_R[j_2+1]+1 = t_R[j_1-1]-t_R[j_2+1]+k$. For instance, in figure 1, the break is $(3,4)$ and the length of the duplication is $3-2+3=4$. However the formula does not hold whenever short insertions occur at the breakpoint or when there is an overlap between the start of the duplication and what follows it in the reference sequence. All those cases can actually be easily dealt with. Each inserted nucleotide will lead to a break length increased by one. Conversely each overlap will lead to a decreased gap length. We note that when no such insertion or overlap occurs, we have $j_2-j_1=k$. Hence we deduce that more generally the duplication size is $t_R[j_1-1]-t_R[j_2+1]+j_2-j_1$. The formula also holds when substitutions occur at $d < k$ nucleotides from the breakpoint. In this case, this will make the break longer by d nucleotides, but this will obviously decrease $t_R[j_1-1]$ by d or increase $t_R[j_2+1]$ by d (depending where the substitution occurs). Thus the computed duplication size won't change compared to the situation where no such substitution occurs. Thus our heuristic is robust to such events.

We are finally able to identify duplications within a read in linear time. In order to limit the false positives, we introduce several strategies to mitigate this risk.

Mitigating false positives

False positive detection of duplication can be encountered because k-mers are short sequences that can occur by chance because of sequencing errors, or other events. We must not detect a duplication whenever a (few) short k-mers occur in the wrong place.

Disambiguating positions

In the read trace, some k-mers may occur at several locations in the reference sequence. Of course only one (at most) is correct, thus only one is kept. Assuming we want to disambiguate $t_R[i]$ (*ie.* $|t_R[i]| > 1$), we follow those rules:

1. First, if $t_R[i-1]$ is non-ambiguous and $t_R[i-1]+1$ appears in $t_R[i]$, then we retain this value.
2. Otherwise, among the possible positions in $t_R[i]$ we choose the one which minimises the distances (*ie.* the asolute values of the subtractions) with the neighbouring positions in t_R (both the original one and the one currently cleaned). The neighbouring positions considered are all the ones that don't imply crossing a break, or a break can be crossed only if position i is just before or just after a break.

Eliminating substitutions

Substitutions are the most frequent mutations and sequencing errors (at least for Illumina sequencers, see Laehnemann *et al.* (2016)). Therefore many breaks will be due to substitutions. To limit the number of breaks to consider, and to have traces with more valid positions, we remove substitutions by correcting them (either mutations or sequencing errors). This also allows to detect duplications in spite of proximate substitutions and it will improve the estimation of the quantification.

The correction is carried out as follows: let a break (j_1, j_2), we extract the k-mer starting at position j_1. Since this k-mer doesn't exist in T, as position j_1 is the first one of a break, we try the three other nucleotides at position $k-1$ in this k-mer. If one of the corrected k-mer occurs at a position p such that $p-1$

occurs in $t_R[j_1 - 1]$, then we correct the k-mer this way and we move on to the next k-mer. In case the correction did not work, the k-mers in the break are not corrected.

Enlarging k-mers

Before reporting a duplication event, we make sure this event is robust by checking that δ consecutive positions in the trace are consistent. Thus if we have a candidate break (j_1, j_2), we report this as an event if and only if $t_R[j_1 - 1 - i] + i = t_R[j_1 - 1]$, for $1 \leq i \leq \delta$, and conversely iff $t_R[j_2 + 1 + i] - i = t_R[j_2 + 1]$, for $1 \leq i \leq \delta$.

Estimating duplication abundance

Identifying a duplication is not sufficient as its abundance is also a prognosis factor. Rather than a raw abundance, a relative abundance (expressed as a VAF (Variant Allele Frequency) or a MT/WT ratio (ratio between the number of duplications and the number of wildtypes observations)) is much more meaningful.

To do so, we use the disambiguated traces and we count the coverage at each position of the reference sequence. This corresponds to the total coverage. More formally when $t_R[i] = j$, we increment the coverage at position j in the reference sequence. We deduce the wildtype coverage by removing all the coverages from the duplications that have been detected.

However with our approach we cannot detect duplications that would occur at the start or at the end of a read. As explained previously, for a duplication to be detected we need to have δ positions in the trace before and after the break. For a duplication creating a break of size b, we will need to have $t = 2\delta + b + k - 1$ nucleotides in the read. That means that we will not be able to detect duplications occurring in the first $t - 1$ or the last $t - 1$ nucleotides of the reads. Thus when quantifying a duplication, the quantification will be under-estimated because $\frac{2(t-1)}{R}$ of the duplications could not be counted, where R is the size of the reads. The counts are thus corrected to take this observation into account, and by assuming that any indel has a uniform probability of occurrence within the read. Let q be the raw quantification, *ie.* the number of reads in which the duplication was identified, the corrected quantification is $q' = q\left(1 + \frac{2(t-1)}{R}\right)$. The VAF or MT/WT ratios are then computed using the corrected counts and the wildtype coverage at the position of the duplication.

Optimising data processing

As specified previously we only search duplications in reads having enough k-mers coming from the reference sequence. However in most cases most of the reads would not come from the region of interest. Thus the most time consuming step in our algorithm is to determine whether the read is coming from that region. To lower this time consumption we use a heuristic: k-mers from the region of interest are stored in a Bloom filter (Bloom (1970)).

The first step is therefore to query the Bloom filter to determine how many k-mers come from the region of interest in the read. If this value is above a threshold then the read goes to the following stages of the algorithm, otherwise it is discarded. By default this threshold is set at 30 %.

Bloom filters do not yield any false negative, therefore we have the guarantee that this heuristic will not prevent us from identifying a read that actually comes from the region of interest. As the region of interest, namely exons 14 to 15 from the *FLT3* gene, is quite short the Bloom filter can be very small and will likely be stored in the CPU cache, which will allow very quick accesses.

Implementation and benchmarking

FiLT3r is implemented in C++ using the GATB library (Drezen *et al.* (2014)) and is freely distributed (`https://gitlab.univ-lille.fr/filt3r/filt3r`).

We will compare FiLT3r with state-of-the-art approaches. We made a preliminary assessment on our cohort with several software (see Supp Table 1). However with the later publication of Yuan *et al.* (2021) we prefered referring to their independent assessment and chose for this paper the best tool they assessed: FLT3-ITD-ext (Tsai *et al.* (2020)). We also added km (Audemard *et al.* (2019)) as it was not evaluated in Yuan et al's benchmark and showed good performances in our preliminary assessment. km is an alignment-free approach already presented in the introduction. We also add that km is not focused on *FLT3* and can more genrally detect variations in raw reads. *FLT3*-ITD-ext relies on a general-purpose aligner (BWA-MEM, Li (2013)) to align reads to the reference sequence (the *FLT3* genomic sequence of exons 14-15).

FiLT3r was launched with a k-mer size of 12, km with a k-mer size of 31 (as advised by km authors) and with a threshold for duplication detections of .01 (thresholds .1 and .001 were attempted but the best results were achieved with a threshold of 0.01), Jellyfish parameters were identical to the ones used in km publication (`-C -L 2 -Q+ -s 1000000`). FLT3-ITD-ext was launched with default parameters. The experiments can be reproduced with a Snakefile available on the Git repository of the software.

All the experiments were performed on a server with 24 Intel® Xeon® CPU E5-2420 and 193GB of RAM. The data was stored on a NAS connected to the server through NFS. The programs were run on a single thread. User times and peak memory consumptions were measured.

