Un pipeline bioinformatique de ré-interprétation d'analyses constitutionnelles d'exome

Alexis Praga

ANNEE 2024 - N° 24 – 027

Un pipeline bioinformatique de ré-interprétation d'analyses constitutionnelles d'exome

THÈSE

présentée et soutenue publiquement

le **18 avril 2024** à 14 h

pour obtenir le Diplôme d'Etat de

DOCTEUR EN MEDECINE

PAR

Alexis PRAGA

Né(e) le 28/07/1987 à Metz (57)

La composition du jury est la suivante :

Président :	Monsieur Jean-Paul FEUGEAS	Professeur
Directeur de la thèse :	Monsieur Alexis Overs	Assistant-Hospitalier
Juges :	Monsieur Paul Kuentz	Professeur
	Monsieur Didier Hocquet	Professeur
	Monsieur Eric Dahlen	Praticien Hospitalier
	Monsieur Rémi Mathevet	Practicien Hospitalier

ANNEE 2024 - N° 24 – 027

Un pipeline bioinformatique de ré-interprétation d'analyses constitutionnelles d'exome

THÈSE

présentée et soutenue publiquement

le **18 avril 2024** à 14 h

pour obtenir le Diplôme d'Etat de

DOCTEUR EN MEDECINE

PAR

Alexis PRAGA

Né(e) le 28/07/1987 à Metz (57)

La composition du jury est la suivante :

Président :	Monsieur Jean-Paul FEUGEAS	Professeur
Directeur de la thèse :	Monsieur Alexis Overs	Assistant-Hospitalier
Juges :	Monsieur Paul Kuentz	Professeur
	Monsieur Didier Hocquet	Professeur
	Monsieur Eric Dahlen	Praticien Hospitalier
	Monsieur Rémi Mathevet	Practicien Hospitalier

UNIVERSITÉ DE FRANCHE-COMTÉ

U.F.R. SCIENCES DE LA SANTE

BESANÇON

DIRECTEUR	**PROFESSEUR THIERRY MOULIN**	
DIRECTEUR ADJOINT	**PROFESSEUR XAVIER BERTRAND**	DOYEN PHARMACIE

RESPONSABLE ADMINISTRATIVE **MME CAROLE COINTEAU**

DEPARTEMENT MEDECINE

DOCTEUR MALIKA BOUHADDI (MCU-PH)	DIRECTRICE DES ÉTUDES
PROFESSEUR JEAN-PAUL FEUGEAS	ASSESSEUR 1ER CYCLE
PROFESSEUR MARIE-FRANCE SERONDE	ASSESSEUR 2EME CYCLE
PROFESSEUR CATHERINE CHIROUZE	ASSESSEUR 3EME CYCLE
PROFESSEUR BENOIT DE BILLY	COORDINATEUR CHIRURGIE
PROFESSEUR BENOIT DINET	COORDINATEUR MEDECINE GENERALE

DEPARTEMENT PHARMACIE

PROFESSEUR XAVIER BERTRAND	DOYEN PHARMACIE
PROFESSEUR LHASSANE ISMAILI	DIRECTEUR DES ETUDES
PROFESSEUR SAMUEL LIMAT	COORDINATEURS 3E CYCLE
PROFESSEUR VIRGINIE NERICH	

DEPARTEMENT MAÏEUTIQUE

MARILIA GIRAULT (SAGE-FEMME)	COORDINATRICE PEDAGOGIQUE

DEPARTEMENT ODONTOLOGIE

PROFESSEUR EDOUARD EUVRARD {PAST)	COORDINATEURS PEDAGOGIQUES
DR SOPHIE PECHOUX {MAST)	
DR JEAN-PIERRE SALOMON (MCU-PH) (Université. LORRAINE)	

DEPARTEMENT SCIENCES DES METIERS DE LA REEDUCATION : ORTHOPHONIE

PROFESSEUR ELOI MAGNIN	COORDINATEURS PEDAGOGIQUES

DEPARTEMENT SCIENCES DES METIERS DE LA REEDUCATION : KINESITHERAPIE

ALEXANDRE KUBICKI {MONTBELIARD) {MCF)	COORDINATEURS PEDAGOGIQUES
YOSHIMASA SAGAWA {MCF)	

DEPARTEMENT SCIENCES DES METIERS DE LA REEDUCATION : ERGOTHERAPIE/PSYCHOMOTRICITE

JULIE LAPREVOTTE {MAST)	COORDINATRICES PEDAGOGIQUES
CLEMENCE VALLIER {MAST)	

DEPARTEMENT SCIENCES INFIRMIERES

ALINE CHASSAGNE (MCF)	COORDINATEURS PEDAGOGIQUES
JEAN MAILLET-CONTOZ (MAST)	
CHRISTINE MEYER (SOINS INFIRMIERS IFSI)	
DR ANTOINE THIERY-VUILLEMIN (MCU-PH)	
PROFESSEUR FABRICE VUILLIER	COORDINATEUR PEDAGOGIQUE **IPA**

DEPARTEMENT DE PEDAGOGIE

PROFESSEUR CLEMENT PRATI	RESPONSABLE
PROFESSEUR SEBASTIEN PILI-FLOURY	CENTRE DE SIMULATION
PROFESSEUR BENOIT DINET	
FRANK VERHOEVEN (MEDECINE)	
MARILIA GIRAULT (MAÏEUTIQUE)	
MARC PUDLO (PHARMACIE)	
YOSHIMATA SAGAWA (REEDUCATION)	
LAURENCE GANDON (INFIRMIER)	

RELATIONS HUMAINES DE L'UFR

PROFESSEUR SYLVIE NEZELOF	ASSESSEUR

COMMISSION SCIENTIFIQUE DE L'UFR

PROFESSEUR VIRGINIE WESTEEL	ASSESSEUR RECHERCHE - PRESIDENTE
PROFESSEUR FREDERIC AUBER	VICE-PRESIDENT

CHARGES DE MISSIONS

FORMATION CONTINUE	MME SYLVIE DEVAUX (MCF)	COORDINATEURS
HISTOIRE DE LA MEDECINE	PROFESSEUR LAURENT TATU	COORDINATEURS
	DOCTEUR PHILIPPE MERCET	COORDINATEURS
RELATIONS INTERNATIONALES	DOCTEUR OLEG BLAGOSKLONOV (MCU-PH)	
		COORDINATEUR
ALUMNI-USB	PROFESSEUR GILLES CAPELLIER (EMERITE)	
	PROFESSEUR GABRIEL CAMELOT (EMERITE)	PRESIDENT HONORAIRE

MÉDECINE

PROFESSEURS DES UNIVERSITÉS

- PRATICIENS HOSPITALIERS

M.	Olivier	**ADOTEVI**	IMMUNOLOGIE
M.	Frédéric	**AUBER**	CHIRURGIE INFANTILE
M.	François	**AUBIN**	DERMATO-VÉNÉRÉOLOGIE
M.	Jamal	**BAMOULID**	IMMUNOLOGIE
Me	Cindy	**BARNIG**	PNEUMOLOGIE
Me	Djamila	**BENNABI**	PSYCHIATRIE ADULTES
M.	Guillaume	**BESCH**	ANESTHESIE REANIMATION
M.	Frédéric	**BIBEAU**	ANATOMIE ET CYTOLOGIE PATHOLOGIQUES
Me	Alessandra	**BIONDI**	RADIOLOGIE ET IMAGERIE MÉDICALE
M.	Christophe	**BORG**	CANCÉROLOGIE
M.	Hatem	**BOULAHDOUR**	BIOPHYSIQUE ET MÉDECINE NUCLÉAIRE
Me	Catherine	**CHIROUZE**	MALADIES INFECTIEUSES
M.	Romain	**CHOPARD**	CARDIOLOGIE
M.	Sidney	**CHOCRON**	CHIRURGIE THORACIQUE ET CARDIOVASCULAIRE
Me	Cécile	**COURIVAUD**	NÉPHROLOGIE
M.	Siamak	**DAVANI**	PHARMACOLOGIE CLINIQUE
M.	Benoît	**DE BILLY**	CHIRURGIE INFANTILE
M.	Eric	**DECONINCK**	HÉMATOLOGIE
M.	Eric	**DELABROUSSE**	RADIOLOGIE ET IMAGERIE MÉDICALE
M.	Thibaut	**DESMETTRE**	MÉDECINE D'URGENCE
M.	Vincent	**DI MARTINO**	HÉPATOLOGIE
M.	Didier	**DUCLOUX**	NÉPHROLOGIE
M.	Jean- Paul	**FEUGEAS**	BIOCHIMIE ET BIOLOGIE MOLÉCULAIRE
M.	Patrick	**GARBUIO**	CHIRURGIE ORTHOPÉDIQUE ET TRAUMATOLOGIQUE
Me	Anne-Sophie	**GAUTHIER**	OPHTALMOLOGIE
M.	Emmanuel	**HAFFEN**	PSYCHIATRIE D'ADULTES
M.	Georges	**HERBEIN**	VIROLOGIE
M.	Bruno	**HEYD**	CHIRURGIE GÉNÉRALE
M.	Didier	**HOCQUET**	HYGIÈNE HOSPITALIÈRE
Me	Katy	**JEANNOT**	BACTÉRIOLOGIE - VIROLOGIE
M.	François	**KLEINCLAUSS**	UROLOGIE
M.	Paul	**KUENTZ**	HISTOLOGIE EMBRYOLOGIE ET CYTOGENETIQUE
M.	Zaher	**LAKKIS**	CHIRURGIE VISCERALE ET DIGESTIVE
M.	Daniel	**LEPAGE**	ANATOMIE
M.	Quentin	**LEPILLER**	VIROLOGIE
M.	Eloi	**MAGNIN**	NEUROLOGIE
Me	Nadine	**MAGY-BERTRAND**	MEDECINE INTERNE
M.	Frédéric	**MAUNY**	BIOSTATISTIQUES, INFORMATIQUE MÉDICALE
M.	Nicolas	**MENEVEAU**	CARDIOLOGIE
M.	Christophe	**MEYER**	CHIRURGIE MAXILLO FACIALE ET STOMATOLOGIE
M.	Fabrice	**MICHEL**	MEDECINE PHYSIQUE ET DE READAPTATION
Me	Laurence	**MILLON**	PARASITOLOGIE ET MYCOLOGIE

M.	Nicolas	**MOTTET**	GYNECOLOGIE OBSTETRIQUE
M.	Thierry	**MOULIN**	NEUROLOGIE
Me	Sylvie	**NEZELOF**	PÉDOPSYCHIATRIE
M.	Laurent	**OBERT**	CHIRURGIE ORTHOPÉDIQUE ET TRAUMATOLOGIQUE
M.	Andréas	**PERROTTI**	CHIRURGIE THORACIQUE ET CARDIOVASCULAIRE
M.	Sébastien	**PILI-FLOURY**	ANESTHÉSIOLOGIE RÉANIMATION
M.	Gaël	**PITON**	MEDECINE INTENSIVE REANIMATION
M.	Clément	**PRATI**	RHUMATOLOGIE
M.	Jean-Luc	**PRETET**	BIOLOGIE CELLULAIRE
M.	Rajeev	**RAMANAH**	GYNÉCOLOGIE - OBSTÉTRIQUE
M.	Simon	**RINCKENBACH**	CHIRURGIE VASCULAIRE
M.	Christophe	**ROUX**	BIOLOGIE ET MÉDECINE DU DÉVELOPPEMENT ET DE LA REPRODUCTION
Me	Lucie	**SALOMON DU MONT**	CHIRURGIE VASCULAIRE
M.	Emmanuel	**SAMAIN**	ANESTHÉSIOLOGIE RÉANIMATION
M.	François	**SCHIELE**	CARDIOLOGIE
Me	Marie-France	**SERONDE**	CARDIOLOGIE
M.	Laurent	**TATU**	ANATOMIE
M.	Laurent	**TAVERNIER**	OTO-RHINO-LARYNGOLOGIE
M.	Thierry	**THEVENOT**	HÉPATOLOGIE
M.	Laurent	**THINES**	NEUROCHIRURGIE
M.	Gérard	**THIRIEZ**	PÉDIATRIE
M.	Pierre	**TIBERGHIEN**	IMMUNOLOGIE
M.	Eric	**TOUSSIROT**	THÉRAPEUTIQUE
M.	Pierre	**VANDEL**	PSYCHIATRIE d'ADULTES
M.	Fabrice	**VUILLIER**	ANATOMIE
Me	Lauriane	**VULLIEZ COADY**	PEDO-PSYCHIATRIE
Me	Lucine	**VUITTON**	GASTRO-ENTEROLOGIE
Me	Virginie	**WESTEEL-KAULEK**	PNEUMOLOGIE

PROFESSEURS EMÉRITES

M.	Régis	**AUBRY**	ÉPISTEMOLOGIE-SOINS PALLIATIFS
M.	Jean-Luc	**BRESSON**	BIOLOGIE ET MÉDECINE DU DÉVELOPPEMENT ET DE LA REPRODUCTION
M.	Gilles	**CAPELLIER**	MEDECINE INTENSIVE-REANIMATION
M.	Jean-Luc	**CHOPARD**	MEDECINE LEGALE
M.	Alain	**CZORNY**	NEUROCHIRURGIE
M.	Bernard	**DELBOSC**	OPHTALMOLOGIE
M.	Gilles	**DUMOULIN**	PHYSIOLOGIE
M.	Dominique	**FELLMANN**	CYTOLOGIE ET HISTOLOGIE
M.	Georges	**MANTION**	CHIRURGIE GÉNÉRALE
Me	Christiane	**MOUGIN**	BIOLOGIE CELLULAIRE
M.	Bernard	**PARRATTE**	ANATOMIE
M.	Patrick	**PLESIAT**	BACTERIOLOGIE - VIROLOGIE
M.	Christophe	**ROUX**	BIOLOGIE ET MÉDECINE DU DÉVELOPPEMENT ET DE LA REPRODUCTION
M.	Daniel	**SECHTER**	PSYCHIATRIE D'ADULTES
Me	Dominique	**VUITTON**	IMMUNOLOGIE
M.	Daniel	**WENDLING**	RHUMATOLOGIE

MAITRES DE CONFÉRENCES DES UNIVERSITÉS - PRATICIENS HOSPITALIERS

Me	Clotilde	**AMIOT**	HISTOLOGIE EMBRYOLOGIE ET CYTOGENETIQUE
Me	Anne-Pauline	**BELLANGER**	PARASITOLOGIE
M.	Matthieu	**BEREAU**	THERAPEUTIQUE
Me	Oxana	**BLAGOSKLONOV**	BIOLOGIE ET MÉDECINE DÉVELOPPEMENT ET DE REPRODUCTION
Me	Sophie	**BOROT**	ENDOCRINOLOGIE, DIABÈTE ET MALADIES MÉTABOLIQUES
Me	Malika	**BOUHADDI**	PHYSIOLOGIE
M.	Kévin	**BOUILLER**	MALADIES INFECTIEUSES
M.	Paul	**CALAME**	RADIOLOGIE ET IMAGERIE MÉDICALE
M.	Yann	**CHAUSSY**	CHIRURGIE INFANTILE
M.	Alain	**COAQUETTE**	VIROLOGIE
Me	Elsa	**CURTIT**	CANCÉROLOGIE
M.	Etienne	**DAGUINDAU**	HEMATOLOGIE
M.	Maxime	**DESMARETS**	EPIDEMIOLOGIE, ECONOMISE DE LA SANTE ET PREVENTION
Me	Julie	**GIUSTINANNI**	PSYCHIATRIE ADULTE-ADDICTOLOGIE
M.	François	**LOISEL**	CHIRURGIE ORTHOPEDIQUE ET TRAUMATOLOGIQUE
Me	Tania	**MARX**	MÉDECINE D'URGENCE
Me	Elisabeth	**MEDEIROS**	NEUROLOGIE
M.	Patrice	**MURET**	PHARMACOLOGIE CLINIQUE
Me	Charlée	**NARDIN**	DERMATOLOGIE
M.	Fabien	**PELLETIER**	DERMATO-VÉNÉRÉOLOGIE
Me	Isabelle	**PLUVY**	CHIRURGIE PLASTIQUE, RECONSTRUCTIVE, ESTHETIQUE
Me	Anaïs	**POTRON**	BACTÉRIOLOGIE - VIROLOGIE
M.	Zoher	**SELMANI**	BIOLOGIE CELLULAIRE
M.	Antoine	**THIERY-VUILLEMIN**	CANCÉROLOGIE
M.	Frank	**VERHOEVEN**	RHUMATOLOGIE
Me	Delphine	**WEIL-VERHOEVEN**	HEPATOLOGIE
M.	Hadrien	**WINISZEWSKI**	MÉDECINE INTENSIVE-RÉANIMATION

ENSEIGNANTS ASSOCIÉS

M.	Rémi	**BARDET**	MÉDECINE GÉNÉRALE (PROFESSEUR)
M.	Francis	**BERTHIER**	ANESTHESIE-REANIMATION (PROFESSEUR)
Me	Anne-Lise	**BOLOT**	MÉDECINE GÉNÉRALE (MCF)
M.	Benoit	**DINET**	MÉDECINE GÉNÉRALE (PROFESSEUR)
Me	Catherine	**GAY**	GYNÉCOLOGIE-OBSTÉTRIQUE (PROFESSEUR)
M	Abdo	**KHOURY**	MÉDECINE D'URGENCE (PROFESSEUR)
Me	Aurore	**LEBEAU-JEUNET**	MEDECINE GENERALE (MCF)
M.	Thierry	**LEPETZ**	MÉDECINE GÉNÉRALE (MCF)
M.	José-Philippe	**MORENO**	MÉDECINE GÉNÉRALE (PROFESSEUR)
M.	Jean-Michel	**PERROT**	MÉDECINE GÉNÉRALE (PROFESSEUR)
M.	Thomas	**RODRIGUEZ**	MÉDECINE GÉNÉRALE (MCF)
Me	Esther	**SZWARC**	SANTE AU TRAVAIL (MCF)
Me	Anne-Lise	**TREMEAU**	MÉDECINE GÉNÉRALE (MCF)

PHARMACIE

PROFESSEURS

M.	Xavier	**BERTRAND**	MICROBIOLOGIE - INFECTIOLOGIE
Me	Céline	**DEMOUGEOT**	PHARMACOLOGIE
Me	Francine	**GARNACHE-OTTOU**	HÉMATOLOGIE
Me	Corine	**GIRARD**	PHARMACOGNOSIE
M.	Yann	**GODET**	IMMUNOLOGIE
M.	Frédéric	**GRENOUILLET**	PARASITOLOGIE-MYCOLOGIE
M.	Yves	**GUILLAUME**	CHIMIE ANALYTIQUE
M.	Lhassane	**ISMAILI**	CHIMIE ORGANIQUE
M.	Samuel	**LIMAT**	PHARMACIE CLINIQUE
M.	Frédéric	**LIRUSSI**	PHARMACOLOGIE - TOXICOLOGIE
M.	Dominique	**MEILLET**	PARASITOLOGIE - MYCOLOGIE
Me	Virginie	**NERICH**	PHARMACIE CLINIQUE
M.	Yann	**PELLEQUER**	PHARMACIE GALENIQUE
M.	Bernard	**REFOUVELET**	CHIMIE ORGANIQUE ET THERAPEUTIQUE
M.	Philippe	**SAAS**	IMMUNOLOGIE
Me	Marie-Christine	**WORONOFF-LEMSI**	PHARMACIE CLINIQUE

PROFESSEUR EMÉRITE

Me	Laurence	**NICOD**	BIOLOGIE CELLULAIRE

MAITRES DE CONFÉRENCES

Me	Aurélie	**BAGUET**	BIOCHIMIE
M.	Arnaud	**BEDUNEAU**	PHARMACIE GALÉNIQUE
M.	Laurent	**BERMONT**	BIOCHIMIE
M.	Oleg	**BLAGOSKLONOV**	BIOPHYSIQUE ET IMAGERIE MÉDICALE
Me	Céline	**BOUVIER-SLEKOVEC**	HYGIENE PREVENTION RISQUES INFECTIEUX
M.	Eric	**CAVALLI**	CHIMIE PHYSIQUE ET MINÉRALE
Me	Anne-Laure	**CLAIRET**	SCIENCES DU MEDICAMENT
M.	Jean-Patrick	**DASPET**	BIOPHYSIQUE
Me	Sylvie	**DEVAUX**	PHYSIOLOGIE
Me	Jeanne	**GALAINE**	SCIENCES BIOLOGIQUES, FONDAMENTALES ET CLINIQUES
Me	Marie	**KROEMER**	SCIENCES DU MEDICAMENT ET AUTRES PRODUITS DE SANTE
Me	Isabelle	**LASCOMBE**	BIOCHIMIE / ISIFC
Me	Carole	**MIGUET ALFONSI**	TOXICOLOGIE
M.	Johnny	**MORETTO**	PHYSIOLOGIE
M.	Brice	**MOULARI**	PHARMACIE GALENIQUE
M.	Frédéric	**MUYARD**	PHARMACOGNOSIE
M.	Marc	**PUDLO**	CHIMIE THÉRAPEUTIQUE
M.	Florian	**RENOSI**	SCIENCES BIOLOGIQUES, FONDAMENTALES ET CLINIQUES
M.	Gwenaël	**ROLIN**	SCIENCES BIOLOGIQUES, FONDAMENTALES (ODONTOLOGIE)
Me	Nathalie	**RUDE**	BIOMATHÉMATIQUES ET BIOSTATISTIQUES
M.	François	**SENEJOUX**	PHARMACOGNOSIE
Me	Perle	**TOTOSON**	PHARMACOLOGIE

ENSEIGNANTS ASSOCIÉS

M.	Lionel	**PAZART**	SANTÉ PUBLIQUE (PROFESSEUR)

ODONTOLOGIE

MAITRES DE CONFÉRENCES DES UNIVERSITÉS

M.	Gwenaël	**ROLIN**	SCIENCES BIOLOGIQUES, FONDAMENTALES (ODONTOLOGIE)
M.	Jean-Pierre	**SALOMON**	DENTISTERIE RESTAURATRICE, ENDODONTIE, PROTHÈSES, FONCTION-DYSFONCTION, IMAGERIE, BIOMATÉRIAUX (DÉLÉGATION DE L'UNIVERSITÉ DE LORRAINE)

ENSEIGNANTS ASSOCIÉS

Me	Hélène	**BIGEARD**	ODONTOLOGIE (MCF)
M.	Edouard	**EUVRARD**	CHIRURGIE ORALE – ODONTOLOGIE (PROFESSEUR)
Me	Sophie	**PECHOUX**	ODONTOLOGIE (MCF)
Me	Sylvie	**ROMAGNA**	ODONTOLOGIE (PROFESSEUR)

PROFESSIONS DE SANTE

Me	Aline	**CHASSAGNE**	SCIENCES INFIRMIERES (MCF)
M.	Jean	**MAILLET-CONTOZ**	SCIENCES INFIRMIERES (MAST)
M.	Alain	**DEVEVEY**	SCIENCES LANGAGE- ORTHOPHONIE (MCF)
M.	Alexandre	**KUBICKI**	SCIENCES REEDUCATION - KINESITHERAPIE (MCF)
M.	Yoshimasa	**SAGAWA JUNIOR**	SCIENCES REEDUCATION - KINESITHERAPIE (MCF)
Me	Emilie	**CERUTTI**	KINESITHERAPIE (MAST)
M.	Charles-Henry	**MAXENCE**	KINESITHERAPIE (MAST)
Me	Marine	**BRIKA**	KINESITHERAPIE
M.	Clément	**GRIESSINGER**	KINESITHERAPIE
Me	Mélanie	**MICHELIN-VAUTIER**	KINESITHERAPIE
Me	Marie-Carole	**PLAY**	KINESITHERAPIE
M.	Jérome	**PLONGERON**	KINESITHERAPIE
M.	Maxime	**WERNER**	KINESITHERAPIE
Me	Anne-Sophie	**RIOU**	ORTHOPHONIE (MAST)
Me	Carine	**PETIT**	ORTHOPHONIE (MAST)
Me	Laurence	**DEFORET**	ORTHOPHONIE (MAST)
Me	Julie	**LAPREVOTTE**	ERGOTHERAPIE (MAST)
Me	Margaux	**GUIMARD**	ERGOTHERAPIE
Me	Clemence	**VALLIER**	PSYCHOMOTRICITE (MAST)
Me	Mélanie	**DIEMER**	PSYCHOMOTRICITE
Me	Lola	**VUILLAUME**	PSYCHOMOTRICITE
M.	Pierrick	**BOYER**	PSYCHOMOTRICITE

AUTRES ENSEIGNANTS

Me	Vanessa	**MARTIN**	PROFESSEUR AGREGEE ANGLAIS
Me	Anna	**MKRTCHIAN**	PROFESSEUR AGREGEE ANGLAIS
M.	Charles Dale	**SANTANA**	PROFESSEUR AGREGE ANGLAIS
M.			PROFESSEUR AGREGE ANGLAIS

Mise à jour 02/10/2023

Remerciements

À notre maître et président du jury, **Monsieur le Professeur Jean-Paul Feugeas**, Professeur des Universités - Praticien Hospitalier et chef du service d'oncobiologie génétique bio-informatique au CHU de Besançon, merci de me faire l'honneur de présider cette thèse dont j'espère qu'elle pourra servir à étayer le développement d'autres pipelines en oncogénétique notamment.

À notre maître et directeur de thèse, **Monsieur le Docteur Alexis Overs**, Assistant Hospitalo-Universitaire au sein du service d'oncobiologie génétique bio-informatique au CHU de Besançon, sans ton travail, cette thèse n'existerait pas. Je te remercie de ton encadrement avec de multiples séances de *brainstorming*, souvent impromptues, et de ta réactivité à mes multiples questions. Ce fut un plaisir de travailler avec toi.

À notre maître et juge, **Monsieur le Professeur Paul Kuentz**, Professeur des Universités - Praticien Hospitalier au sein du service d'oncobiologie génétique bio-informatique au CHU de Besançon, merci de ton encadrement sur nos questions parfois épineuses et d'avoir fourni le terreau intellectuel de cette thèse. J'espère qu'elle contribuera à l'activité diagnostique future en génétique constitutionnelle.

À notre maître et juge, **Monsieur le Professeur Didier Hocquet**, Professeur des Universités - Praticien Hospitalier au sein du service d'hygiène hospitalière au CHU de Besançon, merci d'avoir accepté d'examiner cette thèse dont j'espère qu'elle pourra qu'elle pourra servir à développer l'activité bio-informatique de votre unité.

À notre maître et juge, **Monsieur le Docteur Éric Dahlen**, Praticien Hospitalier au sein du service d'oncobiologie génétique bio-informatique au CHU de Besançon, merci d'avoir accepté de juger ce travail sur lequel ton expertise est dûment appréciée.

À notre maître et juge, **Monsieur le Docteur Rémi Mathevet**, Praticien Hospitalier au sein du service de foetopathologie au CHU de Besançon, merci enfin de ton regard de généticien médical et foetopathologiste sur ce sujet à la frontière de la génétique moléculaire et la bio-informatique.

L'équipe du **méso-centre de Franche-Comté** a été particulièrement aidante pour sa réactivité et avoir osé tenter l'"expérience Nix", pierre angulaire de ce travail.

L'aide des **Drs Eric Dahlen et Virgine Roze** a été également plus qu'appréciée pour la conception des Sangers de cette étude, ainsi que celle du **Dr Mathieu Laeng** lors de l'écriture de script python permettant le téléchargement des données.

Enfin, plus personnellement, je remercie du fond du cœur **Laure**, pour m'avoir toujours soutenu et stimulé tout du long des études de médecine et bien avant, sans qui je ne serai pas là aujourd'hui.

SERMENT D'HIPPOCRATE

En présence des Maîtres de cette École, de mes chers condisciples, je promets et je jure, au nom de l'Être Suprême, d'être fidèle aux lois de l'honneur et de la probité, dans l'exercice de la Médecine.

Je donnerai mes soins gratuits à l'indigent, et n'exigerai jamais un salaire au dessus de mon travail.

Admis dans l'intérieur des maisons, mes yeux ne verront pas ce qui s'y passe, ma langue taira les secrets qui me sont confiés, et mon état ne servira pas à corrompre les mœurs, ni à favoriser le crime.

Respectueux et reconnaissant envers mes Maîtres, je rendrai à leurs enfants l'instruction que j'ai reçue de leurs pères.

Que les hommes m'accordent leur estime si je suis fidèle à mes promesses !

Que je sois couvert d'opprobre et méprisé de mes confrères si j'y manque !

Sommaire

Glossaire

***BAC* – Bacterial Artificial Chromosomes**. 26, 29

***CNV* – Copy Number Variation**: Modification du nombre de copies d'un segment d'ADN 11, 36

***ESE* – exonic splicing enhancer**. 43

***FOSDEM* – Free and Open source Software Developers' European Meeting**. 58

***GA4GH* – Global Alliance for Genomics and Health**. 75

***GATK* – Genome Analysis Toolkit**: Un ensemble de logiciels pour l'analyse de données de NGS développé par le Broad Institute 3, 9, 17

***GIAB* – Genome In A Bottle**. 72, 77, 98

pipeline: En bioinformatique, ensemble d'étapes exécutées dans un ordre prédéfini pour traiter et analyser des données. 2

***GRC* – Genome Reference Consortium**. 26

***GWAS* – Genome-Wide Association Studies**. 37, 38

***HPO* – Human Phenotype Ontology**. 90, 91

***HSat* – Human satellite repeat array**: Grandes régions d'ADN non codant répétees en tandem. Il s'gait du principal composant des centromères 30

***indel* – Insertion-deletion**. 17, 19, 20, 22, 23, 33, 36

***MAF* – Minor allele frequency**: Fréquence de la seconde allèle la plus fréquente. Permet de qualifier les variants de rares (ex: 0.05). Voir https://gatk.broadinstitute.org/hc/en-us/articles/360039984151-DRAGEN-GATK-Update-Let-s-get-more-specific 23, 37, 49

***RCP* – Réunion de Concertation Pluridisciplinaire**. 2

Sensibilité: Vrais positifs divisé par Vrais positifs + faux négatifs 22

***SNP* – Single Nucleotide Polymorphism**. 41

***SNV* – Single Nucleotide Variant**: Modification d'un seul nucléotide 7, 9, 11, 17, 19, 20, 22, 29, 36

***UCSC* – University of California, Santa Cruz**. 54

***VCF* – Variant Call Format**. 73

***VEP* – Variant Effect Predictor**. 36, 37, 38

***VM* – Virtual Machime**. 55, 56, 69

***VPP* – Valeur Prédictive Positive**: Vrais positifs divisé par vrais positifs + faux positifs 22, 48

***VSI* – Variant de Signification Indéterminée**. 2, 3, 9, 85, 95, 96, 98

Introduction

Au sein du service de génétique médicale de Franche-Comté, l'offre de soins proposée pour les patients atteints ou suspects de maladies rares est variée. En effet, la patientèle est hétérogène avec une majorité de consultations pédiatriques, mais également des adultes. On peut estimer à 2/3 les consultations pour troubles du neurodéveloppement, mais il existe aussi une demande pour des pathologies cardiaques, ophtalmologiques par exemple. Bien que de taille restreinte, des consultations en ante et post-natal sont proposées. La génétique joue un rôle crucial pour certaines pathologies avec un diagnostic urgent qui va impacter le traitement pour des patients en réanimations ou avec une amaurose congénitale de Leber.[1] En tant que centre de référence de maladies rares, le service propose donc aux patients franc-comtois une expertise locale contribuant à l'égalité des soins sur le territoire français.

Depuis 2017, l'exome est au cœur du diagnostic des maladies rares dans le service bisontin. Dans le cadre d'un bilan constitutionnel après un bilan préliminaire (ACPA, caryotype). il est souvent prescrit car le service médical rendu est intéressant au vu du recrutement des patients et de son rendement diagnostic, estimé à 32%. Pour des raisons pratiques, le CHU de Besançon a choisi de sous-traiter le séquençage et l'interprétation de cette analyse à un laboratoire médical privé allemand. Cela est au bénéfice du patient, car les délais de rendus sont rapides avec 4 semaines pour les interprétations en urgence et pour le laboratoire, une tarification avantageuse. Le laboratoire sous-traitant est accrédité par le *College of American Pathologists* et selon la norme ISO 15189. Depuis 2022, le laboratoire de génétique de Besançon dispose des données brutes de séquençage pour permettre l'interprétation ou ré-interprétations des résultats de signification indéterminée ou sans cause génétique retrouvée.

Le laboratoire de génétique de Besançon est composé de 2 médecins et d'une ingénieure avec une activité mixe cytogénétique et biomoléculaire. Cette dernière consiste

[1] Il s'agit d'une diminution sévère de l'acuité visuelle, voire une cécité due à une dystrophie rétinienne. Des essais de thérapie génique sont actuellement en cours, mais nécessitent une prise en charge rapide.

d'une part à étudier la transmission d'un variant génétique dans une famille pour aider le diagnostic de VSI (*Variant de Signification Indéterminée*). D'autre part, la prestation de conseil auprès des cliniciens dans de cadre d'une RCP (*Réunion de Concertation Pluridisciplinaire*) à la fois sur l'interprétation des résultats sous-traités ainsi que des analyses menées au sein du laboratoire. Dans les cas d'errance diagnostique, la réinterprétation des données brutes joue un rôle important. Cela passe par l'utilisation d'un pipeline bio-informatique interne développé par le Dr. Alexis Overs.

Ce pipeline de réinterprétation d'exome est donc à l'interface entre le diagnostic et la recherche car il permet d'explorer les diagnostics difficiles ou sans cause retrouvée. La majorité des diagnostics est faite par le laboratoire sous-traitant, mais la réinterprétation a son rôle au vu de l'évolution rapide des connaissances en génétique et de l'évolution clinique des patients, par exemple sur le plan du neurodéveloppement. On peut par exemple ré-interpréter les données pour de nouveaux gènes. En cas de suspicion diagnostique, un séquençage par la technique de référence, ou Sanger, est effectué au sein du CHU qui est accrédité pour cette technique par le COFRAC.

Développer un pipeline interne est une entreprise pluridisciplinaire conséquente demandant des compétences biologiques, médicales et informatiques. En effet, il ne suffit pas de produire du code informatique, mais aussi de s'assurer de sa fiabilité, pertinence et performance. Dans notre cas, les performances analytiques du pipeline ne doivent pas être inférieures à celle du laboratoire sous-traitant. À ce titre, le rôle de l'informaticien permet la maîtrise des complexités logicielles et matérielles, tandis que le médecin apporte le versant clinique avec la compréhension physiopathologique et la correspondance avec la clinique du patient. Enfin, le biologiste offre une expertise sur les versants pré-analytique, analytiques et post-analytiques permettant de comprendre les enjeux, difficultés et limites de l'interprétation. Le pipeline développé initialement et amélioré dans le cadre de cette thèse résulte d'une collaboration entre ces trois domaines.

Les problématiques dans ce contexte sont multiples. L'*interprétation* des données d'exome est délicate pour les patients atteints de maladies rares et nécessite obligatoirement un biologiste car elle ne peut encore être complètement automatisée. Par définition, il y a peu de cas rapportés, ce qui nécessite une recherche bibliographique constante et poussée. Les réseaux d'expertise biologique comme NGS-diag sont également utiles pour échanger sur des cas difficiles. Pour déterminer la causalité entre un variant et les symptômes d'un patient, le biologiste va se baser sur son expertise ainsi que de nombreuses bases de données. Celles-ci ont fortement évoluées au cours des dernières années en nombre et qualité, nécessitant une veille technologique régulières. Il est cependant impératif pour la qualité des soins et l'exactitude de l'interprétation d'utiliser les versions les plus à jour . Il existe bien des études fonctionnelles,

dans lesquelles la conséquence d'un variant génétique est étudiée en laboratoire, par exemple sur des cellules de patients ou sur modèle murin. Ce domaine est moins développé qu'en oncologie, où l'utilisation de la technologie CRISPR est utilisée, par exemple, pour introduire dans des lignées cellulaires des variants de *BRCA1*, testant ainsi leur impact fonctionnel (Kweon et al. 2020). Ce déficit en études fonctionnelles en génétique constitutionnelle souligne donc l'importance d'une revue manuelle par des experts des bases de données pour assurer leur qualité.

S'agissant de biologie médicale, une autre contrainte est lié à l'utilisation en *diagnostic* des résultats. Il y a ainsi une pression importante pour diminuer le délai de rendu, souvent long en génétique constitutionnelle, avec un nombre de demandes importantes. Certains patients sont suivis depuis de nombreuses années et ont bénéficié de multiples explorations sans cause génétique retrouvée. En cas de forte suspicion clinique, il est donc important mais difficile de ré-interpréter les données disponibles en profitant de l'évolution technologique et scientifique.

Une troisième problématique concerne le *volume* de données. L'exome est la partie du génome qui code pour les protéines, ce qui représente environ 1% des 3 milliards de paires de bases du génome répartis sur 20 000 gènes environ[2]. Contrairement aux analyses ciblées, il y a une quantité importante de données que doit traiter et filtrer le pipeline. Si les filtres sont trop permissifs, l'interprétation par le biologiste sera trop compliquée au vu du nombre de variants candidats. À l'inverse, des filtres trop restrictifs peuvent conduire à un diagnostic manqué du fait de l'exclusion du variant par les filtres. La charge de calcul est donc relativement conséquente : notre pipeline s'exécute ainsi en 5h pour un patient une fois optimisé (voir Chapitre 2.3). Le nombre important de variants va donc conduire à de nombreux VSI, d'interprétations délicates. Se posent également des problématiques éthiques concernant la découverte de données secondaires ou incidentes, c'est-à-dire de variants pathogènes pour des maladies non recherchées par le clinicien.

Pour répondre à ces problèmes, le laboratoire peut choisir de sous-traiter l'analyse bio-informatique, la conception du pipeline ou bien le développer en interne. Quelque soit la solution choisie, le laboratoire doit s'assurer de la validité des résultats (Chapitre 3). La solution retenue a été de concevoir notre pipeline à l'aide de composants open-source afin de maîtriser le processus bio-informatique par opposition à un algorithme propriétaire. S'il existe quelques pipelines « prêts à l'emploi », la multiplicité des différents algorithmes rend difficile le choix d'un seul outil. Les bases de données sont nombreuses et le choix est loin d'être évident. Un guide de bonnes pratiques a été publié dans le cadre du GATK (*Genome Analysis Toolkit*) (Auwera et al. 2013) mais

[2] https://www.gencodegenes.org/human/stats.html, accédé le 5 mars 2024.

cela ne concerne pas l'intégralité du pipeline. Notamment, le choix des filtres dépend des objectifs analytiques de la technique et des propriétés du séquençage. À ce titre, il relève de l'expertise du laboratoire et nécessite une attention toute particulière lors de la validation. Enfin, le choix d'une solution dépend fortement des ressources humaines et matérielles. Dans tout le processus, le rôle du biologiste est indispensable, à la fois pour la conception et développement et évidemment pour la validation.

Cette *validation* de méthode est indispensable pour un laboratoire de biologie médicale, car il s'agit de respecter la législation en cours sur les bonnes pratiques imposant au laboratoire d'être accrédité, y compris sur la partie bio-informatique. Ces contraintes sont justifiées par la nécessité de résultats fiables en diagnostic. Dans le contexte bisontin, le séquençage est sous-traité donc seule l'extraction de l'ADN est maîtrisée par le laboratoire. Selon les recommandations nationales et internationales (Chapitre 3), la validation doit se faire sur des échantillons et peut éventuellement être complétée par une validation purement informatique.

Les **apports de cette thèse** sont de concevoir et tester un pipeline d'interprétation de données d'exome. À ce titre, une revue de la littérature des différents algorithmes et bases de données avec une justification des choix retenus sont présentées dans le Chapitre 1. Un effort conséquent a été fait pour rendre l'installation et l'utilisation 100% reproductible et portable avec une exécution possible sur supercalculateur (Chapitre 2). La validation a également été exhaustive avec un échantillon de référence et une validation purement bio-informatique sur des données de références ainsi que sur des données « synthétiques » dans le Chapitre 3. L'impact clinique de la réinterprétation est discutée dans le Chapitre 4.

Chapitre 1

Pipeline

1.1 Introduction

Si le séquençage initial du génome a été initialement long et coûteux avec la technique de référence par Sanger, l'apparition du séquençage haut-débit au début des années 2000 a été une révolution technologique permettant un rendu très rapide et bien plus économique. On peut désormais parler de génome à moins de 1 000$ pour un patient (Wetterstrand, s. d.). On distingue habituellement les technologies selon la taille des fragments générés : de petite (*short read*) ou de grande taille (*long read*). Le séquençage de nouvelle génération peut être divisé en 3 étapes : pré-traitement de l'échantillon, préparation de la librairie, le séquençage proprement dit et l'analyse bio-informatique (Fig. 1). Cette introduction présente un aperçu des défis de ces différentes étapes.