Benchmarking data

We assessed the three software on high-throughput sequencing data of 185 patients aged from 18 to 60 years old, diagnosed with de novo or secondary AML at the hematology laboratory of the University Hospital of Lille. This population was composed of 114 *FLT3*-ITD positive patients and 71 *FLT3*-ITD negative patients, according to the gold-standard fragment analysis (Boissel *et al.* (2002)). The local ethics committee approved the study and all patients signed an informed consent at diagnosis (CHRU de Lille, Tumorothèque du C2RC, approval nos. CSTMT289).

Exons 14-15 of *FLT3* were sequenced in all patients using the following methodology: library was prepared by capture method targeting 62 genes with SureSelect-QXT (AGILENT®) according to the manufacturer's protocol. Samples were sequenced with NextSeq Illumina® ($2×$121bp). The region of interest was sequenced with a mean depth of 2116x (range: 421.32 – 4538.84). The base calling was performed with

(a)

	FiLT3r	Km	FLT3-ITD-ext
True pos.	145	123	144
False neg.	0	22	1
False pos.	12	5	8

(b)

	FiLT3r	Km	FLT3-ITD-ext
True pos.	147	125	144
False neg.	0	22	3
False pos.	0	0	0
Precision	1	1	1
Recall	1	0.85	0.98
F1	1	0.92	0.99

Table 1. (a) Raw results of the three software assessed on the samples of 185 patients. Any result from a software is considered as a false positive as long as its quantification is at least 1% and it is not detected by the reference method. (b) Corrected results, after taking into account two Sanger sequencing to verify the two *FLT3*-ITDs detected by two software, and by considering as false positive only *FLT3*-ITDs detected above 1% by a software in negative samples.

`bcl2fastq2` (v:2.20.0) and fastq were trimmed with `fastp` (v:0.20.0). The number of paired-end reads per sample ranged from 1.8 million to 35 million (median: 4.8 million). The sequences were deposited on SRA (see supplementary file 3).

Results

The software results are compared to what was identified with the gold-standard method (DNA fragment analysis). Any duplication found by a software, at whatever abundance, and identified with the reference method is considered as a true positive. Any duplication found by a software with a ratio mutated/wild type (MT/WT) above .01 and that was not identified by the reference method is for now considered as a false positive. After analysing the results in details, we will change this definition of false positivity as it is too stringent.

The raw results are shown in table 1 (a). FiLT3r and FLT3-ITD-ext show similar performances. FiLT3r was slightly better in terms of true positives, but at the apparent cost of more false positives. This higher number of false positives is mainly due to the fact that we did not apply any filter to FiLT3r's results, to prevent any kind of over-fitting. In the details, 3 false positives are due to short single nucleotide deletions that could easily be filtered out; 2 others are detected by both FiLT3r and km and were further confirmed by Sanger sequencing making them false negatives of the reference method and FLT3-ITD-ext (see supplementary figures 3 and 4); 4 others are also seen either by km or FLT3-ITD-ext; and finally the last 3 appear to be duplications specific to FiLT3r (see supplementary file 2). Interestingly two of them are detected in SRR15006540 and their cumulative lengths (16 and 56) correspond to the length of a duplication that was detected by the conventional method (72). We thus believe that those two duplications are not false positives and are a consequence of our more fine-grained method that is able to distinguish multiple duplications. In supplementary figure 1 we show a dotplot of a read from SRR15006540 against our reference sequence where two duplications appear.

We also argue that the events identified as *false positives* are not real false positives. Apart from one of the two Sanger-confirmed duplications, all the other false positives (for all the software) occur in positive samples where other duplications were detected at higher ratios, apart from the two duplications in SRR15006540 detailed previously. With the conventional method there was no reason to dig into rarer duplications when more abundant ones existed. This is not the case anymore with high-throughput sequencing. Thus we consider that our initial definition of false positive is not well suited and we instead switch to the following, more strict definition: a false positive corresponds to any duplication detected at the threshold of .01 in a negative sample (apart from the sample SRR15006372, initially considered as negative, where the duplication was later confirmed by Sanger sequencing). No software detected a duplication above that threshold among the 71 negative samples. Hence the most important metric in this context is the false negativity. We show a corrected version of the results in table 1 (b) where false positivity is thus at 0 and where the two Sanger-confirmed sequences are moved from the false positives to the true positives for FiLT3r and km (and added to the false negatives for FLT3-ITD-ext). On those 185 samples, FiLT3r displays perfect results as it did not report any false positive nor false negative result.

Figure 3 shows the quantification computed by the three software for the duplications found. Their quantifications are compared to the one found by the reference method. With FiLT3r and FLT3-ITD-ext the quantification is closer to the fragment analysis quantification (respectively $r = .93$ and $r = .88$ for the log-transformed quantifications), compared to km ($r = .75$). It should also be noted that one of the quantification FiLT3r underestimates corresponds to the 72nt duplication we previously detailed (which FiLT3r also detects as two duplications of 46nt and 26nt). Interestingly, the other duplication FiLT3r largely underestimates (in SRR15006376) is also underestimated by a factor of 10 by FLT3-ITD-Ext and was not detected by km.

Regarding time and space consumption, FiLT3r shows the best performances (see Figures 4 and 5). We also show the time and memory usage for counting k-mers, with Jellyfish, which is a required step for km.

In Figure 4 we see that km at thresholds 0.01 is blazingly fast as it systematically took less than 10 seconds to detect the duplications. However this doesn't include Jellyfish time, which must be launched on each sample. While Jellyfish is a very efficient k-mer counter it was 2-3 orders of magnitude slower than km itself. FiLT3r's median user time (65 s) is 6 times quicker than Jellyfish's and 9 times quicker than FLT3-ITD-Ext's. In some cases FLT3-ITD-Ext can be very time consuming, taking several hours while FiLT3r never took more than 10 minutes. The progression margin for FiLT3r is quite low as `gunzip` on the same files took about a third of the time took by FiLT3r (see supplementary figure 2).

For FiLT3r it appears that using a Bloom filter dramatically speeds up the software. We launched FiLT3r by deactivating the usage of the Bloom filter and FiLT3r was an order of magnitude longer in that case (see supplementary figure 3).

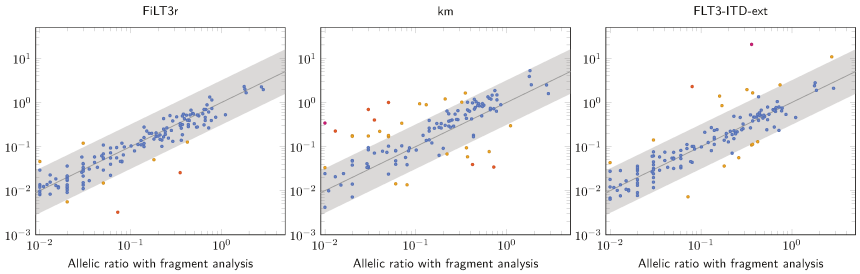

Fig. 3. Quantification of duplications found with FiLT3r, km (threshold .01) and FLT3-ITD-ext compared with the fragment analysis method. The gray straight line corresponds to $y = x$ and ideally the dots should be aligned along that line. The gray area is centered around that line and its width is of one log. The hotter the color of the dots, the higher the quantification error. Only 8 results are not within this gray area for FiLT3r, 27 for km and 16 for FLT3-ITD-ext.

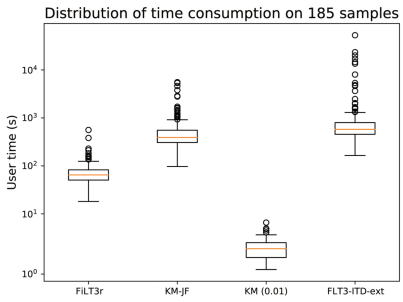

Fig. 4. Time consumption of FiLT3r, km and FLT3-ITD-Ext on the 185 samples analysed. KM-JF is the time taken by Jellyfish (preliminary step required for km), km (0.01) is the time taken by km with the corresponding threshold.

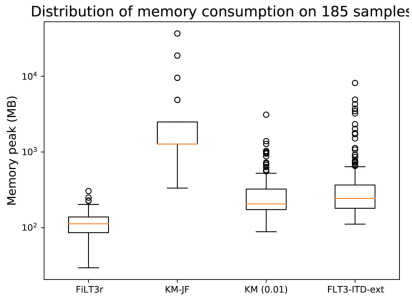

Fig. 5. Memory consumption of FiLT3r, km and FLT3-ITD-Ext on the 185 samples analysed. KM-JF is the memory taken by Jellyfish (preliminary step required for km), km (0.01) is the memory taken by km with the corresponding threshold.

Memory consumption is very limited in FiLT3r as we store a Bloom filter and a hash table of the reference sequence. We also store reads that match the reference sequence in main memory, but there are generally just a few thousands of them. FLT3-ITD-Ext median memory consumption is close to FiLT3r's as it is twice as space consuming. However in some cases the space consumption can reach several GB while FiLT3r used at most 300 MB. Jellyfish memory usage was systematically above that highest memory usage recorded for FiLT3r. We could still improve FiLT3r memory usage by storing on the disk the reads matching the reference sequence. This would however lead to a time penalty.

Conclusions

We introduced FiLT3r, a time and memory efficient algorithm implemented in an open-source C++ software, which shows very good detection and quantification performances compared to state-of-the-art software. FiLT3r demonstrated no false negatives, which was not the case of the other software. Even the reference fragment analysis method, exhibited two false negatives as shown by further Sanger sequencing of the samples. Similarly to the other software, FiLT3r did not detect any duplication in the negative samples. FiLT3r performances are better than that of FLT3-ITD-ext, the best software identified so far, which is alignment-based. This illustrates that alignment-free algorithms can be more efficient than alignment-based algorithms while using a fraction of the resources they need. Beyond the *FLT3*-ITD analysis, FiLT3r can also be used to detect duplications in any gene as soon as the reference sequence is known.

Acknowledgments

We are grateful to É. Audemard for valuable discussions on their publications and software. We are also grateful to the bioinformatics service in CHU Lille for their helpful advices.

References

Audemard,E.O. *et al.* (2019) Targeted variant detection using unaligned RNA-Seq reads. *Life Science Alliance*, **2**.

Bloom,B.H. (1970) Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, **13**, 422–426.

Boissel,N. *et al.* (2002) Prognostic significance of FLT3 internal tandem repeat in patients with de novo acute myeloid leukemia treated with reinforced courses of chemotherapy. *Leukemia*, **16**, 1699–1704.

Cao,T. *et al.* (2019) The flt3-itd mutation and the expression of its downstream signaling intermediates stat5 and pim-1 are positively correlated with cxcr4 expression in patients with acute myeloid leukemia. *Scientific reports*, **9**, 1–10.

Daver,N. *et al.* (2019) Targeting FLT3 mutations in AML: Review of current knowledge and evidence. *Leukemia*, **33**, 299–312.

Döhner,H. *et al.* (2017) Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood, The Journal of the American Society of Hematology*, **129**, 424–447.

Drezen,E. *et al.* (2014) GATB: Genome assembly & analysis tool box. *Bioinformatics*, **30**, 2959–2961.

Hunter,B.D. and Chen,Y.-B. (2020) Current approaches to transplantation for FLT3-ITD AML. *Current hematologic malignancy reports*, **15**, 1–8.

Kim,Y. *et al.* (2015) Quantitative fragment analysis of FLT3-ITD efficiently identifying poor prognostic group with high mutant allele burden or long ITD length. *Blood cancer journal*, **5**, e336–e336.

Kottaridis,P.D. *et al.* (2001) The presence of a FLT3 internal tandem duplication in patients with acute myeloid leukemia (AML) adds important prognostic information to cytogenetic risk group and response to the first cycle of chemotherapy: Analysis of 854 patients from the United Kingdom Medical Research Council AML 10 and 12 trials. *Blood, The Journal of the American Society of Hematology*, **98**, 1752–1759.

Laehnemann,D. *et al.* (2016) Denoising dna deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in bioinformatics*, **17**, 154–179.

Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*.

Lyu,M. *et al.* (2020) The prognosis predictive value of FMS-like tyrosine kinase 3-internal tandem duplications mutant allelic ratio (FLT3-ITD MR) in patients with acute myeloid leukemia detected by GeneScan. *Gene*, **726**, 144195.

Marçais,G. and Kingsford,C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, **27**, 764–770.

Murphy,K.M. *et al.* (2003) Detection of FLT3 internal tandem duplication and D835 mutations by a multiplex polymerase chain reaction and capillary electrophoresis assay. *The Journal of molecular diagnostics*, **5**, 96–102.

Philippe,N. *et al.* (2013) CRAC: An integrated approach to the analysis of RNA-seq reads. *Genome biology*, **14**, 1–16.

Pratz,K.W. and Levis,M.J. (2010) Bench to bedside targeting of FLT3 in acute leukemia. *Current drug targets*, **11**, 781–789.

Rücker,F.G. *et al.* (2021) Molecular landscape and prognostic impact of flt3-itd insertion site in acute myeloid leukemia: RATIFY study results. *Leukemia*, 1–10.

Sakaguchi,M. *et al.* (2020) The sensitivity of the FLT3-ITD detection method is an important consideration when diagnosing acute myeloid leukemia. *Leukemia research reports*, **13**, 100198.

Schlenk,R.F. *et al.* (2014) Differential impact of allelic ratio and insertion site in FLT3-ITD–positive AML with respect to allogeneic transplantation. *Blood, The Journal of the American Society of Hematology*, **124**, 3441–3449.

Schranz,K. *et al.* (2018) Clonal heterogeneity of FLT3-ITD detected by high-throughput amplicon sequencing correlates with adverse prognosis in acute myeloid leukemia. *Oncotarget*, **9**, 30128.

Small,D. (2006) FLT3 mutations: Biology and treatment. *ASH Education Program Book*, **2006**, 178–184.

Stone,R.M. *et al.* (2017) Midostaurin plus chemotherapy for acute myeloid leukemia with a FLT3 mutation. *New England Journal of Medicine*, **377**, 454–464.

Thiede,C. *et al.* (2002) Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: Association with FAB subtypes and identification of subgroups with poor prognosis. *Blood, The Journal of the American Society of Hematology*, **99**, 4326–4335.

Tsai,H.K. *et al.* (2020) Targeted Informatics for Optimal Detection, Characterization, and Quantification of FLT3 Internal Tandem Duplications Across Multiple Next-Generation Sequencing Platforms. *The Journal of Molecular Diagnostics*, **22**, 1162–1178.