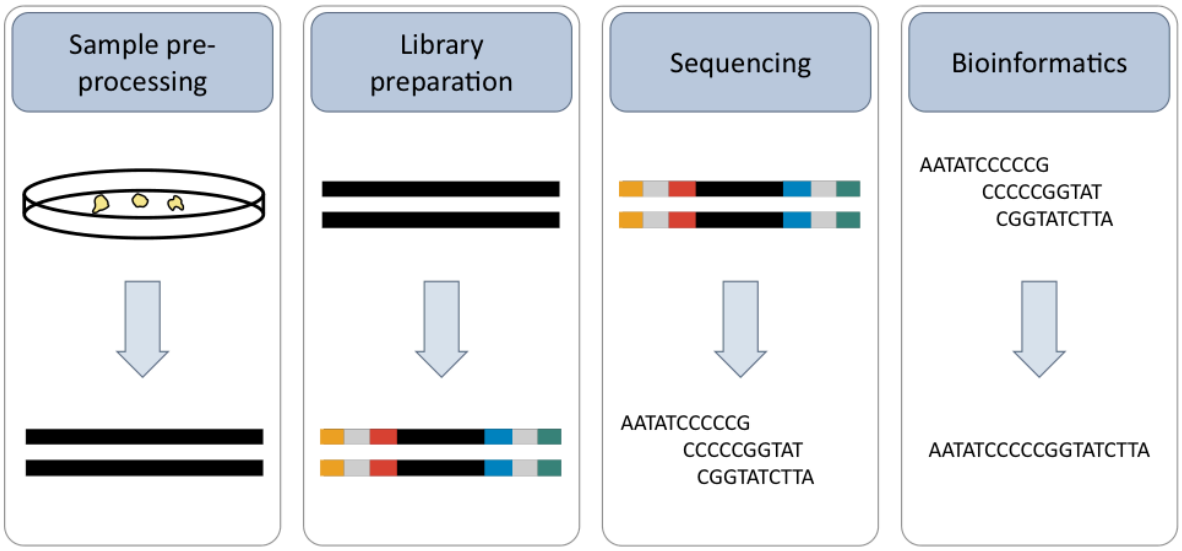


Fig. 1. – Principales étapes d'un séquençage de nouvelle génération (Head et al. 2014).

La préparation des librairies consister à générer des fragments d'ADN de tailles prédéfinies avec des adaptateurs à leurs extrémités 5′ et 3′. Si les protocols sont nombreux (voir en annexe la Fig. 62), on peut identifier 4 étapes communes. Tout d'abord, après extraction de l'ADN, celui-ci est fragmenté de manière, soit mécanique (par ultrason par exemple), soit enzymatique. Les fragments de tailles souhaitées sont ensuite séparés des éléments non désirés, par élution sur bille magnétique, colonne ou gel. Puis une amplification par PCR permet d'ajouter des adaptateurs (voir ci-dessous) aux fragments et d'augmenter la quantité d'ADN disponible. Après une autre étape de lavage, un contrôle de qualité est effectué par le dosage de l'ADN obtenu. La préparation de librairies est une étape délicate du fait de sa complexité, de son coût et d'une possible contamination inter échantillons. De plus, l'amplification par PCR peut conduit à une sous-représentation des régions difficilement amplifiables, comme les régions riches en GC, malgrés des progrès importants.

Une fois les librairies préparées, le séquençage en tant que tel peut être effectué. Les technologies dites *short-read* sont représentées par Illumina avec ses séquenceurs MiSeq, NextSeq, NextSeq etc. et ThermoFischer avec Ion Torrent. Le principe consiste à faire une amplification clonale soit avec des *beads* pour Ion torrent ou sur des *flow cells* pour Illumina. L'amplification se fait par *émulsion PCR* (ion torrent Fig. 2) ou *bridging* (Illumina, voir Fig. 3) qui est une amplification dite en phase solide. Ion torrent a cependant des difficultés à séquencer les homopolymères (voir Fig. 61 pour une illustration d'un hompolymère) mais les taux d'erreur par paires de bases sont inférieurs à 0.1%. Illumina a également un taux d'erreur bas de 0.1% avec des temps de séquençage initialement longs sur la génération des MiSeq mais qui furent diminués

par la suite.[3]. En terme de coût, le coût d'achat pour Nextseq1000 a été estimé à 325 000$ par Hu et al. (2021) mais cela peut varier selon le type de contrat souscrit.

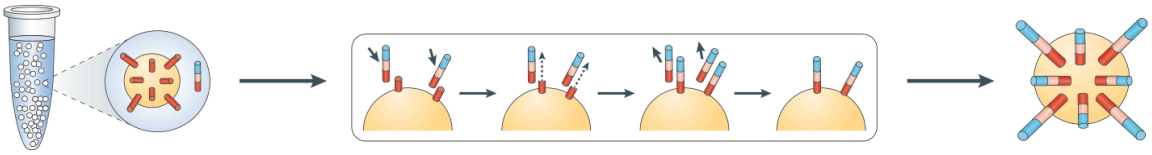


Fig. 2. – Amplification par émulsion pour la technologie Ion Torrent. Les gouttelettes sont d'abord chargées avec des billes recouvertes d'adaptateurs, un ADN matrice, des dNTPs et une polymérase. Puis l'ADN matrice va s'hybrider sur les adaptateurs qui vont servir d'amorces pour la polymérase et élongation d'un brin complémentaire. Enfin l'ADN matrice va se dissocier de la bille. Au final, il y aura entre 100 et 2000 de billes avec des centaines de *templates* fixés sur chacune. (Goodwin, McPherson, et McCombie 2016)

A l'inverse, les séquenceurs *long-read* permettent d'avoir des lectures de plusieurs milliers de paires de bases, mais avec des taux d'erreurs initialement très élevés. Le Pacbio de Pacific Biosciences a une précision de 98% (Hu et al. 2021)[4] avec des fragments de 60k pb . De plus, il n'y a pas d'amplification PCR lors de la préparation de la librairie, ce qui permet d'éviter les biais de PCR. Les temps d'exécutions peuvent aller jusqu'à 20-30h. Le coût du séquenceur est d'environ 25 000$. *A contrario*, Oxford nanopore propose des technologies à très longs fragments, supérieurs à 1mb, avec les mêmes avantages que PacBio mais plus légères. En effet, les différents séquenceurs pèsent entre 450g et 25kg[5], à comparer au 360kg du Pacbio et 140kg du NextSeq1000. En revanche les erreurs d'appel de base montent entre 2 et 15% et, du fait de leur caractère systématique, ne peuvent pas être facilement réduites en augmentant la profondeur.

[3]Voir https://www.illumina.com/systems/sequencing-platforms.html pour une illustration des différents séquenceurs d'Illumina

[4]Avec 99.1% de précision pour les SNV (*Single Nucleotide Variant*), 95.98% pour insertion et délétion et 95.99% sur un génome de référence (voir les références fournies dans (Hu et al. 2021))

[5]Plus une unité d'acquisition de 25kg

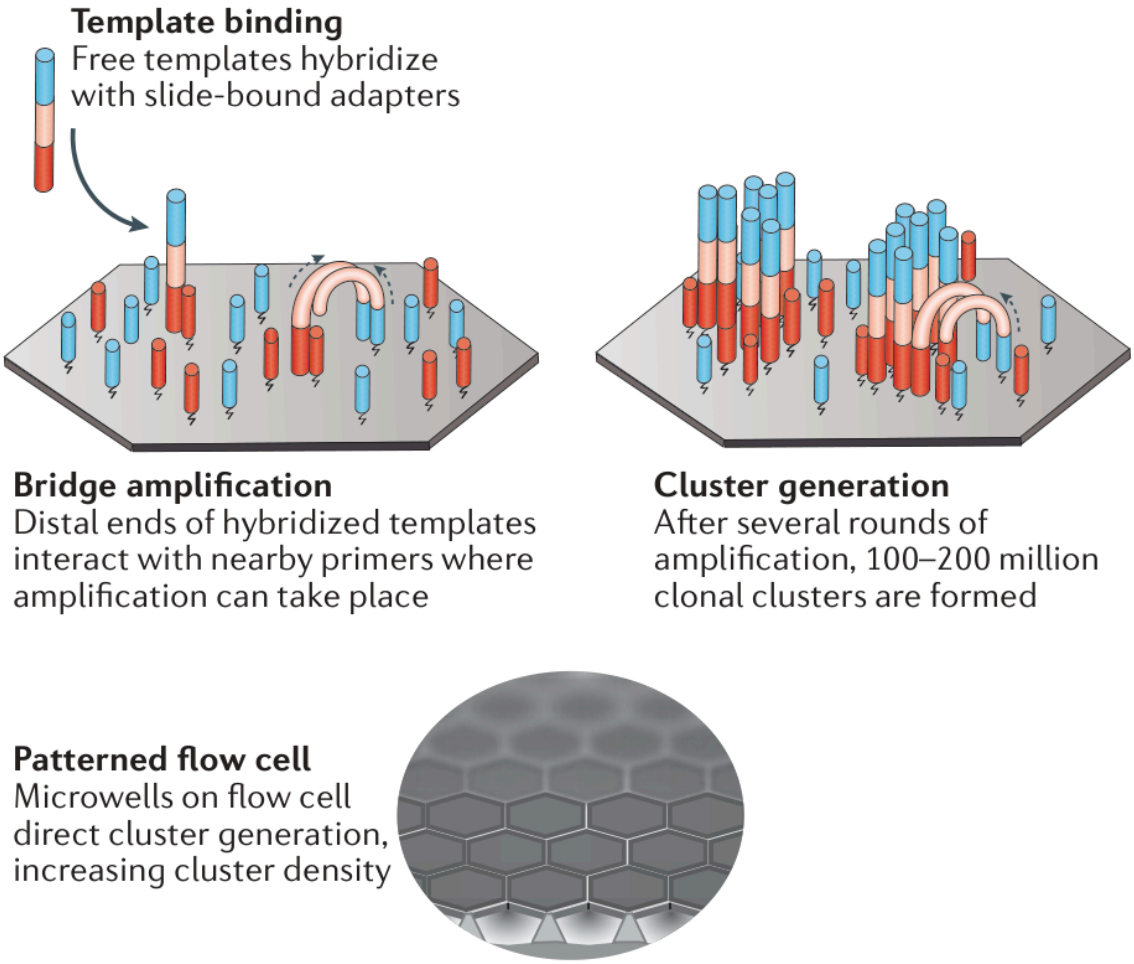


Fig. 3. – Amplification par *bridging* par la société Illumina (Goodwin, McPherson, et McCombie 2016). 1) Des adaptateurs sont attachés à des brins d'ADN. Ces adaptateurs permettent l'hybridation avec des amorces fixées sur un support solide (*flowcell*) et une polymérase génère un brin complémentaire. 2) Son extrémité restant libre, il va former un pont avec une amorce proche et une la polymérase pourra de nouveau générer un brin complémentaire. 3) Après amplification, les brins complémentaires seront supprimés et une des extrémités des « ponts » libérée pour faire le séquençage proprement dit avec insertion de nucléotides fluorescent, résultat en une image. Voir https://www.youtube.com/watch?v=fCd6B5HRaZ8 pour une illustration par Illumina.

Sur le plan bio-informatique, ce chapitre discute des problématiques liées au choix et à l'assemblage des différents composants d'un pipeline. En effet, la multiplicité des différents outils, tous de bonne facture, ne rend pas aisée le choix de l'aligneur et de l'appel de variant. Le choix des bases de données et outils pour annoter les données afin d'aider l'interprétation est complexe. Enfin, la définition de filtres pertinents né-

cessite un soin tout particulier. Il n'existe pas de pipeline général, car les laboratoires peuvent avoir besoin de l'adapter à leur situation, même si certains efforts vont dans ce sens (Chapitre 2.2). Cependant, il existe des « bonnes pratiques » pour l'ordre des étapes jusqu'à l'appel de variant établi par GATK illustrées dans Fig. 4.

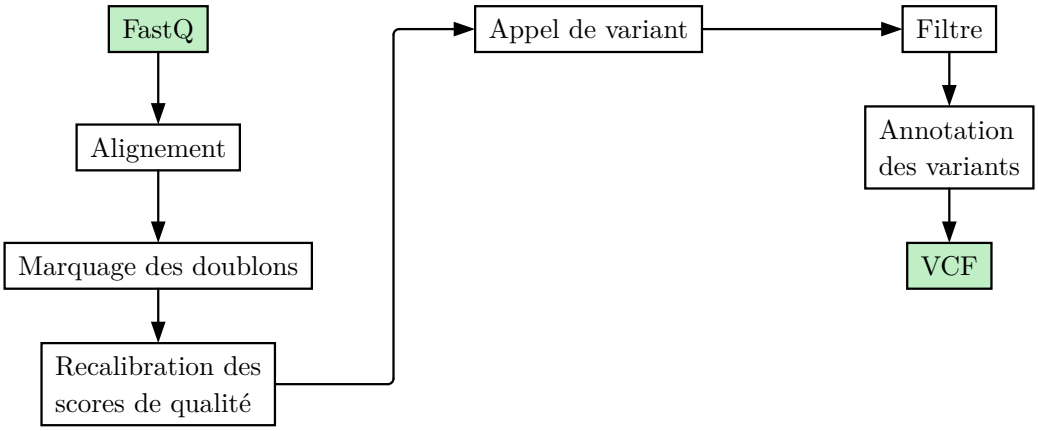


Fig. 4. – Pipeline simplifié selon les bonnes pratiques recommandées par GATK

Dans le cadre des demandes d'exomes au laboratoire de Besançon, , l'ADN est extrait à partir de leucocytes du sang au sein du laboratoire puis envoyé à Centogène, le laboratoire sous-traitant. Celui-ci va réaliser le séquençage de l'exome sur Illumina NovaSeq , avec un kit Twist Exome . La partie bio-informatique est également réalisée par le sous-traitant qui fourni un compte-rendu d'interprétation sous forme de PDFs. Depuis 2022, les données brutes (FASTQ, BAM et VCF) sont disponibles pour le laboratoire de génétique[6].

Afin de ré-interpréter les VSIs ou de trouver de nouveaux diagnostics, un pipeline interne nommé Bisonex a été développé en collaboration et sous la direction du Dr. Alexis Overs. Il vise donc à avoir un rendement diagnostic supérieur ou égal à celui de Centogène, c'est-à-dire ne pas manquer de diagnostics déjà posés. Dans l'état actuel de notre pipeline, les variants rendus sont limités aux SNVs et les insertions et délétions de moins de 10bp. En effet, les CNVs sont recherchés par caryotype ou CGH, qui sont plus fiables et plus adaptées. Enfin, la technologie illumina est ciblée par le pipeline.

Ce chapitre présente les différents algorithmes possibles et justifie le choix retenu de chaque élément . Chaque étape sera discutée dans l'ordre d'exécution du pipeline selon l'approche résumée dans la Fig. 4.

[6]Pendant une durée d'un mois

1.2 Alignement des données

Le séquenceur va générer un ensemble de fragments (appelés *reads* dans la suite) qui doivent être regroupés. Il existe 2 approches. La première va utiliser des séquences communes aux différents *reads* pour les regrouper : il s'agit d'un assemblage *de novo*, illustré sur la Fig. 5. La seconde va utiliser un génome dit de référence sur lequel on cherche l'emplacement le plus probable pour un *read* (Fig. 5). Le nombre de *reads* étant de l'ordre de la dizaine de millions, deux constats s'imposent. D'abord, tester toutes les combinaisons possibles est trop coûteux (approche par « force brute »). De plus, l'assemblage *de novo* n'est pas encore utilisé en diagnostic du fait de la complexité du génome, de la petite taille des *reads*[7] conduisant à des temps de calculs trop longs. La méthode de référence est donc d'aligner sur un génome de référence.

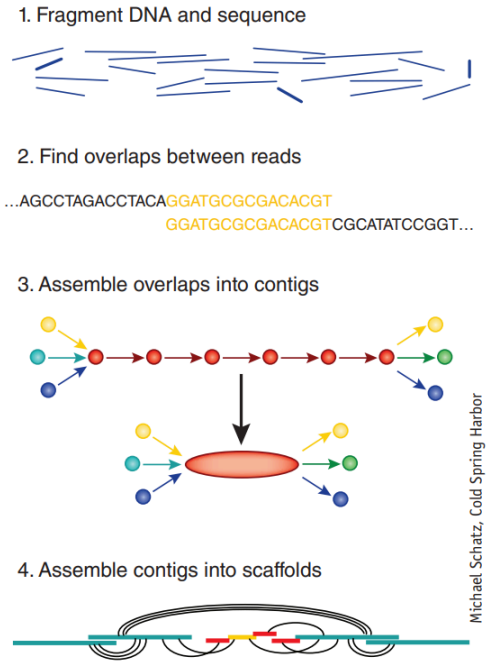


Fig. 5. – Alignement *de novo* en regroupant les reads en se basant sur les séquences communes (*contig*). En exploitant le fait que les reads sont en tandem, on peut trier ces *contig* en *scaffold* avec des intervalles vides (Baker 2012)

[7]L'utilisation de *long read* ne permet pas de s'affranchir de ces problèmes

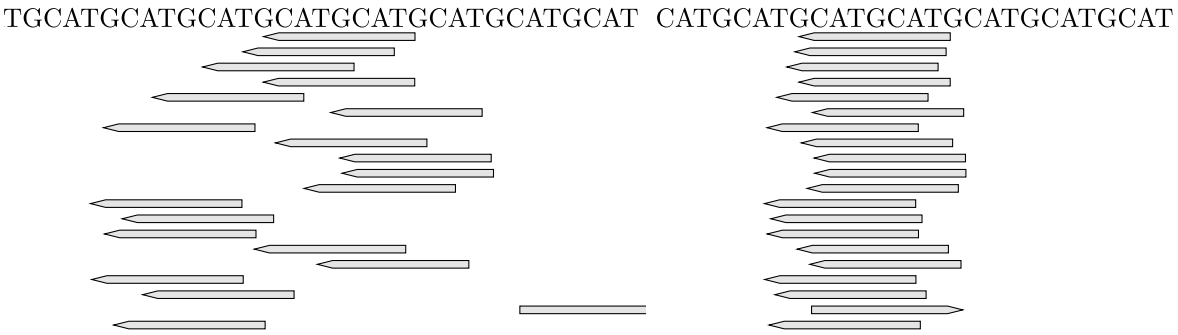


Fig. 6. – Principe de l'alignement de reads sur un génome de référence (en haut). À gauche, reads non alignés. À droite, les reads ont été alignés

Il existe différents problèmes lors de l'alignement. D'une part, le caractère incomplet du génome peut conduire à des reads mal ou non alignés. S'il existe un génome de référence, celui-ci n'est pas dépourvu de problèmes, comme abordé dans Chapitre 1.4. De plus, la gestion des régions répétées est difficile : un read pouvant être placé sur différentes régions, comment choisir la bonne ? BWA utilise un algorithme pseudo-aléatoire pouvant conduire à ce qu'un read de placement ambigu soit aligné à des endroits différents (Firtina et Alkan 2016) ! Il faut aussi pouvoir gérer les SNVs mais aussi les variants structuraux[8]. Enfin, pour des reads en tandem, il faut gérer les brins sens et anti-sens pour éviter les « biais de brins » pouvant conduire à des génotypes différents.

1.2.1 Bibliographie

Il existe de très nombreux algorithmes pour aligner des *reads* sur un génome de référence (voir en annexe le Tableau 18). On peut les décomposer en 3 étapes. L'*indexation* du génome va permettre de chercher une séquence rapidement. Celle-ci peut se faire par 2 approches. La première consiste à définir des courtes séquences qui vont servir de « marqueur » (*seed* en anglais). Pour chacun de ces marqueurs, la liste de ses positions dans le génome est stockée dans une structure adaptée (dite table de hachage). Ainsi, pour trouver l'emplacement d'un read, on extrait plusieurs marqueurs et on recherche leur position dans cette table de hachage (Fig. 7). Une seconde possibilité est de modifier la structure de stockage en combinant les séquences communes sous la forme d'un arbre (Fig. 7). Si la première approche par table de hachage permet une recherche très efficace avec un coût d'indexage faible, l'index généré est de grande taille. L'utilisation d'un arbre permet de chercher une correspondance partielle d'une séquence au détriment d'une recherche lente et d'un coût d'indexage élevé. La plupart des outils utilisant un arbre ont une approche modifiée diminuant

[8]Délétions, insertions, translocations, duplications CNVs (*Copy Number Variation*)

le coût en mémoire comme BWA-mem qui utilise un *FM-index*, lui-même une version modifiée de la transformée dite de Burrows-Wheeler.

En pratique, les performances des 2 approches ont été étudiées par Alser et al. (2021) mais de manière non parallélisée. Les auteurs concluent que la différence en temps d'exécution entre les 2 approches n'est pas statistiquement différente mais que le stockage en mémoire est plus important[9]. 3 outils couramment utilisés, BWA, bowtie et bowtie2 sont proches sur le temps d'exécution et le coût en mémoire.

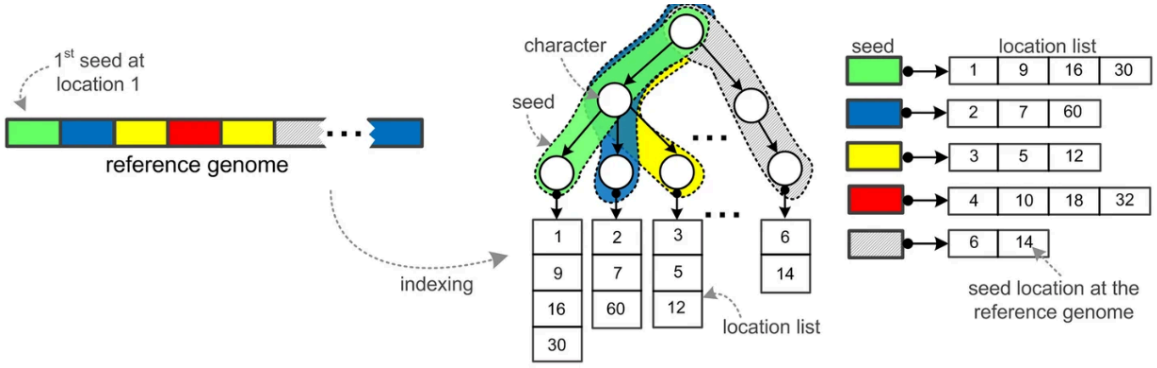


Fig. 7. – Sur le génome, on va définir des « marqueurs » (*seed*) (image de gauche). L'emplacement de ceux-ci sera stocké sous forme d'arbre (milieu) ou d'une table de hachage (droite) (Alser et al. 2021)

La seconde étape consiste à *lister les alignements possible* d'une séquence. On va donc prendre quelques marqueurs et générer une liste de positions possible dans le génome (Fig. 8). À noter que des marqueurs trop courts vont résulter en un trop grand nombre de candidats, nombre qu'il faudra diminuer avec des stratégies expérimentales. L'utilisation de marqueurs de plus grandes tailles pourrait sembler une alternative, mais en pratique, cela fait diminuer la sensibilité. La majorité des outils utilisent des marqueurs de tailles fixes. À noter que BWA-mem utilise une approche en deux temps pour ces marqueurs afin de gérer les mauvais alignements causé par des marqueurs absents.

[9]Ce résultat est attendu car une table de hachage est plus gourmande en mémoire

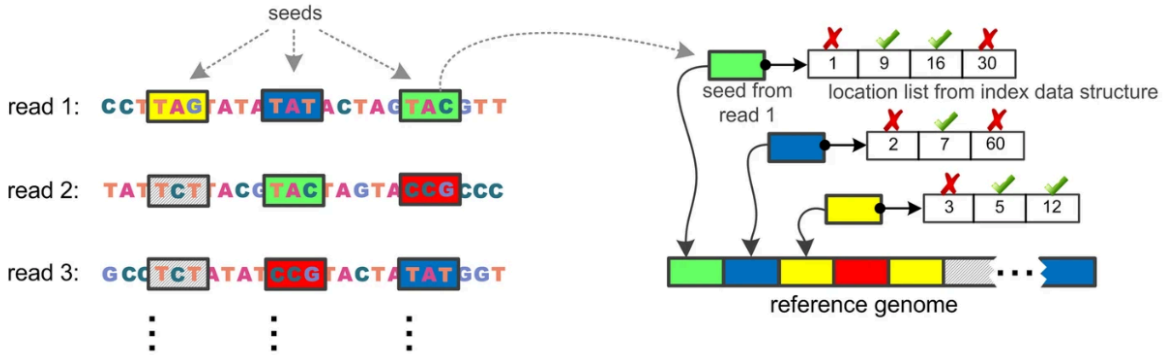


Fig. 8. – Extraction des « marqueurs » des différents reads (gauche) et recherche dans l'index du génome (droite)(Alser et al. 2021)

La dernière étape calcule la *similarité* entre le *read* et la portion du génome. Ce calcul peut se faire selon une approche *dynamique*[10] ou *non-dynamique*. Pour les substitutions, insertion, délétion, l'approche dynamique est privilégiée[11]. Le calcul de similarité reste une étape très coûteuse en temps et en stockage malgré de nombreuses recherches dans ce domaine.

Quelle seront les prochaines améliorations des aligneurs ? On peut citer le projet DRAGMAP qui vise à une collaboration entre le pipeline propriétaire DRAGEN s'exécutant sur FGPA pour améliorer les performances et GATK[12]. L'objectif est de proposer un aligneur open source nommé DRAGMAP[13] avec des fonctionnalités équivalentes à BWA mem. Dans l'état actuel, cet aligneur semble plus lent que BWA-mem (l'aligneur retenu pour Bisonex, voir ci-dessous) mais avec de meilleures performances sur une courbe ROC.

Un autre aligneur, Winnowmap2, s'attaque au problème des duplications segmentaires. Le principe est de comparer une partie du read avec la séquence la plus similaire et la seconde plus similaire. Cela semble diminuer le nombre de faux positifs et faux négatifs. Dans le même cadre, Duplomap va faire une liste *a priori* des différences entre les duplications segmentaires et calcule la plus longue séquence commune. Les résultats semblent être améliorés pour le Hifi de PacBio et ONT.

[10]Le principe est de décomposer ce problème en problèmes plus petits et plus rapides à résoudre. La solution finale est assemblée à partir de ces sous-résultats

[11]Pour les variants structurels, une correspondance partielle suffit donc on utilise alors une correspondance locale et non globale

[12]https://gatk.broadinstitute.org/hc/en-us/articles/360039984151-DRAGEN-GATK-Update-Let-s-get-more-specific

[13]https://gatk.broadinstitute.org/hc/en-us/articles/4410953761563-Introducing-DRAGMAP-the-new-genome-mapper-in-DRAGEN-GATK

1.2.2 Comparatif

Une des difficultés pour choisir un aligneur est que les différents articles comparant les pipelines ne s'attardent pas sur les aligneurs. Nous présentons ici les articles qui nous ont semblé les plus pertinents.

Pour évaluer la *popularité*, Alser et al. (2021) a calculé le nombre de citations de l'article initial. (Fig.). Les auteurs se sont arrêtés en 2020 mais les plus populaires sont BLAST, Star, Bowtie (1 et 2), BWA. Pour reproduire et mettre à jour les résultats, nous avons gardé cette métrique en utilisant le nombre de citations sur Pubmed. Les plus populaires sont par ordre décroissant Bowtie2, Bowtie, BWA et une augmentation récente de HISAT2 et minimap2.

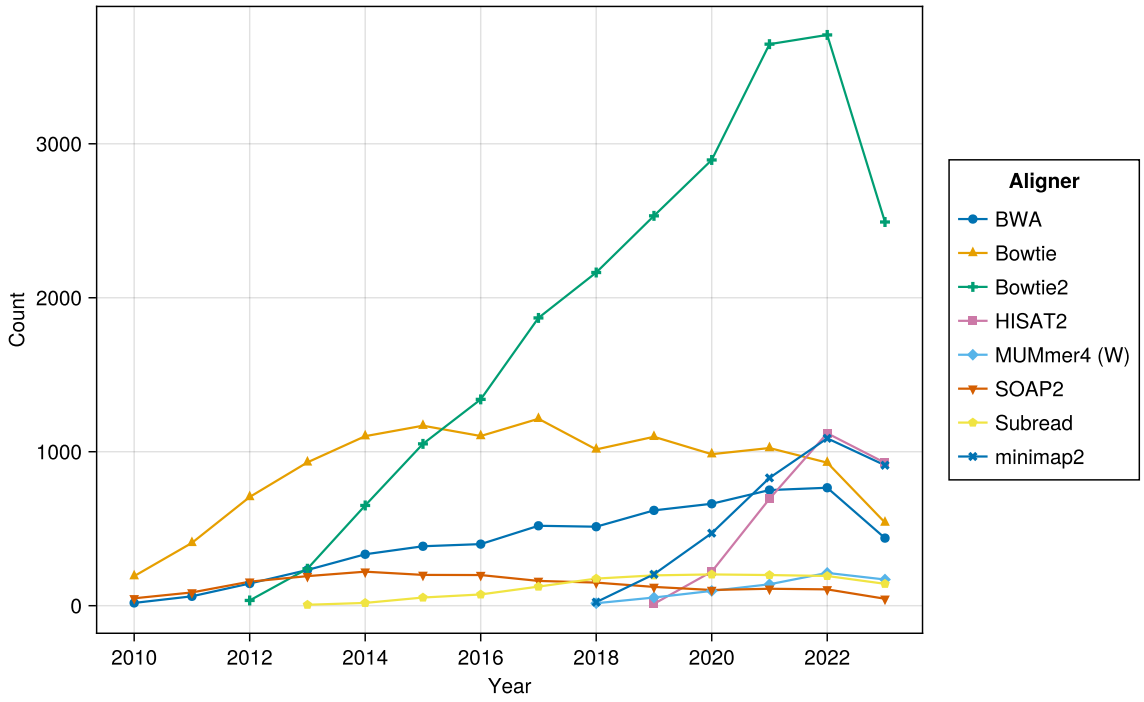


Fig. 9. – Nombre de citation par année selon Pubmed. Nous avons gardés les 8 plus populaires.

La qualité a été mesurée par Donato et al. (2021) en comparant 17 aligneurs sur données simulées d'humain sur le génome de référence GRCh38.p13 et de souris[14], mais aussi sur des données réelles avec 1 génome sur Ion torrent, 1 exome sur illumina HiSeq 2500. Sur les données simulées, les performances ont été comparables. Sur les données réelles, le plus grand nombre de reads alignés a été obtenu avec CLC, BWA-

[14]Les reads ont été générés avec une erreur aléatoire selon un modèle statistique basé sur la couverture, l'erreur de séquençage, la distribution des variants et le contenu en GC

MEM, GEM and Magic-BLAST. Les plus rapides[15] ont été subread puis minimap2, les plus lents Segemelhl, Tophat2 et Novoalign (Fig. 10). Les auteurs recommandent Segemehl and DNASTAR mais constatent qu'il n'y a pas d'aligneur idéal. Cet article est malheureusement limité par le faible nombre d'échantillons et l'absence de définition de la précision.

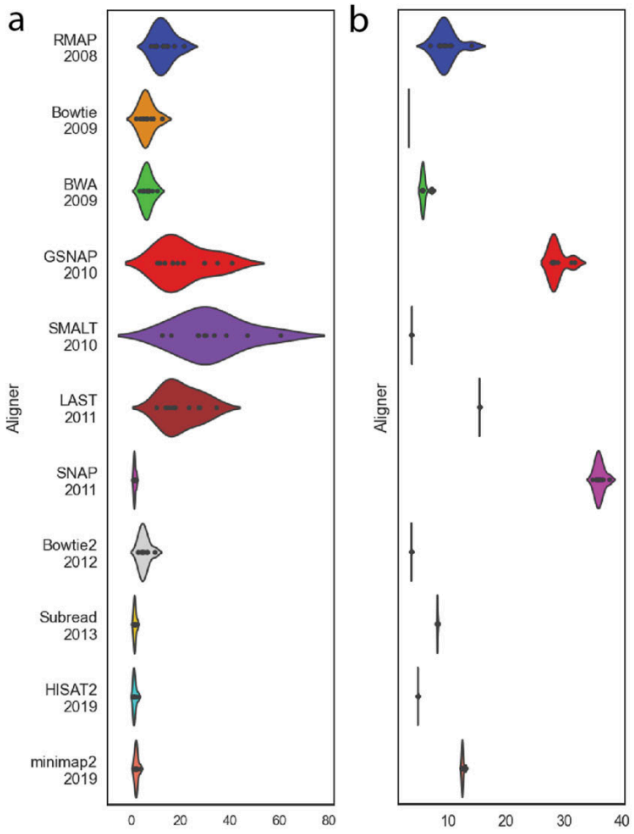


Fig. 10. – Temps CPU en heures (gauche) et coût mémoire en GB (droite) (Alser et al. 2021)

Pour les *performances* Alser et al. (2021) a étudié 10 génomes conférant ainsi une meilleure reproductibilité que Donato et al. (2021). Il retrouve les mêmes résultats que ce dernier pour les exomes avec Subread et Minimap2 parmi les plus rapides en temps cpu et BWA dans la moyenne. Les plus performants en mémoire sont BWA mem, HISAT2, Bowtie2 (Fig. 10). Pour la précision, BWA était également dans la moyenne.

Musich, Cadle-Davidson, et Osier (2021) a comparé BWA, Bowtie2, MUMmer4, STAR, HiSAt2 sur un génome de champignon en RNAseq. Concernant le taux d'alignement BWA et bowtie2 ont obtenus les meilleurs scores. Le temps d'exécution

[15]Sur un processeur i7 avec 6 cœurs et 32Gb de mémoire

par read a été le plus court pour HISAT2 avec des performances intermédiaires pour BWA, Bowtie2, star. Les performances en parallèle[16] ont été excellentes pour Bowtie2 avec une mention honorable pour BWA

Hatem et al. (2013) a aussi utilisé des données génomiques simulées[17] et des données issues de RNA-seq pour un génome de souris après avoir vérifié que les performances étaient comparables entre des génomes de différents animaux et un génome humain.[18] Concernant le taux de reads aligné correctement, bowtie a été le plus performant avec 93%. En seconde place, BWA et e troisième bowtie2 avec 85%. Le gain en performances avec un nombre de *threads*[19]. Les auteurs signalent que bowtie a le meilleur « débit » (*throughput*) , BWA est plus performant pour des reads plus longs. Mais surtout, ils concluent également qu'il n'y a pas d'aligneur idéal.

1.2.3 Choix retenu: BWA-mem

En conséquent, aucun aligneur ne semble s'imposer vraiment. Nous avons donc retenu BWA-mem qui est populaire, fiable et éprouvé par de nombreuses années d'utilisation. En terme de performances, il présente un bon compromis. On notera que l'alignement avec des séquences alternatives (ou ALT) du génome est possible avec l'ajout d'un programme supplémentaire et qu'il utilise un algorithme pseudo-aléatoire pouvant rendre la reproductibilité délicate. À noter, qu'il est recommandé par Regier et al. (2018) (voir ci-dessous) Enfin, les aligneurs semblent avoir moins d'impact que l'appel de variant, comme nous allons le voir dans la partie suivante.

1.3 Appel de variant

Cette partie propose une revue des différentes approches pour l'appel de variants, partie qui consiste à déterminer les variants des données de séquençage en les comparant à un génome de référence. Après une présentation des principaux algorithmes, nous détaillerons différents articles proposant une comparaison. Il est difficile d'échapper à un effet « catalogue » car ce sujet a été étudié de nombreuses fois, notamment dans des revues prestigieuses, mais il est difficile d'aboutir à un consensus.

[16]Sur un processeur Dual Xeon E5-2643 (six cœurs et 12 threads pour chaque cœurs) et 512 GB of RAM

[17]Utilisant une erreur de distribution uniforme) et d'autres simulant des données Illumina (ART))

[18]Les aligneurs suivants ont été testés : Bowtie, Bowtie2, BWA, SOAP2, MAQ, RMAP, GSNAP, Novoalign, et mrsFAST (mrFAST)

[19]Voir Chapitre 2 pour ces notions

1.3.1 Algorithmes

Haplotypecaller est l'outil développé par GATK dont le principe est résumé dans la Fig. 11. Dans un premier temps, les intervalles « candidats » sont déterminées à partir des données alignées. Cela est fait en déterminant une probabilité pour chaque position par comparaison à l'allèle de référence et en tenant compte de différents paramètres (SNV, indel (*Insertion-deletion*), soft-clip). Toutes les positions de probabilité supérieure à un seuil vont alors définir un intervalle[20]. Puis, pour chaque intervalle, les différentes combinaisons locales (*haplotypes*) possibles sont générées sous forme d'un graphe[21]. Dans un troisième temps, la probabilité pour chaque haplotype pour un read donné est calculée selon un modèle dit *Hidden Markov* par paires (*paired Hidden Markov Model*). Ce modèle permet de limiter les erreurs liées à la PCR si les données alignées ont été traitées selon les « bonnes pratiques » de GATK. À partir de cette probabilité, la probabilité d'un génotype est estimée[22]. Enfin, un modèle Bayésien est utilisé pour corriger cette probabilité *a posteriori* (Poplin et al. 2017).

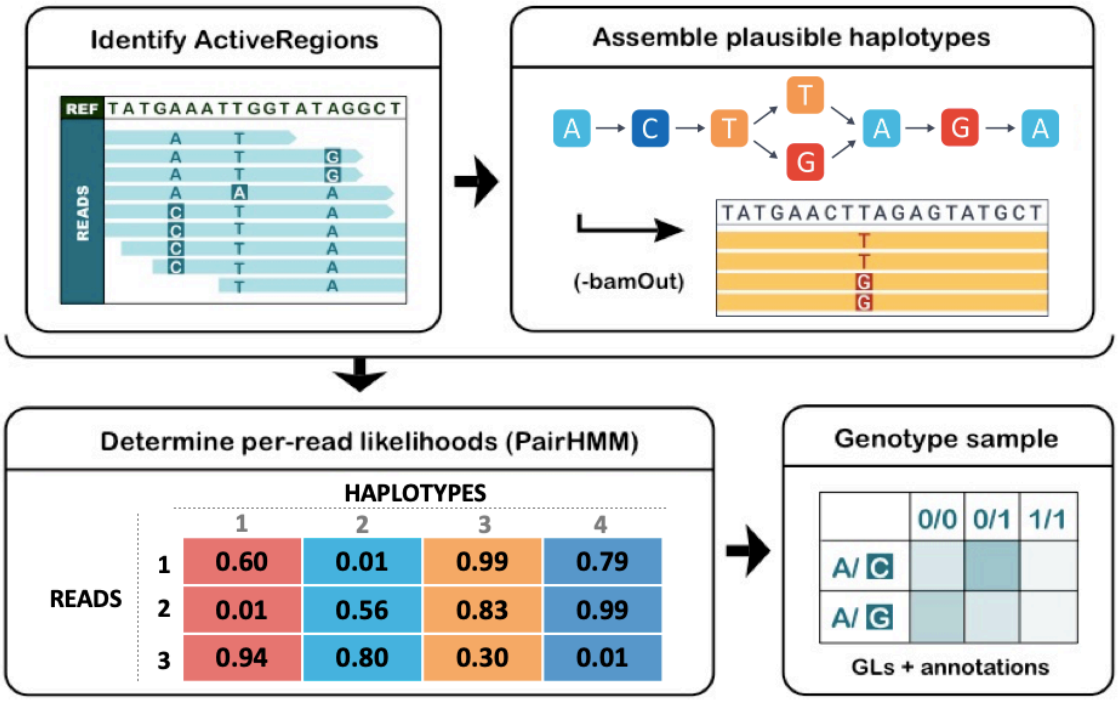


Fig. 11. – Ilustration des différentes étapes d'HaplotypeCaller selon https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell

[20]Les reads considérés par la suite sont ceux situés dans cette région avec une marge de 100 bp

[21]À noter qu'un haplotype qui ne « revient » pas sur la séquence de référence est supprimé des possibilités

[22]En prenant le produit sur tous les reads de la probabilité moyenne des allèles d'un génotype.