Xuan,L. *et al.* (2020) Sorafenib maintenance in patients with FLT3-ITD acute myeloid leukaemia undergoing allogeneic haematopoietic stem-cell transplantation: An open-label, multicentre, randomised phase 3 trial. *The Lancet Oncology*, **21**, 1201–1212.

Ye,K. *et al.* (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

Yuan,D. *et al.* (2021) Comprehensive review and evaluation of computational methods for identifying FLT3-internal tandem duplication in acute myeloid leukaemia. *Briefings in Bioinformatics*.

Zielezinski,A. *et al.* (2017) Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biology*, **18**, 186.

6 Références

1. Shallis RM, Wang R, Davidoff A, Ma X, Zeidan AM. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. Blood Reviews. 1 juill 2019;36:70-87.

2. Song X, Peng Y, Wang X, Chen Y, Jin L, Yang T, et al. Incidence, Survival, and Risk Factors for Adults with Acute Myeloid Leukemia Not Otherwise Specified and Acute Myeloid Leukemia with Recurrent Genetic Abnormalities: Analysis of the Surveillance, Epidemiology, and End Results (SEER) Database, 2001-2013. Acta Haematol. 2018;139(2):115-27.

3. Marceau-Renaut A, Duployez N, Ducourneau B, Labopin M, Petit A, Rousseau A, et al. Molecular Profiling Defines Distinct Prognostic Subgroups in Childhood AML: A Report From the French ELAM02 Study Group. HemaSphere. févr 2018;2(1):e31.

4. Ostgård LSG, Nørgaard JM, Severinsen MT, Sengeløv H, Friis L, Jensen MK, et al. Data quality in the Danish National Acute Leukemia Registry: a hematological data resource. Clin Epidemiol. 2013;5:335-44.

5. Smith A, Howell D, Patmore R, Jack A, Roman E. Incidence of haematological malignancy by sub-type: a report from the Haematological Malignancy Research Network. Br J Cancer. 22 nov 2011;105(11):1684-92.

6. Linet MS, Gilbert ES, Vermeulen R, Dores GM, Yin S-N, Portengen L, et al. Benzene Exposure Response and Risk of Myeloid Neoplasms in Chinese Workers: A Multicenter Case-Cohort Study. J Natl Cancer Inst. 01 2019;111(5):465-74.

7. Stenehjem JS, Kjærheim K, Bråtveit M, Samuelsen SO, Barone-Adesi F, Rothman N, et al. Benzene exposure and risk of lymphohaematopoietic cancers in 25 000 offshore oil industry workers. Br J Cancer. 28 avr 2015;112(9):1603-12.

8. Rinsky RA, Smith AB, Hornung R, Filloon TG, Young RJ, Okun AH, et al. Benzene and leukemia. An epidemiologic risk assessment. N Engl J Med. 23 avr 1987;316(17):1044-50.

9. Johnson GT, Harbison SC, McCluskey JD, Harbison RD. Characterization of cancer risk from airborne benzene exposure. Regul Toxicol Pharmacol. déc 2009;55(3):361-6.

10. Pyatt D. Benzene and hematopoietic malignancies. Clin Occup Environ Med. août 2004;4(3):529-55, vii.

11. Tsushima H, Iwanaga M, Miyazaki Y. Late effect of atomic bomb radiation on myeloid disorders: leukemia and myelodysplastic syndromes. Int J Hematol. mars 2012;95(3):232-8.

12. Preston DL, Kusumi S, Tomonaga M, Izumi S, Ron E, Kuramoto A, et al. Cancer incidence in atomic bomb survivors. Part III. Leukemia, lymphoma and multiple myeloma, 1950-1987. Radiat Res. févr 1994;137(2 Suppl):S68-97.

13. Kuznetsova IS, Labutina EV, Hunter N. Radiation Risks of Leukemia, Lymphoma and Multiple Myeloma Incidence in the Mayak Cohort: 1948-2004. PLoS One. 2016;11(9):e0162710.

14. Ugai T, Matsuo K, Sawada N, Iwasaki M, Yamaji T, Shimazu T, et al. Smoking and subsequent risk of leukemia in Japan: The Japan Public Health Center-based Prospective Study. J Epidemiol. juill 2017;27(7):305-10.

15. Lichtman MA. Cigarette smoking, cytogenetic abnormalities, and acute myelogenous leukemia. Leukemia. juin 2007;21(6):1137-40.

16. Rund D, Ben-Yehuda D. Therapy-related leukemia and myelodysplasia: evolving concepts of pathogenesis and treatment. Hematology. juin 2004;9(3):179-87.

17. Graubert T. Therapy-related MDS: Models and Genetics. Biol Blood Marrow Transplant. janv 2010;16(1 Suppl):S45-7.

18. Yeasmin S, Nakayama K, Ishibashi M, Oride A, Katagiri A, Purwana IN, et al. Therapy-related myelodysplasia and acute myeloid leukemia following paclitaxel- and carboplatin-based chemotherapy in an ovarian cancer patient: a case report and literature review. Int J Gynecol Cancer. déc 2008;18(6):1371-6.

19. Brown CA, Youlden DR, Aitken JF, Moore AS. Therapy-related acute myeloid leukemia following treatment for cancer in childhood: A population-based registry study. Pediatr Blood Cancer. 2018;65(12):e27410.

20. Goncalves I, Burbury K, Michael M, Iravani A, Ravi Kumar AS, Akhurst T, et al. Characteristics and outcomes of therapy-related myeloid neoplasms after peptide receptor radionuclide/chemoradionuclide therapy (PRRT/PRCRT) for metastatic neuroendocrine neoplasia: a single-institution series. Eur J Nucl Med Mol Imaging. août 2019;46(9):1902-10.

21. Andreotti G, Koutros S, Hofmann JN, Sandler DP, Lubin JH, Lynch CF, et al. Glyphosate Use and Cancer Incidence in the Agricultural Health Study. J Natl Cancer Inst. 01 2018;110(5):509-16.

22. Lichtman MA. Obesity and the risk for a hematological malignancy: leukemia, lymphoma, or myeloma. Oncologist. 2010;15(10):1083-101.

23. Lazarov D, Waldron HA, Pejin D. Acute myeloid leukaemia and exposure to organic solvents--a case-control study. Eur J Epidemiol. mars 2000;16(3):295-301.

24. Poynter JN, Richardson M, Roesler M, Blair CK, Hirsch B, Nguyen P, et al. Chemical exposures and risk of acute myeloid leukemia and myelodysplastic syndromes in a population-based study. Int J Cancer. 1 janv 2017;140(1):23-33.

25. Albin M, Björk J, Welinder H, Tinnerberg H, Mauritzson N, Johansson B, et al. Acute myeloid leukemia and clonal chromosome aberrations in relation to past exposure to organic solvents. Scand J Work Environ Health. déc 2000;26(6):482-91.

26. Soulier J. Fanconi anemia. Hematology Am Soc Hematol Educ Program. 2011;2011:492-7.

27. Alter BP, Giri N, Savage SA, Rosenberg PS. Cancer in dyskeratosis congenita. Blood. 25 juin 2009;113(26):6549-57.

28. Montalban-Bravo G, Garcia-Manero G. Myelodysplastic syndromes: 2018 update on diagnosis, risk-stratification and management. Am J Hematol. 2018;93(1):129-47.