FreeBayes est une autre approche détectant les haplotypes à partir des reads mais présente deux particularités. Outre le fait qu'elle se base sur plusieurs échantillions, elle utilise les séquences des reads elles-même et non les alignements. Selon les auteurs[23], cela permet d'éviter les problèmes des séquences qui s'alignent à plusieurs endroits. Selon le modèle décrit dans Garrison et Marth (2012), on part de la distribution *a priori* d'allèles dans des individus d'une population donnée selon le modèle de la loi d'Ewens. Combinée à la qualité de séquencage, on peut calculer la probabilité d'un génotype (théorème de Bayes).[24]

En pratique, on va générer une liste préliminaire d'haplotypes en fusionnant les allèles suffisamment proches . Puis, on va définir un intervalle sur lequel des haplotypes locaux vont être recalculés. Cet intervalle est défini en déterminant l'allèle de représentation la plus grande possible, par exemple une insertion de plusieurs paires de base. L'intervalle sera ensuite construit de manière dynamique en incluant tous les haplotypes recouvrant partiellement cet intervalle initial, puis en augmentant sa taille afin de contenir complètement ces nouveaux haplotypes. Le processus sera répété tant qu'il reste des haplotypes de qualité suffisante[25].

Sur cet intervalle, les probabilités d'avoir un polymorphisme pour chaque position est estimée selon le modèle bayésien décrit plus haut. Enfin le génotype sera déterminé en calculant parmi tous les génotypes possibles celui qui maximise la probabilité *a posteriori*.

Strelka2 est un modèle qui va combiner ces 2 approches en utilisant le ré-assemblage local proposé par HaplotypeCaller et l'algorithme de FreeBayes (Kim et al. 2018). La valeur ajoutée de Strelka2 va se faire sur 3 points. D'abord, un filtre adapté est ajouté pour diminuer les perturbations engendrées par le décalage de phase dans les synthèses de nucléotides. De plus, ils évitent le coût de l'étape *pair Hidden Markov model* d'HaplotypeCaller en évitant de faire tous les alignements locaux possibles. Enfin, les auteurs proposent un nouveau score sur le variant à l'aide de machine learning, qui peut être utilisé pour le filtrer ou le prioriser à l'interprétation.[26]

Platypus utilise également une approche basée sur un assemblage local des données (Rimmer et al. 2014). Une liste de variants candidats est générée à partir des données

[23] https://github.com/freebayes/freebayes

[24] À noter que les formules présentées dans cet article font intervenir plusieurs plusieurs génotypes (donc plusieurs échantillons)

[25] Déterminés à partir d'un nombre minimum d'allèles alternatives et de la qualité des paires de bases

[26] Ce score est basé sur la probabilité du génotype, la qualité de l'alignement, le biais de brin, la complexité de l'alignement, comme la longueur des homopolymères. Le modèle a été entraîné sur le génome du patient NA12878.

alignées, puis la construction d'un grande de possibilités va permettre la création d'haplotype candidats. Une matrice de probabilité pour chaque read et chaque haplotype est calculée[27]. Ensuite, les probabilités des haplotypes sont initialement définies en fonction des probabilité estimée dans une population. Pour chaque haplotype, les génotypes et probabilité de chaque variant sont calculés mais en tenant compte des variants de la région. Enfin, les variants ne répondant pas à des critères de qualité[28] sont filtrés.

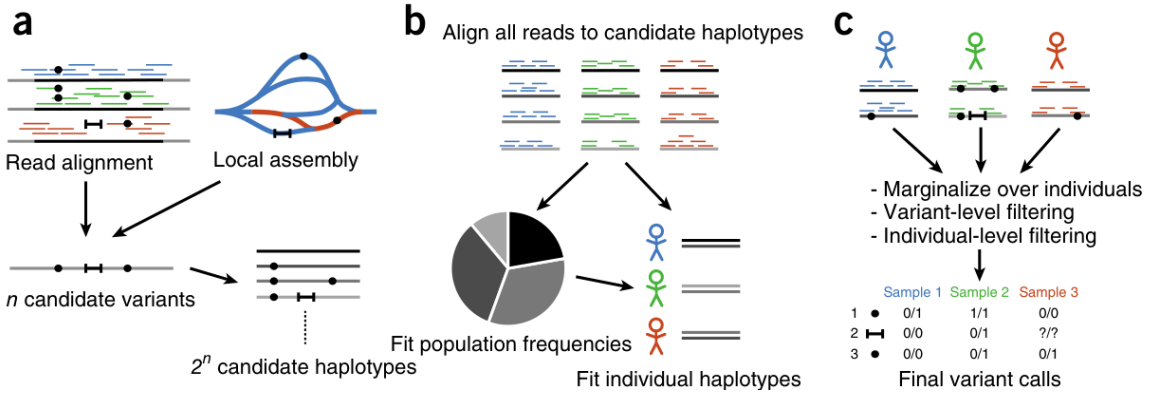


Fig. 12. – Algorithme de Platypus (Rimmer et al. 2014)

Selon les auteurs de l'algorithme de *DeepVariant*, HaplotypeCaller est plutôt orienté pour les séquenceurs d'Illumina mais peuvent être difficiles à adapter sur d'autres technologies (Poplin et al. 2018). Ils proposent donc une approche basée sur le *deep learning*. Le modèle est initialement entrainé sur des génotypes connus. Puis, les SNVs et indels candidats sont générés initialement avec les techniques précédentes (sensible mais peu spécifique). Puis, en utilisant l'accumulation (*pileup*) des alignements à chaque position, une probabilité est calculée puis le modèle de machine learning va donner un variant. Selon ces auteurs, HaplotypeCaller suppose que les erreurs de reads sont indépendantes alors que le réseau de neurones proposé ici prend en compte les dépendances complexes pour améliorer l'appel de variant.

[27]En utilisant un alignement avec un modèle d'erreur donnant la probabilité d'une indel en fonction du séquenceur, mais aussi des probabilités d'un SNV en fonction de la qualité du read

[28]Notamment en fonction de l'allèle, du biais de brin, de la qualité de l'alignement et de la paire de base, du contexte

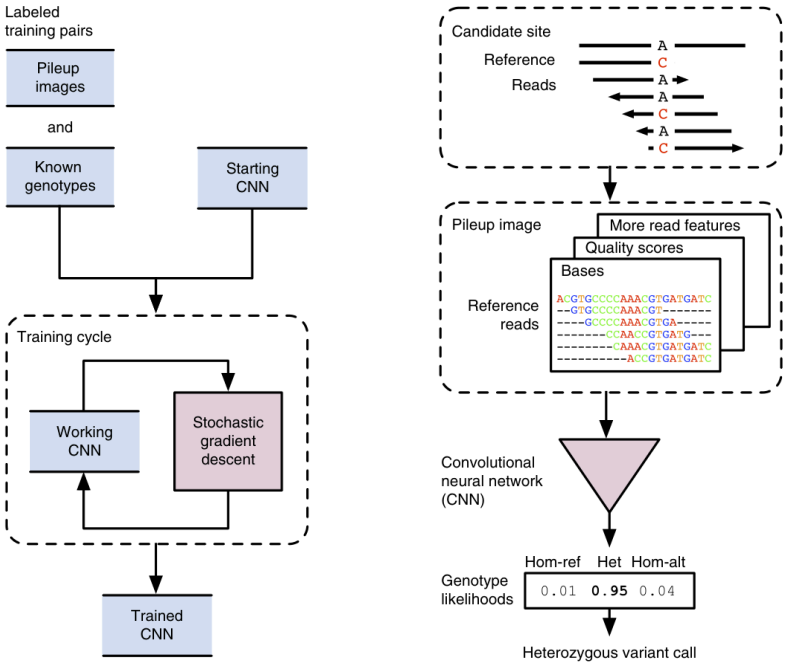


Fig. 13. – Après entraînement du modèle (gauche), l'appel de variant de Deepvariant est effectué à partir de variants candidats et d'une probabilité à chaque locus (Poplin et al. 2018)

1.3.2 Comparaison

Contrairement aux aligneurs, il ne nous a pas été possible de comparer la popularité des outils d'appels de variants car certains articles sont sur Arxiv. Le compte de citation est donc difficile et cela concerne Freebayes et HaplotypeCaller. Les articles les plus pertinents pour notre cas d'usage, à savoir les exomes en constitutionnels, sont présentés ici.

Hwang et al. (2015) a comparé 13 pipelines sur des données d'exomes pour le patient NA12878 (voir Chapitre 3.2). Plusieurs échantillons pour diminuer biais d'échantillonage ont été utilisés : Hiseq200 (7 données), Hiseq2500 (4 données) et ion proton (1 seul). Les pipelines correspondent aux combinaisons de 3 aligneurs (Bwa-mem, Bowtie2, Novoalign) et 3 outils d'appel de variant (HaplotypeCaller, samtools, Freebayes[29]. Leurs résultats montrent que la combinaison de BWA-mem et de samtools est le meilleur, mais avec une variabilité des résultats suivant les données brutes. Pour les indel, n'importe quel aligner combiné à HaplotypeCaller est le meilleur choix. Qu'en est-il du rôle respectif de l'aligneur et appel de variant ? Selon cette étude, l'appel de variant a un impact plus important que l'aligneur pour les SNVs et indels et surtout ces derniers. De plus, il y a une très bonne concordance entre les 3 outils

[29]Ainsi que l'appel de varant avec Ion proton mais non détaillé ici

d'appels de variants (Fig. 14). Notamment, les données d'Illumina ont un bon score de 92%, ce qui est en contradiction avec 2 autres études pour le séquenceur HiSeq2000. Les auteurs suggèrent que cette différence peut venir de la version des logiciels ou de pipelines non strictement identiques. Cela marque en tout cas l'important de la reproductibilité, objet de cette thèse. On note également une variabilité importante des données brutes avec un recouvrement variant de 82 à 97% !

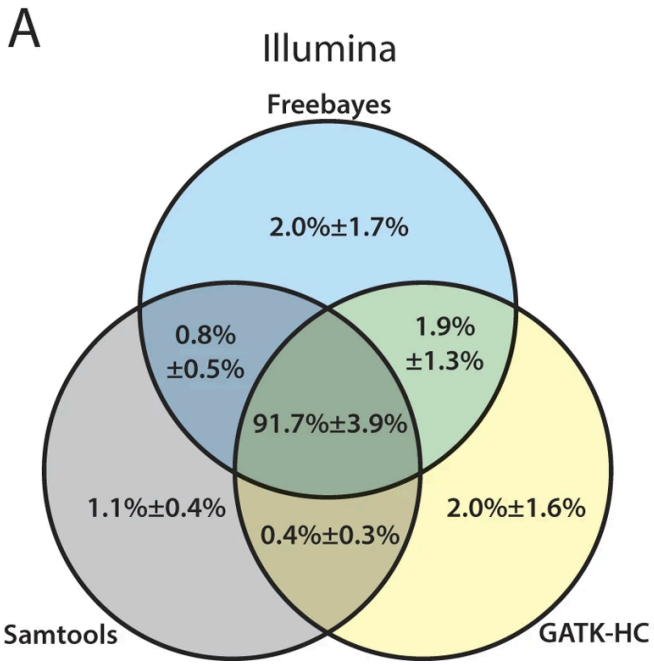


Fig. 14. – Concordance de 3 outils pour l'appel de variant (Hwang et al. 2015)

D'où viennent les erreurs par rapport à la référence ? La moitié des variants en réalité homozygotes sont rendus hétérozygotes car HaplotypeCaller a tendance à classifier en hétérozygote en excès[30]. Il faut donc prêter attention aux variants hétérozygotes avec HaplotypeCaller selon cet article. Au contraire, 36% des variants sont hétérozygotes mais rendus homozygotes car Freebayes a tendance à classifier en homozygote en excès surtout sur Illumina.[31]. Enfin, la couverture n'affecte pas significativement les résultats.

Cet article comporte plusieurs limites pour notre perspective. D'une part, le génome de référence est GRCh37. De plus, la normalisation des variants et la comparaison est faite respectivement avec vcflib et de manière « manuelle », sans les outils de référence développés après la publication et détaillé dans Chapitre 3.1. Il a possible-

[30]Probabilité entre 0 et 0.016 (mais très proche de 0 surtout !)

[31]Remarque supplémentaire : l'article note qu'il faut surveiller les variants de faible qualité (< 20) car ils peuvent faussement diminuer le nombre de faux positifs. Cependant, cela n'est pas applicable à HaplotypeCaller, car il ne considère que ceux de qualité > 30

ment une mauvaise gestion des variants complexes pouvant avoir des représentaions synonymes.[32]

Kumaran, Subramanian, et Devarajan (2019) ont testés 20 pipelines sur données d'exome (patient NA12878, NA24385, NA24631 et données simulées) en combinant 4 appels de variants (deepvariant, samtools, freebayes, gatk) et 5 aligneurs (bowtie, bwa, novoalign, SOAP,mosaik). Pour le patient NA12878, les données ont été également alignées en GRCh38, contrairement à l'article précédent. Les résultats pour tous les patients montrent que les pipelines basés sur deepvariant sont supérieurs à ceux basés sur HaplotypeCaller. Pour les indels sur le patient NA12878, BWA-mem et HaploteCaller occupe la troisième position. Les performances ont pu être améliorées comme attendu en fusionnant des 4 meilleurs pipelines avec 99-98% précision pour les SNVs et 96-98% pour les indels.

Concernant les facteurs pouvant influencer les résultats, la plupart des SNVs sont détectés à une profondeur de 150x et la profondeur a un profil similaire SNV et indel. L'augmentation du Genotype quality (GQ) améliore les performances. Les faux négatifs représentent moins de 1.5% des SNVs, moins de 4% des indels et sont dans des zones de profondeur < 30x et de GQ < 10[33] Les résultats sont inférieurs pour les indels, ce qui est attendu pour des données d'exome, possiblement à cause d'indels de taille importante. Les performances sont globablement améliorées en GRCh38 par rapport à GRCh37 avec une diminution du nomber de faux négatifs, de -8% et -20% pour SNV et indel respectivement.

Hwang et al. (2019) ont étudié 70 combinaisons[34] avec 7 aligneurs et 10 outils d'appel de variant sur un génome européen et surtout un génome africain. En revanche, les données restent en GRCh37. Si les articles précédents utilisait GIAB pour les variants de référence, cet article utilise Platinum Genome pour NA12878 et 1000 génomes pour les 2 patients. Nous avons inclus cet article comparant des génomes, car il porte plusieurs messages très intéressants du fait de la diversité de ces 2 génome et du grand nombre de comparatifs. D'une part, la majorité des variants sont appelé par la plupart des pipelines. Il y a peu de variabilité pour HaplotypeCaller probablement grâce à son algorithme de ré-alignement local. De plus, le type de variant (SNP vs indel) influe plus que les individus.

[32]Enfin, on note que la métrique utilisée est l'aire sous la courbe VPP (*Valeur Prédictive Positive*) - Sensibilité.

[33]On note également que le ratio heterozygote/homozygote est supérieur pour SNV que pour indele (1.6-1.5 vs 1.2-1.3) et que le ratio transitio/transversion: 3.4-3.3

[34]En pratique, cela a été représenté sous forme d'un tableau montrant pour chaque combinaison la « dissimimarité » avec les atures pipeline avec la distance dite de Jaccard

D'où viennent alors les différences ? Comme l'article précédent, la profondeur n'impacte pas les résultats[35]. Pour les indels seuls, la balance allélique joue un rôle. Le génome du patient d'origine africaine a entraîné plus de discordances entre les pipelines, surtout pour les variants rares ou de faible fréquence[36] et avec un impact fonctionne prédit par VEP important. Pour les deux patients, les zones répétées, la faible couverture et le contenu en GC sont des facteurs également.

On notera l'impact du choix des variants de référence. En prenant HaplotypeCaller comme exemple, les performances ont été améliorées avec l'utilisation de GIAB par rapport à 1000 génome, possiblement, car ce dernier jeu de données dispose de plus de variants. De même, combiner plusieurs outils d'appel de variant n'est pas plus performant avec GIAB ou Platinum genomes comme référence, ce qui va à l'encontre de l'étude précédente et d'autres. Notamment, BWA-mem et HaplotypeCaller est au moins aussi bon qu'une combinaison de pipeline pour les deux premières références.

Les limites de cet article sont également le problème d'harmonisation de représentations de variants qui sont ici seulement alignés à gauche. Il ne semble pas y avoir eu d'optimisation de paramètres des modèles. Enfin, on note des caractéristiques techniques différentes pour les 2 génomes (profondeur 49 et 72, longeur des reads 101 vs 250).

Au contraire, *Chen et al. (2019)* a testé un seul patient (NA12878) sur les plateformes BGI et Illumina en exome et génome (voir Tableau 19 pour les détails). avec BWA comme aligneurs et 3 outils d'appel de variants (HaplotypeCaller, Strelka, samtools-varscan2). En terme de qualité des exomes 95% ont moins de 1% d'erreur et 89% ont moins de 1%, (92% et 93% pour les génomes, respectivement). Strelka2 est supérieur pour les exomes mais HaploteCaller est parfois au moins aussi précis ponctuellement[37]. Des résultats similaires sont retrouvés pour les génomes avec 94.22% SNV détecté par au moins 10 combinaisons de pipeline et 90.63% des indels par au moins 13. Strelka2 est aussi le plus précis et avec le meilleur F-score. En terme de temps d'exécution[38] : Strelka2 est le plus rapide (facteur 6-8 par rapport à HaplotypeCaller pour les exome et 42-45x pour les génomes).

La difficulté de reproduire un pipeline a bien été prouvée par Regier et al. (2018) où différents laboratoires ont du retrouver des variants identiques sur 14 génomes de patients de référénces avec BWA-mem et HaplotypeCaller. Avant harmonisation des pratiques, il y avait une variabilité importante, résolue en forçant les même arguments

[35]sauf pour les indels homozygotes de NA19240

[36]MAF (*Minor allele frequency*) $< 0.5\%$ et entre 0.5 et 5% respectivement

[37]Indel sur Hiseq400 et Novaseq

[38]88GB mémoire, 24 cpus

des différents outils et, surtout, le même génome de référence qui était la source la plus importante de variabilité (Fig. 15).

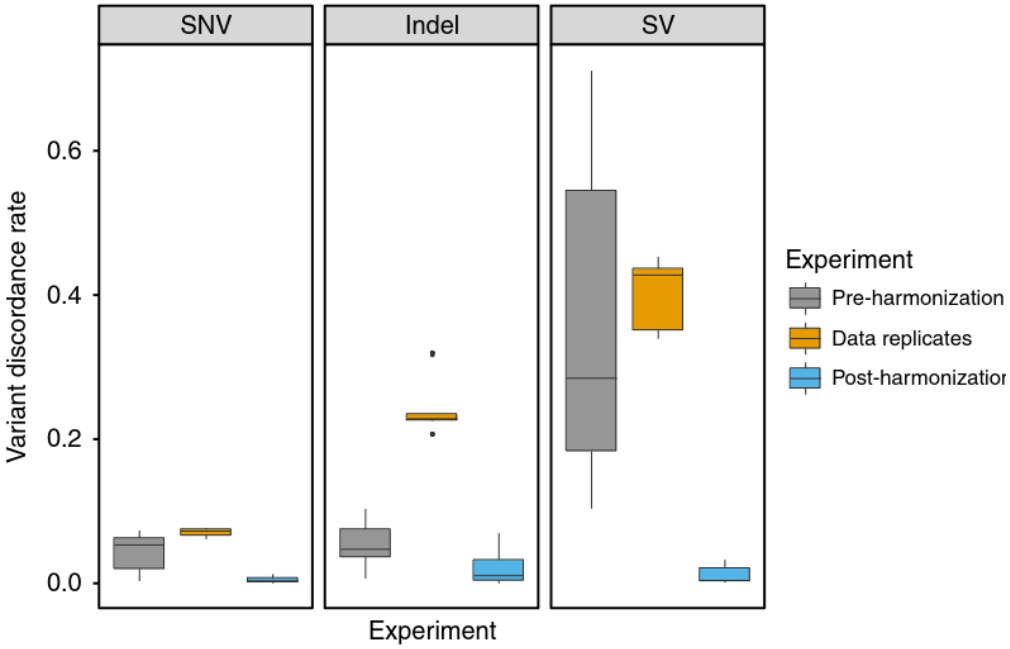


Fig. 15. – Importance de la reproductibilité démontrée par (Regier et al. 2018)

Barbitoff et al. (2022) est la revue la plus récente et utilise 7 patients de référence en exome et génome, à savoir les patients GIAB (NA12878, trio chinois, trio ashkenazi). Cet article étudie également l'utilisation de filtres. Pour HaplotypeCaller, des filtres en « dur » ont été utilisés[39], ainsi qu'avec un modèle de réseaux de neurones pré-entraîné (Convolutional Neural Network)], contrairement aux autres. Cependant, l'alignement se fait sur GRCh37. 5 aligneurs sont testés (BWA MEM, Bowtie2, Novoalign, Isaac) et 7 outils d'appel de variants (FreeBayes, GATK HaplotypeCaller, Strelka2, DeepVariant, Clair3, Octopus)

Ils confirment que sur les régions codant pour des protéines (CDS), l'appel de variant a un impact plus important que l'aligneur et l'ont confirmé avec un test statistique. La combinaison BWA-mem et DeepVariant a le meilleur score, mais de manière générale, DeepVariant a les meilleures performances quelque soit l'aligneur pour les SNPs et indels. La sensibilité varie peu selon le pipeline, mais la VPP est plus variable. Les auteurs concluent que le taux de faux positifs influence peu les résultats.

[39]Normalisation de la qualité en cas de profondeur importante (QualByDepth < 2), qualité < 30, estimation du biais de brin > 3, probabilité de biais de brin (basé sur score Phred) > 60, qualité d'alignement < 40, meilleure qualité des reads supportant l'allèle de référence que les reads supportant l'allèle alternative. Source: https://github.com/bioinf/caller_benchmark/blob/main/pipelines/call_hc.sh

Enfin, l'effet du filtre est variable selon le type de données et l'aligneur choisi. Pour HaplotypeCaller, les filtres basés sur le réseau de neurones ont permis d'améliorer ses performances, contrairement aux filtres « en dur ».

Les facteurs influençant les résultats sont d'une part, le choix du type de séquençage. Les exomes ont des performances équivalentes aux génomes pour les SNPs mais sont bien plusu mauvais pour les indels. En effet, en dehors des régions codant pour les protéines (à 50pb), il y a une diminution faible de la précision pour les SNPs et plus importante pour les indel[40]. Une couverture faible va bien sûr diminuer les performances, mais surprenamment, une couverture très importante peut l'impacter également pour certains appels de variants[41] Lec contenu en GC a au final peu d'effet, sauf en cas de contenu très riche ou très pauvre, qui va alors diminuer fortement la couverture de l'exome. Enfin, les algorithmes utilisant des méthodes semblables à HaplotypeCaller sont moins impacté par les régions difficiles à séquencer que les outils utilisant du machine learning.

Enfin, il faut reconnaître aux auteurs le choix de compléter ces tests[42] par l'utilisation de 3 génomes d'individus nigériens (Yoruba) et 3 exomes d'individus d'origine russe. Les performances ont été identiques pour ces données.

Pour finir, on peut mentionner l'étude de *Zhao et al. (2020)* qui a utilisé le patient de référence NA12878, mais surtout des données de références établies à partir de 2 môles hydatiformes haploïdes (Li et al. 2018), ainsi que des données de génome simulées. Les auteurs ont appliqué l'outil de référence pour la comparaison de VCFs que nous exploiterons par la suite. Malheureusement, les données ont été alignées en GRCh37[43]. La concordance entre les pipeline a été ici également bonne[44]. DRAGEN a été plus performance qu'HaplotypeCaller mais pas nécessairement que DeepVariant. Le temps d'exécution a été nettement en faveur de Dragen autres en total avec un facteur 5.

1.3.3 Choix retenu : GATK

Comme illustré ici, le choix de l'outil pour appeler les variants est complexe. Bien que DeepVariant soit un candidat intéressant, le consensus est difficile à obtenir. Nous avons privilégié un outil libre (*open-source*), développé par un consortium connu et

[40]Même à partir de 25pb

[41]Comme HaplotypeCaller avec les filtres basés sur les réseaux de neurones.

[42]En se limitant à BWA pour l'aligneur et en excluant HaplotypeCaller avec le filtre de réseau de neurones à cause de sa « forte sensibilité à la couverture » (sic).

[43]Avec 4 pipeline: bwamem+GATK, bwamem+deepvariant, Dragen+dragen, dragen+Deepvariant.

[44] 91.7-9.6% des SNVs ont été retrouvés par tous les pipelines et 83.5-99.4% des indel

qui puisse s'exécuter sur un processeur. Pour ces deux dernières raisons, DeepVariant ne convient pas. En effet, sa vitesse d'exécution est surtout applicable sur carte graphique et c'est un outil développé par Google. Si HaplotypeCaller, développé par GATK, ne semble désormais pas être le plus performant – bien que cela se discute comme illustré plus haut (Regier et al. 2018) – les différentes de performances observées n'ont pas d'impact clinique par rapport au nombre total de variants, notamment avec les algorithmes plus récents (voir par exemple Alganmi et Abusamra (2023) qui a testé le successeur de BWA-mem et HaplotypeCaller).

1.4 Génome de référence

Comme illustré précédemment, les différents reads des données séquencées sont alignées sur un génome de référence. Le choix de ce dernier pose de multiples problèmes. Comment proposer une référence qui soit représentative de la diversité génétique ? Prendre un patient unique est une solution difficilement acceptable. Des SNVs rares présents uniquement chez ce patient pourraient conduire à des SNVs inexistants chez le patient étudié. Peut-on comparer un génome d'une personne d'origine africaine avec une autre d'origine européenne ? De plus, il faut tenir compte des limitations techniques pour établir une référence, mais également de l'applicabilité en diagnostic car les délais de rendus doivent être acceptables. Nous présentons ici le génome humain de référence encore en application et illustrons les futurs remplacements en discutant leur pertinence pour notre cas d'usage.

1.4.1 Genome Reference Consortium (GRC)

Le génome humain a été séquencé dans sa quasi-intégralité en 2003, avec ce qui fut un tour de force technologique. 2 approches furent employées : un consortium international (le NIH américan) à l'aide de BACs (*Bacterial Artificial Chromosomes*) (voir plus loin) et une entreprise privée, Céléra, en utilisant un séquençage dit *shotgun*. Le génome actuel de référence est produit par le GRC (*Genome Reference Consortium*). Par rapport aux autres assemblages, il se différencie par une longue *contig* et *scaffold*, une précision importante par base pair et une représentation robuste des régions répétées et duplications segmentaires[45] (Schneider et al. 2017). Ce génome a été établié avec des BACs, qui sont des segments d'ADN clonés dans des bactéries qui sont ensuite séquencées afin d'obtenir des fragments de taille $>$ 150kbp. Ces fragments

[45]Régions dupliquées entre 1 et 200kb.

sont ensuite ordonnés et orientés avec des techniques complémentaires[46]. En utilisant plusieurs donneurs, le groupe cherche à représenter la diversité génétique.

En 2009, la version version GRCh37, encore appelée ou hg19, est diffusée afin de mieux représenter la diversité génétique et les variations structurelles (Genomes Project Consortium 2015). Elle est encore utilisée par de nombreux laboratoires en diagnostic à ce jour. L'ajout de *scaffold* pour des locus alternatifs a permis une représentation alternative pour la région fortement variable sur /MHC/ ainsi que les haplotypes divergents aux locus *MAPT* et *UGT2D*. Cet ajout permet de conserver la linéarité de la représentation des chromosomes.

Après 13 modifications mineures (*patches*), une nouvelle version majeure a été diffusée, nommée GRCh38 (ou hg38) en 2013. À l'heure de l'écriture de ce manuscrit, la prochaine version majeur a été « reportée indéfiniment »[47]. Les améliorations sont résumées par Schneider et al. (2017). Tout d'abord, GRCh38 continue à être une mosaïque d'haplotypes, mais la place du donneur principal, probablement africain-européen, est légèrement diminuée. 1000 problèmes ont été résolus comme illustré sur Fig. 16. Il y a 75mb de nouvelles séquences, soit 2.3% du total, avec une suppression de 5Mb (Fig. 17). Désormais, plus de 95% du génome est couvert, dont 98% des séquences non centromériques. Il y a eu une augmentation des « trous » mais cela est dû au remplacement des centromères, à l'origine un trou, en un scaffold. Le nombre de gènes a été augmenté à 41722. L'alignement de transcrits a été amélioré.

[46]On peut citer le *Radiation Hybrid Mapping*, où le chromosome est fragmenté par radiation. Le principe est que, si 2 marqueurs sont éloignés, ils seront probablement sur des fragments différents. Ou encore les études d'héritabilité (*Genetic Linkage Maps*), la comparaison d'empreinte (*fingerprint*) sur les clones qui, si elles sont similaires, signifient qu'elles viennent vraisemblablement de régions qui se recouvrent.

[47]« https://www.ncbi.nlm.nih.gov/grc », consulté le 2024-02-20

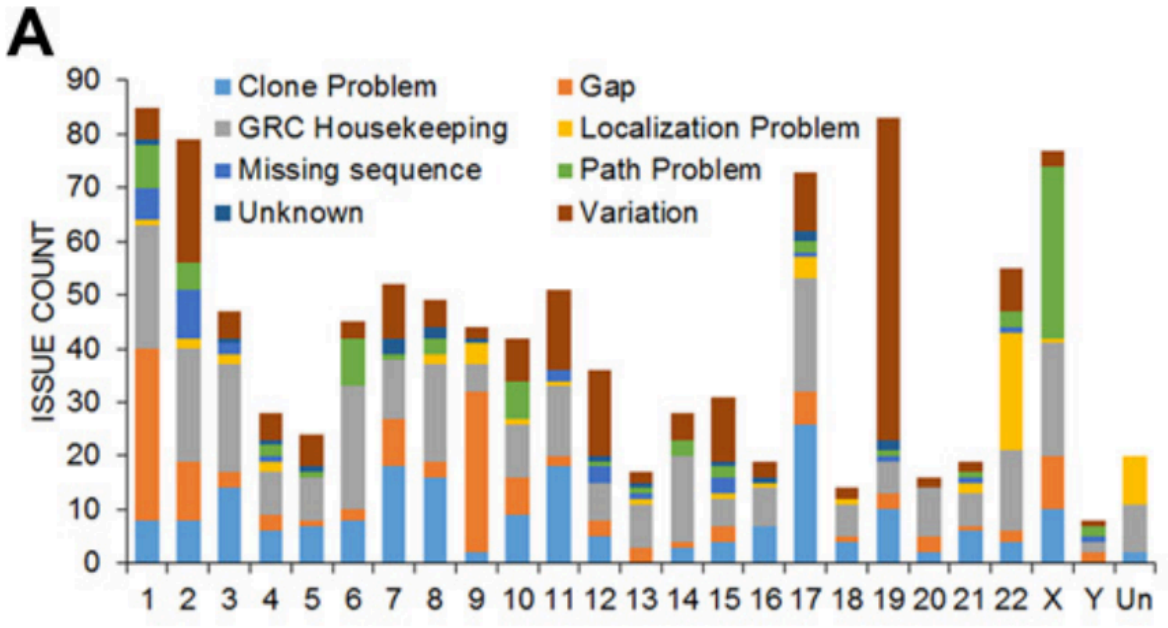


Fig. 16. – Corrections des problèmes entre GRCh37 et GRCh38 (Schneider et al. 2017)

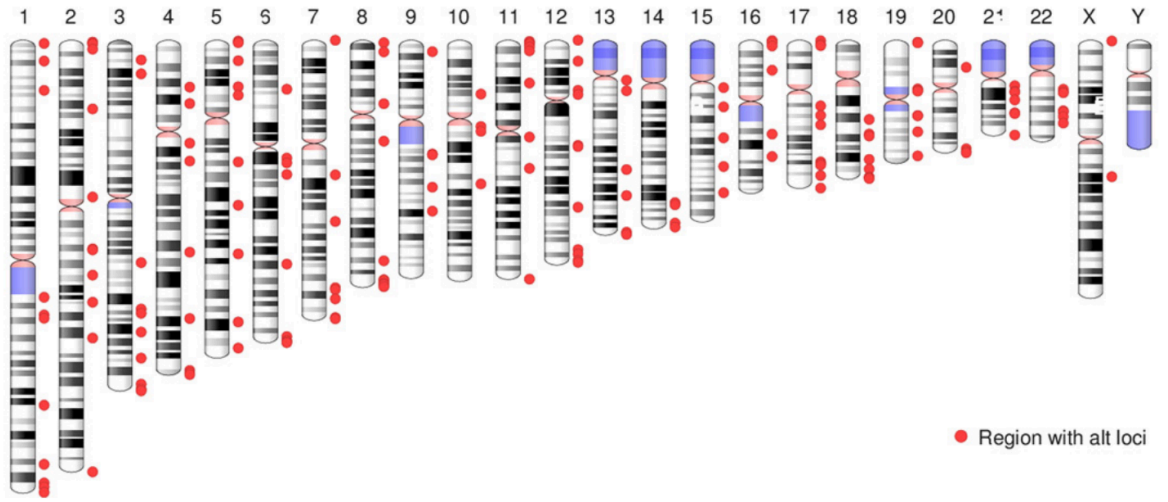


Fig. 17. – Emplacements des corrections entre GRCh37 et GRCh38 (Schneider et al. 2017)

Ce même article illustre les tests faits sur ce génome. Dans un premier temps, plus de 1000 gènes perte de fonction pour les troubles du spectre autistiques et un kit d'exome avec de plus 4600 gènes[48], malheureusement abandonné depuis quelques années. La majorité de ces gènes ne sont plus associés avec des erreurs d'alignement de transcrits, suggérant une augmentation de la qualité. Un second test sur la base de données

[48] « https://www.genomeweb.com/sequencing/emory-chop-harvard-develop-medical-exome-kit-complete-coverage-5k-disease-associ »

Clinvar (voir Chapitre 1.6) a montré que, sur 113 000 variants, environ 200 ont été sélectionnés qui appartiennent uniquement à GRCh37. Les régions sur lesquelles ils sont situés avaient été modifiées afin d'ajouter des séquences manquantes. L'impact du changement est illustré par un variant faux-sens de *KCNE1*, initialement supposé pathologique. L'ajout d'une séquence paralogue[49] a montré qu'il pouvait être aligné sur plusieurs endroits différents. Le variant est actuellement classifié comme « conflictuel » sur Clinvar[50] mais est probablement bénin.

Un troisième test a consisté a utiliser les données du patient NA24143 en les alignants avec BWA-mem sur GRCH37 et GRCh38. Parmi tous les emplacements Clinvar avec une qualité d'alignement inférieure à 20, 10 emplacements correspondaient à des variants seulement en GRCh37. Toutes ces régions avaient été également modifiées pour enlever des séquences redondantes ou corriger des haplotypes dans la version GRCh38. Cela devrait permettre d'appeler des variants qui auraient pu être manqués sur ces régions en GRCh37. De plus, 135 emplacements ne sont plus couverts en GRCh38 à cause de mise à jour manuelle par ajout de séquence paralogues ou alléliques, suggérant la possibilité de faux positifs sur ces régions. Enfin, pour ce patient, l'impact sur l'alignement a été mis en évidence avec environ 99.2% des reads alignés en GRCh38 et surtout 64% reads qui ne l'étaient pas sont désormais alignés. Parmi ceux restants, 23% sont alignés sur les régions alternatives.

1.4.2 T2T

Cependants, GRCh38 a de nombreuses limites. D'une part, il est toujours incomplet, même si le nombre de « trous » est passé de 150 000 à 995 (Jarvis et al. 2022). De plus, l'utilisation de BACs sous-représente les séquences répétées. Il ne représente pas non la diversité du génome du fait de la composition à 70% d'un seul individu et le restant est constitué de 19 autres. Les difficultés ne sont également pas faciles à résoudre concernant les duplications de taille supérieure à celle des reads, les SNVs et variants de structure mal capturés par la technologie short-read.

La fusion de différents haplotypes en une représentation haploïde[51] introduit des erreurs. Par exemple, le regroupement de variants de différents haplotypes en un seul pseudo-haplotype, l'existence de fausses duplications[52]. De plus, la plupart des « trous » sont insolubles à cause de polymorphismes de structures aux extrémités et de nombreux éléments répétés ou polymorphes qui sont non finis ou mal assemblés.

[49]Paralogue = duplication d'un gène, avec séquence et fonction pouvant différer.

[50]https://www.ncbi.nlm.nih.gov/clinvar/variation/13479

[51]C'est-à-dire avec une seule copie de chaque chromosome.

[52]Pour des haplotypes qui sont plus « divergents » que le reste du génome et sont alors représentés sous la forme de (faux) paralogues

Il existe environ 151Mbp de séquences inconnues qui sont péri- ou sous-télomérique, ou des duplications segmentaires ou de l'ADN ribosomal. Les bras courts des 4 chromosomes dits acrocentriques[53] et les HSats (*Human satellite repeat array*), On notera avec intérêt que certaines régions sont complètement artificielles, comme dans le centromère où les zones dites alpha satellite[54] sont représentées sous forme d'un modèle informatique[55] (Fig. 18). Enfin, certaines régions comme le bras court du chr21 sont mal assemblées.

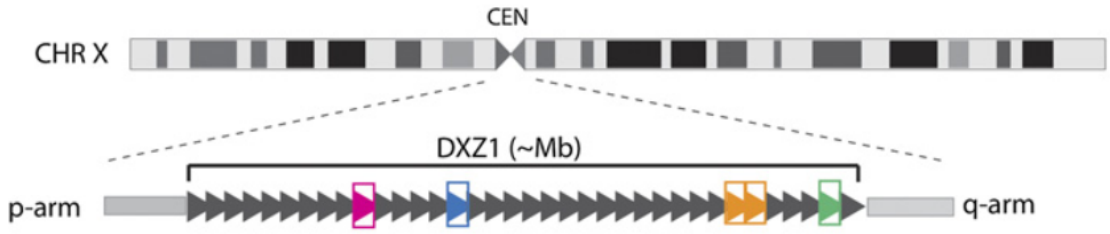


Fig. 18. – Illustration d'une zone *alpha-satellite* avec de nombreuses répétitions (flèches noires), et quelques variants (en couleurs) (Miga et al. 2014)

Pour tenter de résoudre ces problèmes, le consortium T2T a assemblé un nouveau génome de référence à partir d'une môle hydatiforme appelé T2T-CHM13 (Nurk et al. 2022). Le génome de ces môles hydatiformes est issu de la perte du complément maternel avec duplication paternelle après la fertilisation et sont donc homozygotes de caryotype 46XX. Les auteurs ont utilisé du séquençage ultra-long read pour les régions répétées (nanopore) et du séquençage Pacbio hifi pour des read 20kb[56] ainsi que d'autres technologies[57].

Ce nouveau génome n'a pas plus de variants perte de fonction ni d'allèles isolées que les échantillons du projet 1000 génome. De plus, si dans GRCh38, l'individu contribuant pour 72.6% du génome est à 56% d'origine africaine et 28.1% européenne, le second contributeur est 5.5% d'origine majoritairement de l'est de l'Asie. *A contrario*, T2T-CHM13 est principalement d'origine européenne. Ces 2 génomes possèdent tous les deux des « introgressions » de Néandertal.

Les changements dans cette version sont illustrées sur les figures 19, 20 et concernent 8% du génome avec un ajout de 238Mbp au total, soit 4.5% dont 182

[53]Où le centromère est proche d'une extremité, à savoir les chromosomes 13,14,15,21,22

[54]Répétition d'unitée de 171bp en tandem

[55]Séquencage génome puis génération d'une forme linéaire.

[56]Ce qui présente un certains compromis car les reads sont de 20kb mais avec un taux d'erreur de 0.1%

[57]Illumina pcr-free, Illumina hi-c pour les interactions de la chromatine, de la cartographie optiquedu avec Bionano et du *Single-cell DNA template strand sequencing*.

exclusif T2T. T2T-CHM13 est une avancée remarquable car il ne comporte aucun de trous et par conséquent pas de contig. Le nombre de gène a augmenté de 5%. Des tests sur les génomes du projet 1000 génomes ont montré que les faux positifs et négatifs ont diminués. Enfin, le chromosome Y a été ajouté dans une mise à jour (Rhie et al. 2023).

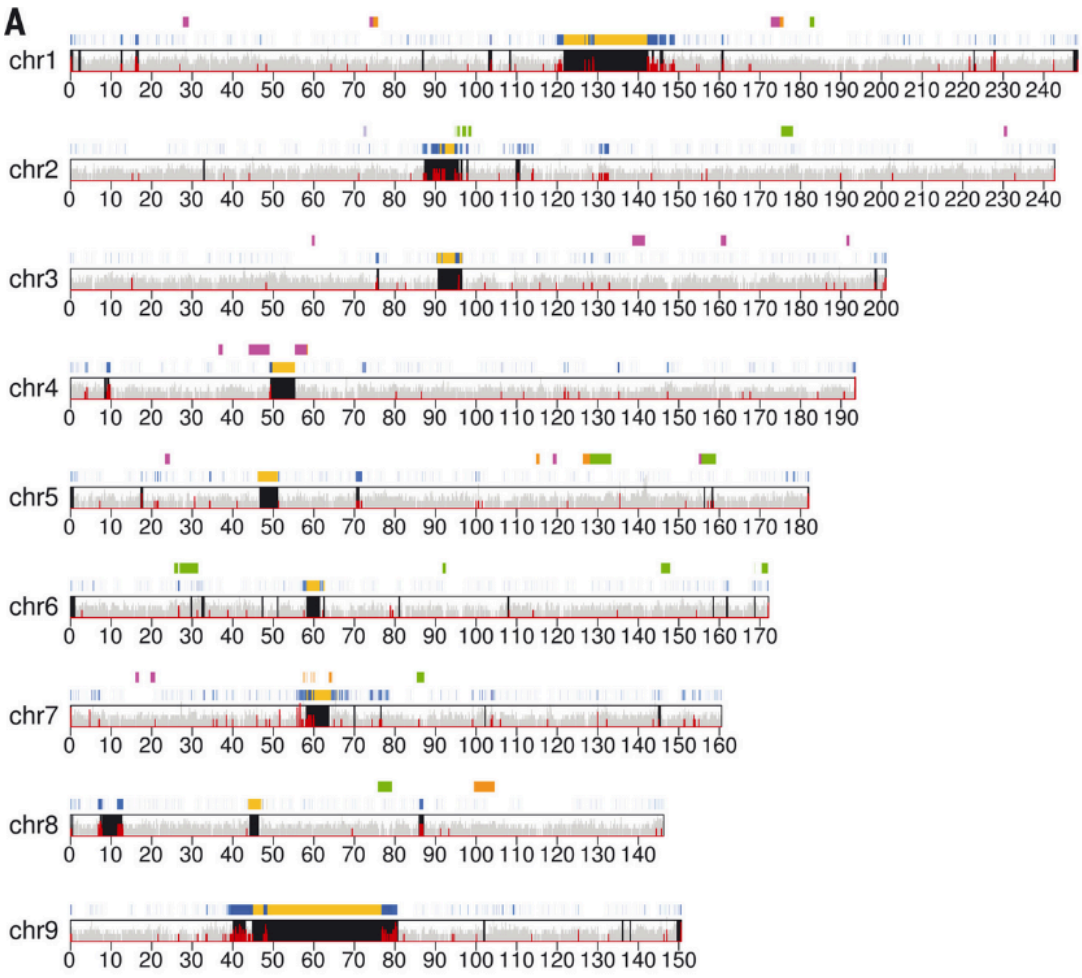


Fig. 19. – Changements de T2T-CHM (chromosomes 1 à 9)

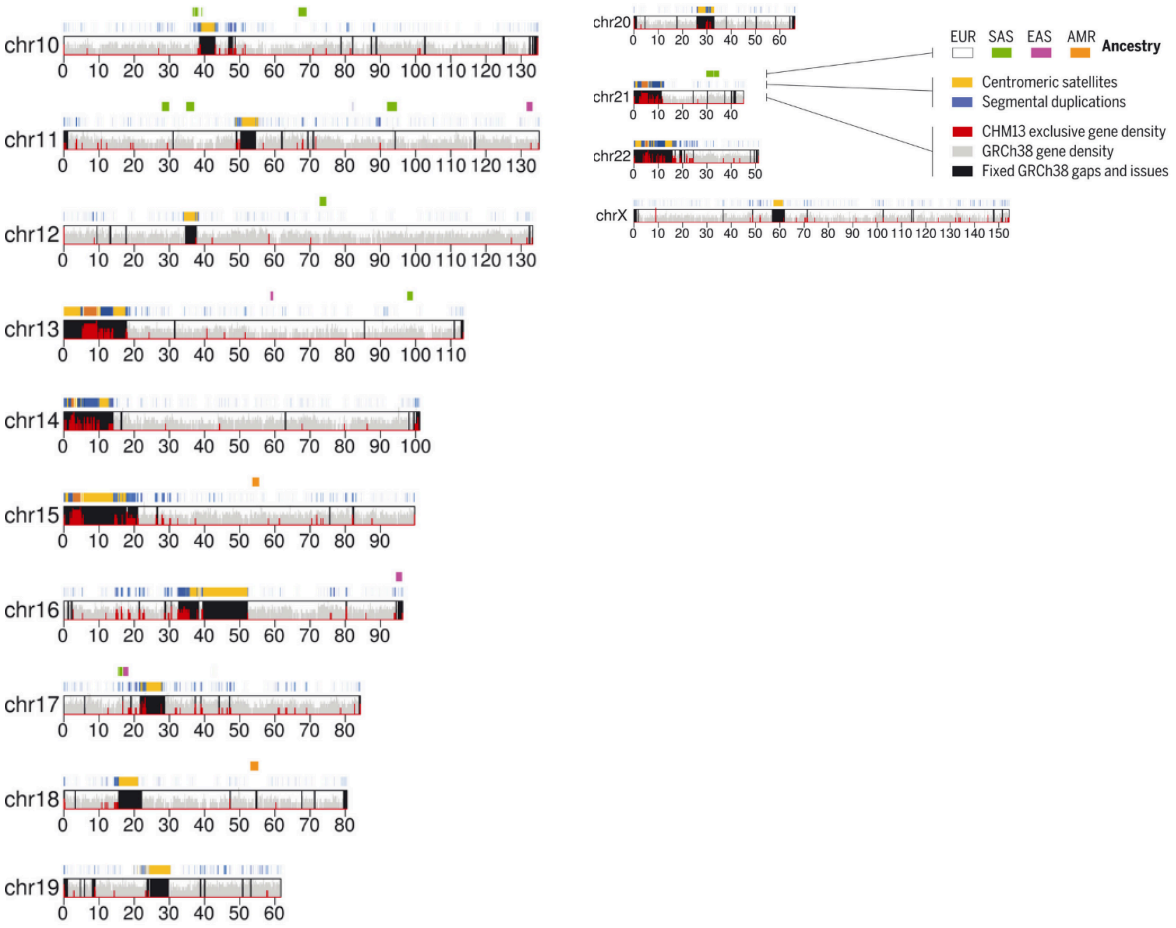


Fig. 20. – Changements de T2T-CHM (chromosome 10 et plus)

L'apport pour le séquençage a été démontré par Aganezov et al. (2022). Ils ont supposé que, par la nature d'assemblage de GRCh38, des haplotypes de structures anormales existent aux bornes des clones utilisés. Cela pourrait donc conduire à des combinaisons d'allèles rares, voire absente de la population. Après vérification, de nombreuses transitions d'haplotypes ont été trouvées aux bornes des clones BAC en GRCh38 et non présentes dans les données 1000 génomes. Ces transitions sont bien plus rares en T2T-CHM13.

Pour identifier les erreurs provenant de régions complexes comme les duplications segmentaires, les auteurs ont exploité le caractère homozygote de T2T-CHM13. En effet, les variants hétérozygotes proviennent alors d'erreurs de séquençage, d'alignement des reads ou de mutations accrues sur la lignée cellulaires. En considérants des régions avec des regroupements de variants hétérozygotes, ils ont constaté que ces régions potentiellement problématiques sont associées à des duplications segmentaires, centromères et problèmes de GRCh38. Du fait de la concentration augmentée en variants marqués par gnomAD comme « beaucoup trop hétérozygotes », il est possible

que ces régions correspondent à des faux positifs. Après conversion de ces régions en T2T-CHM13, environs 8Mb (203 emplacements) ont été identifé avec des copies manquantes dans GRCh38[58]. Ces zones correspondent à 48 gènes codants (dont 14 complètement inclus dans ces régions ont été identifiés, notamment *DUSP22* lié à la régulation immunitaire ou *KMT2C* relié au syndrome de Kleefstra. Ce nouveau génome permet donc d'appeler des variants dans ces zones précédemment inaccessibles.

L'alignement a été également amélioré avec 0.97% de reads additionnels en utilisant BWA-mem, le nombre de discordances par read est diminué et la couverture est plus uniforme. On peut noter que les sujets d'origine africaine ont toujours le plus grand nombre de discordances par read. Concernant l'appel de variant, testé avec HaplotypeCaller, il y a moins de variants par échantillon qu'en GRCh38[59]. Cela est plus marqué pour les sujets non-africains[60]. Les auteurs ne conseillent pas de faire l'appel de variant en T2T et « lifter » en GRCh38 car cela peut conduire à des problèmes, par exemple si l'allèle de référence n'est pas dans l'échantillon. Des analyses statistiques ont montré qu'il y a moins de variants de mauvaise qualité, moins de variant discordants au niveau mendelien[61]. Enfin, l'apport des régions non résolues précédemment concerne surtout la correction de fausses duplications surtout.

L'impact clinique a été évalué au travers de 3 tests. D'une part, les auteurs ont voulu voir la présence de variants délétères (perte de fonction) dans ce nouveau génome, ce qui pourrait impacter l'interprétation de variants. 210 variants potentiels ont été mis en évidence sur 189 gènes dont 31 avec un impact clinique[62]. 158 variants ont été retrouvés chez des individus du 1000 genomes project avec une balance allélique moyenne de 0.47, suggérant qu'ils sont tolérés au niveau fonctionnel. Les autres variants sont des indels plus grands.

Dans un second temps, 4964 gènes d'intérêt clinique ont été examinés (Wagner et al. 2022). Il en ressort que 28 gènes sont sur des régions non résolues ou non colinéaires. Il y a deux fois plus gènes en GRCh38 avec des allèles rares ou des erreurs structurelles dont la plupart semblent corrigées en T2T. Ainsi, le gène *TNNT3* impliqué dans l'arthrogrypose était supposé affecté par un remaniement complexe[63].

[58]En utilisant des estimations du nombre de copies à partir du Simons Genome Diversity Project

[59]Peut-être par une diminution des allèles rares, diminution des erreurs de consensus et de structure

[60]70% de gRCh38 vient d'un individu d'ascendance afro-européenne et les sujets d'origine africaines ont plus de variants rares. Il est alors possible que les sujets non-africains aient donc plus de variants identifiés avec un génome de référence contenant plus de variants rares

[61]C'est-à-dire présent chez les enfants et absents chez les parents ou homozygote chez les parents, mais pas chez les enfants

[62]Ce résultat est en accord avec une étude montrant que le génome humain contient en moyenne 450 variants perte de fonction sur 200 à 300 gènes

[63]Avec une inversion et délétion en amont du gène

L'étude en T2T-CHM13 a corrigé cette région (Fig. 21) avec 2 impacts : d'une part, l'interprétation sur la régulation du gène est modifiée et d'autre part, cela rapproche le gène de son « voisin » au niveau génétique TNNI2. De plus, De plus, 17 gènes avec un impact en clinique sont sur des régions probablement fusionnées ou dupliquées par erreur. Par exemple, *KCNE1* est lié à une fausse duplication et a une couverture plus faible que la normale en GRCh38 car des reads sont alignés sur *KCNE1B*. Un autre exemple est *KCNJ18*, le résultat probable d'une duplication fusionnée en GRCh38, ce qui augmente la profondeur. Enfin, des tests maisons ont été créés par les auteurs en GRCh38 et T2T-CHM13 comprenant 269 gènes d'intérêt clinique étiquettés « difficiles ». Le nouveau génome conduit à une diminution des faux positifs et faux négatifs.

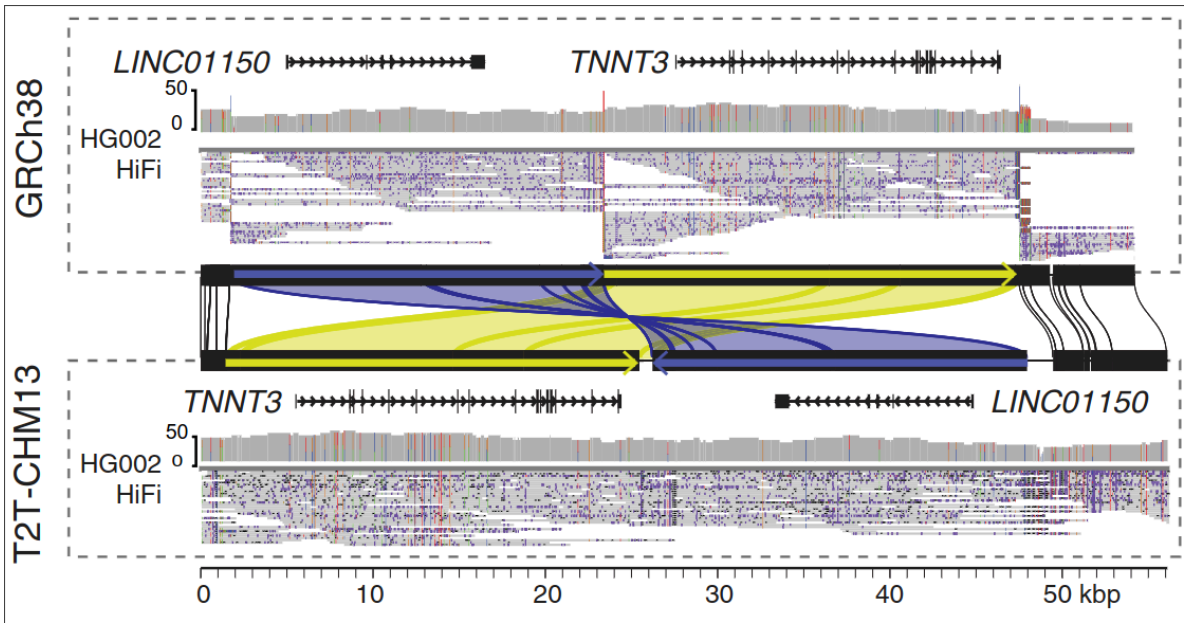


Fig. 21. – Correction sur le gène *TNNT3* entre GRCh38 et T2T-CHM13 (Aganezov et al. 2022)

Qu'en est-il des prochaines versions du génome ? Actuellement, l'assemblage de T2T est encore très manuel et a été constitué sur une môle hydatiforme. Jarvis et al. (2022) propose une méthode semi-automatisée pour l'assemblage d'un génome T2T à partir du patient de référénce HG002. Il est également possible que le futur génome soit en fait un pangénome, qui permet d'avoir le génome de plusieurs individus sous forme d'un graphe (Fig. 22). Une version préliminaire a été publiée Liao et al. (2023) .

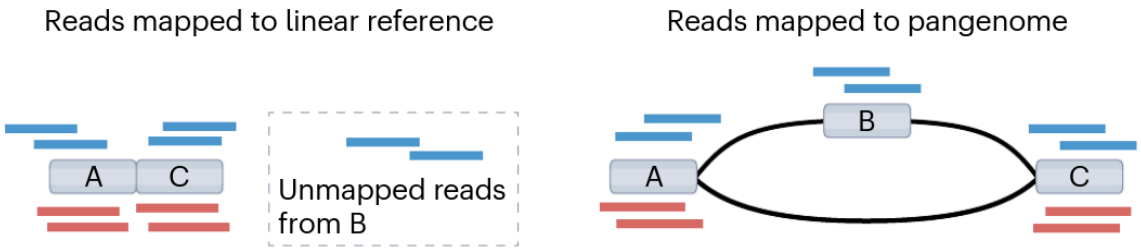


Fig. 22. – Comparaison de l'alignement sur un génome linéaire et sur un pangénome (Olson et al. 2023)

1.4.3 Choix retenu pour bisonex

Nous avons suivi les recommandations de Zverinova et Guryev (2022) en utilisant la version « pipeline-ready » de GRCh38, version majeure sans les versions mineures. Il s'agit donc d'une référence « primaire » contenant les chromosomes, ADN mitochondrial et contig non placés et non localisés[64]. Ces derniers permettent d'éviter le mauvais alignement de ces régions donc diminue les faux positif. À noter que pour l'alignement, bwa-mem peut utiliser ces régions alternatives[65], mais la version actuelle du pipeline ne l'implémente pas encore. Enfin, le pipeline est également compatible avec T2T-CHM13.

1.5 Annotation

L'interprétation de variants se base notamment sur son impact sur le transcrit ou la protéine. En conséquent, elle va dépendre de l'annotation qui associe à un variant un ou plusieurs transcrits. Il existe deux sources principales d'annotation. *GENCODE*[66] vise à représenter toutes les isoformes pour tous les tissus et étapes du développement, ce qui donne en moyenne 4 transcrits par gène codant avec parfois plusieurs dizaines de transcrits pour un variant. À noter qu'elle dépend du génome de référence. Une autre ressource est *Refseq* qui ne dépend pas du génome de référence. Cela se traduit par la présence de certains variants dans Refseq qui sont absents du génome de référence, mais aussi par une plus grande difficulté d'aligner un variant sur les transcrits Refseq[67]. Une des causes de confusion est l'existence d'haplotypes alternatifs dans

[64] À ntore que l'inclusion de zones hypervariables telles que le locus d'histocompatibilité HMC ou la région pseudoautosomique du chromosome Y peut conduire à une perte de l'alignement donc diminuer la sensibilité

[65] https://github.com/lh3/bwa/blob/master/README-alt.md

[66] À noter qu'Ensembl et GENCODE sont identiques https://www.gencodegenes.org/pages/faq.html

[67] https://genome.ucsc.edu/FAQ/FAQgenes.html#ensRefseq

le génome de référence entraînant des représentations différentes d'un même variant (McLaren et al. 2016), comme illustré sur Fig. 23. La notation HGVS des variants est de plus basées sur les transcrits, ce qui peut également entraîner des confusions et multiplie le nombre d'annotations.

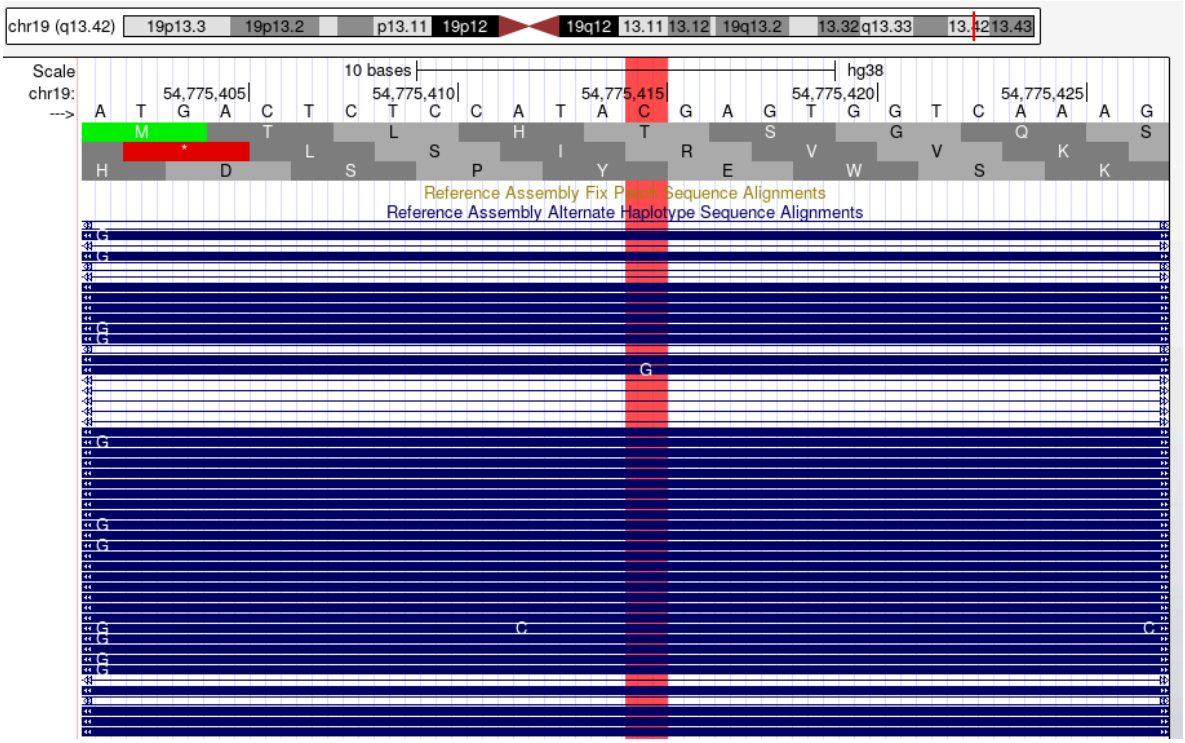


Fig. 23. – Exemple de variant (en rouge) s'alignant sur une zone avec de nombreuses réference alternative (en bleu). https://www.ncbi.nlm.nih.gov/snp/rs150580082 C>G ou C>T ne donnera un codon stop que sur l'un d'entre eux.

Nous détaillons dans cette partie 3 outils populaires pour l'annotation de variants en détaillant les informations fournies, le tout suivi d'une comparaison avec le choix retenu dans Bisonex.

1.5.1 VEP (*Variant Effect Predictor*)

L'outil est décrit par McLaren et al. (2016). Il permet d'annoter les SNVs, indels, substitutions sur plusieurs paires de bases, les microsatellites, les répétions en tandem et les variants structurels de taille supérieur à 50bp, notamment les CNVs. Les génomes de références supportés sont GRCh37 et 38, ainsi qu'en version préliminaire T2T-CHM13. Il donne l'effet d'un variant sur le transcrit, la protéine et les régions régulatrices ansi que, si le variant est de plus connu, la fréquence allélique et des informations telles que le phénotype associé au gène concerné. De manière générale,

VEP est utilisable dès qu'un génome avec ensemble de gènes annotés est disponible. Il est de plus open source et libre d'utilisation.

L'annotation de transcrit est détaillée dans le Tableau 22 avec GENCODE et Refseq. Le format de sortie est un tableau dans lequel chaque allèle et chaque caractéristique du génome (transcrit, zone régulatrice…) est une ligne. L'annotation protéique est détaillée dans le Tableau 23. À noter que les différents scores sont pré-calculés pour toutes les combinaisons et mis à jour quand cela est nécessaire. D'autres scores, comme CADD, FATHMM, mutationtaster…, sont disponibles à l'aide de plugins. Les variants non codant sont annotés s'ils ont un impact sur les éléments régulateurs de la transcription ou traduction. À ce titre, VEP permet l'annotation des variants impactant les ARNs non codants, les zone régulatrices génomiques ou les motifs d'attache pour facteur de transcription (*transcription factor binding motif*)[68].

VEP va utiliser la base de données d'Ensembl, qui contient notamment dbSNP (voir Chapitre 1.6), COSMIC pour les variants somatiques, Human Gene Mutation Database, les CNVs et les variants de structures de la Database of Genomics Variants. La fréquence des variants est extraites à partir des projets 1000 genomes et ExAC. Le phénotype est donnée par OMIM, Orphanet et les données issue des GWAS (*Genome-Wide Association Studies*).

Il existe 2 approches pour filtrer les données. À l'exécution, l'utilisateur peut restreindre les variants choisissant seulement une conséquence par variant[69] ou en filtrant les variants « communs », c'est-à-dire co-localisés avec un variant connu de MAF $>$ 1%. Une autre approche est d'utiliser via le script `filter_vep.pl` une combinaison de toutes les données fournies par VEP. Par exemple, les variants co-localisés avec ceux qui sont fréquents chez les populations africaines, mais rares chez les populations européennes peuvent être filtrés avec `AFR >0.1 AND EUR <0.05`. Bien évidemment, VEP ne se substitue pas à une validation de méthodes pour le choix des filtres, ce qui est l'objet du Chapitre 3. Enfin, les performances sont illustrées sur Tableau 1, montrant que VEP est plus lent que ses concurrents. Cependant, il faut préciser que VEP est écrit en Perl, alors que SnpEff est développé en Java, ce qui donne un avantage au second. De plus, annovar dispose de moins d'annotations.

[68]Basé sur l'Ensembl regulator build consistant en ENCODE, BLUEPRINT et la NIH epigenomics roadmap.

[69]Par défaut, les priorités sont de manière décroissante : transcrit MANE select, transcript MANE Plus Clinical, transcrit canonique, isoform APPRIS, *transcript support level*, biotype du transcript, status CCDS, conséquence selon http://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences, longueur du transcrit ou de la caractéristique

Type	Temps d'exécution
Annovar	21min50 s (3415 v/s)
SnpEff	46min39 s (1598 v/s)
SnpEff (threaded)*	10min28 s (7121 v/s)
VEP	62min9 s (1200 v/s)

Tableau 1. – Performances de VEP selon McLaren et al. (2016)

1.5.2 SnpEff

SnpEff est un autre outil, présenté par Cingolani et al. (2012), permettant d'annoter les SNPs, indels et polymorphismes sur plusieurs nucléotides. Il donne des informations sur la position : intronique, UTR, amont ou aval du gènes, site d'épissage, intergénique. Les détails de l'annotation sont précisés dans le Tableau 24. Les effets pour les variants sur des gènes codants sont donnés (synonyme ou non, gain ou perte d'un codon start ou stop, décalage de phase). Pour les variants sur des gènes non codants, une annotation est fournie avec le biotype si cette information est disponible. Par rapport à annovar, il dispose de plus de versions du génome et semble plus rapide (Tableau 1).

1.5.3 Annovar

Les fonctionnalités de ce troisième outil sont détaillées par Wang, Li, et Hakonarson (2010) et la documentation en ligne (ANNOVAR 2023). Plus précisément, il existe 3 types d'annotations. Par gène, tout d'abord, en utilisant par défaut Refseq[70]. Son emplacement ensuite : exonique ou intronique. Dans le premier cas, on disposera de son rôle sur le plan fonctionnel (synonyme, décalage de phase……) et du changement d'acides aminés. En intronique, les deux gènes les plus proches seront précisés.

Un second type d'annotation concerne les régions, ce qui est utile pour savoir si un variant est dans une région d'intérêt. Par exemple, les régions conservées au sein de plusieurs espèces, les sites de prédiction de liaison pour les facteurs de transcription, les bandes cytogénétiques. Mais aussi les variants impactant les microARN et *small nucleolar ARN*, ou les sites d'accroches prédits pour les microARN. Ou encore les duplications, les variants structurels, les variants publiés dans les GWAS, les zones dans ENCODE. En intronique, on aura l'impact sur les *enhancer*, répresseur ou promoteur d'un gène[71].

[70]mais ENSEMBLE, UCSC, GENCODE ou un GFF3 fourni par l'utilisateur sont également possibles.

[71]Prédit par chromHMM par exemple.

Enfin, pour filtrer les données, Annovar propose des bases de données de fréquence en génome et exome (1000 Genomes, gnomAD…) et pour les populations moins représentées dans ces bases de données[72]. De nombreux scores de prédictions sont également accessibles en génome (Gerp++, CADD, FATHMM…) ou en exome (SIFT, Polyphen2…) ainsi que pour les variants d'épissage (dbscSNV, spidex). Le tout est complété par Clinvar, et des bases de données somatiques (COSMIC, International Cancer Genome Consortium, NCI-650).

Les limites sont détaillées par Yang et Wang (2015). D'une part, les annotations peuvent être trop chargées pour l'utilisateur avec, par exemple, 10 scores différents de pathogénicité. À ce titre, un nouveau méta-score a été développé pour les SNVs non synonymes, appelé MetaSVM. Les auteurs recommandent de manière générale l'utilisation d'un outil courant de prédiction comme SIFT et un méta-score comme ce nouveau score. De la même manière pour les variants non-codants, un score comme PhyloP et un méta-score comme CADDs peuvent être suffisants. De plus, les annotations fonctionnelles ne disposant pas d'un format standard de représentation, ce qui complique la comparaison des différents outils. Enfin, Annovar ne gère par les variants structuraux complexes et les translocations. Il peut cependant annoter les indels de moins de 50bp.

1.5.4 Comparaison

Une comparaison VEP, SnpEff et Variant reporter a été effectuée par Yen et al. (2017) afin de vérifier les résultats selon la nomenclature HGVS. Les auteurs ont utilisé 126 variants validés manuellement et difficiles à annoter selon leur expérience[73]. Dans un second temps, ils ont comparé VEP et SnpEff sur plus de 100 000 variants extraits[74] et plus de 2 millions de variants dans COSMIC[75]. Ils ont considéré deux annotations comme identiques si les transcrits étaient identiques et que la syntaxe était la même en cordonnées codantes et protéiques[76].

Sur leur jeu de 126 variants[77]. VEP et SnpEff ont des résultats majoritairement concordants (92.6%) pour les coordonnées codantes, mais sont plus différents pour l'annotation protéique. La précision était globalement bonne pour les différents outils (Fig. 24) avec également plus de variabilité au niveau protéique. Pour les variants

[72]ajews pour les Juifs Ashkenazes, TMC-SNPDB pour les populations indiennes

[73]50 variants provenaient de Clinvar, dbSnp, COSMIC, Mycancer genome, LOVD, Emory genetics lboratory. 76 variants ont été générés à l'aide de Mutalizer et de Variant viewer.

[74]84% de SNVs, 10% de deletions, 3.3% de duplications, 1% d'insertions et 1% d'indel.

[75]Après suppressions des doublons. 93.53% étaient des SNVs.

[76]L'annotation était « équivalente » en cas de coordonnés synonymes, par exemple `c.482del` et `c482delT`

[77]Mais seulement 118 ont pu être comparés par les 3 modèles.

avec 2 nucléotides, les performances ont été également bonnes. L'effet prédit a été globalement concordant entre les différents outils[78]. Sur les données extraites de Clinvar, la concordance entre SnpEff et VEP est très bonne pour les SNVs. Les insertions et délétions ont en revanche une plus mauvaise performance : 16.6% des insertions sont incorrectes car ne respectent pas la règle 3′ de la nomenclature ou non écrites comme des duplications. Mais surtout la syntaxe pour l'impact protéique a des performances assez mauvaises (16-78%). La majorité de ces différences correspond à une erreur systématique et corrigeable automatiquement[79].

En conclusion, les performances des outils sont bonnes pour SNV mais une attention particulière doit être apportées aux insertions ou délétions avec potentiellement une vérification manuelle. Sur cet article, malgré la variabilité des résultats, VEP semble être un meilleur choix que SnpEff.

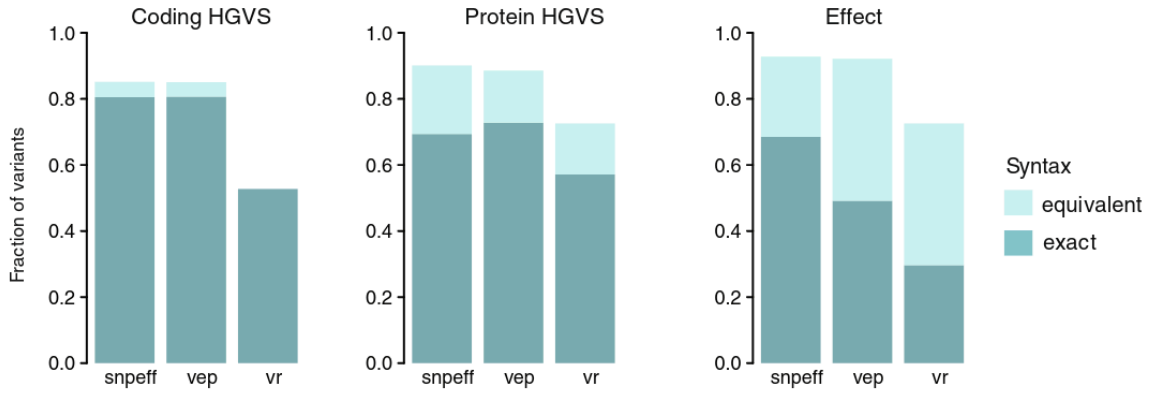


Fig. 24. – Comparaison de la précision sur 126 variants de référence selon (Yen et al. 2017)

1.5.5 Choix retenu : VEP

Pour le pipeline, nous avons choisi VEP qui a l'avantage d'être plus ancien, donc plus testé, avec des performances paraissant meilleure que SnpEff. On pourrait discuter de leurs performances respectives, mais il ne s'agit pour l'instant pas d'une étape critique (Chapitre 2.3). Les deux outils proposent une interface facile à utiliser pour les scripts. VEP dispose de nombreux plugins pour étendre ses fonctionnalités et permet d'annoter sur T2T-CHM13. En revanche, l'architecture logicielle est complexe, ce qui a conduit à plusieurs difficultés pour rendre son installation reproductible (voir Chapitre 2.1).

[78]En cas de plusieurs effets prédits, il suffisait que l'un d'entre eux corresponde.

[79]Ex: substitution au lieu d'indel p.AspAla2625GluPro versus p.Asp2625_Ala2626delins GluPro pour rs267606668. Les délétions sont mal annotées, par exemple pIle55fs au lieu de p.Ile55Ter.

1.6 Bases de données

Clinvar représente un ensemble d'interprétations pour un variant correspondant à sa signification clinique ou fonctionnelle, soumise par les utilisateurs (Landrum et al. 2016). Son spectre est large, car elle concerne les variants constitutionnels et somatiques, tous types de variants, de toutes les tailles (SNVs, CNVs, anomalies cytogénétiques) dans toutes les régions du génome. L'objectif est d'aider à l'interprétation dans un contexte de recherche ou diagnostic clinique. Les ressources externes utilisées sont OMIM (voir ci-dessous), GeneReviews, dbSNP, si les variants ont été soumis avec des informations cliniques, ainsi que des variants soumis par un petit nombre de laboratoires de biologie médicale, des équipes de recherche et depuis UniProt. Une équipe dédiée va vérifier la nomenclature HGVS, la relation gène-maladie, mais pas l'interprétation. Pour cela, des panels d'experts vont procéder à une revue manuelle des interprétations, ce qui conduit ClinVar à être une ressource majeure dans l'interprétation des variants.

Il s'agit également d'une archive, ce qui permet de suivre l'évolution de la classification d'un variant au cours du temps. Les personnes pouvant soumettre des nouveaux variants ou nouvelles interprétations sont les laboratoires cliniques, les chercheurs, les base de données existants et les groupes d'experts[80]. Le résultat final sera donc une agrégation des différentes interprétations avec une évaluation globale selon les recommandations ACMG en précisant si les différentes soumissions sont en accord. Il s'agit enfin d'une ressource gratuite et sans barrière d'accès. Des statistiques sont données dans le Tableau 20.

dbSNP propose une autre approche en se focalisation sur les SNPs (*Single Nucleotide Polymorphism*). S'agissant des variants les plus courants, il existe donc un besoin pour un catalogue comprenant les études d'association, fonctionnelles, pharmacogénomiques et autres… La majorité sont des substitutions (99.77%) mais il existe également des petits indels , régions invariantes, répétions de microsatellites… (Sherry 2001). Il est important de noter que dbSNP contient des variants pathogènes. Les données concernent seulement les *Homo Sapiens* depuis 2017 avec plus de 3 milliards d'entrées[81]. Selon un livre datant de 2002[82], l'analyse fonctionnelle va fournir une annotation basée sur les séquences adjacentes au variant pour déterminer s'il est codant, intronique, sur un site d'épissage, etc. En cas de changement d'acide aminé, la position en 3 dimensions dans la protéine est donnée à partir de la base de données

[80]Comme InSiGHT, CFTR23, ENIGMA ou l'American College of Medical Genetics pour des variants de *CFTR*

[81]https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi, accédé 17 décembre 2023

[82]https://www.ncbi.nlm.nih.gov/books/NBK21088/

PDB et en utilisant un alignement avec BLAST pour identifier la position. De plus, la diversité d'un variant au sein de différentes populations est estimée avec le calcul d'une « hétérozygosité moyenne » à partir de la fréquence allélique des différentes soumissions[83], qui permet d'estimer la probabilité que les 2 allèles soit dans un individu diploïde. D'autres données sont également intéressantes comme le compte dans une population, la fréquence d'un génotype et les probabilités selon la loi d'Hardy-Weinberg. Selon cette même source, les soumissions proviennent du consortium SNP, de variants extraits de génomes humains et de contributions de laboratoires.

OMIM est une autre base de données incontournable en génétique constitutionnelle dont l'objectif est de cataloguer les corrélations entre gènes et phénotypes (Amberger et al. 2015). Elle s'inscrit dans la continuité des travaux d'A. McKusick dans son livre *Mendelian INheritance in Man* publié en ligne 1987. Pour alimenter cette base, une veille scientifique sur les journaux biomédicaux peer-reviewed et sur Pubmed est effectuée. Chaque entrée dans OMIM est formatée de la même manière et propose un lien vers des bases connexes comme le génome Ensembl, la clinique avec GeneReviews, les variant avec Clinvar etc. De fait, les identifiants OMIM sont largement utilisés dans la communauté et la littérature scientifique[84]. Les informations sur le gène sont classiques : élément régulateur, gène codant… Concernant les phénotypes[85], OMIM nous renseigne sur le caractère monogénique, sur les phénotypes, la sensibilité à certains médicaments ou à une infection ou à un cancer, mais aussi sur les syndromes de délétion récurrente ou duplication comme le syndrome de Smith-Magenis[86]. Des statistiques sur la base de données sont dans le Tableau 21.

Le sous-ensemble « OMIM morbid » est tout particulièrement intéressant, car il liste les gènes dont les variants sont reliés de manière causale avec une maladie. Pour établir une relation entre un gène et un phénotype, il faut que le variant se retrouve chez plusieurs individus non apparentés, ségrège avec le phénotype dans plusieurs familles, ou qu'il apparaisse *de novo* dans un nombre statistiquement significatif de familles. Enfin, la relation sera « confirmée » (*qualified*) si une famille avec plusieurs variants dans un gène a pu être trouvée et si et les variants ségrègent avec le phénotype avec de plus une preuve fonctionnelle de la causalité[87].

[83] https://www.ncbi.nlm.nih.gov/SNP/Hetfreq.html

[84] Il comporte 6 chiffres : le premier est 1,2,6 si la transmission est autosomique, 3 si elle est lié à l'X, 4 si elle est lié à l'Y et 5 si elle est mitochondriale. S'il s'agit d'un gène, un astérisque est ajouté devant l'identifiant.

[85] Qui ont # pour préfixe si la base moléculaire connue, % si la transmission est mendélienne, mais sans base moléculaire connue.

[86] https://www.omim.org/entry/182290

[87] Un seul patient peut suffire si l'argumentaire est convaincant.

1.7 Scores d'épissage

Le génome ne se limite pas aux régions codantes pour le diagnostic et les régions non-codantes, ont de fait un impact sur l'expression des gènes. De fait, environ 4% des variants dans la base de données Exome Aggregation Consortium (ExaC) altèrent l'épissage (Cheug 2019). Elles peuvent avoir un rôle régulateur notamment et, à ce titre, peuvent avoir un impact dans la pathogenèse des maladies rares. Il existe différentes atteintes de l'épissage possible, résumées dans la Fig. 25. Même si les données d'entrées de notre pipeline sont des séquençages d'exomes, il peut donc y avoir des atteintes de l'épissage. À défaut de pouvoir vérifier l'impact du variant, par exemple par RNAseq, il existe des prédicteurs informatiques permettant de prédire leur impact et le quantifier avec un score. Nous présentons ici deux outils très populaires et utilisés dans notre pipeline.