29. Rampal R, Ahn J, Abdel-Wahab O, Nahas M, Wang K, Lipson D, et al. Genomic and functional analysis of leukemic transformation of myeloproliferative neoplasms. Proc Natl Acad Sci U S A. 16 déc 2014;111(50):E5401-5410.

30. DeBoer R, Garrahy I, Rettew A, Libera R. Transformation of CMML to AML presenting with acute kidney injury. J Community Hosp Intern Med Perspect. 2 août 2020;10(4):353-7.

31. Mathew RA, Bennett JM, Liu JJ, Komrokji RS, Lancet JE, Naghashpour M, et al. Cutaneous manifestations in CMML: Indication of disease acceleration or transformation to AML and review of the literature. Leukemia Research. 1 janv 2012;36(1):72-80.

32. Röllig C, Ehninger G. How I treat hyperleukocytosis in acute myeloid leukemia. Blood. 21 mai 2015;125(21):3246-52.

33. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. Br J Haematol. août 1976;33(4):451-8.

34. Bennett JM, Catovsky D, Daniel MT, Flandrin G, Galton DA, Gralnick HR, et al. Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French-American-British Cooperative Group. Ann Intern Med. oct 1985;103(4):620-5.

35. Hong M, He G. 2016 Revision to the WHO Classification of Acute Myeloid Leukemia. J Transl Int Med. 30 juin 2017;5(2):69-71.

36. Arber DA, Stein AS, Carter NH, Ikle D, Forman SJ, Slovak ML. Prognostic impact of acute myeloid leukemia classification. Importance of detection of recurring cytogenetic abnormalities and multilineage dysplasia on survival. Am J Clin Pathol. mai 2003;119(5):672-80.

37. Reikvam H, Hatfield KJ, Kittang AO, Hovland R, Bruserud Ø. Acute myeloid leukemia with the t(8;21) translocation: clinical consequences and biological implications. J Biomed Biotechnol. 2011;2011:104631.

38. Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. Blood. 26 2017;129(4):424-47.

39. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 19 2016;127(20):2391-405.

40. Vardiman JW, Thiele J, Arber DA, Brunning RD, Borowitz MJ, Porwit A, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. Blood. 30 juill 2009;114(5):937-51.

41. Martelli MP, Sportoletti P, Tiacci E, Martelli MF, Falini B. Mutational landscape of AML with normal cytogenetics: biological and clinical implications. Blood Rev. janv 2013;27(1):13-22.

42. Cancer Genome Atlas Research Network, Ley TJ, Miller C, Ding L, Raphael BJ, Mungall AJ, et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 30 2013;368(22):2059-74.

43. Welch JS, Ley TJ, Link DC, Miller CA, Larson DE, Koboldt DC, et al. The origin and evolution of mutations in Acute Myeloid Leukemia. Cell. 20 juill 2012;150(2):264-78.

44. Metzeler KH, Group on behalf of the AS, Herold T, Group on behalf of the AS, Rothenberg-Thurley M, Group on behalf of the AS, et al. Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. Blood. 4 août 2016;128(5):686-98.

45. Vardiman JW, Harris NL, Brunning RD. The World Health Organization (WHO) classification of the myeloid neoplasms. Blood. 1 oct 2002;100(7):2292-302.

46. Stone RM, Mandrekar SJ, Sanford BL, Laumann K, Geyer S, Bloomfield CD, et al. Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. New England Journal of Medicine. 3 août 2017;377(5):454-64.

47. Daver N, Schlenk RF, Russell NH, Levis MJ. Targeting FLT3 mutations in AML: review of current knowledge and evidence. Leukemia. 2019;33(2):299-312.

48. Cao T, Jiang N, Liao H, Shuai X, Su J, Zheng Q. The FLT3-ITD mutation and the expression of its downstream signaling intermediates STAT5 and Pim-1 are positively correlated with CXCR4 expression in patients with acute myeloid leukemia. Scientific Reports. 21 août 2019;9(1):12209.

49. Pratz KW, Levis MJ. Bench to Bedside Targeting of FLT3 in Acute Leukemia. Curr Drug Targets. juill 2010;11(7):781-9.

50. Kazi JU, Rönnstrand L. FMS-like Tyrosine Kinase 3/FLT3: From Basic Science to Clinical Implications. Physiol Rev. 1 juill 2019;99(3):1433-66.

51. Wu M, Li C, Zhu X. FLT3 inhibitors in acute myeloid leukemia. J Hematol Oncol. 04 2018;11(1):133.

52. Rosnet O, Bühring HJ, Marchetto S, Rappold I, Lavagna C, Sainty D, et al. Human FLT3/FLK2 receptor tyrosine kinase is expressed at the surface of normal and malignant hematopoietic cells. Leukemia. févr 1996;10(2):238-48.

53. Karsunky H, Merad M, Cozzio A, Weissman IL, Manz MG. Flt3 Ligand Regulates Dendritic Cell Development from Flt3+ Lymphoid and Myeloid-committed Progenitors to Flt3+ Dendritic Cells In Vivo. J Exp Med. 21 juill 2003;198(2):305-13.

54. Lyman SD, Jacobsen SEW. c-kit Ligand and Flt3 Ligand: Stem/Progenitor Cell Factors With Overlapping Yet Distinct Activities. Blood. 15 févr 1998;91(4):1101-34.

55. Horiuchi K, Morioka H, Takaishi H, Akiyama H, Blobel CP, Toyama Y. Ectodomain shedding of FLT3 ligand is mediated by TACE. J Immunol. 15 juin 2009;182(12):7408-14.

56. Lyman SD. Biology of flt3 ligand and receptor. Int J Hematol. août 1995;62(2):63-73.

57. Solanilla A, Grosset C, Lemercier C, Dupouy M, Mahon FX, Schweitzer K, et al. Expression of Flt3-ligand by the endothelial cell. Leukemia. janv 2000;14(1):153-62.

58. Knapper S. FLT3 inhibition in acute myeloid leukaemia. Br J Haematol. sept 2007;138(6):687-99.

59. Rusten LS, Lyman SD, Veiby OP, Jacobsen SE. The FLT3 ligand is a direct and potent stimulator of the growth of primitive and committed human CD34+ bone marrow progenitor cells in vitro. Blood. 15 févr 1996;87(4):1317-25.

60. Gilliland DG, Griffin JD. The roles of FLT3 in hematopoiesis and leukemia. Blood. 1 sept 2002;100(5):1532-42.

61. Hubbard SR. Theme and variations: juxtamembrane regulation of receptor protein kinases. Mol Cell. sept 2001;8(3):481-2.

62. Griffith J, Black J, Faerman C, Swenson L, Wynn M, Lu F, et al. The structural basis for autoinhibition of FLT3 by the juxtamembrane domain. Mol Cell. 30 janv 2004;13(2):169-78.

63. Breitenbuecher F, Schnittger S, Grundler R, Markova B, Carius B, Brecht A, et al. Identification of a novel type of ITD mutations located in nonjuxtamembrane domains of the FLT3 tyrosine kinase receptor. Blood. 23 avr 2009;113(17):4074-7.

64. Nakao M, Yokota S, Iwai T, Kaneko H, Horiike S, Kashima K, et al. Internal tandem duplication of the flt3 gene found in acute myeloid leukemia. Leukemia. déc 1996;10(12):1911-8.

65. Borrow J, Dyer SA, Akiki S, Griffiths MJ. Terminal deoxynucleotidyl transferase promotes acute myeloid leukemia by priming FLT3-ITD replication slippage. Blood. 19 déc 2019;134(25):2281-90.

66. Kelly LM, Liu Q, Kutok JL, Williams IR, Boulton CL, Gilliland DG. FLT3 internal tandem duplication mutations associated with human acute myeloid leukemias induce myeloproliferative disease in a murine bone marrow transplant model. Blood. 1 janv 2002;99(1):310-8.

67. Gale RE, Party on behalf of the MRCALW, Green C, Party on behalf of the MRCALW, Allen C, Party on behalf of the MRCALW, et al. The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. Blood. 1 mars 2008;111(5):2776-84.

68. Kottaridis PD, Gale RE, Frew ME, Harrison G, Langabeer SE, Belton AA, et al. The presence of a FLT3 internal tandem duplication in patients with acute myeloid leukemia (AML) adds important prognostic information to cytogenetic risk group and response to the first cycle of chemotherapy: analysis of 854 patients from the United Kingdom Medical Research Council AML 10 and 12 trials. Blood. 15 sept 2001;98(6):1752-9.

69. Thiede C, Steudel C, Mohr B, Schaich M, Schäkel U, Platzbecker U, et al. Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. Blood. 15 juin 2002;99(12):4326-35.

70. Schlenk RF, Group for the G-AAS, Kayser S, Group for the G-AAS, Bullinger L, Group for the G-AAS, et al. Differential impact of allelic ratio and insertion site in FLT3-ITD–positive AML with respect to allogeneic transplantation. Blood. 27 nov 2014;124(23):3441-9.

71. Stirewalt DL, Kopecky KJ, Meshinchi S, Engel JH, Pogosova-Agadjanyan EL, Linsley J, et al. Size of FLT3 internal tandem duplication has prognostic significance in patients with acute myeloid leukemia. Blood. 1 mai 2006;107(9):3724-6.

72. Kayser S, Schlenk RF, Londono MC, Breitenbuecher F, Wittke K, Du J, et al. Insertion of FLT3 internal tandem duplication in the tyrosine kinase domain-1 is associated with resistance to chemotherapy and inferior outcome. Blood. 17 sept 2009;114(12):2386-92.

73. Schwartz GW, Manning B, Zhou Y, Velu P, Bigdeli A, Astles R, et al. Classes of ITD Predict Outcomes in AML Patients Treated with FLT3 Inhibitors. Clin Cancer Res. 15 janv 2019;25(2):573-83.

74. Gallipoli P, Huntly BJP. Prognostic Models Turn the Heat(IT)up on FLT3ITD - Mutated AML. Clin Cancer Res. 15 janv 2019;25(2):460-2.

75. Ponziani V, Gianfaldoni G, Mannelli F, Leoni F, Ciolli S, Guglielmelli P, et al. The size of duplication does not add to the prognostic significance of FLT3 internal tandem duplication in acute myeloid leukemia patients. Leukemia. nov 2006;20(11):2074-6.

76. Warren M, Luthra R, Yin CC, Ravandi F, Cortes JE, Kantarjian HM, et al. Clinical impact of change of FLT3 mutation status in acute myeloid leukemia patients. Mod Pathol. oct 2012;25(10):1405-12.

77. Schlenk RF, Frech P, Weber D, Brossart P, Horst H-A, Kraemer D, et al. Impact of pretreatment characteristics and salvage strategy on outcome in patients with relapsed acute myeloid leukemia. Leukemia. mai 2017;31(5):1217-20.

78. Wattad M, Weber D, Döhner K, Krauter J, Gaidzik VI, Paschka P, et al. Impact of salvage regimens on response and overall survival in acute myeloid leukemia with induction failure. Leukemia. 2017;31(6):1306-13.

79. Hunter BD, Chen Y-B. Current Approaches to Transplantation for FLT3-ITD AML. Curr Hematol Malig Rep. févr 2020;15(1):1-8.

80. Pratcorona M, Brunet S, Nomdedéu J, Ribera JM, Tormo M, Duarte R, et al. Favorable outcome of patients with acute myeloid leukemia harboring a low-allelic burden FLT3-ITD mutation and concomitant NPM1 mutation: relevance to post-remission therapy. Blood. 4 avr 2013;121(14):2734-8.

81. Ho AD, Schetelig J, Bochtler T, Schaich M, Schäfer-Eckart K, Hänel M, et al. Allogeneic Stem Cell Transplantation Improves Survival in Patients with Acute Myeloid Leukemia Characterized by a High Allelic Ratio of Mutant FLT3-ITD. Biol Blood Marrow Transplant. mars 2016;22(3):462-9.

82. Sakaguchi M, Yamaguchi H, Najima Y, Usuki K, Ueki T, Oh I, et al. Prognostic impact of low allelic ratio FLT3-ITD and NPM1 mutation in acute myeloid leukemia. Blood Adv. 19 oct 2018;2(20):2744-54.

83. Straube J, Ling VY, Hill GR, Lane SW. The impact of age, NPM1mut, and FLT3ITD allelic ratio in patients with acute myeloid leukemia. Blood. 8 mars 2018;131(10):1148-53.

84. Dombret H, Gardin C. An update of current treatments for adult acute myeloid leukemia. Blood. 7 janv 2016;127(1):53-61.

85. Perl AE, Martinelli G, Cortes JE, Neubauer A, Berman E, Paolini S, et al. Gilteritinib or Chemotherapy for Relapsed or Refractory FLT3-Mutated AML. N Engl J Med. 31 2019;381(18):1728-40.

86. Dhillon S. Gilteritinib: First Global Approval. Drugs. févr 2019;79(3):331-9.

87. Xuan L, Wang Y, Huang F, Fan Z, Xu Y, Sun J, et al. Sorafenib maintenance in patients with FLT3-ITD acute myeloid leukaemia undergoing allogeneic haematopoietic stem-cell transplantation: an open-label, multicentre, randomised phase 3 trial. The Lancet Oncology. 1 sept 2020;21(9):1201-12.

88. Kindler T, Lipka DB, Fischer T. FLT3 as a therapeutic target in AML: still challenging after all these years. Blood. 9 déc 2010;116(24):5089-102.

89. Scholl S, Fleischmann M, Schnetzke U, Heidel FH. Molecular Mechanisms of Resistance to FLT3 Inhibitors in Acute Myeloid Leukemia: Ongoing Challenges and Future Treatments. Cells. 17 nov 2020;9(11):2493.

90. Tarlock K, Alonzo TA, Gerbing RB, Raimondi SC, Hirsch BA, Sung L, et al. Gemtuzumab Ozogamicin Reduces Relapse Risk in FLT3/ITD Acute Myeloid Leukemia: A Report from the Children's Oncology Group. Clin Cancer Res. 15 avr 2016;22(8):1951-7.

91. Godwin CD, Gale RP, Walter RB. Gemtuzumab ozogamicin in acute myeloid leukemia. Leukemia. sept 2017;31(9):1855-68.

92. Fenwarth L, Fournier E, Cheok M, Boyer T, Gonzales F, Castaigne S, et al. Biomarkers of Gemtuzumab Ozogamicin Response for Acute Myeloid Leukemia Treatment. Int J Mol Sci. 6 août 2020;21(16):E5626.

93. Gottardi M, Simonetti G, Sperotto A, Nappi D, Ghelli Luserna di Rorà A, Padella A, et al. Therapeutic Targeting of Acute Myeloid Leukemia by Gemtuzumab Ozogamicin. Cancers (Basel). 11 sept 2021;13(18):4566.