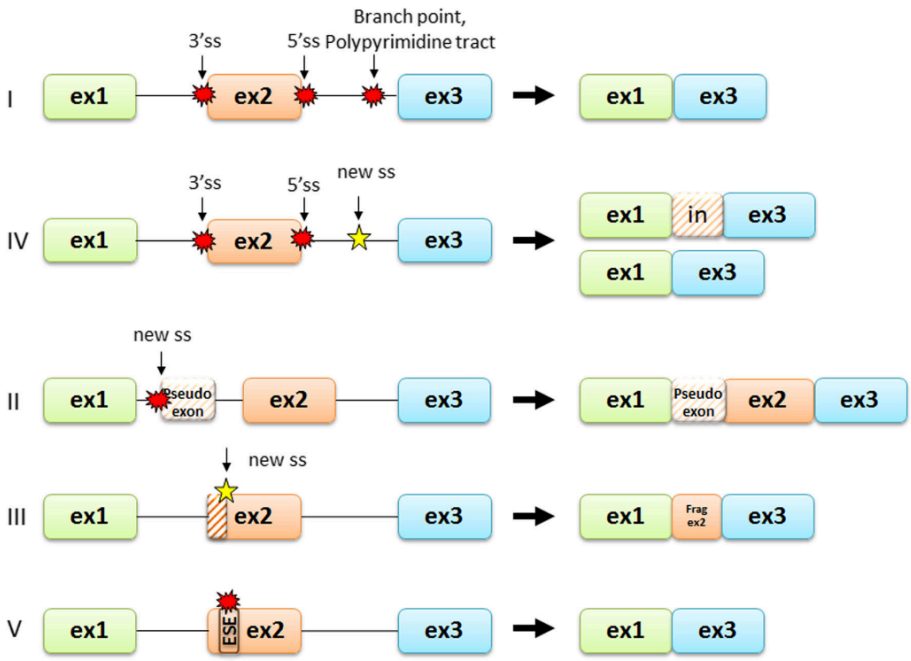
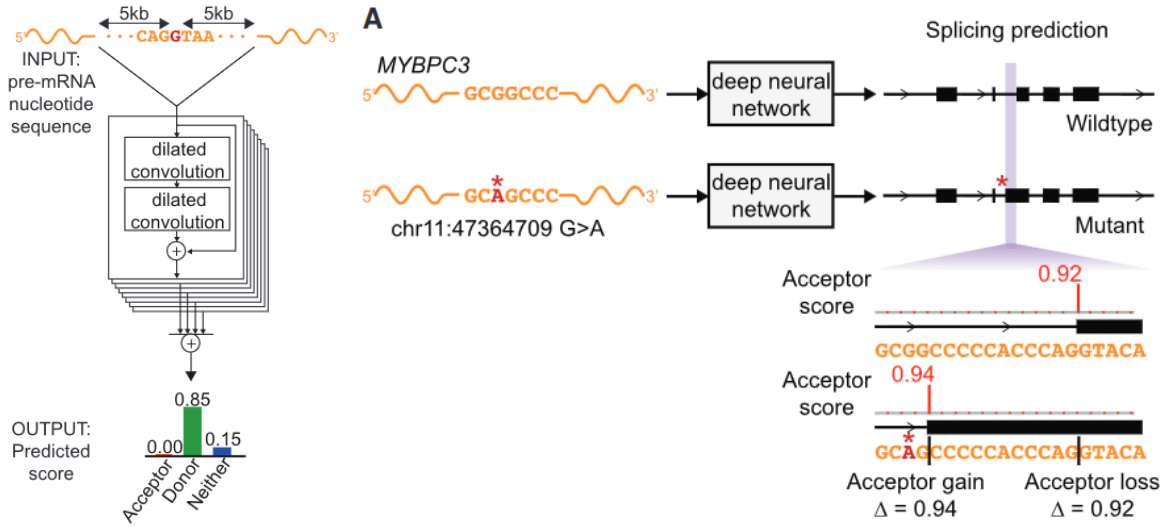


Fig. 25. – Les différentes catégories d'atteinte de l'épissage résumée par Anna et Monika (2018) I. Modifications des sites canoniques conduisant à un saut d'exon. IV. Modification de sites canonique mais conduisant à l'utilisation d'un site « cryptique » avec perte d'un fragment d'exon ou inclusion d'un fragment d'intron (illustré ici). II. Variant intronique profond conduisant à la création d'un nouveau site d'épissage avec l'inclusion d'un pseudo-exon. III. Variant dans l'exon consuidant à la perte partielle de l'exon. V. Variant dans l'exon conduisant à un saut d'exon (le plus souvent par modification d'un ESE (*exonic splicing enhancer*)).

1.7.1 SpliceAI

Ce modèle utilise un réseau de neurones pour estimer si une position dans l'ARN pré-messager est un site donneur, accepteur ou ni l'un ni l'autre. Il donne donc 4 scores : gain ou perte pour un site donneur et gain ou perte pour un site accepter. A titre d'exemple, la Fig. 26 montre un variant conduisant très probablement à un nouveau site accepteur d'épissage en amont du site initial. À noter qu'une visualisation des résultats a été récemment ajoutée par Sainte Agathe et al. (2023), qui est notamment utile pour les scores faibles (< 0.2) comme illustré sur la Fig. 27.



a) Principe de SpliceAI b) Validation par comparaison à des variants

Fig. 26. – Illustration a) du principe de SpliceAi selon Jaganathan et al. (2019) et b) de la procédure de validation où le variant représenté ici induit très probablement la création d'un nouveau site accepteur.

Le modèle a été entrainé sur les séquences d'ARN message précurseur[88]. Autour d'une position donnée, SpliceAI va utiliser les séquences à +/- 5kb. Une partie des chromosomes[89] a été utilisée pour entraîner le modèle, le reste pour le tester. Concernant ses résultats, SpliceAI a une précision de 95%[90]. Cette précision diminue à 84% pour les longs ARN non codants, mais cela est interprété positivement par les auteurs au vu du manque d'annotation des transcrits non codant. En utilisant l'atlas GTEx (Gene and Tissue Expression), les exons qui étaient sautés ou ajoutés ont un score de 0 et 1 respectivement. Ceux qui ont un épissage alternatif non négligeable ont, comme on peut s'y attendre, un score intermédiaire entre 10 et 90%. Par défaut,

[88] À partir des annotations fournies par GENCODE qui donne les exons , 5'UTR etc.

[89] Sans précisions dans l'article.

[90] Mesurée par la fraction de sites d'épissage correctement prédits où le nombre de sites d'épissage prédits est exactement le nombre de sites dans la base de tests.

SpliceAI considère les séquences dans un intervalle de 10kbp autour du nucléotide. Pour étudier l'impact de cet intervalle, les auteurs ont entrainé le modèle avec un intervalle de 80bp et un intervalle de 10kbp. Dans le cadre d'un petit intervalle, le modèle va privilègier les sites de jonctions autour des exons ou introns de petite ou grande taille. Pour un grand intervalle, ce seront au contraire les exons ou introns de taille moyenne (150 et 10 000 resp.) qui auront un meilleur score. Selon les auteurs, le premier résultat est en accord avec un article précédent. Mais l'utilisation d'un grand intervalle permet de voir les influences à grande distance et rendrait le modèle plus spécifique[91]. Enfin, les auteurs suggèrent que le modèle utilise implicitement des informations sur la chromatine[92].

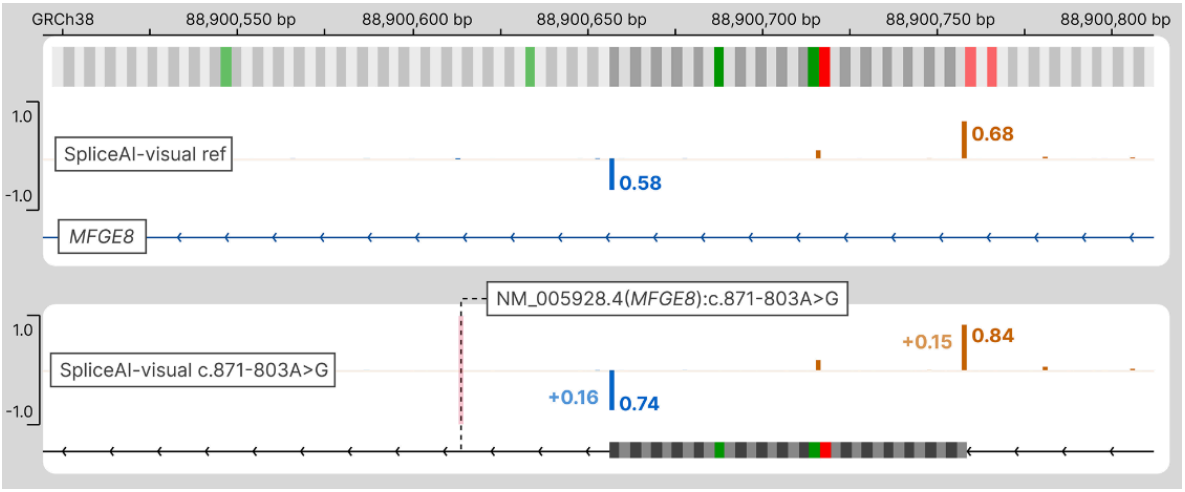


Fig. 27. – Illustration de l'intérêt de spliceAI visual (Sainte Agathe et al. 2023): le variant pathologique de *MFGE8* entraîne la création d'un exon en intronique profond: augmentation de la probabilité site donneur (bleu) et site accepteur (marron)

De plus, les résultats ont été comparés à des données de génomes et RNA-seq sur 164 individus (cohorte GTEx) en utilisant des variants rares (1 seul patient porteur)[93]. Les variants prédits comme touchant l'épissage se retrouvent principalement sur des nouveaux sites d'épissage ou aux frontières de sauts d'exons, ce qui suggère que le score est bien corrélé à un impact fonctionnel. Cet effet a été quantifié[94], résultant

[91]Selon eux, ce résultat est en accord avec l'expérimentation. Dans le cas de longs introns ininterrompus, l'élongation rapide de ARN polymérase peut avoir moins de temps pour reconnaître certains motifs.

[92]En examinant des sites accepteurs et donneurs « optimaux », ceux qui sont prédits avec une création d'exone sont situés sur des zones avec un fort taux d'occupation des nucléosomes.

[93]Le modèle n'a pas été entraîné sur des variants mais seulement sur les séquences d'ARN prémessager.

[94]En utilisant le maximum entre diminution du nombre de reads sur le site d'épissage touché et l'augmentation du nombre de reads sautant l'exon.

en trois quarts de validation des scores ≥ 0.5. Ceux avec des scores < 0.5 ont plus d'épissage alternatif. Les auteurs concluent à une sensibilité à 71% près des exons et 41% en intronique profond. Par comparaison aux outils existants à l'heure de la publication[95], SpliceAi est plus performants[96]. L'impact de facteurs confondants[97]a été également éliminé.

Des tests d'impact en clinique ont été conduits. En utilisant Exac et en comparant à la fréquence attendue du variant[98], ils ont trouvé que ceux avec un score $\geq 0.8\%$ sont soumis à une pression négative importante[99]. En intronique profond, soit à plus de 50bp des bornes d'exomes, ils étaient diminués de 56%, probablement lié à l'augmentation de la difficulté de la prédiction. Enfin, sur la cohorte DDD contenant des patients atteints de déficience intellectuelle et Simons Simplex (troubles du spectre autistiques), 28 variants *de novo* prédits comme atteignant l'épissage ont été sélectionnés. Parmi ceux-ci, 21 ont été confirmés en RNA-seq, soit un taux de validation de 75%[100].

1.7.2 SPiP

SpliceAI est l'un des rares outils à prendre en compte toutes les altérations d'épissage et semble avoir les meilleurs performances comparé aux autres outils. Cependant, son modèle a été entrainé sur des séquences « sauvages » et non sur des variants impactant l'épissage. Pour répondre à ce problème, SPiP a été développé par Leman et al. (2022) avec une approche de *random forest*.

Leur méthode consistait à rassembler 4 616 variants dont près de 2 000 modifiant l'épissage à partir d'étude ARN (4114 publiés, 502 non publiée) avec 95 000 variant servant de témoin[101]. Ces variants vont alors servir à entraîner le modèle (voir ci-

[95]Donc sans SPiP détaillé ci-dessous

[96]Outils testés : Genesplicer, Nnpplice, Maxentscan.

[97]Variants privés/courants, création de nouveaux sites canonique, variants dans zones plus distales, impact du choix des chromosomes pour l'entraînement du modèle.

[98]Le raisonnement est variants « courants » (>= 0.1% population humaine) ont moins d'effet fonctionnel, car ils n'ont pas été filtrés par la sélection naturelle donc on peut supposer que la pénétrance quasi-complète.

[99]78% de diminution pour les variants synonymes et introniques, 82% pour décalage du cadre de lecture, codonst stop ou atteinte d'un site canonique d'épissage.

[100]Les auteurs suggèrent que les 7 variants non retrouvés en RNA-seq ont peut-être une expression variable suivant le tissu considéré. Pour ces patients ont été séquencé des lignées lymphoblastoides cellulaires à partir du sang

[101]Le nombre de variants nécessaire a été estimé par un modèle bayesien. Si l'on estime à 2% le nombre de variants modifiants l'épissage, il faut donc 100 000 variants pour en avoir 1 924. Comme le nombre de variants dans un *assay* est insuffisant, des variants témoins ont donc été utilisé, supposé sans atteinte de l'épissage. En effet, les variants courants (MAF $> 5\%$) sont supposés de pas atteindre l'épissage.

dessous). Une fois celui-ci entraîné, l'algorithme se déroule en 2 temps : 1) le score SPiP est calculé. S'il est supérieur à un seuil fixé, le type d'altération va être calculé en utilisant 4 scores existants ou un nouveau métascore proposé par les auteurs pour prédire la création d'un nouveau site d'épissage (Fig. 28). Cette seconde étape est illustrée avec plus de détails en annexe (Fig. 68).

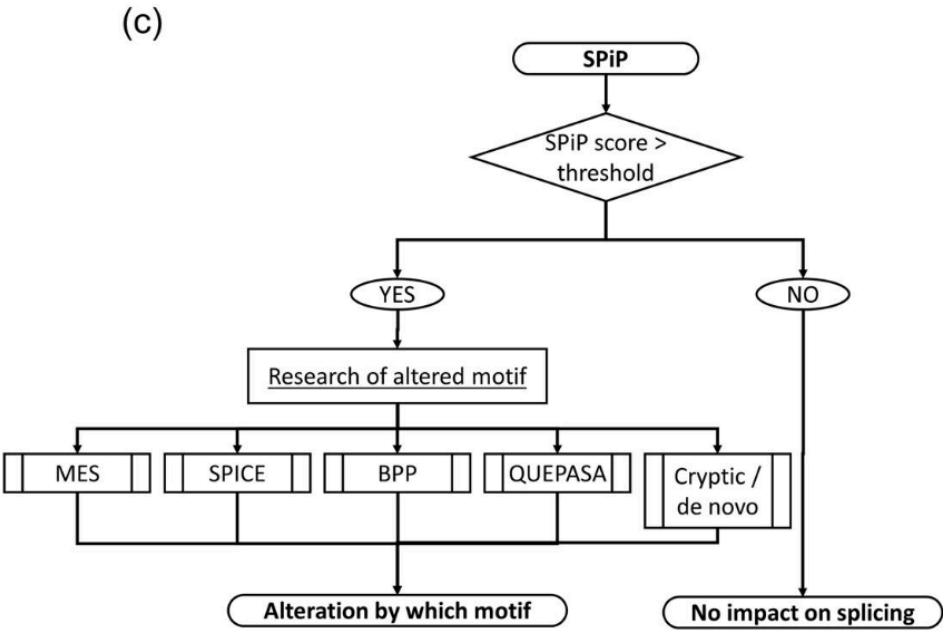


Fig. 28. – Algorithme pour SPIP (Leman et al. 2022). L'utilisation de modèles existants va dépendre de l'intervalle considéré, comme représenté sur la Fig. 29.

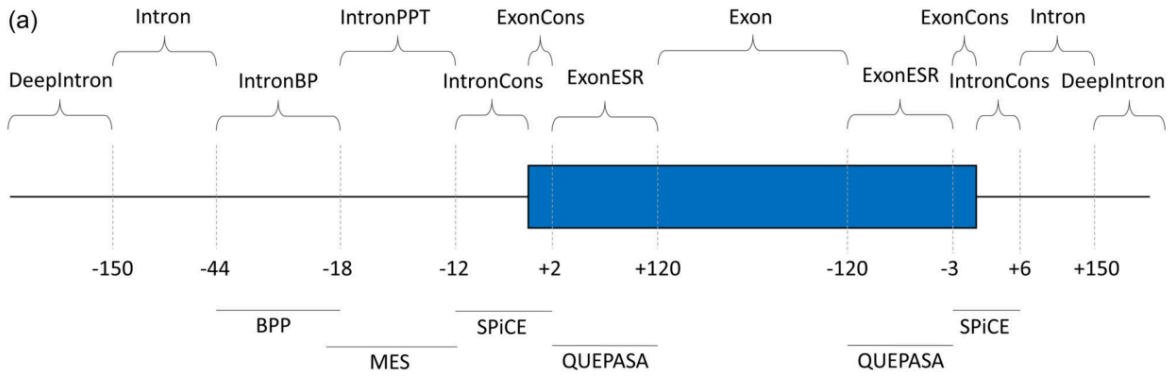


Fig. 29. – Régions introniques pour SPIP avec l'intervalle d'utilisation des différents modèles (Leman et al. 2022)

Il y a donc deux apports par Leman et al. Le premier est le score SPiP lui-même, modélisé avec une approche dite *random forest*[102]. Les prédicteurs utilisés sont le type

[102]En combinant une centaine d'arbres de décisions qui classifient le variant selon un nombre aléatoire de variables. Le score est le ratio du nombre d'arbres qui le classe en positif

du variant, les scores existants si le site est déjà connu, la taille des exons et introns, la distance au site d'épissage « naturel » le plus proche, s'il est sur un *branch point* prédit etc. Pour ce calcul, le modèle a été entraîné sur la moitié des données présentées plus haut. Le second apport est celui de la création d'un métascore estimant la création d'un nouveau site d'épissage[103]. Ce métascore a été entraîné (2/3 données) et validé (1/3) à partir de positifs extraits à partir des transcrits Ensembl et de négatifs définis comme des motifs AG et GT aléatoire en dehors de vrais site d'épissage[104]. Enfin, le métascore a été calculé comme le score maximum des sites potentiels autour du variant.

D'un point de vue pratique, SPiP donne donc une probabilité d'atteinte d'épissage et le motif impacté. Les scores sont pré-calculés. Le programme se présente sous la forme d'un script dans le langage R et parallélisé.

Parmi les variants ayant servi à valider leur modèle, le score SPiP est très élevé pour les sauts d'exon ou la création d'un site d'épissage (0.94 et 0.97 resp.). La création d'exon a en revanche de très mauvais scores, du à la rareté de ces évènements selon les auteurs. Pour les sites canoniques d'épissage et les PPTs[105], le score tendait vers 1. En cas d'altération de motifs *branch points* ou *exonic splicing regulators*, les scores étaients intermédiaires. Enfin, le score était assez bas en cas de création d'un nouveau site d'épissage ou d'altération complexe avec plusieurs motifs touchés. Les auteurs ont comparé leur modèle à SliceAI ainsi qu'à SQUIRLS, qui utilise une méthode similaire à SPiP. La comparaison a été faite en utilisant tous les variants ayant servi à l'entraînement et la validation.En calculant l'aire sous la courbe ROC[106], SPiP est à 0.986, SpliceAI à 0.965 et SQUIRLS à 0.766. Les données sont similaire pour une courbe VPP - sensibilité. Ils notent cependant que SpliceAI est meilleur en intronique et notamment en intronique profond[107]. Enfin, le temps par variant a été estimé au détriment de SpliceAI avec à 0.3s pour SPiP et 7.6s pour SpliceAI car ce dernier est vraiment performant sur carte graphique et ces résultats sont sur CPU.

1.7.3 Autres

Un autre candidat intéressant pour le calcul de scores d'épissages est Pangolin (Zeng et Li 2022). Ce modèle utilise du *deep learning* et propose un score sur 4 tissus dif-

[103] Par un régression logistique à partir des scores des sites consensus (avec MaxENtSCan (MES) et matrix position « SpliceSite Finder » (SSF)-like) et motifs régulateurs (Exon Skipping Ratio)

[104] À des positions ou il y environ 380 fois plus de faux site d'épissages que de vrais

[105] Polypyrimidine tract, voir Fig. 29

[106] 1-spécificité vs sensibilité.

[107] 5' + 6 bp et −44bp.

férents, à savoir le coeur, foie, cerveaux et testicule. Il a de plus été entraîné sur 4 espèces : l'homme, le macaque rhésus, le rat et la souris.

La comparaison à SpliceAI (mais pas SPiP) a été faite sur 27 000 variants d'Exac pour lesquels les modèles devaient estimer s'ils modifiaient l'épissage. Les résultats ont été comparés à du RNA-seq (Sort-Seq). Au final, Pangolin a une meilleure aire sous la courbe VPP-sensibilité (0.56 vs 0.476). Les auteurs ont testé la capacité de Pangolin a prédir des variants perte de fonction sur *BRCA1*. Les performances étaient médiocres sur tous les SNVs, ce qui était attendu, mais en excluant les faux-sens et nonsense, les résultats étaient très bons (aire sous la courbe à 0.95). Un autre test sur 800 SNVs de clinvar a montré qu'il pouvait identifier correctement des variants bénins et pathologiques (aire de 0.90 et 0.87), avec une meilleur performance que SpliceAI. Les variants modifiant l'épissage avaient une aire de 0.99.

1.7.4 Choix retenu: SpliceAI et SPiP

En conclusion, il est difficile de choisir un modèle parmi les deux, qui sont très populaires. Pour des raisons de support du génome de référence, SpliceAI est utilisé dans le pipeline en GRCh38 et SPiP pour T2T-CHM13.

1.8 Filtres

La partie peut-être la plus délicate dans le pipeline consiste à choisir les filtres pour rendre la sortie interprétable par un humain. Trop filtrer peut conduire à manquer des diagnostics, mais ne pas filtrer assez rend l'interprétation très fastidieuse. Il n'existe pas de consensus sur les filtres. En revanche, les recommandations sont très claires, que ce soit le COFRAC (COFRAC 2019) ou l'ACMG (Souche et al. 2022; Matthijs et al. 2015) : le laboratoire doit les tester, ce qui est le sujet du Chapitre 3. Il y a de nombreuses possibilités, comme des filtres « en dur » sur la qualité des variants, selon les bases de données utilisés, en fonction de l'annotation ou bien avec des critères spécifiques au laboratoire qui peut par exemple supprimer les variants liés à un biais techniques connu.

Les choix retenus dans Bisonex sont les suivants. Un premier filtre est effectué après appel de variants qui exclue les variants de dbSNP « non-rares »[108] et non pathologiques selon Clinvar. Puis l'on filtre également les variants avec une profondeur ≤ 30 ou ceux portés par 10 reads ou moins. Cela est fait en amont de l'annotation

[108]C'est-à-dire définis dans dbSNP par une MAF ≥ 0.01 dans l'une des 5 populations selon les données de 1000 génomes phase 3. Les 3 populations sont : African (AFR) , Admixed American (AMR) , East Asian (EAS) , European (EUR) , South Asian (SAS).

pour diminuer la quantité de données annotées. Un second filtre est mis en place après l'annotation où sont enlevés les variants non codants, intergéniques, en UTR, introniques, sur un pseudogène, ou sur un micro ARN (voir Fig. 30) *sauf* si le score d'épissage est supérieur à un seuil fixé. SpliceAI est utilisé pour ce critère en GRCh38, SPiP en T2T-CHM13. Le tout est résumé sur la Fig. 31.

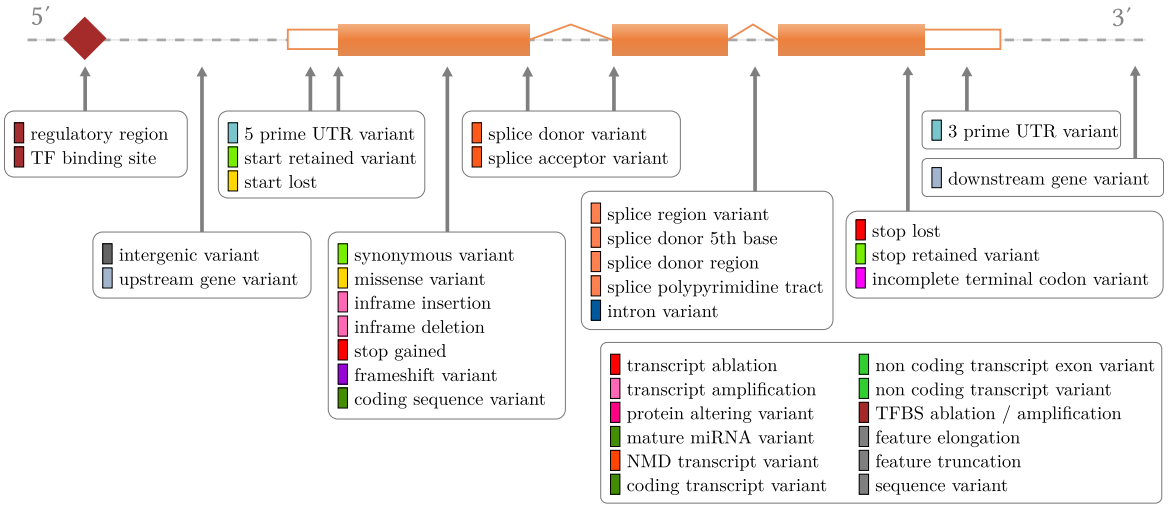


Fig. 30. – Conséquences fonctionnelles fournies par VEP selon http://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

1.9 Conclusion

Pour le pipeline, nous avons donc retenu des outils performants, validés et open-source dont l'imbrication est résumée sur la Fig. 31. La version actuelle n'utilise qu'un outil pour chaque étape mais il serait intéressant d'imiter d'autres pipeline qui exécutent plusieurs outils différents pour une même étape afin de cumuler les forces de chacun, comme l'appel de variant ou l'annotation.

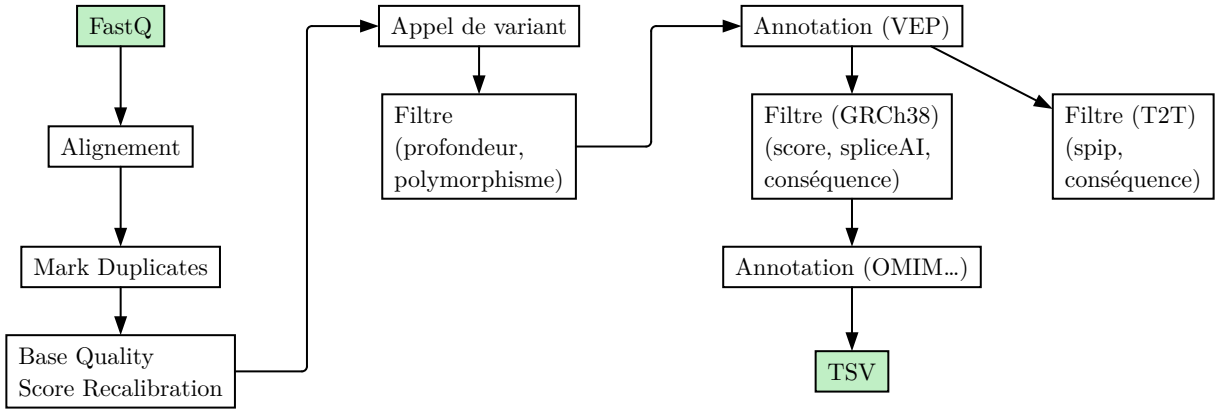


Fig. 31. – Vue d'ensemble du pipeline. La sortie est un fichier texte pour augmenter la portabilité et permettre une ré-inteprétation facile à l'aide de l'outil `grep`.

Chapitre 2

Reproductibilité, portabilité, performance

La reproductibilité d'une expérience scientifique consiste à utiliser la même méthodologie pour retrouver les mêmes résultats. C'est actuellement un vrai défi en science, car il est le souvent difficile, parfois impossible, de reproduire des expériences scientifiques à cause de protocoles imprécis ou d'outils statistiques utilisés de manière inappropriée par exemple. En bio-informatique, pour un code donné avec un jeu de donnés fixé, le calcul devrait donner le même résultat à chaque fois. Malheureusement, le code ou les données publiées ne sont pas toujours accessibles. De plus, installer exactement la même version d'un logiciel est un problème non trivial. Il s'agit cependant d'un point bien précisé par le COFRAC dans le SH-GTA-16 : la documentation doit mentionner les différentes versions des outils[109], ces versions doivent être bloquées en phase préliminaire au dossier de qualification[110], précisées dans le compte rendu[111]. Enfin, toute modification, y compris une mise à jour d'un logiciel, voire le changement d'une ligne de code, nécessite une nouvelle qualification[112] (COFRAC 2019).

[109]Section 4.3 Maitrise des documents : « Dans ce cadre, la structure devrait documenter : [..]: la définition des différents outils bio-informatiques ainsi que leur « versionnage » et la documentation fournisseur associée. »

[110]Annexe 3 « A l'issue de cette phase [préliminaire au dossier de qualification]:, les configurations retenues (versions de logiciel, paramètres…) devraient être verrouillées en vue d'entreprendre la qualification et la vérification/validation de méthode. »

[111]5.8 Compte rendu des résultats « les recommandations des sociétés savantes mentionnent les informations à communiquer aux utilisateurs [..]: dans les comptes-rendus, par exemple : […]:la version du pipeline NGS (ou des logiciels utilisés : pour l'alignement, l'appel de variants, etc.), la base de données, l'interface d'interprétation le cas échéant »

[112]Processus analytique « Toute modification d'une étape de l'analyse, notamment du pipeline bio-informatique ou des changements de versions des bases de données ou du génome de référence

Pour exécuter un code informatique avec des besoins en calculs importants, nous avons identifié 3 problématiques. D'abord, l'*installation* des différentes composantes du pipeline avec leur dépendences. Un outil souvent utilisé pour cela est un gestionnaire de paquets (comme Conda) qui propose une approche centralisée en mettant à la disposition des utilisateurs une liste de logiciels et les outils pour installer[113]. Un second problème concerne l'*exécution* du programme : comment garantir une exécution correcte du programme quelque soit l'ordinateur utilisé ? « Correcte » veut dire sans conflits avec les autres programmes en cours, par exemple ceux d'autres utilisateurs, mais aussi reproductible d'une machine à une autre. Une solution courante consiste à utiliser des « conteneurs » (Docker, Singularity et machines virtuelles) permettant de contrôler l'environnement pendant l'exécution[115]. Dans un troisième temps, comment *ordonner* les différentes étapes d'un pipeline afin que l'exécution se fasse dans l'ordre adéquat mais puisse redémarrer en cas d'erreur depuis la dernière étape qui a échoué ? Et comment gérer le lancement du pipeline pour des dizaines de patients ? Cette notion est celle du *workflow* (ensemble de tâches à effectuer) pour laquelle il existe des outils dédiés dont les plus populaires sont Snakemake (Köster et Rahmann 2012), Nextflow (Di Tommaso et al. 2017), WDL[116][117]. Le principe est de décomposer le pipeline en étapes élémentaires correspondant à l'exécution d'un exécutable sur un[118] fichier d'entrée produisant une sortie[119]. Le programme se chargera alors de l'ordonnancement et de l'exécution sur des multiples architectures (*cloud*, supercalculateur…). Enfin, les pipelines bio-informatiques d'analyse de données de séquençage nécessitent souvent des ressources non négligeables[120] et à ce titre, doit souvent s'exécuter sur des supercalculateurs ou architectures sur le *cloud*.

Les apports de cette thèse sont à ce titre triple. D'une part, nous avons exploité l'outil Nix pour proposer une installation complètement reproductible avec de nombreuses contributions open-source (Chapitre 2.1). La portabilité a été réalisée avec le logiciel Nextflow, permettant une exécution sur supercalculateur et ordinateur de

doit faire l'objet d'une nouvelle qualification avec des jeux de données-tests connus, des tests de non-régression associés dictés par une analyse de risques. »

[113]Il existe des dépôts supplémentaires selon les domaines comme bioconda[114] en bio-informatique

[114]https://bioconda.github.io/

[115]À noter qu'il existe des solutions plus simples et moins coûteuses en ressources comme des *namespace*, c'est-à-dire des espaces d'exécution réservés. Les conteneurs créent au contraire un environnement complet dédié au programme dans un style de « boîte à sable ».

[116]https://github.com/openwdl/wdl

[117]Mais il en existe de nombreux autres : Toil, Cromwell, Ruffus, Rubra, Ccwdl récemment avec Guix etc.

[118]Ou plus

[119]Ou plus

[120]Une exécution prend plusieurs heures sur une architecture adaptée, mais on est loin de mois de calculs de certains modèles de climat par exemple.

bureautique (Chapitre 2.2). Le tout afin d'être performant en exploitant la puissance de calcul d'un supercalculateur (Chapitre 2.3).

2.1 Reproductibilité avec Nix

Les difficultés d'installation d'un pipeline reproductible sont résumées sur la Fig. 32. Plus précisément se pose tout d'abord la question des dépendances, c'est-à-dire les autres programmes nécessaires à la bonne exécution du pipeline. L'exemple le plus parlant, et le plus complexe, du pipeline est celui de l'outil d'annotation VEP. Écrit en Perl, il va donc dépendre de l'interpréteur Perl et d'un ensemble de paquets également en Perl. L'un de ces paquets, `Bio::DB::Bigfile`, nécessite des outils développés par l'UCSC (*University of California, Santa Cruz*), qui sont eux écrits en C. Installer tout cela peut donc prendre plusieurs heures ou jours. Un second problème concerne la cohérence des versions. En continuant sur l'exemple de VEP, il nécessite une ancienne version de `Bio::DB::Bigfile` et donc des logiciels de l'UCSC[121]. On voit ainsi que les mises à jour successives peuvent devenir difficiles pour assurer la compatibilité de tous les outils du pipeline. Installer manuellement les différentes dépendances conduit souvent à l'existence sur un même système de différentes versions du même outil, augmentant d'autant le risque d'erreur. De plus, comment gérer un même outil installé avec des fonctionnalités différentes ? Par exemple, la version « standard » de l'aligneur `bwa mem` et une version installée manuellement supportant les reads alternatifs du génome. S'assurer que la « bonne » version est utilisée pour analyser les données d'un patient devient délicat. Enfin, comment disposer du logiciel sous forme exécutable ? Faut-il compiler depuis le code source ou le récupérer directement l'exécutable ? La seconde approche a l'avantage de la simplicité et semble être une solution facile : distribuer le même exécutable à tout le monde ne supprimerait-il pas toutes ces difficultés de reproductibilité ? Outre des problèmes de sécurité[122], il faut déjà générer un exécutable par architecture (Linux 32 bits, Windows 64 bits…). De plus, il est impossible de choisir les options fournies par cet exécutable, ni d'exploiter les performances de certaines architectures, ce qui peut conduire à des pertes de performances[123].

[121] Les développeurs de VEP n'ont pas effectué la mise à jour depuis des années, car cela n'est « pas une priorité » https://github.com/Ensembl/ensembl-vep/issues/1412

[122] Il est difficile de savoir exactement ce que fait un exécutable, car il s'agit de données binaires

[123] Ce dernier point est extrêmement important sur les supercalculateurs qui disposent de compilateurs propriétaires optimisés pour ces machines. Il est donc plus que souhaitable de compiler en fonction de l'architecture pour exploiter tout leur potentiel.

On peut distinguer 4 grandes approches pour installer un logiciel. La première consiste à proposer une approche centralisée via un *gestionnaire de paquets* depuis lequel l'utilisateur peut installer tous les programmes. Cela est couramment utilisé pour les systèmes d'exploitation et le *Microsoft store* en est un bon exemple. Cependant, les logiciels proposés et leurs mises à jour sont indépendantes de la volonté de l'utilisateur. Deuxièmement, il existe de bonnes solutions techniques pour certains langages, comme `poetry` en Python ou `cargo` en Rust mais, par définition, ne conviendront pas pour d'autres langages. Une troisième alternative consiste à récupérer le code source et à le compiler. Comme détaillé précédemment avec l'exemple de VEP, cette approche devient particulièrement coûteuse lorsqu'il faut installer chaque dépendance manuellement et rend les mises à jour hasardeuses et délicates. On peut mentionner également l'utilisation de conteneurs ou de VM (*Virtual Machime*) pour disposer d'un environnement séparé mais cela ne résout que certains problèmes. Les limites de ces approches sont résumées sur le Tableau 2.

Enfin, et c'est le choix retenu pour cette thèse, des outils ont été récemment développés proposant des installations inter-plateformes, indépendantes du langage et 100% reproductible (Nix, Guix). Ces deux outils proposent des approches similaires. Les différences sont principalement entre le langage utilisé[124], le nombre de logiciels disponibles et la communauté d'utilisateurs. Ces deux derniers points sont en faveur de Nix avec plus de 3 fois le nombre de paquets disponibles dans Nix[125]. Sur le plan bioinformatique, la situation est assez hétérogène[126]. Le choix a été fait *in fine* après une discussion avec les administrateurs du centre de calcul de Franche-Comté (mésocentre), qui ont décidé de tenter l'expérience avec Nix et que nous remercions pour leur collaboration.

[124]Nix utilise son propre langage, Guix utilise Guile.

[125]On notera cependant que Guix semble être plus populaire dans la communauté des supercalculateurs, privilégie les licences libres et a bénéficié récemment d'un article dans Nature (Vallet, Michonneau, et Tournier 2022).

[126]Guix propose une ancienne version de VEP (absent de Nix avant cette thèse) mais n'a pour l'instant que la version 3 de GATK https://github.com/guix-science/guix-science/issues/28

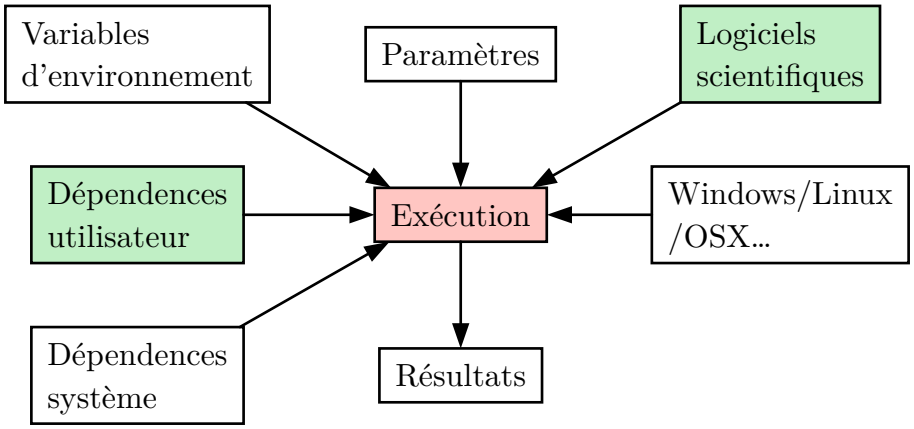


Fig. 32. – Illustration du probème de reproductibilé de manière générale , adapté de (Devresse, Delalondre, et Schürmann 2015). En vert, la partie à laquelle cette thèse contribue

Contrainte	Linux	NetBSD	MacOS	Docker	VM
Reproductible				✓	✓
Portable	✓	✓	✓	✓	
Tracabilité			partiel	partiel	partiel
Plusieurs utilisateurs		✓	✓	✓	✓
Pluseurs version				✓	✓
Binaires	✓			✓	✓
Dépôts centralisés	✓			✓	✓
Compilation isolée	✓			✓	✓
Exécution isolée				✓	✓

Tableau 2. – Limites des approches par gestionnaire de paquets ou conteneurs : Linux (Redhat/Debian avec respectivement RPM et DPKG), NetBSD (pkgsrc), conteneur (Docker), VM (adapté de Devresse, Delalondre, et Schürmann (2015)). « Plusieurs versions » signifie que plusieurs versions différentes d'un même logiciel peuvent coexister. Le terme « isolé » pour la compilation ou l'exécution signifie que cette étape dispose d'un espace dédié pour éviter les interférences avec d'autres programmes.

L'approche adoptée par nix est détaillée dans Dolstra, Jonge, et Visser (2004), Kowalewski et Seeber (2022). Son modèle est simple mais répond efficacement aux problématiques ci-dessus. Tout d'abord, chaque paquet est identifié de manière unique et, une fois installé, ne peut être modifié. Son identifiant dépend à la fois des dépendances mais aussi du code source donc mettre à jour le code va générer un nouvel identifiant. Chaque paquet est installé à un emplacement unique dépendant du nom et de l'iden-

tifiant dans un dossier réservé à Nix[127]. Cela permet donc de disposer de plusieurs versions installées de manière simultanée et sans risque d'erreur (Tableau 3). Chaque paquet est créé en compilant le code source correspondant[128] dans un environnement dédié de type « bac à sable ». Par défaut, un répertoire central propose, sous réserve d'un accès internet, à l'intégralité des paquets disponibles. Il est également possible de définir localement ses propres paquets (*packaging*). Enfin, cette approche se prête bien aux difficultés des environnements des supercalculateurs, pour lesquels les solutions existantes sont présentées dans le Tableau 4.

Logiciel	Version	Emplacement
BWA	0.7.17	/nix/store/a97smky0zjwfgllfsj57yrzhjs14f1ks-bwa-0.7.17
BWA	latest	/nix/store/ajrsypxjh811bbcmws3fa6vz8mrv081g-bwa-unstable-2022-09-23

Tableau 3. – Exemple de versions différentes de `bwa` installées en parallèle avec Nix.

Contrainte	Nix	Classical	Spack	Singularity	Modules
Mult. versions	✓		✓	✓	✓
Transferable	✓		partiel	✓	
Reproducible	✓	partiel		partiel	
Composable	✓	partiel	partiel		partiel
Customizable	✓		✓	✓	✓
Multiuser	✓		✓	✓	✓

Tableau 4. – Limites des solution sur supercalculateurs, selon Kowalewski et Seeber (2022). À comparer avec le Tableau 2

Nous avons identifié deux difficultés pouvant empêcher une adoption plus large en bio-informatique. D'une part, l'apprentissage d'un nouveau langage de programmation est une barrière supplémentaire. D'autre part, malgré de nombreux paquets disponibles, certains logiciels clés ne sont pas disponibles. L'un des apports de cette thèse est donc la contribution de plusieurs paquets indispensables pour un pipeline de génétique constitutionnelle. Ceux-ci ont été soumis à la communauté pour être facilement accessible. Si certains sont disponibles dans la dernière version de Nix, d'autres sont en cours d'adoption comme illustré sur la Fig. 33. Certains de ces pa-

[127] `/nix/store` sous Linux.

[128] À noter que les dépendances sont stockées sur un serveur sous forme d'exécutable pour économiser du temps de calcul aux utilisateurs. Mais grâce à l'architecture de Nix, les exécutables sont parfaitement reproductibles et, d'autre part, il est possible de changer les options des dépendances.

quets ont nécessité un travail non négligeable, principalement VEP avec ses dépendances multiples et complexes. S'agissant d'une communauté open-source et victime de sa popularité[129], les délais avant que les soumissions ne soient acceptées peuvent être conséquents comme le montre la figure. Au final, toutes les dépendances, illustrées sur la Fig. 34, sont accessibles depuis Nix. Ces contributions ont été preséntés à la conférence FOSDEM (*Free and Open source Software Developers' European Meeting*) 2024 dans la salle dédié à Nix.

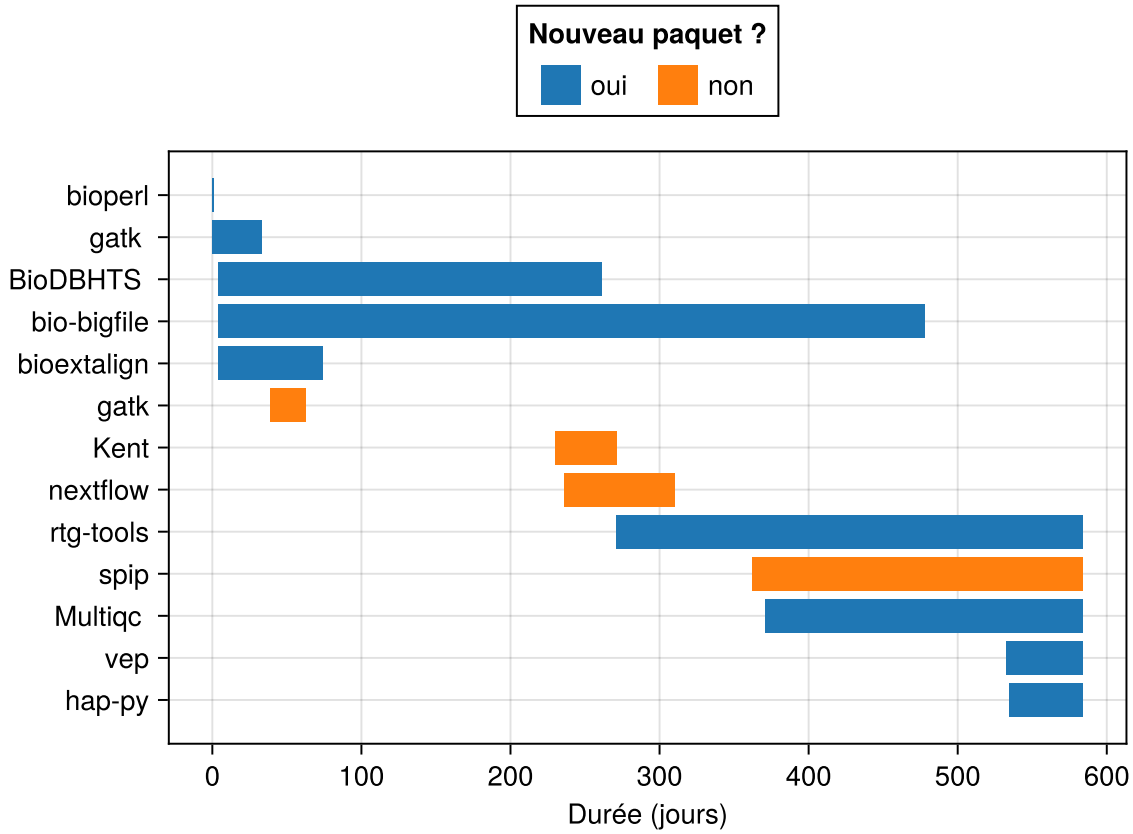


Fig. 33. – Contributions de nouveaux paquets bio-informatiques dans Nix sur Github lors de cette thèse.

[129]On compte en moyenne 5 000 demandes de contributions en cours (*pull requests*).

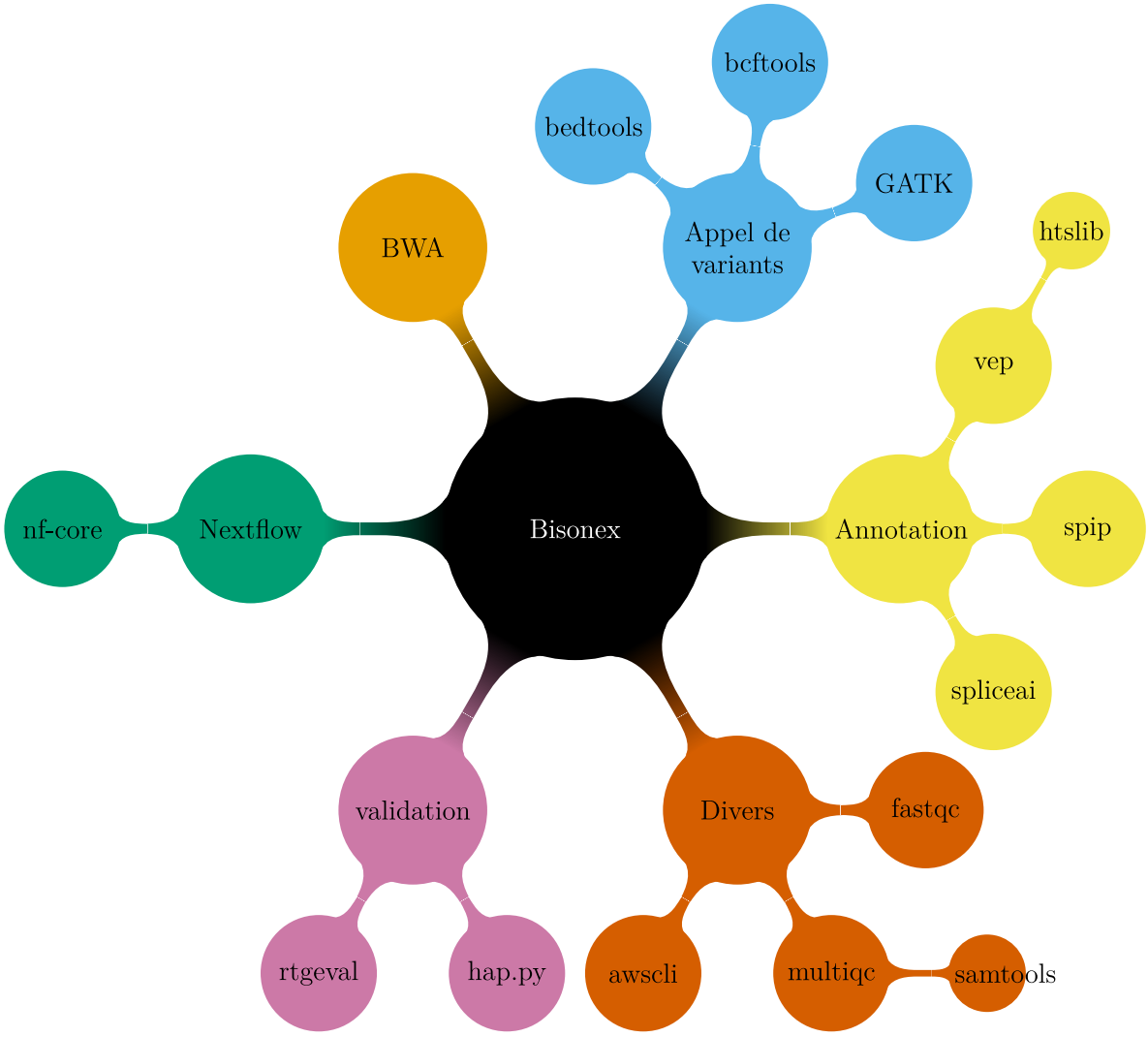


Fig. 34. – Dépendances du pipeline Bisonex.

2.2 Portabilité avec Nextflow

Pour concevoir un pipeline, au lieu d'une approche monolithique avec un seul programme comportant différentes étapes, une alternative est de le décomposer en tâches unitaires, qui seront chaînées ensemble par un outil. Ce paradigme permet de s'affranchir de l'architecture, ce qui permet d'avoir un code tournant sur différents supercalculateurs, un ordinateur de bureau ou encore sur le *cloud*. Il permet également de gérer l'ordonnancement des différentes étapes. Cette approche est devenue très attractive en bio-informatique avec de nombreux logiciels implémentant ce paradigme. Parmi les plus populaires, on peut citer Nextflow, Snakemake (Köster et Rahmann

2012), Cromwell[130], WDL[131]. À côté de ceux-ci, il existe une myriade d'outils alternatifs. L'un d'eutre eux, Bionix s'appuie sur nix (Bedő, Di~Stefano, et Papenfuss 2020).

Nous avons choisi Nextflow comme outil du fait de sa popularité et d'une communauté active et en pleine croissance, surtout en bio-informatique. Il supporte de nombreuses architectures pour l'exécution du pipeline, notamment Slurm qui est utilisé sur le centre de calcul de Franche-Comté. Enfin, et non des moindres, la communauté a développé des outils communs pour les pipelines sous forme de « modules »[132] et de pipelines prêts à l'emploi[133]. Les premiers permettent d'utiliser rapidement des outils existants dans un pipeline, comme les aligneurs ou outils d'appels de variants les plus connus[134]. Les seconds sont nombreux, mais selon une règle simple : un seul pipeline par type d'analyse, qui sert de référence. Une comparaison de Nextflow aux autres outils est détaillée en annexe (Tableau 25).

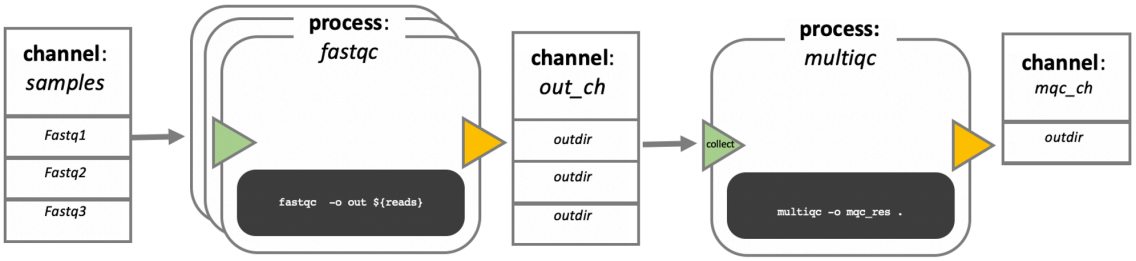


Fig. 35. – Un processus nextflow est l'unité de base d'un pipeline (source: https://carpentries-incubator.github.io/workflows-nextflow/aio.html)

Le principe est donc de définir dans un langage spécifique à Nextflow mais inspiré de Java des unités élémentaires effectuant une action sur un ensemble de fichiers, principe illustré dans la Fig. 35. Ces unités sont connectées avec une syntaxe simple définissant l'ordonnancement ce qui définit le pipeline. Deux exemples, détaillés ci-dessous, sont donnés dans la Fig. 36. Nextflow nous permettra alors d'être exécuté sur des architectures :

- Cloud (AWS, Azura, Google Cloud, Google Life Science)
- Supercalculateur: Bridge, Flux, HyperQueue, LSF, Moab, OAR, PBS, SGE, Slurm
- Divers: HTCondor, Kubernetes Task Execution Schema par GA4GH, Ignite, NQSII,

[130] https://github.com/broadinstitute/cromwell (GATK)
[131] https://support.terra.bio/hc/en-us/sections/360007274612
[132] https://nf-co.re/modules
[133] https://nf-co.re/pipelines
[134] Le revers de la médaille est que ces modules sont en constante évolution. Ayant choisi Nix, une version a été figée pour ces modules

Cette liste vise à illustrer la complexité d'avoir un pipeline portable et de ce fait l'intérêt de Nextflow. Deux autres fonctionnalités importantes sont fournies par cet outil : en cas d'échec de l'un des composants du pipeline, il sera redémarré à partir de la dernière étape réussie. Enfin, si des composants du pipeline peuvent être exécuté en parallèle, comme l'alignement du génome de plusieurs patients, cela sera fait automatiquement.

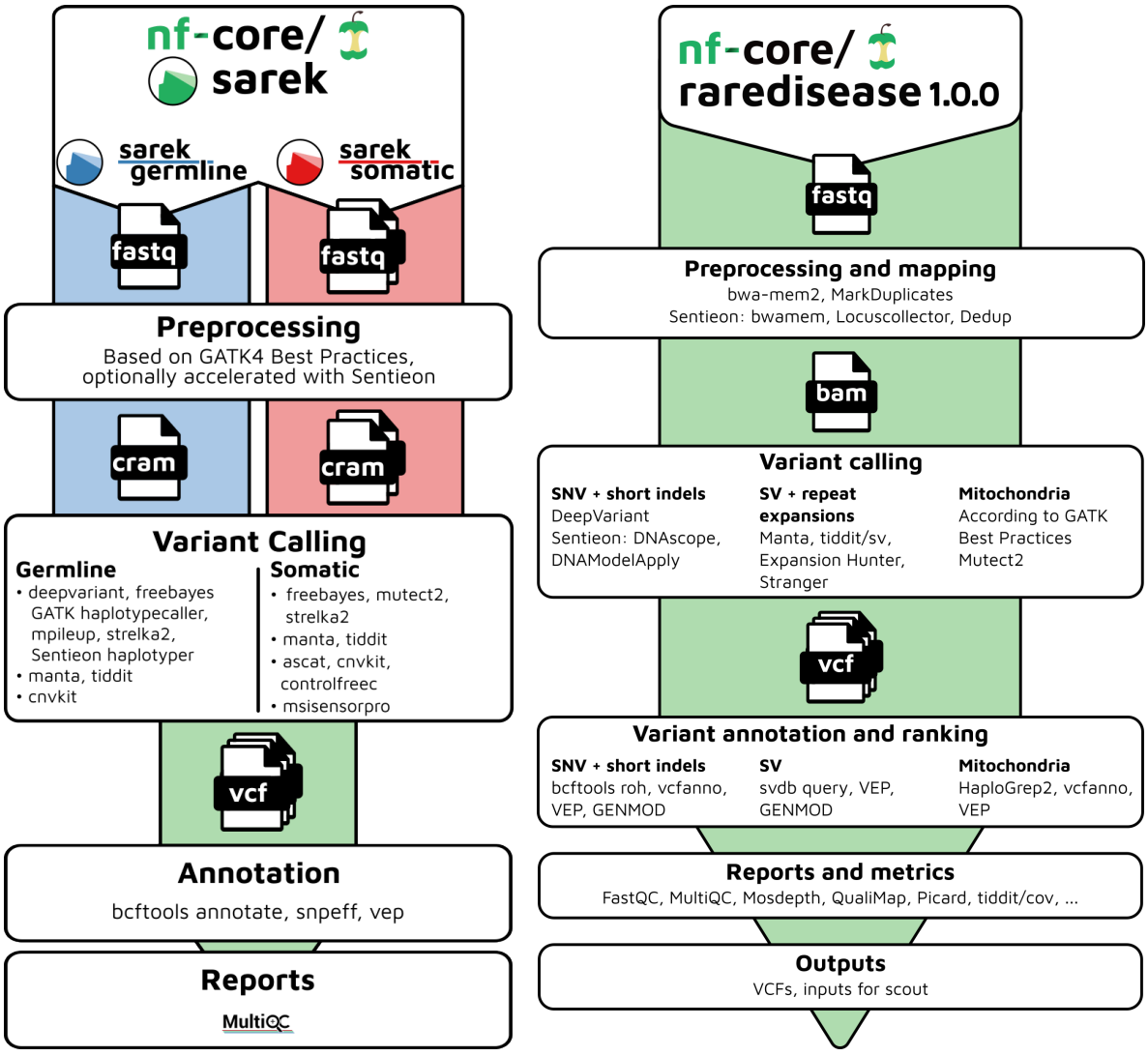


Fig. 36. – Les 2 pipelines dans *nf-core* pouvant analyser des données de génome ou d'exome : `sarek` à gauche (somatique et constitutional) et `raredisease` (constitutionnel)

Il existe deux principales limites à Nextflow. Tout d'abord, il faut arriver à comprendre et maîtriser le modèle d'exécution de Nextflow[135]. Deuxièmement, *nf-core*

[135]Le langage lui-même n'est pas vraiment un obstacle, car la syntaxe est assez claire

a choisi d'éviter la multiplication de pipeline pour une même tâche. En pratique, il faut adapter le pipeline existant à l'usage du laboratoire. À l'heure d'écriture de cette thèse, il existe deux pipelines permettant de faire jusqu'à l'appel de variant en génétique constitutionnelle (Fig. 36). Nous avons choisi d'écrire notre version afin de l'adapter à notre usage. Au début de cette thèse, seul `sarek` existait et proposait la fonctionnalité en somatique et constitutionnel. Durant sa conception, deux choix avaient été faits, qui ne convenait pas à notre usage. Tout d'abord, l'utilisation du nouveau format de fichier CRAM, qui est pour l'instant déconseillé par GATK. Ensuite, ce pipeline utilise pour le génome de référence et les autres bases de données d'annotation iGenomes, hébergée par Illumina[136]. Ce site propose pour les humains et d'autres organismes les versions majeures. Nous avons préféré garder la possibilité d'utiliser des versions plus récente, comme les mises à jour mineures du génome. `raredisease` n'étais pas disponible au début de cette thèse et délègue à l'utilisateur le choix des réferences[137].

Les spécificités de notre pipeline sont, d'une part, contrairement aux différents pipeline de *nf-core*, l'installation des dépendances avec Nix. L'approche de *nf-core* est d'utiliser un conteneur[138] ou en dernier recours, `conda`. L'avantage de notre approche est d'être reproductible pour les différents programmes et d'avoir une meilleure portabilité. Une seconde différence, abordée plus haut, est que le pipeline vérifie la présence des différentes sources d'annotation et les téléchargent si elles n'existent pas. De plus, nous avons porté directement sous Nextflow la première version de notre pipeline, qui est donc adapté à l'usage bisontin. Enfin, les variants sont filtrés et annotés pour avoir un résultat prêt à être interprété.

2.3 Performances

Une manière d'augmenter la vitesse d'exécution d'un calcul informatique consiste à le décomposer en tâches indépendantes et les exécuter en parallèle. Pour cela, il y a deux conditions importantes. Premièrement, le problème doit se prêter à une telle décomposition. Si les calculs nécessitent de nombreuses communications entre les différentes tâches, beaucoup de temps sera passé à attendre les échanges de messages. Deuxièmement, les composant et les programmes de l'ordinateurs doivent permettre d'exécuter des tâches en parallèle.

[136] https://emea.support.illumina.com/sequencing/sequencing_software/igenome.html
[137] https://nf-co.re/raredisease/1.1.1/docs/usage
[138] Docker, Singularity, Podman, Shifter, Charliecloud, Apptainer, Conda

Il existe 2 types principaux d'unité de calcul : les processeurs (CPUs), très utilisés en cas de calcul intensifs, et les cartes graphiques (GPUs), initialement conçues pour afficher en parallèles de nombreux pixels à l'écran. Les architectures actuelles sur supercalculateurs sont mixtes, car elles proposent à la fois des CPUs regroupés en *noeuds* de calculs ou des GPUs. Le pipeline actuel s'exécute sur CPUs seuls. Dans ce cadre, il est important de noter la différence entre l'unité de calcul élémentaire physique du processeur, nommé un cœur (ou *core* en anglais), et l'unité de calcul logicielle qu'est le *thread* (voir Fig. 37)[139]. On peut découper la charge de calcul en 100 000 threads, mais si le CPU ne contient que 4 coeurs, il n'y aura au maximum que 4 calculs peuvent s'exécuter en même temps. C'est pour cela que les codes nécessitant des temps de calculs longs s'exécutent souvent sur des supercalculateurs possédant plus de processeurs

Notre pipeline a 2 niveaux de parallélisation du code : il traite plusieurs patients en parallèle et il parallélise les étapes les plus coûteuses d'un même patient. En pratique, les deux étapes les plus coûteuses sont l'alignement et l'appel de variant. Avec le choix de BWA-mem comme aligneur, l'algorithme est « trivialement » parallélisable, car les reads sont alignés indépendamment Les performances de cette étape peuvent être grandement améliorées par le centre de calcul de Franche-Comté sous Nextflow. En revanche, l'appel de variant effectue des ré-alignements rendant les tâches inter-dépendantes. Il existe une version d'HaplotypeCaller exploitant une architecture spécifique, Spark[140], mais elle est encore actuellement en version de test et déconseillée par le GATK. Cependant, une partie de l'algorithme (PairHHM) peut être optimisée en utilisant une spécificité de certains processeurs du supercalculateur[141].

Une fois ces améliorations ajoutées, nous avons mesur le gain de la parallélisation pour un patient. En pratique, le nombre de *threads* est augmenté en mesurant à chaque fois le temps d'exécution. On s'attend à ce que le gain soit de moins en moins intéressant à partir d'un seuil pour les raisons mentionnées plus haut (charge de calcul par *thread* peu importante). Une métrique classique utilisée est le *speedup* que l'on peut définir comme le ratio entre le temps de calcul pour 1 *thread* et n *threads* :

$$S = \frac{T_1}{T_n}$$

Idéalement, le temps de calcul de n *threads* est n fois plus rapide que celui d'une *thread* donc le *speedup* idéal est n. Les résultats pour l'aligneur et l'appel de variant sont

[139] Pour fixer les idées, les ordinateurs de bureaux contiennent désormais des processeurs avec au moins 4 cœurs.

[140] https://spark.apache.org/docs/latest/cluster-overview.html

[141] Jeux d'instruction AVX

présentés à la Fig. 38. On observe de bonnes performances pour BWA-mem jusque 32 threads. En revanche, HaplotypeCaller n'est pas amélioré par le nombre de threads, ce qui était attendu.

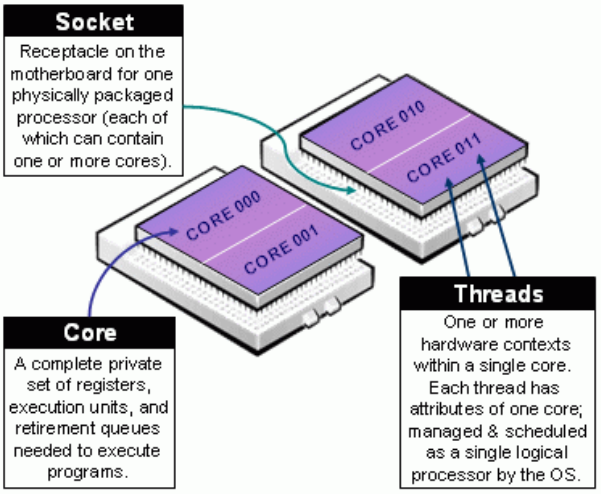


Fig. 37. – Différence entre une *thread* et un coeur (*core*). Source: https://slurm.schedmd.com/mc_support.html

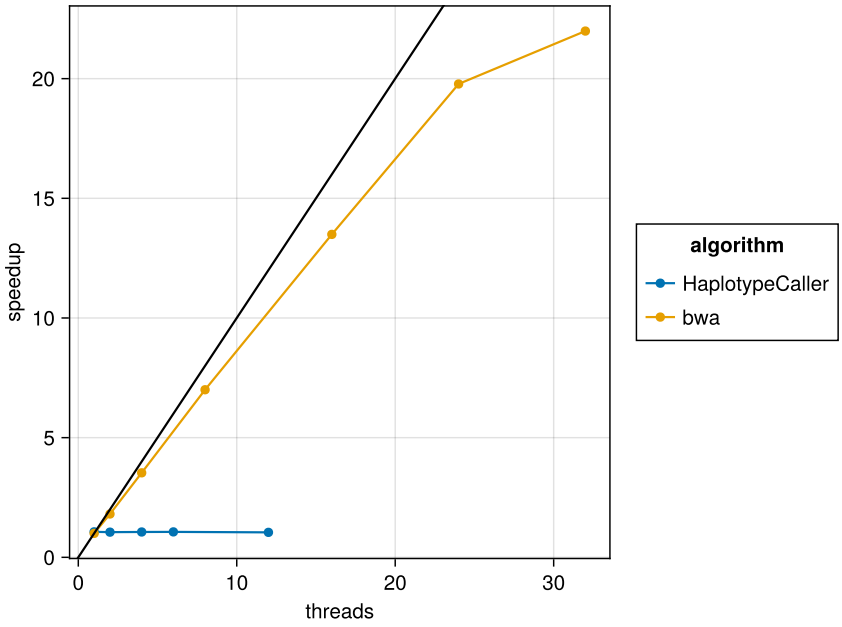


Fig. 38. – Speedup pour les deux programmes les plus coûteux : BWA-mem et HaplotypeCaller. Le speedup idéal correspond à une relation linéaire (en noir).

Une analyse des différentes métriques pour chaque étape d'une exécution du pipeline a montré que l'alignement est l'étape la plus demandeuse en termes de ressource

processeur (Fig. 69) et qu'elle est la plus coûteuse en mémoire avec le marquage des doublons après l'alignement (Fig. 70). L'appel de variant est de loin le processus écrivant le plus de données, suivi par l'alignement (Fig. 71) alors que l'écriture prédomine pour l'aligneur (Fig. 72). Enfin, grâce à la parallélisation, l'étape la plus longue en temps écoulé est l'appel de variant (Fig. 39).

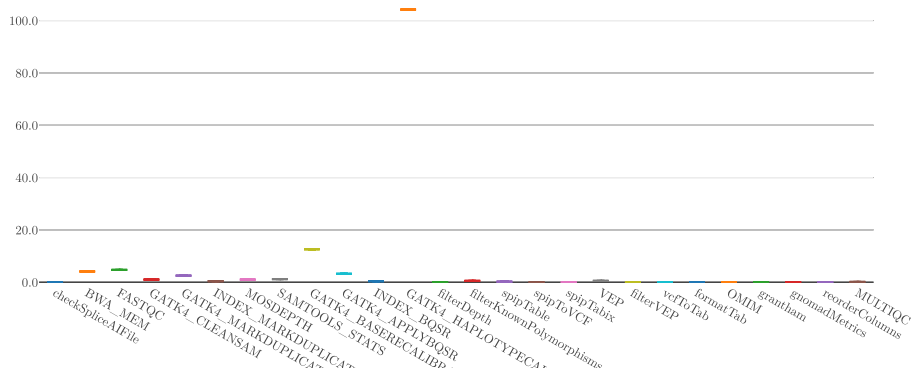


Fig. 39. – Temps effectif écoulé (en secondes) de chaque étape du pipeline pour une exécution.

D'un point de vue pratique, la version initiale du code s'exécutant sur ordinateur portable s'exécutait en 12h environ. La version parallèle permet d'analyser un exome de patient[142] en 5h30 environ par patient, ce qui permet d'analyser les données en moyenne 20 patients par jour. Pour diminuer encore ce temps, une future du version du pipeline pourrait ajouter la parallélisation par chromosome de l'appel de variant[143].

2.4 Qualité des données

Lors de l'exécution du pipeline, plusieurs outils permettent d'obtenir des métriques sur la qualité des différentes étapes. À l'aide de `mosdepth`, ces données sont collectées dans un rapport disponible à la fin de l'exécution. Les fichiers FastQ contiennent un score de qualité sur le séquencage par paire de base, le score Phred[144]. En analysant plusieurs patients, on peut ainsi vérifier la qualité moyenne du séquencage en fonction de la position dans le reads (Fig. 40.), la qualité par paire de base ainsi que le

[142]Hors temps d'interprétation biologique.

[143]Chaque chromosome est donc considéré comme indépendant pour l'appel de variants. Outre les contraintes que cela induit, la taille très différentes de certains chromosomes résultera probablement en une charge de calcul très inégalement répartie entre les différentes tâches.

[144]Qui est relié à la probabilité P d'identifier incorrectement une paire de base $Q = -10 \log_{10} P$.

contenu en GC (Fig. 42). Ici, le séquencage est de bonne qualité. Après alignement, on peut calculer la couverture pour différents patients avec `mosdepth`. Comme précisé dans les compte-rendus, celle-ci est de 96% à 20x (Fig. 41). Cette couverture est également homogène sur les autosomes. Après l'annotation des variants, nous pouvons voir, comme attendu, que la majorité des variants sont des SNVs, faux-sens, d'impact fonctionnel faible selon 2 scores bionformatiques (SIFT, PolyPhen) comme le montre la Fig. 43. Enfin, nous avons étudié la distribution des variation selon le type de séquences assemblées (chromosomes 1 à 22, X, Y et mitochrondrial), non localisées (c'est-à-dire sur un chromosome mais d'ordre ou d'orientation inconnu) et non placé. On note une distribution assez homogène avec cependant quantité importante de variants situés sur des séquences non localisées sur le chromosome 14 et une très faible proportion des contig alternatifs .

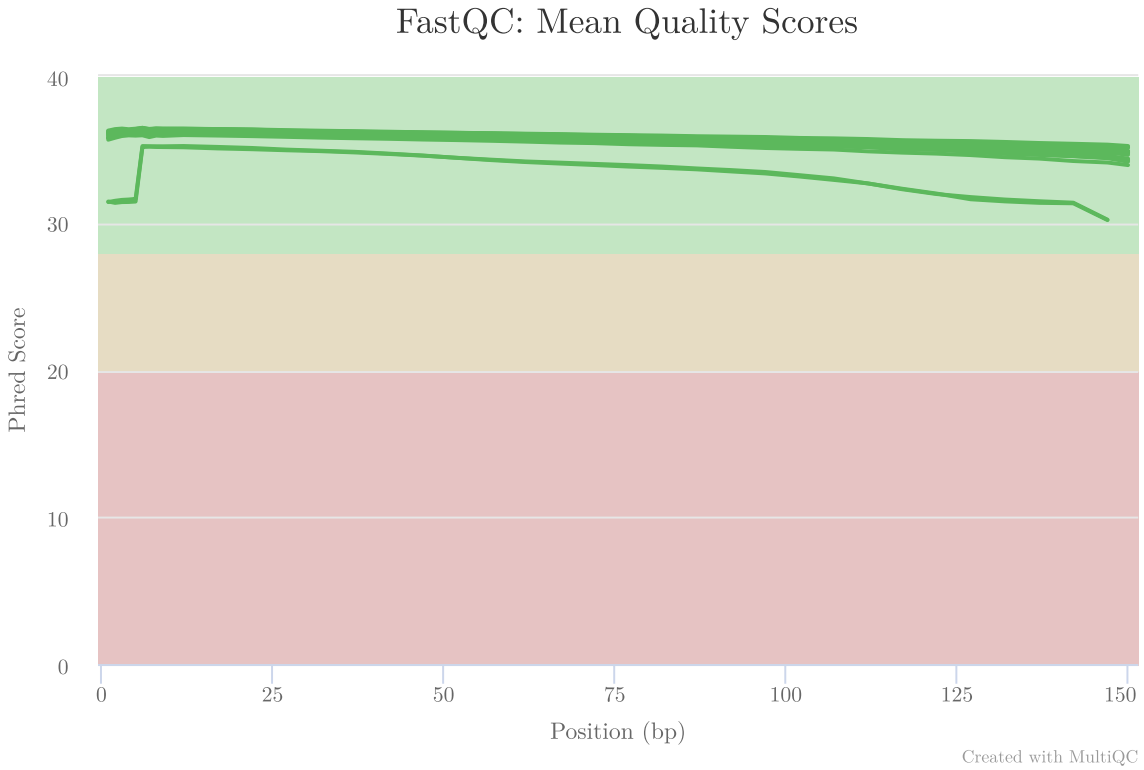


Fig. 40. – Qualité moyenne par position estimée par le score Phread. Un score à 30 signifie une précision de 99.9% et un score à 40 99.99%. Une légère décroissance en fin de read est retrouvée mais il s'agit d'un résultat attendu. Une ligne correspond à un patient. Programme utilisé : `fastqc`

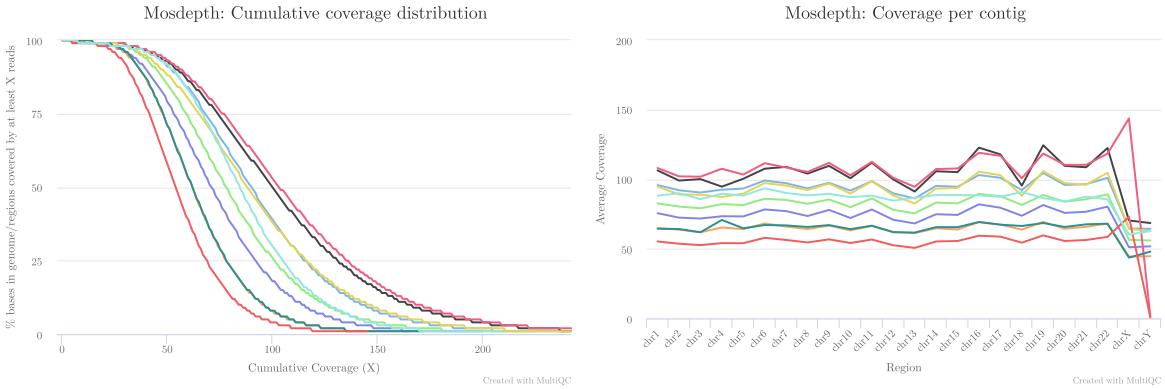


Fig. 41. – Couverture après alignement pour plusieurs patients (à gauche) et par chromosomes (à droite): hormis les gonomosomes, celle-ci est homogène.

)

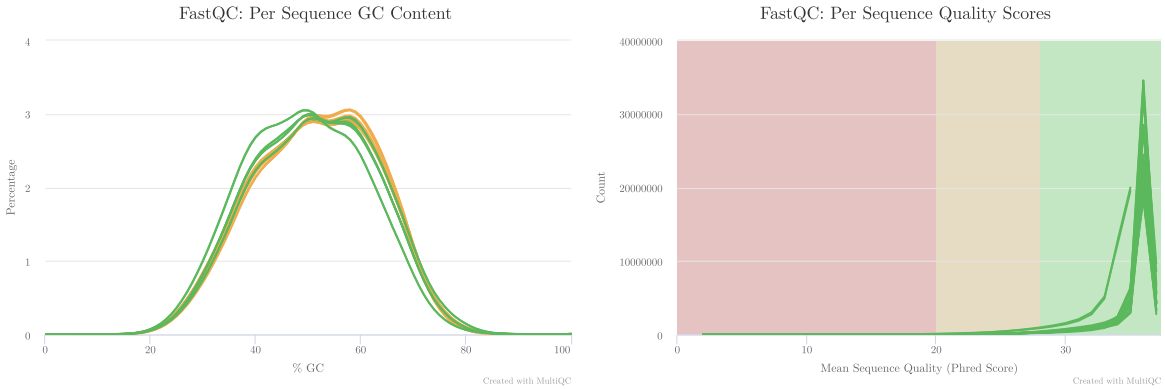


Fig. 42. – Qualité des données brutes estimée par le contenu en GC, qui doit être proche de 0.5 (gauche), et la qualité par paire de base (droite). La majorité est de bonne qualité (score Phred élevé).

Fig. 43. – Métriques fournies par `vep` montrant que la majorité des variants annotés sont des faux-sens (en haut à gauche), SNVs (en haut à droite) avec peu d'impact fonctionnel (scores PolyPhen et SIFT en bas à gauche et droite respectivement).

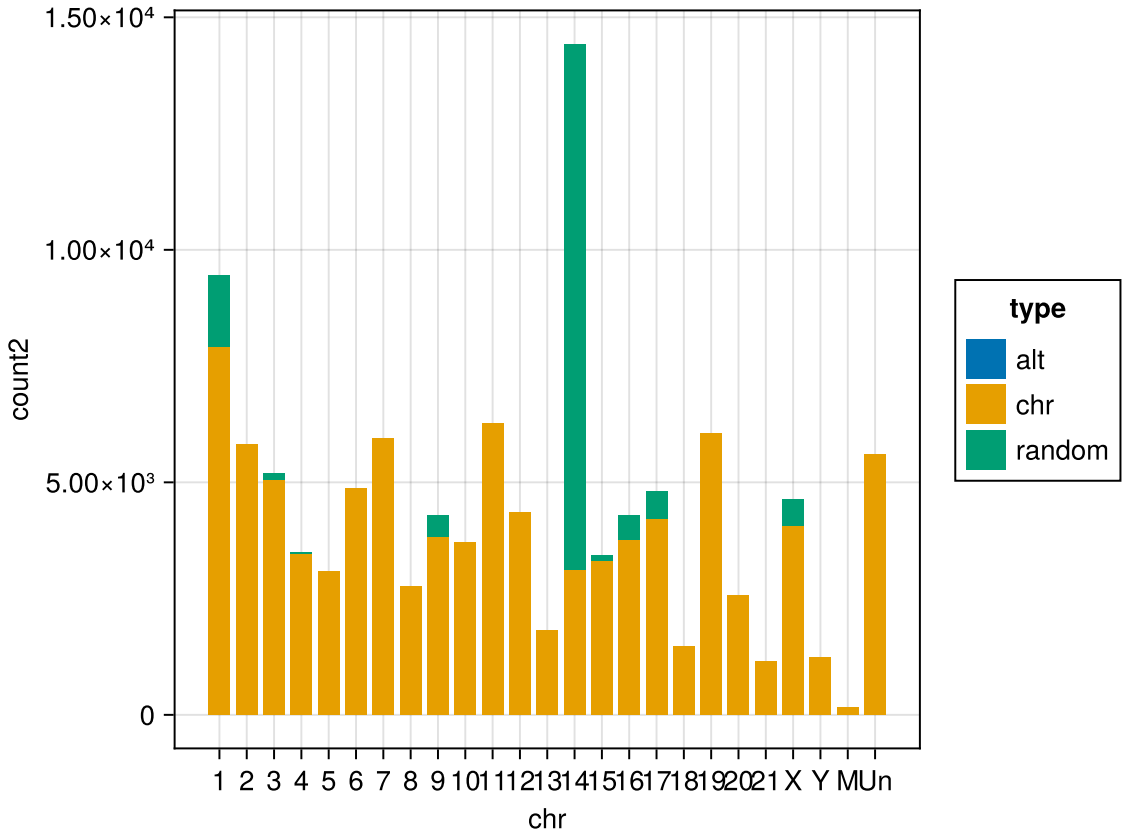


Fig. 44. – Distribution des variants selon le type de séquence : chromosomes 1-22, X,Y mitochondriale en orange, contig alternatifs en bleu et séquences sur un chromosome mais d'orientation non spécifiée en vert. Le chromosome *Un* correspond aux régions non placées.

2.5 Infrastructure

Comme décrit dans la section précédente, le choix d'une architecture informatique performante est un facteur essentiel. Les résultats présentés en termes de performances ont été effectués sur le mésocentre, centre de calcul de Franche-Comté avec la configuration suivante. Ses caractéristiques sont résumées sur le Tableau 5 et avec plus de détails techniques en annexes (Tableau 26). Si les chiffres sont plus impressionnants pour le mésocentre, il faut noter qu'il s'agit d'un environnement de *High Performance Computing* pour la recherche avec de nombreux utilisateurs. À l'inverse, un environnement de production est en cours de déploiement pour un nombre plus restreints d'utilisateurs en utilisant une VM dont les caractéristiques sont précisées

également dans la Tableau 5. Celui-ci présente l'avantage d'être interne à l'hôpital et donc permet une gestion des risques en collaboration avec la DSI.