94. Walter RB, Appelbaum FR, Estey EH, Bernstein ID. Acute myeloid leukemia stem cells and CD33-targeted immunotherapy. Blood. 28 juin 2012;119(26):6198-208.

95. Griffin JD, Linch D, Sabbath K, Larcom P, Schlossman SF. A monoclonal antibody reactive with normal and leukemic human myeloid progenitor cells. Leuk Res. 1984;8(4):521-34.

96. Dinndorf PA, Andrews RG, Benjamin D, Ridgway D, Wolff L, Bernstein ID. Expression of normal myeloid-associated antigens by acute leukemia cells. Blood. avr 1986;67(4):1048-53.

97. Burnett AK, Russell NH, Hills RK, Kell J, Freeman S, Kjeldsen L, et al. Addition of gemtuzumab ozogamicin to induction chemotherapy improves survival in older patients with acute myeloid leukemia. J Clin Oncol. 10 nov 2012;30(32):3924-31.

98. Burnett AK, Hills RK, Milligan D, Kjeldsen L, Kell J, Russell NH, et al. Identification of patients with acute myeloblastic leukemia who benefit from the addition of gemtuzumab ozogamicin: results of the MRC AML15 trial. J Clin Oncol. 1 févr 2011;29(4):369-77.

99. Fournier E, Duployez N, Ducourneau B, Raffoux E, Turlure P, Caillot D, et al. Mutational profile and benefit of gemtuzumab ozogamicin in acute myeloid leukemia. Blood. 20 févr 2020;135(8):542-6.

100. Castaigne S, Pautas C, Terré C, Raffoux E, Bordessoule D, Bastie J-N, et al. Effect of gemtuzumab ozogamicin on survival of adult patients with de-novo acute myeloid leukaemia (ALFA-0701): a randomised, open-label, phase 3 study. Lancet. 21 avr 2012;379(9825):1508-16.

101. Renneville A, Abdelali RB, Chevret S, Nibourel O, Cheok M, Pautas C, et al. Clinical impact of gene mutations and lesions detected by SNP-array karyotyping in acute myeloid leukemia patients in the context of gemtuzumab ozogamicin treatment: results of the ALFA-0701 trial. Oncotarget. 28 févr 2014;5(4):916-32.

102. Ehninger A, Kramer M, Röllig C, Thiede C, Bornhäuser M, von Bonin M, et al. Distribution and levels of cell surface expression of CD33 and CD123 in acute myeloid leukemia. Blood Cancer J. juin 2014;4(6):e218.

103. Linenberger ML, Hong T, Flowers D, Sievers EL, Gooley TA, Bennett JM, et al. Multidrug-resistance phenotype and clinical responses to gemtuzumab ozogamicin. Blood. 15 août 2001;98(4):988-94.

104. Walter RB, Raden BW, Hong TC, Flowers DA, Bernstein ID, Linenberger ML. Multidrug resistance protein attenuates gemtuzumab ozogamicin–induced cytotoxicity in acute myeloid leukemia cells. Blood. 15 août 2003;102(4):1466-73.

105. Middeldorf I, Galm O, Osieka R, Jost E, Herman JG, Wilop S. Sequence of administration and methylation of SOCS3 may govern response to gemtuzumab ozogamicin in combination with conventional chemotherapy in patients with refractory or relapsed acute myelogenous leukemia (AML). Am J Hematol. juill 2010;85(7):477-81.

106. Lamba JK, Pounds S, Cao X, Downing JR, Campana D, Ribeiro RC, et al. Coding polymorphisms in CD33 and response to gemtuzumab ozogamicin in pediatric patients with AML: a pilot study. Leukemia. févr 2009;23(2):402-4.

107. Murphy KM, Levis M, Hafez MJ, Geiger T, Cooper LC, Smith BD, et al. Detection of FLT3 Internal Tandem Duplication and D835 Mutations by a Multiplex Polymerase Chain Reaction and Capillary Electrophoresis Assay. J Mol Diagn. mai 2003;5(2):96-102.

108. Lyu M, Liao H, Shuai X, Jin Y, Su J, Zheng Q. The prognosis predictive value of FMS-like tyrosine kinase 3-internal tandem duplications mutant allelic ratio (FLT3-ITD MR) in patients with acute myeloid leukemia detected by GeneScan. Gene. 5 févr 2020;726:144195.

109. Sakaguchi M, Nakajima N, Yamaguchi H, Najima Y, Shono K, Marumo A, et al. The sensitivity of the FLT3-ITD detection method is an important consideration when diagnosing acute myeloid leukemia. Leuk Res Rep. 2020;13:100198.

110. Kim Y, Lee GD, Park J, Yoon J-H, Kim H-J, Min W-S, et al. Quantitative fragment analysis of FLT3-ITD efficiently identifying poor prognostic group with high mutant allele burden or long ITD length. Blood Cancer J. août 2015;5(8):e336.

111. Engen C, Hellesøy M, Grob T, Al Hinai A, Brendehaug A, Wergeland L, et al. FLT3-ITD mutations in acute myeloid leukaemia - molecular characteristics, distribution and numerical variation. Mol Oncol. sept 2021;15(9):2300-17.

112. Spencer DH, Abel HJ, Lockwood CM, Payton JE, Szankasi P, Kelley TW, et al. Detection of FLT3 Internal Tandem Duplication in Targeted, Short-Read-Length, Next-

Generation Sequencing Data. The Journal of Molecular Diagnostics. 1 janv 2013;15(1):81-93.

113. Bibault J-E, Figeac M, Hélevaut N, Rodriguez C, Quief S, Sebda S, et al. Next-generation sequencing of FLT3 internal tandem duplications for minimal residual disease monitoring in acute myeloid leukemia. Oncotarget. 2 juin 2015;6(26):22812-21.

114. Au CH, Wa A, Ho DN, Chan TL, Ma ESK. Clinical evaluation of panel testing by next-generation sequencing (NGS) for gene mutations in myeloid neoplasms. Diagn Pathol. 22 janv 2016;11(1):11.

115. Wang T-Y, Yang R. ScanITD: Detecting internal tandem duplication with robust variant allele frequency estimation. Gigascience. 1 août 2020;9(8):giaa089.

116. Blätte TJ, Schmalbrock LK, Skambraks S, Lux S, Cocciardi S, Dolnik A, et al. getITD for FLT3 -ITD-based MRD monitoring in AML. Leukemia. oct 2019;33(10):2535-9.

117. Tsai HK, Brackett DG, Szeto D, Frazier R, MacLeay A, Davineni P, et al. Targeted Informatics for Optimal Detection, Characterization, and Quantification of FLT3 Internal Tandem Duplications Across Multiple Next-Generation Sequencing Platforms. The Journal of Molecular Diagnostics. 1 sept 2020;22(9):1162-78.

118. Audemard EO, Gendron P, Feghaly A, Lavallée V-P, Hébert J, Sauvageau G, et al. Targeted variant detection using unaligned RNA-Seq reads. Life Sci Alliance. août 2019;2(4):e201900336.

119. Chiba K, Shiraishi Y, Nagata Y, Yoshida K, Imoto S, Ogawa S, et al. Genomon ITDetector: a tool for somatic internal tandem duplication detection from cancer genome sequencing data. Bioinformatics. 1 janv 2015;31(1):116-8.