		Noeuds	Coeurs	Mémoire(To)	Stockage (To)
Mesocentre	Maximum	28	1024	5.88	500
	Utilisé par Bisonex	1	24	32	2
Production	Maximum	NA	8	0.064	1

Tableau 5. – Configurations du mésocentre avec les ressources utilisées par le pipeline, ainsi que celles du futur environnement de production. L'architecture du mésocentre est partagée entre de nombreux utilisateurs. Le stockage contient les différentes bases de données

Outre les performances, le stockage et la circulation des données sont des facteurs importants dans le cadre de données patients, qui doivent respecter la législation en vigueur. Le processus final est illustré sur la Fig. 45 où les données brutes (FASTQ ou BAM) sont téléchargées depuis un serveur externe du sous-traitant, archivée au sein de l'hôpital sur une architecture gérée par la DSI avec duplication de données. En cas de demande de ré-analyse, les données sont importées sur le serveur de calcul et analysée par le pipeline. Les fichiers intermédiaires (VCF) et finaux (tableur au format .TSV, soit du texte brut) sont également archivés au sein de l'hôpital. Actuellement, le service d'hébergement fourni par l'hôpital n'est pas accrédité comme hébergeur de santé. Cependant, l'accès à ces données de génétique constitutionnelle est limité aux biologistes (4 personnes en dehors de la DSI).

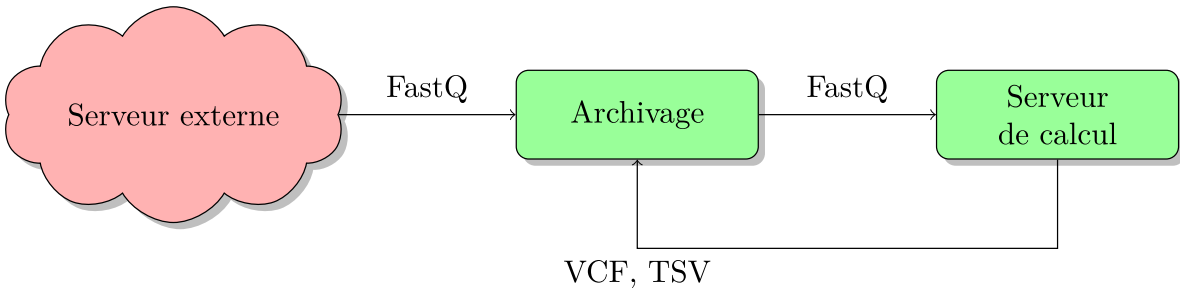


Fig. 45. – Circulation des données dans le futur processus en production. En vert, la partie interne à l'hôpital.

Chapitre 3

Validation

La validation analytique d'un pipeline analysant des données de génétique constitutionnelle est délicate. Les données fournies au pipeline sont bien évidemment dépendantes de la préparation de la librairie et du type de séquenceur. De plus, le nombre de nucléotides examinés par un exome est de l'ordre de plusieurs dizaines de millions. Avoir des échantillons avec toutes les combinaisons de SNVs, et à plus forte raison tous les combinaisons d'indels est donc imppossible Pour aider les laboratoires, des recommandations ont été émises par l'association des biologistes moléculaires américains et le collège des anatomo-pathologistes américains (*Association for Molecular Pathology* et *College of American Pathologists* respectivement) (Roy et al. 2018). La liste des critères est disponible en annexe, traduite par nos soins, mais peut être résumée comme suit.

La validation doit correspondre à l'usage clinique attendu, au type d'échantillon et de variant. Il faut tenir compte de la profondeur, balance allélique et des difficultés techniques du NGS. Ce sont les régions difficiles à séquencer, comme celles riches en GC ou avec des homopolymères comme illustré sur la Fig. 46, mais aussi des variants, variants complexes en *cis* ou en *trans* dont un exemple est fourni par la Fig. 47. Une attention toute particulière doit être apportée à la manière dont sont calculées la profondeur et couverture. En effet, la profondeur minimale sur des positions prédéfinies peut être plus utile pour la sensibilité que la profondeur moyenne. La couverture, par exemple 95% des bases ont 30x de profondeur, peut masquer certaines régions critiques pour le patient. Il est donc recommandé de mentionner les zones mal couvertes pertinentes dans le compte-rendu. Un nombre minimum de variants est estimé par les auteurs selon la formule $r^n = \alpha$ où r est la fiabilité, n le nombre d'échantillons et α la probabilité d'une erreur de type I (faux positif), soit :

$$n = \frac{\ln(\alpha)}{\ln(r)}$$

Ainsi une confiance de 95% ($\alpha = 0.05$) et une fiabilité de 95% ($r = 0.95$) conduisent à un nombre minimal de 59 SNVs et autant d'indels. Si la méthode de référence est d'utiliser des échantillons représentatifs, en obtenir suffisamment peut être une gageure. C'est pourquoi il existe, en complément seulement, la possibilité de créer des

variants *in silico*, c'est-à-dire purement en manipulant les données brutes (Duncavage et al. 2023).

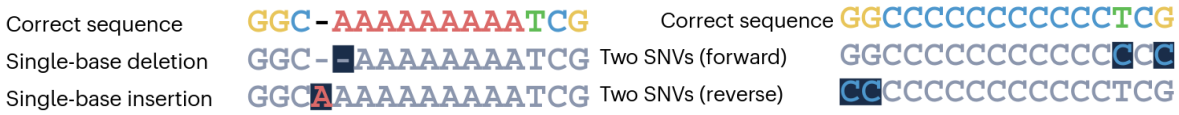


Fig. 46. – Exemple d'erreur de séquençage d'homopolymère : insertion et délétion systématique (IonTorrent, Nanopore, Pacbio) sur la figure de gauche. Sur Illumina, les erreurs ont tendance à être des SNV ou indel en fin de séquence (à droite sur le brin sens et à gauche en antisens) sur la figure de droite (Olson et al. 2023).

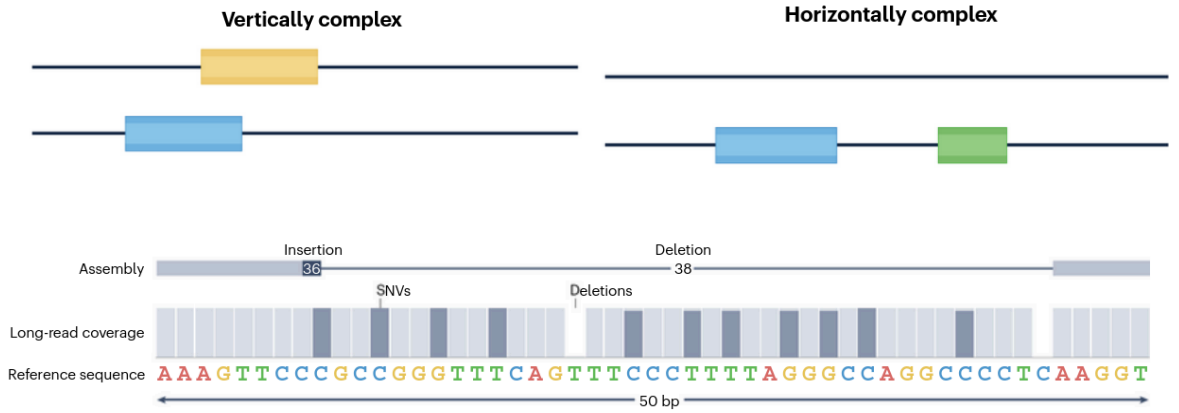


Fig. 47. – Exemple de variants complexes en *trans* et en *cis*. En bas, un exemple de variant représenté par une insertion de 36bp et une délétion de 38bp dans la référence, mais par un ensemble de SNV et de délétion de 1bp en Hi-Fi (Olson et al. 2023)

À ce titre, nous présentons dans cette partie les résultats des tests de validations menés sur des patients de références ainsi que sur des données *in silico* comme détaillé ci-dessous. Les autres points forts du pipeline selon les recommandations citées plus haut sont l'utilisation d'un gestionnaire de version permettant la traçabilité du code, l'utilisation de Nextflow nous permet d'être assez proche de l'environnement final de production (Chapitre 2.2). Grâce à Nix, la version des dépendances est immuable (Chapitre 2.1). Enfin, le pipeline a été supervisé par du personnel qualifié sur le plan biologique et bio-informatique.

Dans ce chapitre, nous rappelons d'abord la nécessité d'utiliser des outils adaptés pour comparer des résultats de pipeline puis détaillons l'utilisation de variants de référence au travers de l'utilisation du *gold standard* fourni par le consortium GIAB (*Genome In A Bottle*) (Chapitre 3.2). Le séquençage et le pipeline sont testés avec l'échantillon d'un patient de référence, puis nous examinerons plusieurs autres patients de référence sur la partie bio-informatique. Dans un second temps, des données *in silico* sont utilisées en modifiant les données d'un vrai patient ainsi qu'en générant

des données de séquençage afin de tester un ensemble de variants validés en Sanger (Chapitre 3.3).

3.1 Outils de comparaison

3.1.1 Définitions

Grâce au travail du consortium GIAB, nous disposons d'un ensemble de variants constitutionnels de références pour plusieurs patients (voir Chapitre 3.2). Outre ces variants, GIAB fournit également des intervalles sur lesquels seuls leurs variants doivent être retrouvés. Cela permet ainsi de définir des vrais positifs, faux positifs et faux négatifs sur ces intervalles. Cependant, il est difficile de définir des vrais négatifs, surtout autour de variants complexes limitant l'utilisation de la spécificité comme métrique d'évalutation. On utilise donc les métriques suivantes pour mesurer les performances d'un pipeline:

- la valeur prédictive positive (VPP) est la probabilité qu'un variant soit bien présent s'il est appelé. En notant FP les Faux Positifs (variants appelés à tort) et VP les Vrais Positifs (variants appelés à raison) :

$$\mathrm{VPP} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FP}}$$

- la sensibilité S est la probabilité qu'un variant soit appelé s'il est présent. En notant FN les Faux Négatifs (variants non-appelés à tort) :

$$S = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}}$$

- la VPP et la sensibilité ne s'interprètent pas séparement donc on peut également utiliser leur moyenne harmonique F_1

$$F_1 = \frac{2}{\frac{1}{\mathrm{VPP}} + \frac{1}{S}}$$

3.1.2 Le problème de la représentation des variants

Le format VCF (*Variant Call Format*) est devenu un standard international pour représenter les variants par sa souplesse et, à ce titre, est supporté par la majorité des outils de bio-informatique. La comparaison de pipeline peut donc se réduire à la comparaison de fichiers VCF. Toutefois, elle se heurte à un problème sur la représentation ambigüe de certains variants, comme illustré sur la Fig. 48. En particulier, une indel dans une zone répéte ou un homopolymère peut avoir plusieurs positions à cause d'artefact d'alignement ou de choix différents par rapport au 5′ et 3′ de réference. Les

variants complexes et les polymorphismes sur plusieurs nucléotides souffrent encore plus de ce problème.

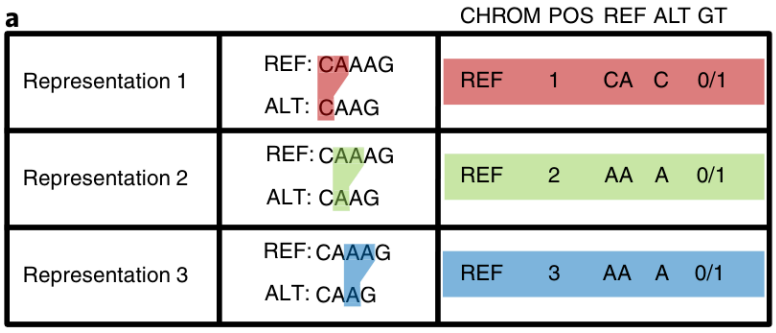


Fig. 48. – Une délétion peut être représentée de plusieurs manière au format VCF (Krusche et al. 2019).

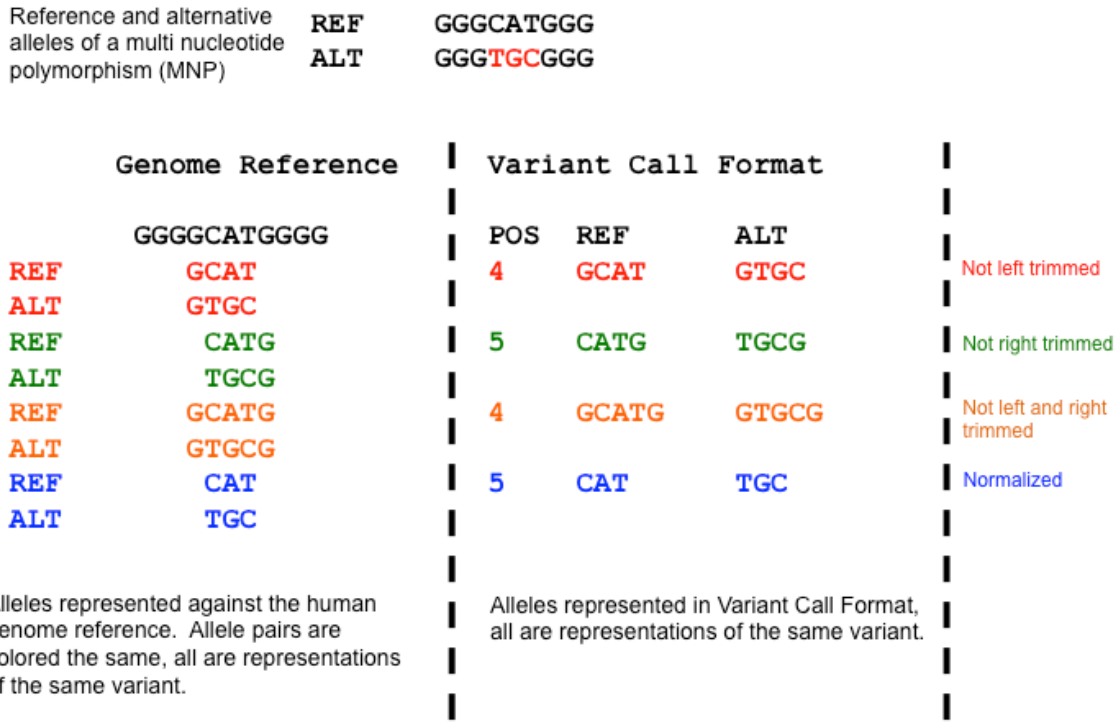


Fig. 49. – Illustration de la normalisation et l'alignement à gauche' variant (https://genome.sph.umich.edu/wiki/Variant_Normalization.)

Certains de ces problèmes peuvent être résolus en normalisant les variants pour qu'ils soient le plus à « gauche » (au sens d'un VCF) et de taille la plus petite possible en

avec une opération de normalisation et d'alignement à gauche comme précisé sur sur la Fig. 49. Cette approche s'est standardisée mais échoue pour les variants complexes, représentés sur plusieurs lignes dans un VCF (Fig. 50). Nous présentons ici les deux outils principaux permettant de corriger ce problème à notre connaissance.

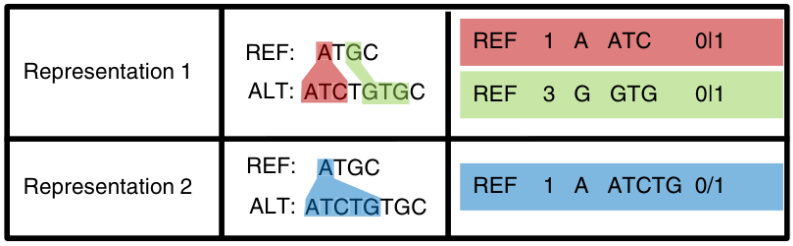


Fig. 50. – Une insertion peut être représentée par plusieurs lignes dans un VCF (Krusche et al. 2019).

Le principe de *vcfeval* est de chercher parmi toutes les combinaisons des variants donné par l'appel de variant et des variants de référence celle qui assure la meilleure correspondance. Cela permet donc de maximiser le nombre de vrais positifs et minimiser le nombre de faux positifs et faux négatifs. Pour une possibilité donné, les variants non présents dans la référence sont donc des faux négatifs, alors que ceux absence des variants à comparer sont catégorisés comme des faux positifs (voir Fig. 51). Différentes techniques algorithmiques sont utilisées pour éviter que le coût calculatoire n'explose[145]. En pratique, le temps d'exécution est très court avec moins de 30 secondes pour comparer 90 000 variants à 155 variants de référence.

Une approche semblable a été choisie par le groupe GA4GH (*Global Alliance for Genomics and Health*) dans le cadre de leurs recommandations pour la comparaison de variants constitutionnels avec l'introduction de l'outil *Hap.py* (Krusche et al. 2019). Une standardisation des métriques est également proposée en séparant une correspondance sur le génotype, où l'allèle et le génotype doivent correspondent pour un vrai positif, d'une correspondance partielle sur l'allèle, donc sans tenir compte du génotype et d'une correspondance locale où tout site avec un variant de référence dans un voisinage proche[146]. Un exemple est présenté sur le Tableau 6. À noter que les vrais négatifs ne sont pas rendus pour les raisons évoquées plus haut.

[145]Le nombre de chemins est théoriquement exponentiel mais les auteurs utilisent une programmation dynamique avec un parcours incrémental des haplotypes pour chacun des chemins et la décision d'inclure un variant est retardée le plus longtemps possible. De plus, les diverses représentations possibles sont pondérées par le nombre de variants de référence divisé par le nombre de variants appelés sur un intervalle entre 2 points de synchronisation.

[146]La correspondance sur le génotype est pour la clinique mais la discordance sur l'allèle ou le génotype permet de faciliter l'amélioration du pipeline

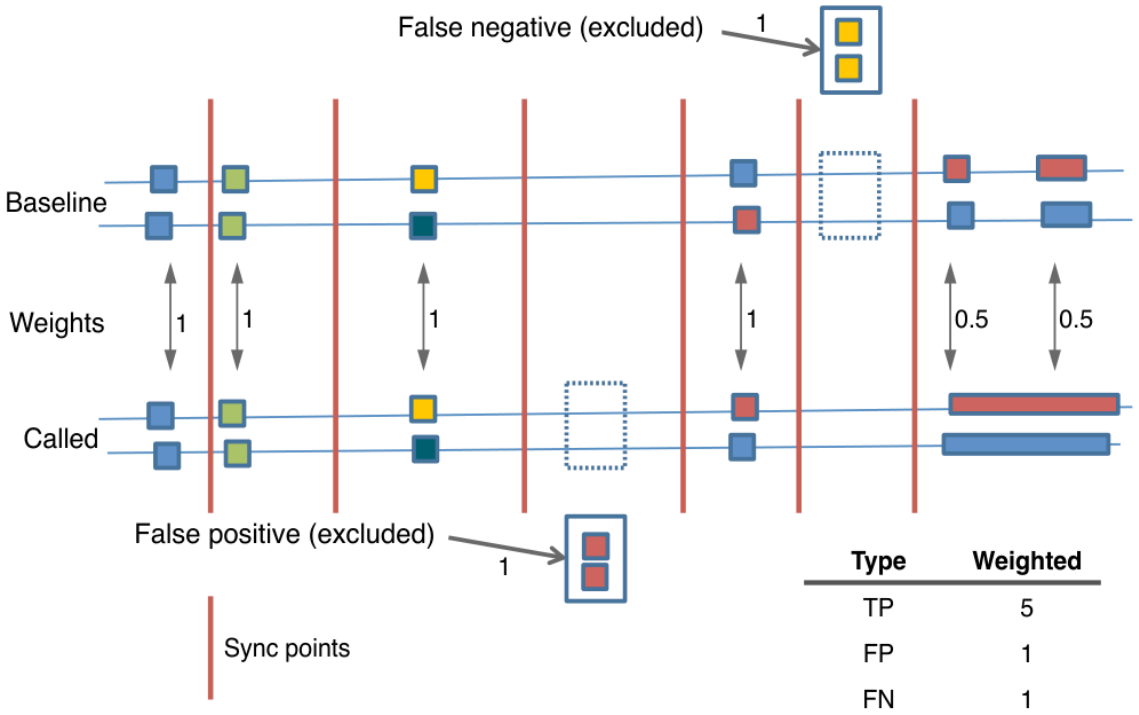


Fig. 51. – Algorithme de *vcfeval* où des variants (*called*) sont comparés à des variants de référence (*baseline*). Les lignes verticales sont les points de synchronisation (Cleary et al. 2015).

Référence	Gold standard	Variants	Compté comme
A/A	C/C	C/C	Vrai Positif
		A/A	Faux Négatif
		A/C	Faux Positif sur le génotype (*)
		G/G	Faux Positif sur l'allèle (*)
	A/A	C/C	Faux Positif

Tableau 6. – Illustration des métriques de Krusche et al. (2019) pour des variants où un génome de référence contient A à cette position. (*) sont également comptés comme faux positifs.

Hap.py propose donc un autre algorithme pour l'harmonisation de la représentation des variants, une classification plus précise de la comparaison ainsi qu'une stratification selon différentes régions. Celles-ci peuvent être marquées comme répétitions en tandem, homopolymère ou avec un contenu important (>85%) en GC. À titre d'exemple, Krusche et al. (2019) montrent qu'une amplification par PCR a une VPP et sensibilité diminuées par rapport à une approche « sans PCR » (*PCR-free*) pour les indel et cela est surtout dû aux homopolymères et répétition en tandem.

En terme d'exécution, *hap.py* est plus lent que *vcfeval*[147] et, selon les auteurs eux-mêmes, moins performant pour les zones compliquées, ce qui peut avoir un impact non négligeable en utilisant les données de GIAB[148]. Il est cependant possible d'utiliser *vcfeval* comme outil de comparaison avec *hap.py* pour la classification et stratification des résultats. C'est l'approche qui a été retenue ici, en accord avec la comparaison à grande échelle proposée par le défi *precisionFDA* pour la première (Olson et al. 2020) et seconde version (Olson et al. 2022).

3.2 Patients de réference

3.2.1 Données GIAB

Le consortium GIAB a analysé les génomes de 7 patients : une patiente « pilote » mormon (HG001/NA12878)[149], un trio père/mère/fils juif Ashkenaze et un trio père/mère/fils de l'ethnie Han (Chine). Les données ont été séquencées sur plusieurs technologies contenant du *short read* et *linked read*[150] (Zook et al. 2016).

La méthode pour établir leur *gold standard* est détaillée par Zook et al. (2019). Pour chaque séquençage, une première étape consiste à définir les variants et régions d'intérêt en excluant les régions difficiles et les variants structurels. Un variant est considéré comme « consensuel » s'il est retrouvé par au moins 2 technologies sans discordance. Ces variants sont ensuite utilisés pour entraîner un modèle marquant les variants « peu fiables ». Ces deux types de variants sont utilisés pour gérer les discordances entre les données de séquençage et le résultat final consiste en l'union des régions validées en enlevant les variants peu fiables. L'*analyse en trio* a été utilisée pour ajouter la phase de certains variants et étudier les variants ne correspondant pas à une transmission mendélienne. Parmi ceux-ci, plus de 54% ont été classifiés en possiblement *de novo*, avec un variant hétérozygote chez le fils mais dont les deux parents sont homozygotes. Parmi les autres variants, la majorité (91%) sont des indels avec des erreurs sur des régions répétés. Cette dernière catégorie a été exclue du résultat final (Supplementary Note 1). Les auteurs ont estimé à 16 SNVs et 150 indels pour 1 million de variants.

[147] Avec des temps d'exécution inférieurs à 30 minutes

[148] https://github.com/Illumina/hap.py

[149] On notera qu'Illumina propose des génomes pour la famille étendue de cette patiente avec pas moins de 17 apparentés dans le cadre du Platinum Genome (Eberle et al. 2017)

[150] HiSeq2500 en paired-end (300x pour cas index, 100x pour parents, 45x, 15x mate-pair), 100x complete genomics , IonPi exome 1000x, Solid5500W, 10xgenomics chromium

Les auteurs n'ont pas confirmé les variants par Sanger, mais seulement par une revue manuelle des résultats. Selon eux, une discordance avec les résultats d'un Sanger n'augmente probablement pas la fiabilité de variants retrouvés par plusieurs technologies et l'utilisation de *long read* (PacBio, Moleculo) a été utilisé pour valider les données de manière manuelle. Ils citent 23 variants mal classés qui étaient dans des homopolymères, en répétitions en tandom ou indel hétérozygote composite pour les quel un séquençage Sanger n'aurait pas aidé à classifier.

Cette étude est donc extrêmement utile pour la validation de pipeline car elle apporte plusieurs échantillons, les données brutes, les variants et zones de « haute confiance » avec des données mises à jour régulièrement. Une des limite de la publication initiale est de ne pas avoir pu établir de *gold standard* dans les régions difficiles à séquencer.

3.2.2 Résultats sur le patient pilote

Dans un premier temps, nous avons voulu tester l'intégralité du processus de séquençage et d'analyse bioinformatique. Pour cela, nous avons utilisé un échantillon du patient « pilote » HG001 (ou NA12878) en commandant 25µg d'ADN auprès de l'institut Coriell[151]. L'ADN a été ensuite envoyé à Centogène où il a été séquencé sans analyse bio-informatique suite à notre demande. Les données brutes au format FASTQ ont été ensuite analysée par notre pipeline. Le kit de capture utilisé pour les tests est Twist Exome Core, qui s'approche le plus du kit du sous-traitant[152]. Les résultats après l'appel de variants, sans filtre, ont été comparé à la dernière version du *gold standard* de GIAB (variants et intervalles de références) en utilisant *hap.py* avec *vcfeval* comme algorithme de comparaison.

Type	Sensibilité	VPP	Indéterminé (fraction)	F_1
Indel	0.954	0.775	0.428	0.855
SNV	0.983	0.965	0.160	0.976

Tableau 7. – Résultat du séquencage et du pipeline pour le patient de référence HG001/NA12878. La fraction de variants indéterminés correspond au ration entre les variants en dehors des intervalles de références divisés par le nombre de variants.

On trouve de bons résultats pour les SNVs avec, comme on pouvait s'y attendre, des performances inférieures pour les indels(Tableau 7), due à un nombre relativement important de faux positifs (Tableau 8). Ceux-ci ne semblent pas liés à une discordance

[151] https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=NA12878&Product=DNA

[152] Initialement, le kit utilisé était le Twist Human Exome mais nous n'avons pas pu obtenir le nouveau kit de capture.

sur le génotype ou à l'allèle (voir Tableau 6 pour rappel) donc sont probablement des « vrais » faux positifs. Le rapport transition/transversion[153] sont dans les bornes attendues pour l'exome : 0.30 pour les données GIAB et 2.86 pour Bisonex (SNVs) avec une médiane à 2.81 pour les Européens (Wang et al. 2015). Le rapport hétérozygote/homozygotes est également correct pour les SNVs à 1.85 pour une médiane à 1.6 pour les Européens mais trop élevé pour les indels (2.90). Les statistiques données par *hap.py* montrent que la majorités des indels sont de petite taille, sans indel complexe, et que les petites délétions contribuent le plus aux faux positifs de manière hétérozygote comme détaillé sur le Tableau 9.

	référence			variants				
Type	Total	VP	FN	Total	FP	Indéterminé	FP génotype	FP allèle
Indel	328	313	15	722	93	309	10	9
SNV	19198	18966	232	23381	683	3738	50	73

Tableau 8. – Les variants marqués comme indéterminés sont en dehors des intervalles de références. VP: vrais positifs, FN: faux négatifs, FP: faux positifs.

Type	Taille	Variants de référénce	Variants	FP	FP.het	FP.homalt	Indéterminé
complexe	Tous	0	0	0	0	0	0
deletion	≥ 16	13	54	0	0	0	41
	1-5	125	285	56	55	1	108
	6-15	36	63	8	6	2	21
insertion	≥ 16	9	43	1	1	0	34
	1-5	116	223	27	27	0	84
	6-15	33	55	1	1	0	22

Tableau 9. – Stratification des faux positifs en fonction des indels. La proportion la plus importante est surlignée.

Il serait extrêmement utile de pouvoir comparer les performances de notre pipeline d'analyse d'exome à des pipelines d'autres laboratoires. Il existe bien un projet, le « precisionFDA Truth Challenge », comparant en aveugle les pipelines de différents laboratoires sur le patient HG002. Malheureusement, les résultats de ce défi en 2016, accessibles en ligne[154], ne concernent que des données de génomes.

[153]Les transitions sont définies par le changement $A \leftrightarrow G$ ou $C \leftarrow T >$ alors qe les transversions correspondent à $A \leftrightarrow \{C, T\}$, $G \leftrightarrow \{C, T\}$

[154]https://precision.fda.gov/challenges/truth/results

3.2.3 Résultats sur les autres patients

Afin de comparer la partie bio-informatique seules, nous avons cherché quelles données brutes (FAST, BAM) étaient disponibles pour ces patients de GIAB. Sur le dépôt Github du consortium, il existe des données pour chaque patient[155], mais celles d'exomes sont restreintes et surtout n'ont pas la même configuration (séquenceur, kit de capture). Les données sont également quelque peu datées et certains de ces kits sont encore en GRCh37, nécessitant une étape de conversion (*liftover*), qui peut induire des erreurs. À notre connaissance, la meilleure source de données brutes pour les patients est fournie par Google dans un *preprint* (Baid et al. 2020). Ils fournissent pour les 7 patients[156] les données de séquençage de génome et d'exome. Les exomes ont été réalisés avec 4 kits (Agilent v7, IDT, Nextera, et Truseq), sur 2 séquenceurs Illumina (NovaSeq et HiSeqX). Les données sont disponibles avec plusieurs profondeurs (100x, 75x, and 50x)[157]. Au total, il y a donc 168 possibilités. Les auteurs ont mis à disposition le résultat après appel de variants par 3 outils pour chaque séquencage.

En pratique, nous avons utilisé les données disponible avec une couverture de 50x, soit 42 exécutions au total. Le résultat de Bisonex a donc été comparé sur les mêmes données avec les mêmes aligneurs et 3 outils d'appel de variants dont 1 identique à celui de Bisonex. Les résultats après appel de variants sont représentés pour les SNVs et indels séparément sur les figures suivantes en étudiant l'impact de 3 facteurs : le choix du kit de capture, du patient et du séquenceur Illumina. Sur les données de Baid et al. (2020), les différents kit de capture représentent des zones distinctes pour les indels et SNVs avec une meilleure performance du kit IDT-xgEn (Fig. 52). Nous n'observons pas de différence entre les kits ni entre les patients sur notre pipeline (Fig. 53). Enfin, il existe une discordance sur les performances du séquenceur NovaSeq le plus récent, entre nos résultats et les données des auteurs (Fig. 54). Enfin, il est suprenant de constater les meilleures performances sur les SNVs et indels de notre pipeline, même quand les outils d'alignements et d'appels de variants sont identiques (HaplotypeCaller). La seule différence semble être l'utilisation d'un génome de réference avec les *contig* alternatives pour Bisonex, pouvant améliorer les performances. Cette constatation souligne l'importance de la reproductibilité de la validation des différents pipelines.

[155] https://github.com/genome-in-a-bottle/giab_data_indexes

[156] Ainsi que les 2 parents du patient HG001, soit 9 patients au total mais non étudiés ici.

[157] Obtenues par sous-échantillonnage.

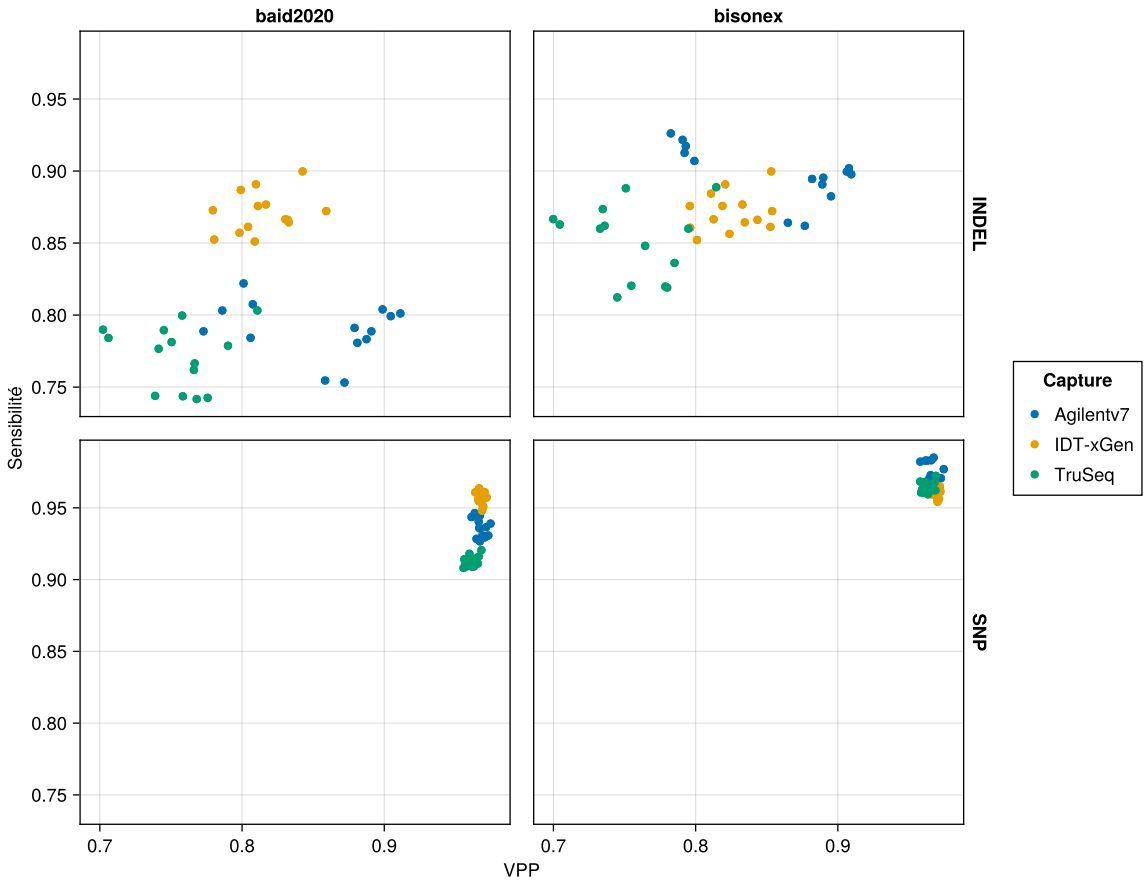


Fig. 52. – Impact du kit de capture sur les patients GIAB avec les données de séquencage de Baid et al. (2020).

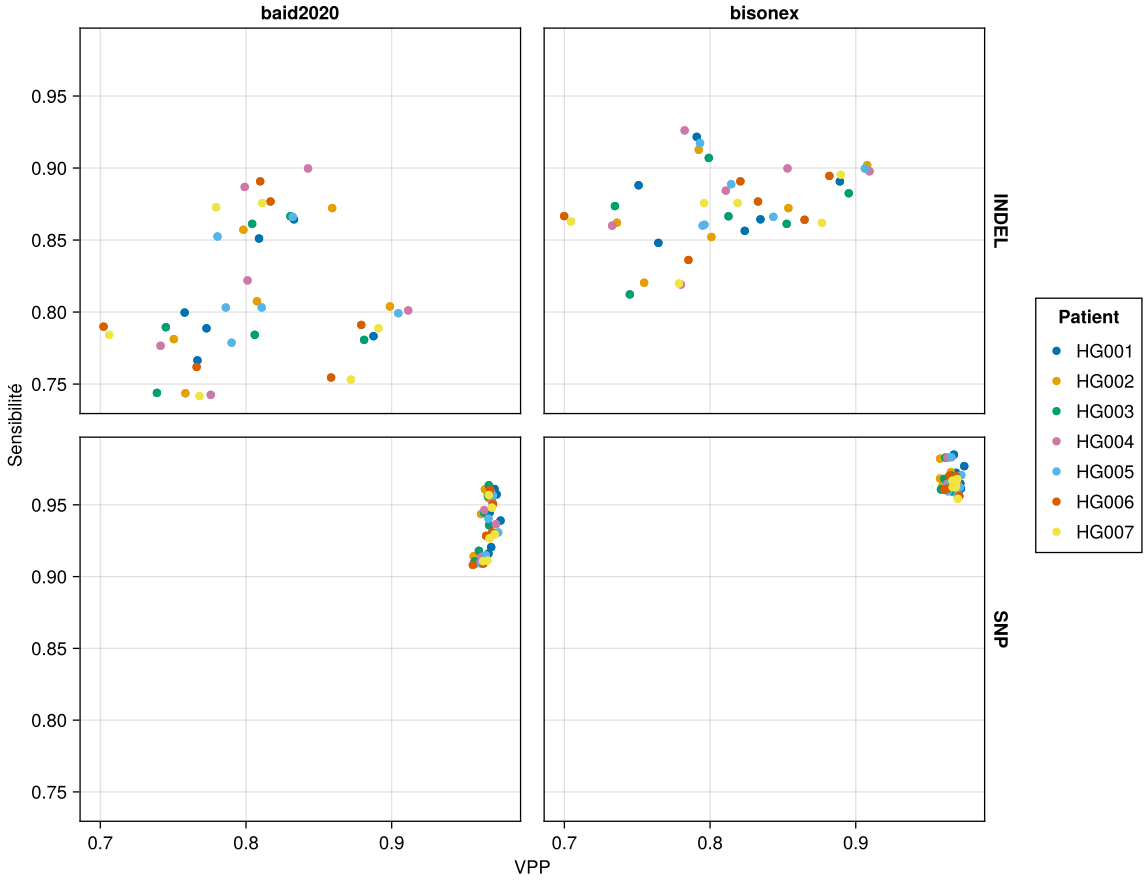


Fig. 53. – Impact du choix du patient GIAB avec les données de séquencage de Baid et al. (2020).

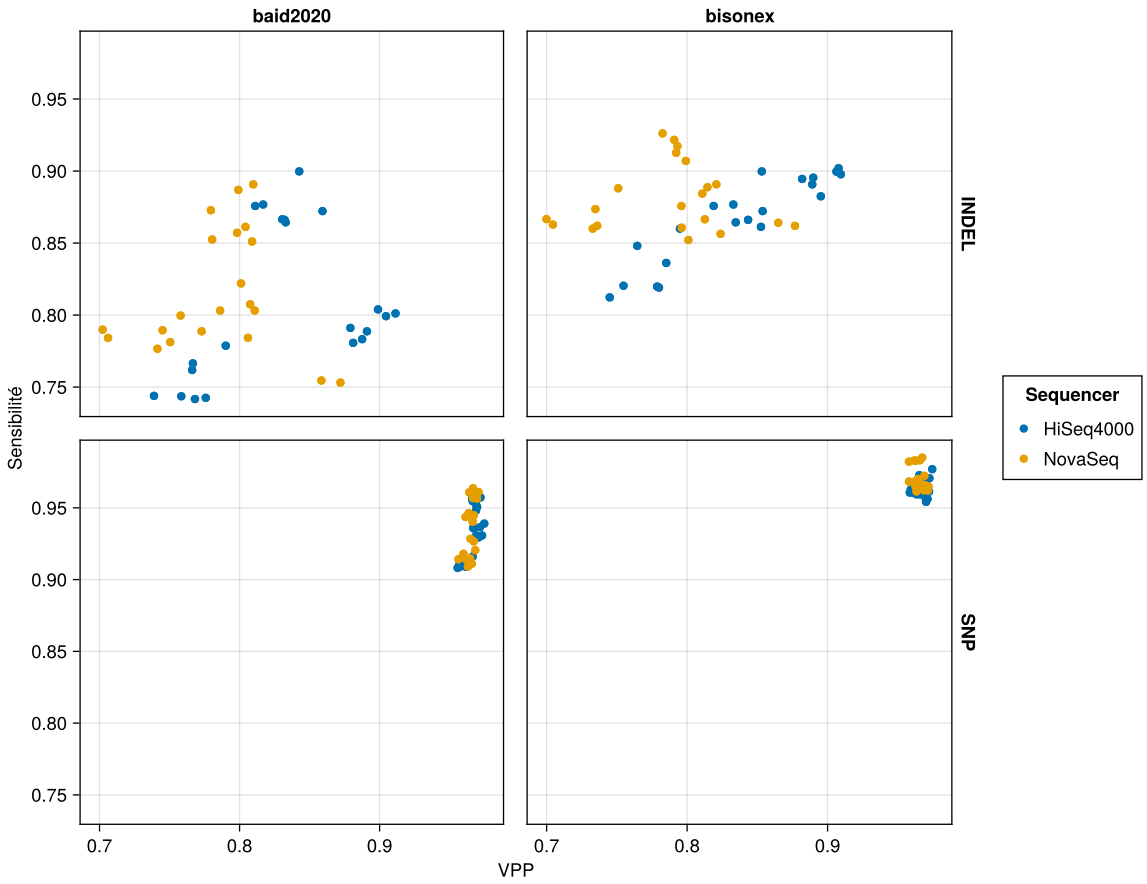


Fig. 54. – Impact du sequenceur sur les patients GIAB avec les données de séquencage de Baid et al. (2020).