120. Rustagi N, Hampton OA, Li J, Xi L, Gibbs RA, Plon SE, et al. ITD assembler: an algorithm for internal tandem duplication discovery from short-read sequencing data. BMC Bioinformatics. déc 2016;17(1):1-8.

121. Yuan D, He X, Han X, Yang C, Liu F, Zhang S, et al. Comprehensive review and evaluation of computational methods for identifying FLT3-internal tandem duplication in acute myeloid leukaemia. Brief Bioinform. 13 avr 2021;bbab099.

122. Pindel () [Internet]. [cité 27 sept 2021]. Disponible sur: https://gmt.genome.wustl.edu/packages/pindel/user-manual.html

123. Au T. tommyau/itdseek [Internet]. 2019 [cité 8 déc 2020]. Disponible sur: https://github.com/tommyau/itdseek

124. ht50. FLT3_ITD_ext [Internet]. 2021 [cité 10 août 2021]. Disponible sur: https://github.com/ht50/FLT3_ITD_ext

125. iric-soft/km [Internet]. iric-soft; 2020 [cité 29 sept 2020]. Disponible sur: https://github.com/iric-soft/km

126. Genomon-ITDetector/testout at master · ken0-1n/Genomon-ITDetector [Internet]. GitHub. [cité 10 août 2021]. Disponible sur: https://github.com/ken0-1n/Genomon-ITDetector

127. Kim B, Kim S, Lee S-T, Min YH, Choi JR. FLT3 Internal Tandem Duplication in Patients With Acute Myeloid Leukemia Is Readily Detectable in a Single Next-Generation Sequencing Assay Using the Pindel Algorithm. Ann Lab Med. mai 2019;39(3):327-9.

128. Schranz K, Hubmann M, Harin E, Vosberg S, Herold T, Metzeler KH, et al. Clonal heterogeneity of FLT3 -ITD detected by high-throughput amplicon sequencing correlates with adverse prognosis in acute myeloid leukemia. Oncotarget. 10 juill 2018;9(53):30128-45.

129. Thol F, Kölking B, Damm F, Reinhardt K, Klusmann J-H, Reinhardt D, et al. Next-generation sequencing for minimal residual disease monitoring in acute myeloid leukemia patients with FLT3-ITD or NPM1 mutations. Genes Chromosomes Cancer. juill 2012;51(7):689-95.

130. Nair R, Salinas-Illarena A, Baldauf H-M. New strategies to treat AML: novel insights into AML survival pathways and combination therapies. Leukemia. févr 2021;35(2):299-311.

131. Blätte TJ. tjblaette/getitd [Internet]. 2020 [cité 29 sept 2020]. Disponible sur: https://github.com/tjblaette/getitd

132. Reese MG, Eeckman FH, Kulp D, Haussler D. Improved splice site detection in Genie. J Comput Biol. 1997;4(3):311-23.

133. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Research. 16 déc 2014;42(22):13534-44.

134. Rücker FG, Du L, Luck TJ, Benner A, Krzykalla J, Gathmann I, et al. Molecular landscape and prognostic impact of FLT3-ITD insertion site in acute myeloid leukemia: RATIFY study results. Leukemia. 28 juill 2021;1-10.

135. Schmalbrock LK, Dolnik A, Cocciardi S, Sträng E, Theis F, Jahn N, et al. Clonal evolution of acute myeloid leukemia with FLT3-ITD mutation under treatment with midostaurin. Blood. 3 juin 2021;137(22):3093-104.

136. Tao S, Wang C, Chen Y, Deng Y, Song L, Shi Y, et al. Prognosis and outcome of patients with acute myeloid leukemia based on FLT3-ITD mutation with or without additional abnormal cytogenetics. Oncol Lett. déc 2019;18(6):6766-74.

137. Ragon BK. FLT3-ITDos and FLT3-ITDon'ts: navigating maintenance therapy in FLT3-ITD-positive acute myeloid leukemia following stem cell transplantation. Bone Marrow Transplant. août 2021;56(8):1774-6.

138. Bazarbachi A, Bug G, Baron F, Brissot E, Ciceri F, Dalle IA, et al. Clinical practice recommendation on hematopoietic stem cell transplantation for acute myeloid leukemia patients with FLT3-internal tandem duplication: a position statement from the Acute Leukemia Working Party of the European Society for Blood and Marrow Transplantation. Haematologica. juin 2020;105(6):1507-16.

Université de Lille

FACULTE DE PHARMACIE DE LILLE

DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

Année Universitaire 2020/2021.

Nom : Boudry
Prénom : Augustin
Titre de la thèse : Développement d'un nouvel algorithme pour la détection et la quantification des duplications en tandem de *FLT3* dans les leucémies aigues myéloïdes
Mots-clés : Leucémies aiguës myéloïdes ; *FLT3*-ITD ; NGS ; Marqueurs moléculaires ; Bio-informatique ;

Résumé : Les duplications en tandem dans le gène *FLT3* (*FLT3*-ITD) sont retrouvées dans 20 à 30 % des leucémies aiguës myéloïdes. *FLT3*-ITD constitue un marqueur pronostique et une cible thérapeutique et est donc systématiquement recherchée et quantifiée à l'aide de l'analyse de fragments, conformément aux recommandations de l'European Leukemia Net. Bien que robuste, cette technique présente plusieurs limites : limite de détection élevée (1-3%), approximation de la taille de l'ITD, elle ne fournit pas la séquence nucléotidique de l'ITD ni son site d'insertion, multiplexage non réalisable. La détection de ces ITDs par séquençage de nouvelle génération (NGS) pourrait, à terme, remplacer cette technique et corriger l'ensemble de ses limites. *FLT3*-ITD, de par son hétérogénéité en termes de taille et de site d'insertion, est difficile à mettre en évidence par NGS avec les outils bio-informatiques classiques.
Ce travail a pour but de développer un algorithme de détection et de quantification pour *FLT3*-ITD, nommé FiLT3r, et de le comparer avec les algorithmes existants à la méthode de référence.
Il s'agissait d'une étude rétrospective multicentrique randomisée sur plusieurs cohortes : une d'entrainement (BIG-1), et trois de validation (ALFA-0701, ALFA-0702, ALFA-1200) incluant au total 1173 patients. Les différents algorithmes ont été testés et comparés à la méthode de référence en utilisant les échantillons analysés par NGS.
Toutes les ITD détectées par l'analyse de fragments des 1173 patients ont été confirmées par FiLT3r, y compris chez les patients présentant plusieurs ITDs. Les ratios calculés avec notre algorithme ont démontré une corrélation avec la méthode de référence estimée à 0,87, correspondant au coefficient de corrélation le plus élevé de tous les algorithmes testés (intervalle : 0,36 - 0,87). FiLT3r a donc fourni les résultats les plus fidèles par rapport à la technique de référence, avec un score F_1 de 1 (intervalle : 0,7 - 1) et aucun faux positif.
L'algorithme FiLT3r représente une approche prometteuse pour la détection sensible, robuste et informative des mutations *FLT3*-ITD par NGS.

Membres du jury :

Président :	Dupont Annabelle, PU-PH à la faculté de pharmacie de Lille
Assesseur(s) :	Quesnel Bruno, PU-PH à la faculté de médecine de Lille
	Duployez Nicolas, MCU-PH à la faculté de médécine de Lille
	Figeac Martin, IR à la faculté de médecine de Lille

Directeur de thèse : Preudhomme Claude, PU-PH à la faculté de médecine de Lille