3.3 *In silico*

En complément d'une validation sur des échantillons avec des variants connus et représentatifs, il est possible d'utiliser des méthodes purement bio-informatiques pour lesquelles des recommandations ont été récemment émises par l'*Association for Molecular Pathology*, l'*Association for Pathology Informatics* et le *College of American Pathologists* (Duncavage et al. 2023). Les données complètement *simulées* sont générées informatiquement à partir de paramètre prédéfinis comme la longeur des reads et avec un profil d'erreur donné (Fig. 55, à gauche). Cette approche a l'avantage de permettre de simuler facilement de nombreux variants de types différents. Elle est particulièrement utile pour des variants complexes ou pour lesquels des échantillons sont difficiles à se procurer. En revanche, les modèles de biais ou de bruits ne peuvent égaler les données réelles, ce qui peut conduire à une surestimation des performances.

Une approche est alors de partir de données de patients réels et d'introduire des variants à une fréquence allélique connue. Comme précédemment, elle permet de tester de nombreux variants mais conserve les biais de séquençage (Fig. 55, à droite). Une troisième technique est de modifier directement le génome de réference mais cela teste surtout les variants homozygotes. Elle est donc plus adaptée pour les chromosomes X et Y en dehors des régions pseudo-autosomiques. Enfin, on peut citer le sous-échantillonnage d'un FASTQ existant pour simuler des variants avec une faible profondeur, et le mélange de plusieurs échantillons différents pour estimer la limite de détection sur la fréquence allélique (Fig. 56). Selon un sondage réalisé dans cette même étude, seulement 36% des laboratoires médicaux utilisent des donnes *in silico* avec près de 60% pour la modification de fichiers BAM et plus de 40% pour le mélange de plusieurs fichiers.

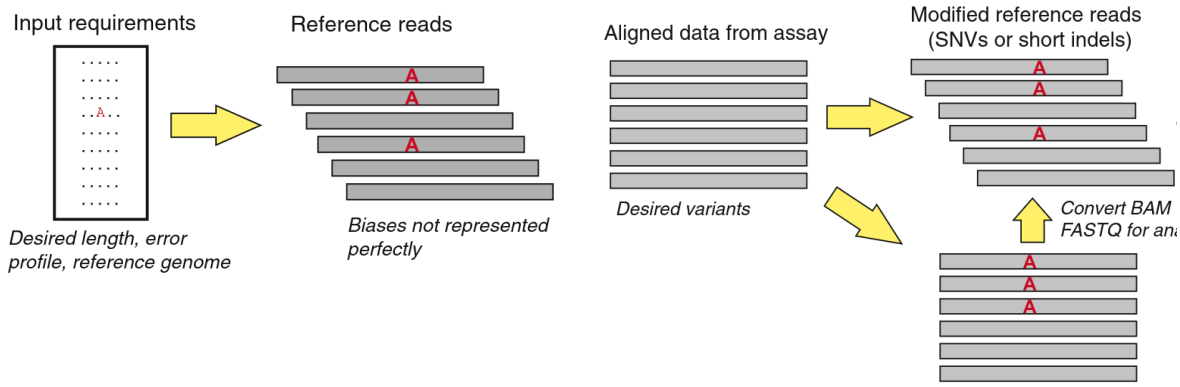


Fig. 55. – Illustration de la génération de *reads* (gauche) et de la modification de données existantes (droite) (Duncavage et al. 2023).

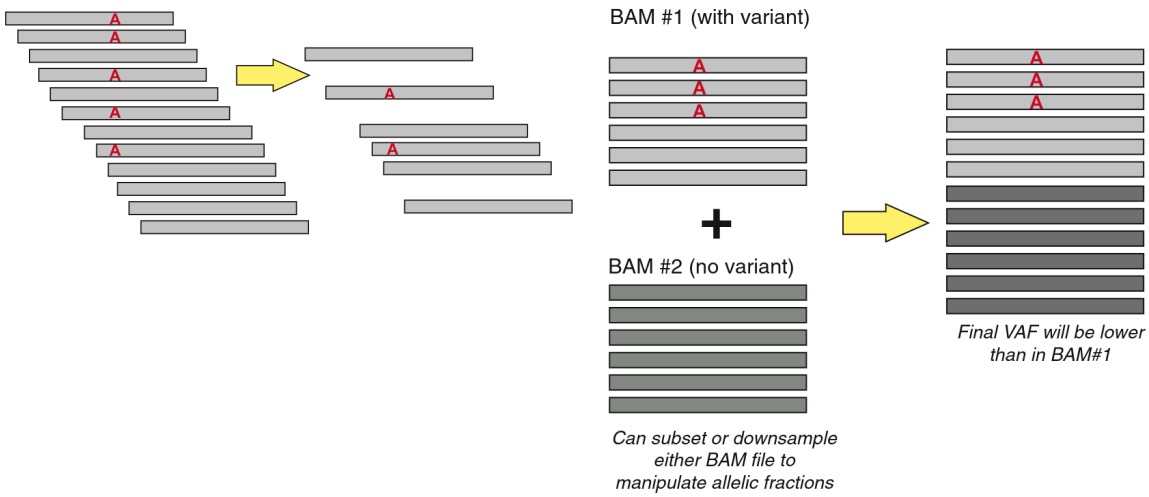


Fig. 56. – Sous-échantillonage d'un FASTQ (gauche) et mélange de BAM différent (droite) selon Duncavage et al. (2023).

3.3.1 Patient de synthèse

Un premier test fut de prendre les données brutes d'un patient, en l'occurrence HG001, pour lequel nous disposions d'un fichier BAM récent. Il existe deux outils pour insérer des variants : *bamsurgeon* (Ewing et al. 2015) et *varben* (Li et al. 2021) . Ces deux outils ont été validé sur des données somatiques et publié dans la littérature scientifique Leur approche est similaire : à une position données, des *reads* alignés du fichier BAM sont sélectionnés et leur séquence est éditée pour refléter l'occurrence du variant. Lors de nos tests, *bamsurgeon* a eu des difficultés à traiter nos fichiers BAM, nous conduisant à passer à *varben* dont l'algorithme pour l'insertion de SNVs est illustré sur la Fig. 57.

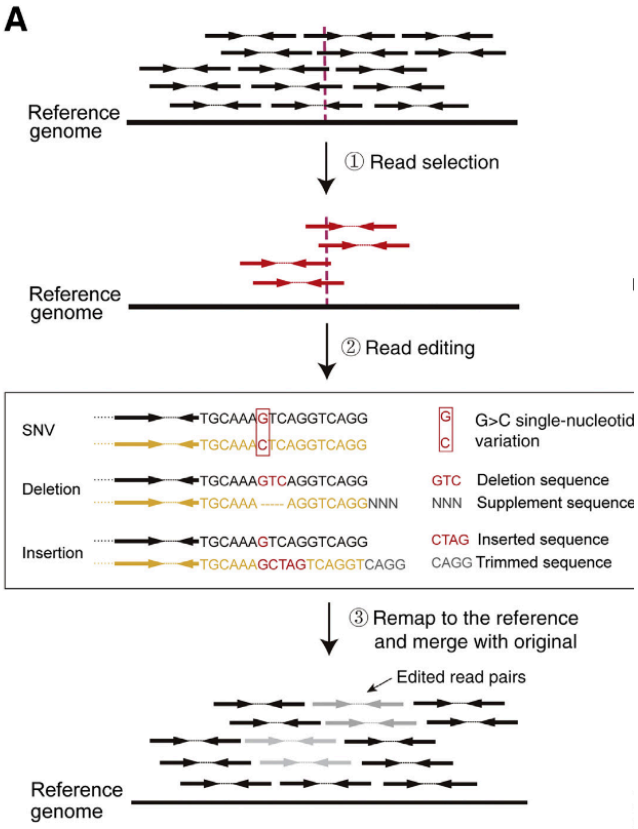


Fig. 57. – Algorithme de VarBen pour simuler des SNVs dans un fichier BAM existant (Li et al. 2021).

Pour valider le pipeline, nous avons sélectionné 163 variants retrouvés en exome puis confirmés en Sanger, dont les caractéristiques sont résumées sur le Tableau 10. On retrouve sans surprise une majorité de SNV et de VSI, hétérozygotes. À noter qu'aucune insertion n'a été identifiée. Pour comparer les VCFs émis par le pipeline à ces variants tests, nous avons utilisé `vcfeval` car la stratification proposée par `hap.py` était peu utile. Au final, Les résultats de nos tests montrent qu'avec notre outil, tous

les SNVs sont bien retrouvés après l'appel de variant. Pour utiliser Varben, nous avons défini une balance allélique aléatoire autour de la balance allélique cible, soit [0.4,0.6] pour les variants hétérozygotes et [0.8,1] pour les homozygotes. Une fois des variants inséres et analysés par Bisonex, nous avons trouvé qu'en sortie finale du pipeline :

- un seul variant n'a pas pu être inséré pour cause d'un nombre de reads insuffisant à cette position,
- un variant a été filtré, car sa profondeur de 21 était inférieure au seuil de 30,
- tous les autres variants ont été retrouvés, mais 13 homozygotes ont été catégorisés par erreur comme hétérozygotes. Pour ce dernier point, nous n'avons pas d'explication sur le comportement d'HaplotypeCaller.

Type	Nombre
hemizygote	17
heterozygote	131
homozygote	15

Classif. ACMG	Nombre
5	15
4	12
3	133
NA	3

Type	Nombre
SNV	152
del	7
dup	4

Tableau 10. – Caractéristiques des 126 variants confirmés en Sanger utilisés pour la validation

3.3.2 Données de synthèse

Pour la génération de *reads*, il existe une pléthore d'outils, dont un aperçu est donnée par le Tableau 11. Cependant, de nombreux logiciels ne permettent que des simulations de génomes. Parmi celles simulant des données d'exomes, il n'est pas toujours possible d'insérer des variants connus, comme *capsim*, ou avec des temps d'exécutions très longs, de l'ordre de plusieurs heures. Lors de notre revue, nous avons testé *NEAT*, qui s'est révélé être très lent et *ReSeq*, pour lequel la génération d'exome a été difficile. Simuscop (Yu et al. 2020) est en revanche un bon candidat pour les données Illumina, car il est rapide, avec une exécution de l'ordre de 15 minutes pour 100x, permet de générer des données d'exomes et peut insérer une liste de SNVs, indels ou variants structuraux. Le principe est dans un premier temps de générer un profil d'erreur à partir d'un VCF et d'un BAM pour apprendre les biais du séquenceur. Puis les données sont générées depuis ce profil et une liste de variants. Les variants sont marqués comme hétéro- ou homozygotes mais l'utilisateur n'a pas le contrôle sur la balance allélique.

Simulateur	Pair-End	sortie	Langage	SNV	CNV	Indel	Biais GC	Posi-tion	Contexte	Séqencage
ART	SE, PE	FQ, SAM	C++, Perl					X		G
Grinder	SE, PE	FQ, FA	Perl					X		G
pIRS	PE	FQ	C++, Perl	X	X		X	X		G
GemSIM	SE, PE	FQ, SAM	Python	X				X	X	G
Wessim	SE, PE	FQ, SAM	Python				X	X	X	E
NeSSM	SE, PE	FQ	C, Perl				X	X		G
BEAR	SE, PE	FQ	Perl, Python					X		G
FASTQSim	SE	FQ	Python					X		G
SInC	PE	FQ	C	X	X	X		X		G
SCNVSim	SE, PE	FQ	Java	X	X	X		X		G
NEAT	SE, PE	FQ	Python	X	X		X	X		G, E
IntSIM	SE, PE	FQ	C++, Perl, R	X	X	X	X	X		G
Pysim-sv	SE, PE	FQ	Python	X	X	X	X	X		G
InSilicoSeq	PE	FQ	Python				X	X		G
SimuSCoP	SE, PE	FQ	C++	X	X	X	X	X	X	G, E

Tableau 11. – Revue non exhaustive des simulateurs de reads par Yu et al. (2020). « G » dénote les génomes et « E » les exomes

Comme précédemment, les variants confirmés en Sanger (Tableau 10) ont été fournis à Simuscop et les VCFs des différentes étapes ont été comparés à la liste initiale de variants avec *vcfeval*. En sortie du pipeline, nous avons retrouvé sur les 162 variants :

- 4 faux négatifs confirmés qui sont tous des SNVs. 1 de ces variants n'était inséré que sur 3 reads sur 34 et n'a donc pas été appelé. 1 des variants aurait du être appelé par HaplotypeCaller. Enfin, 2 variants étaient en dehors des zones du kit de capture utilisé pour les tests, probablement lié à un kit mis à jour par le sous-traitant mais non communiqué,
- 7 variants ont été étiquetés comme hétérozygotes au lieu d'homozygotes. Tous, sauf 1, étaient portés par 100% des reads. Là, encore, nous n'expliquons pas ce résultat,

Pour finir, des tests préliminaires ont été effectués sur les variants Clinvar. Nous avons sélections tous les variants pathogènes ou probablement pathogènes contenus dans le kit de capture et distant de 10bp. Les indels de taille ≥ 10 ont été exclus, résultant en 132 069 variants. On notera que *vcfeval* n'a pu analyser de nombreuses petites régions car trop complexes. Pour valider les résultats, nous avons utilisé `bedtools intersect`, qui ne résout pas les différences de représentations entre les variants dans le format VCF. Avec ces deux outils et en sortie de pipeline, nous retrouvons 67% des variants dans ces premiers résultats.

3.4 Améliorations possibles

Les axes d'améliorations concernent la documentation, qui consiste actuellement en un fichier résumant l'installation et l'utilisation du pipeline ainsi que ce manuscrit. L'intégrité des données n'est pas actuellement vérifiée correctement : seul une vérification sur la taille est effectuée lors du téléchargement et une corruption lors de l'exécution du pipeline ne sera pas détectée. Cependant, l'archivage est géré par la DSI, qui garantit à ce stade l'intégrité des données. Les différents composants n'ont pas encore été testé individuellement, car seul le pipeline dans son ensemble a été examiné. Un autre point à améliorer consiste en l'identification des données de manière unique en tenant compte de patient, échantillon (aliquot) et exécution. Enfin, certaines métriques de qualité sont disponibles (Chapitre 2.4), mais certaines sont manquantes selon celles données en annexe (Tableau 16).

Concernant la validation, on peut citer des améliorations récentes au sein du consortium GIAB avec l'apport du *long-read* pour des gènes difficiles à séquencer[158] (Wagner et al. 2022), l'ajout de grand indels supérieurs à 50bp (Zook et al. 2020). Un partenariat a également été entrepris avec le consortium T2T conduisant à l'établissement d'un benchmark préliminaire utilisant les données du patient HG002 à l'aide du génome CHM13-T2T proposant les répétitions en tandem pour indel et variants structurels de taille supérieure à 50bp, ainsi que les variants de petite taille sur les chromosomes X et Y.

[158]SNVs de plus de 30kbp SNV, indel de 50kbp.

Chapitre 4

Ré-interprétation

Malgré les avancées techniques et scientifiques, le taux de diagnostics de l'exome est hétérogène selon la clinique, le séquençage en solo ou trio. Des méta-analyses l'estime entre 42 (Clark et al. 2018) et 52% (Tan et al. 2017), par exemple. Il existe donc un besoin de réanalyser les données pour les patients sans diagnostic, en exploitant la découverte de nouveaux gènes pathologiques, de nouveaux phénotypes ou de nouveaux outils bio-informatiques. Cependant, l'analyse de variants est un processus coûteux en termes de main-d'œuvre et de temps nécessaire pour cadrer et planifier ces réanalyses. Des recommandations récentes basées sur une revue de la littérature ont été émises (Dai et al. 2022). Le rendement diagnostic a ainsi été estimé à 10% avec une hétérogénéité importante. Les raisons pour une ré-analyse étaient principalement liées à des mises à jour de la littérature ou de bases de données cliniques. En effet, plus de 120 000 soumissions sont faites à Clinvar chaque année et 250 nouveaux gènes sont décrits dans des articles scientifiques. À ce titre, les auteurs suggèrent que les réanalyses soient espacées d'au moins 2 ans. Ils notent également qu'une automatisation partielle grâce à l'intelligence artificielle n'a pas diminué statistiquement le rendement diagnostic (8%).

Dans cette partie, nous décrivons tout d'abord le recrutement des patients ayant bénéficié d'un exome, puis discutons de la place du génome CHM13-T2T dans le pipeline. Après avoir comparé les résultats de notre pipeline sur 254 exomes à celui de Centogène, nous présentons quelques résultats préliminaires d'une ré-interprétation.

4.1 Recrutement

Pour mieux cerner le type de patients pour lequel un exome a été réalisé, nous avons procédé à une analyse automatique des compte-rendus au format PDF. 1 226 compte-rendus étaient disponibles soit 1 056 individus. Il s'agit d'une population majoritairement pédiatrique (Fig. 58). Lors des demandes d'exomes, le clinicien précise le phénotype du patient à l'aide de termes selon la nomenclature HPO (*Human Phenotype Ontology*), ce qui permet une standardisation des phénotypes. Un champ libre est également fourni pour préciser. L'extraction des termes HPOs des compte-rendus montre que les troubles du neurodéveloppement représentent une part importante de la clinique, comme détaillé sur le Tableau 12. Une illustration est proposée en Annexe (Fig. 73).

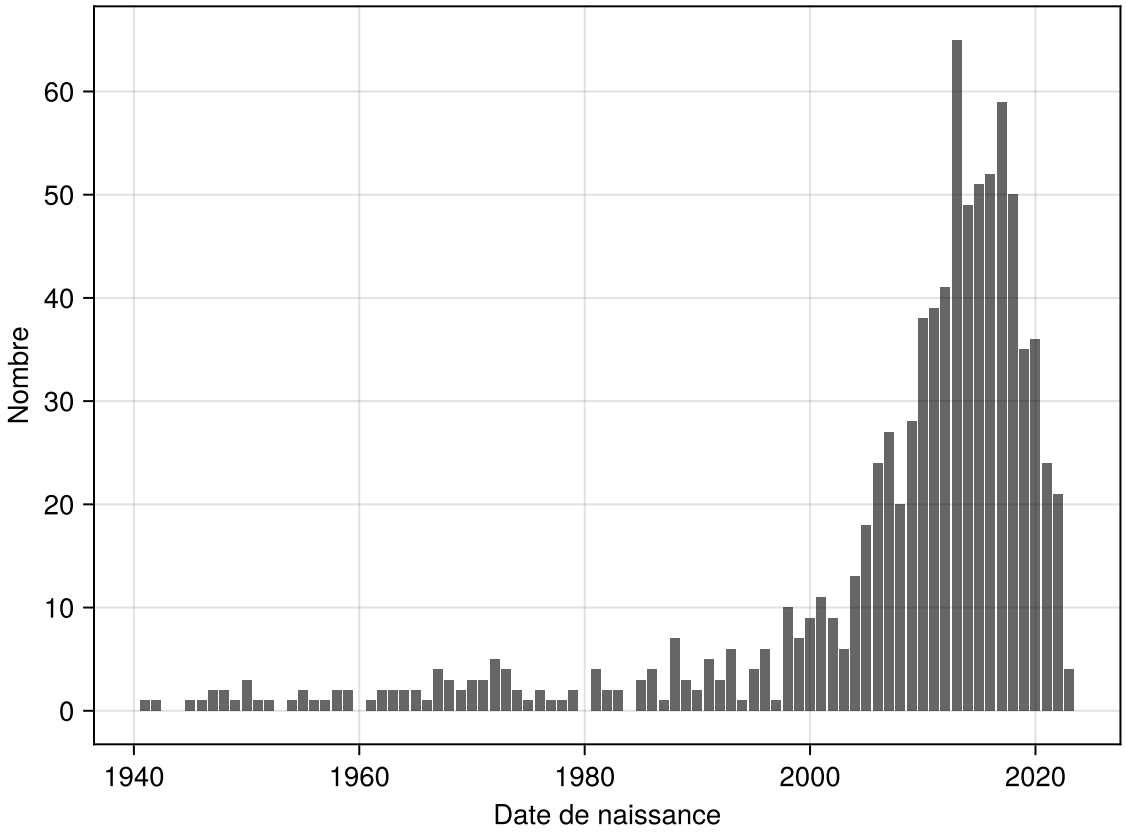


Fig. 58. – Âge des patients ayant bénéficié d'un exome à Besançon par extraction de la date de naissance des compte-rendus.

Terme HPO	Nombre
Macrocephaly	51
Neurodevelopmental delay	52
EEG abnormality	54
Intrauterine growth retardation	58
Hypotonia	67
Autistic behavior	68
Abnormal facial shape	69
Feeding difficulties	71
Short stature	73
Growth delay	75
Seizure	78
Specific learning disability	79
Strabismus	84
Autism	105
Microcephaly	111
Cognitive impairment	120
Motor delay	167
Intellectual disability	261
Delayed speech and language development	389
Global developmental delay	405

Catégorie	Nombre
Growth delay	60
Abnormal eye morphology	76
Abnormality of the vasculature	78
Abnormality of the genital system	81
Abnormality of the urinary system	82
Abnormality of digestive system physiology	109
Abnormality of the skin	115
Abnormality of body height	118
Abnormal cardiovascular system morphology	140
Abnormality of body weight	145
Abnormal eye physiology	243
Abnormality of the musculature	294
Abnormality of the skeletal system	300
Abnormal nervous system morphology	312
Abnormality of the head	679
Abnormal nervous system physiology	2554

Tableau 12. – Phénotypes les plus rapportés (à gauche) avec les grandes catégories les plus citées (à droite) pour les occurrences supérieures à 50. La clinique a été extraite des compte-rendus, et vérifiée qu'elle correspondait bien à la nomenclature HPO actuelle.

Sur le plan moléculaire, 600 exomes ont été rendus négatifs. Parmi les résultats positifs, 562 variants uniques ont été identifiés, dont la majorité sont des SNVs de signification indéterminée hétérozygotes (voir le Tableau 13). Il s'agit donc d'un profil classique pour des résultats d'exomes.

Classif. ACMG	Nombre
Pathogène	77
Probablement patho.	100
Indéterminé	370
N/A	15

Type	Nombre
SNV	478
Délétion	57
Duplication	19
Insertion	8

Type	Nombre
hémizygote	43
hétérozygote	460
homozygote	46
Indéterminé	13

Tableau 13. – Caractéristiques des variants rendus. La classification et le caractère homo/hétérozygote n'ont pu être extraits pour certains variants.

Le pipeline a été exécuté avec les versions suivantes des bases de données détaillées dans le Tableau 14 et les versions logicielles précisées en Annexe (Tableau 17):

Base de données	Version
génome	GRCh38 version non patchées avec ALT
dbSNP	GRCH38.p14
Clinvar	1.6 - GRCH38
CADD	GRCH38
SpliceAI	GRCH38
VEP	110

Tableau 14. – Versions des bases de données utilisées dans Bisonex.

4.2 T2T

Le nouveau génome CHM13-T2T, présenté dans le Chapitre 1.4.2, permet d'accéder la quasi-totalité du génome humain. Selon une première étude (Aganezov et al. 2022), il permettrait d'aligner environ 1% de reads supplémentaire, de diminuer de nombre de faux positifs et d'appeler des variants dans des zones précédemment non résolues mais en diminuant le nombre de variants appelés. Avec la diffusion d'une version stable (version 2) et l'intégration dans VEP, il nous a donc semblé opportun de réaliser des premiers tests. Le pipeline actuel est complètement compatible avec CHM13-T2T, sous réserve de l'utilisation d'une version modifiée de SPiP au lieu de SpliceAI. En effet, SpliceAI ne propose pour l'instant pas ses scores pour T2T. Notre version de SPiP permet un calcul de score d'épissage mais qui est pour le moment incorrect sur les positions en amont de 18 à 44bp du site accepteur (communication personnelle).

Pour valider l'appel de variants, 144 SNVs ont été insérés dans les données brutes d'un patient de synthèse à l'aide de notre outil maison. Tous ces variants ont été re-

trouvés après l'appel de variants. Sur des tests préliminaires sur un patient, le nombre de variants appelés a diminué d'environ 11% par comparaison avec GRCh38, ce qui semble compatible avec les résultats de Aganezov et al. (2022). Cependant, malgré l'usage de filtres avec la version modifiée de SPiP, le nombre de variants pour l'interprétation a été multiplié par 3 (6 000 vs 2 000). Ceux-ci sont principalement des faux-sens et ne sont pas liés à l'ajout de zones précédemment non résolues dans T2T pour plus de 99%. Il est probable que l'annotation fonctionnelle avec VEP ne soit donc pas encore utilisable en diagnostic. Le reste de ce chapitre utilise donc la version GRCh38 du génome.

4.3 Comparaison avec Centogène

Pour vérifier que notre pipeline n'est pas moins inférieur que celui du sous-traitant, nous avons comparé un ensemble de variants retrouvés par leur pipeline pour lesquels nous disposions des données brutes. La démarche et résultats sont résumés sous la forme d'un diagramme de flux dans la Fig. 59. Parmi tous les variants rendus, plus de 95% ont été retrouvés par notre pipeline (94/98). Ceux-ci sont majoritairement des SNVs (82/94) avec 7 délétions, 4 duplications et 1 insertion. Aucun de ces variants n'a été initialement confirmé ou infirmé en Sanger. À l'aide de l'équipe du Dr Dahlen, ces 4 variants ont pu être confirmés en Sanger. Parmi les 4 variants, 2 étaient sur des régions difficiles pour la conception d'amorce (*CHD3* et *PITX3*), cette dernière étant riche en GC avec un homopolymère de 7G au niveau de la délétion. Les variants ont été filtrés par le pipeline en raison soit d'une profondeur insuffisante, soit d'un nombre insuffisant de reads porteurs du variant comme détaillé sur le Tableau 15. Ce tableau montre les mêmes paramètres sur les variants rendus par Centogène en hg19. On peut constater que 2 variants ont moins de reads avec notre pipeline et sont donc filtrés (*CHD3*, *RRAS2*), que pour 1 variant, il suffirait d'abaisser le seuil de reads à 29 pour ne pas le filtrer et que pour le dernier variant, il suffirait de diminuer le seuil du nombre de reads porteurs du variant à 8. Cette comparaison illustre donc l'important des filtres et surtout de leur validation.

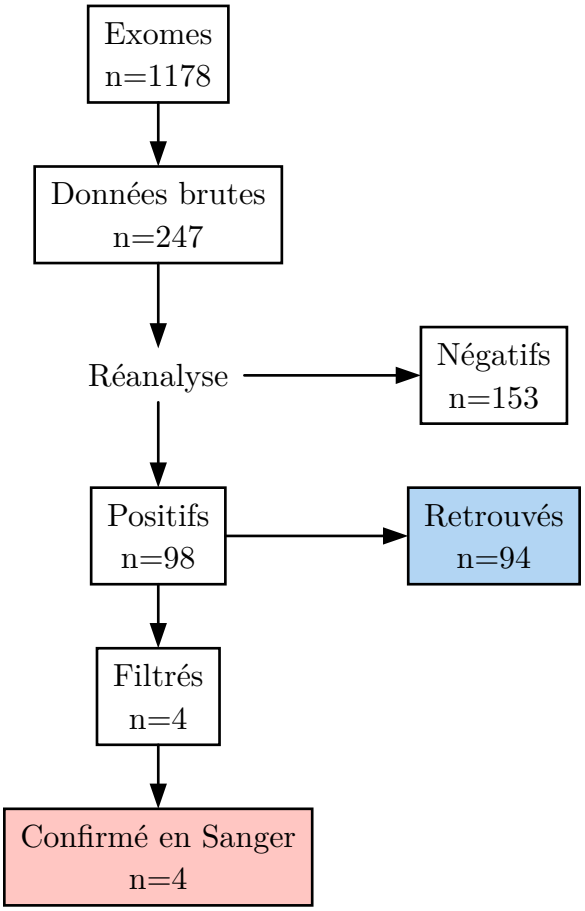


Fig. 59. – Ré-interprétation des données brutes d'exomes en comparant aux résultats du laboratoire sous-traitant. La configuration actuelle du pipeline filtre 4 variants par excès

Gène	Profondeur		Reads porteur du variant	
	Bisonex	Centogène	Bisonex	Centogène
CHD3	**27**	29	22	21
PITX3	34	45	**8**	8
GABRA5	**15**	34	6	42
RRAS2	**29**	29	14	14

Tableau 15. – Variants filtrés par notre pipeline mais rendu par Centogène . En gras, la raison pour laquelle il a été filtré. Pour rappel, les reads de profondeur ≤ 30 et les variants avec ≤ 10 reads porteurs de la variation sont exclus de l'analyse.

4.4 Nouveaux phénotypes

La ré-analyse des données brutes est une stratégie qui permet de profiter de la mise à jour du pipeline, notamment des bases de données OMIM et Clinvar, pouvant conduire à 10 à 30% de nouveau diagnostics (voir la revue faite par Tran Mau-Them et al. (2023)). La découverte de nouveaux diagnostics peut être due à la découverte de nouveaux gènes impliqués en génétique constitutionnelle, mais aussi de nouvelles associations phénotypiques avec des gènes déjà connus. Cependant, il existe un délai avant que ces nouveaux gènes soient ajoutés dans les différentes bases de données. Il est alors intéressant de faire une mise à jour régulière de la bibliographie pour de nouveaux diagnostics. Enfin, on notera que Centogène utilise encore la version GRCh37 du génome pour rendre les résultats, ce qui rend d'autant plus intéressant une ré-analyse dans la version GRCh38 du génome.

Dans cette partie, nous avons recherché les nouveaux phénotypes OMIM depuis le 1er janvier 2021 associés à des gènes connus. Comme la sortie du pipeline est simplement un tableau au format texte, nous avons utilisé l'outil `grep`[159] en recherchant les gènes associés à ces nouveaux phénotypes. Deux variants faux-sens ayant un intérêt potentiel ont été identifiés sur le gène *AFF3* qui a été récemment rattaché au syndrome Kinsship[160]. Le premier variant, sur l'exon 4, a quelques scores bio-informatiques en faveur d'une pathogéncitié : CADD à 23.60, SIFT et Polyphen pour les faux-sens (Fig. 60). Il n'est pas présent dans gnomAD. La clinique correspond partiellement, mais de manière non spécifique car ce patient a une déficience intellectuelle et un strabisme. Centogène a déjà rendu un VSI sur le gène *BCORL1* en lien avec des troubles du neurodéveloppement (syndrome Shukla-Vernon). Un variant sur l'exon 21 a été retrouvé pour le second patient. Le seul score bio-informatique en faveur est celui de SPiP, suggérant que l'épissage est altéré, mais en contradiction avec SpliceAI, qui ne prédit pas d'atteinte de l'épissage. Il est rapporté une fois dans clinvar, de signification indéterminée. La clinique est plus spécifique et correspond partiellement car ce patient souffre également d'une encéphalopathie et ventriculomégalie. Néanmoins, sa paraparésie spastique n'est pas rapportée dans la description clinique du syndrome sur OMIM. L'exome a été rendu négatif. Pour vérifier ces variants, nous avons confirmé qu'ils étaient bien dans appelés par le pipeline de Centogène.

[159] Plus exactement, `ripgrep` qui est une ré-écriture de cet outil Linux permettant un usage sous Windows

[160] https://www.omim.org/entry/619297

Sur le plan moléculaire, ces deux variants semblent donc être des VSI, qu'il faudrait confronter à la clinique avant d'envisager une exploration plus poussée.

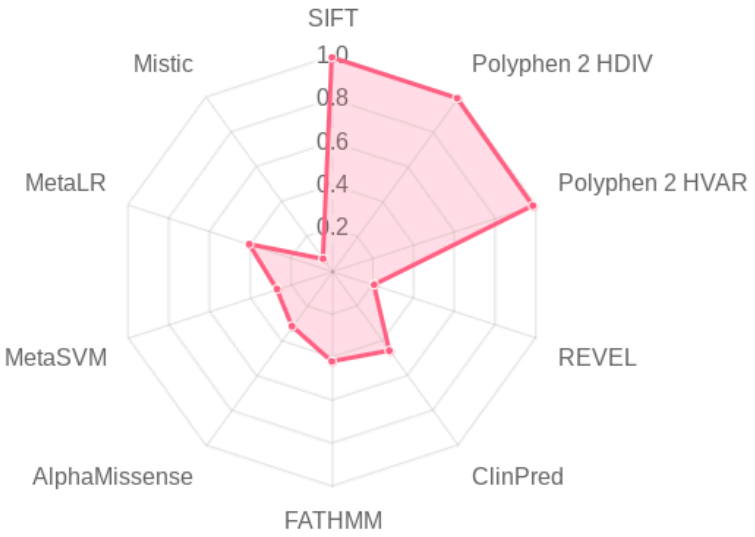


Fig. 60. – Scores bioinformatiques pour les variants faux-sens du patient 1 ré-analysé.

Conclusion

Grâce au développement du séquencage de nouvelle génération, il est désormais possible de proposer à l'ensemble de la population française des analyses de l'exome, soit 1% environ du génome, voire du génome entier. Le service de consultation de génétique constitutionnelle de Besançon propose en première intention un exome, qui a l'avantage d'avoir un délai de rendu bien inférieur à celui du génome, un rendement diagnostic intéressant et un coût raisonnable. Cependant, la quantité de données générée par l'exome est conséquente et nécessite une analyse bioinformatique dédiée et performante pour extraire une liste de variants et les filtrer afin de la rendre interprétable par un biologiste expert. De plus, près de la moitié des résultats rendus d'exome sont rendus négatifs. Pour améliorer le service médical rendu, le laboratoire de génétique de Besançon a donc demandé au laboratoire sous-traitant l'exome la mise à disposition des données brutes pour les ré-analyser. C'est dans ce cadre que cette thèse s'inscrit en proposant le développement d'un tel pipeline. Nous avons également procédé à une validation et discutons des premiers résultats sur la ré-interprétation des résultats.

Sur le plan informatique, les différents constituants d'un tel pipeline existent déjà, mais choisir la meilleure combinaison parmi les différentes possibilités est loin d'être évident. Nous avons donc dans un premier temps réalisé une étude de la bibliographie pour sélectionner les logiciels les plus adaptés à notre besoin. Du fait de la multiplicité des outils, assurer la reproductibilité du pipeline sur différentes architectures est un problème difficile. Un des apports de cette thèse est donc l'utilisation d'un outil proposant une solution à ce problème, appelé Nix, qui assure une reproductibilité à 100% sous Linux. Nous avons incorporé plusieurs logiciels indispensables à un pipeline de génétique constitutionnel dans Nix et ces apports ont été soumis à la communauté pour permettre une plus grande adoption. De plus, l'utilisation de l'outil Nextflow permet de développer un pipeline de manière indépendante de l'architecture sur laquelle il s'exécutera. Autrement dit, le pipeline peut s'exécuter sur un supercalculateur comme sur un ordinateur personnel sans changement dans le code source. À

ce titre, les performances ont été validées sur le supercalculateur de Franche-Comté, permettant une analyse de 20 patients par jour.

Sur le plan biologique, il est indispensable de valider les résultats, notamment sur les filtres choisis afin d'éviter des résultats faussement négatifs ou une interprétation impossible. Pour cela, les données fournies par le consortium GIAB ont permis l'analyse complète d'un échantillon d'un patient de référence et la validation de la partie bio-informatique seule sur 7 patients de référence. Pour augmenter le nombre de variants testés, une approche *in silico* avec des variants confirmés en Sanger a été conduite en modifiant les données d'un patient connu, dit « de synthèse », ainsi qu'en générant des données de séquençage de manière purement informatique.

Sur le plan médical, 254 exomes ont été ré-analysés par notre pipeline. Pour vérifier la non-infériorité avec les résultats de Centogène, nous avons recherché tous les variants rendus initialement. Sur les 98 résultats positifs, 94 ont été également retrouvés par notre pipeline. Les 4 restants ont été filtrés pour cause de profondeur insuffisante ou d'un manque de reads portant ces variations. Une étude par Sanger de ces 4 variants a confirmé leur existence. Ce résultat illustre donc l'importance d'un processus de validation de méthode. Enfin, une analyse préliminaire sur les nouveaux phénotypes associés à des gènes connus dans la base de données OMIM a mis en évidence 2 VSI.

Ce travail pourra utilement être complété par les évolutions actuelles qui s'annoncent passionnantes. D'une part, l'apport du nouveau génome CHM13-T2T, bien que pour l'instant non utilisable en diagnostic, est une piste pour la découverte de nouveaux variants ou la correction de variants complexes. Sur le plan de la validation, des versions préliminaires de données de référence ont été publiées en utilisant le génome d'un patient assemblé par le consortium T2T. Cela permettrait de s'affranchir des biais de séquençage des technologies *short read* pour les données de réference actuelles. Enfin, en Franche-Comté, il reste un travail important de ré-analyse en exploitant les nouveaux gènes décrit dans la base de données OMIM, voire ceux récemment publiés. Il serait également très intéressant de ne pas se limiter à l'étude d'un seul gène mais d'étudier la pathogénicité de deux gènes pour certaines maladies.

Bibliographie

Aganezov, Sergey, Stephanie M. Yan, Daniela C. Soto, Melanie Kirsche, Samantha Zarate, Pavel Avdeyev, Dylan J. Taylor, et al. 2022. « A complete reference genome improves analysis of human genetic variation ». *Science* 376 (6588). https://doi.org/10.1126/science.abl3533

Agathe, Jean-Madeleine de Sainte, Mathilde Filser, Bertrand Isidor, Thomas Besnard, Paul Gueguen, Aurélien Perrin, Charles Van Goethem, et al. 2023. « SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation ». *Human Genomics* 17 (1). https://doi.org/10.1186/s40246-023-00451-1

Alganmi, Nofe, et Heba Abusamra. 2023. « Evaluation of an optimized germline exomes pipeline using BWA-MEM2 and Dragen-GATK tools ». *PLOS ONE* 18 (8): e288371. https://doi.org/10.1371/journal.pone.0288371

Alser, Mohammed, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, et al. 2021. « Technology dictates algorithms: recent developments in read alignment ». *Genome Biology* 22 (1). https://doi.org/10.1186/s13059-021-02443-7

Amberger, Joanna S., Carol A. Bocchini, François Schiettecatte, Alan F. Scott, et Ada Hamosh. 2015. « OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders ». *Nucleic Acids Research* 43 (D1): D789–D798. https://doi.org/10.1093/nar/gku1205

Anna, Abramowicz, et Gos Monika. 2018. « Splicing mutations in human genetic disorders: examples, detection, and confirmation ». *Journal of Applied Genetics* 59 (3): 253-68. https://doi.org/10.1007/s13353-018-0444-7

ANNOVAR. 2023. « ANNOVAR documentation ». https://doi.org/10.1093/nar/gkz923/5603227

Auwera, Geraldine A. Van der, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. « From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline ». *Current Protocols in Bioinformatics* 43 (1). https://doi.org/10.1002/0471250953.bi1110s43

Baid, Gunjan, Maria Nattestad, Alexey Kolesnikov, Sidharth Goel, Howard Yang, Pi-Chuan Chang, et Andrew Carroll. 2020. « An Extensive Sequence Dataset of

Gold-Standard Samples for Benchmarking and Development ». Cold Spring Harbor Laboratory. 2020. https://doi.org/10.1101/2020.12.11.422022

Baker, Monya. 2012. « De novo genome assembly: what every biologist should know ». *Nature Methods* 9 (4): 333-07. https://doi.org/10.1038/nmeth.1935

Barbitoff, Yury A., Ruslan Abasov, Varvara E. Tvorogova, Andrey S. Glotov, et Alexander V. Predeus. 2022. « Systematic benchmark of state-of-the-art variant calling pipelines identifies major factors affecting accuracy of coding sequence variant discovery ». *BMC Genomics* 23 (1). https://doi.org/10.1186/s12864-022-08365-3

Bedő, Justin, Leon Di~Stefano, et Anthony T Papenfuss. 2020. « Unifying package managers, workflow engines, and containers: Computational reproducibility with BioNix ». *GigaScience* 9 (11). https://doi.org/10.1093/gigascience/giaa121

Cameron, Daniel L., Jonathan Baber, Charles Shale, Jose Espejo Valle-Inclan, Nicolle Besselink, Arne van Hoeck, Roel Janssen, Edwin Cuppen, Peter Priestley, et Anthony T. Papenfuss. 2021. « GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing ». *Genome Biology* 22 (1). https://doi.org/10.1186/s13059-021-02423-x

Chen, Jiayun, Xingsong Li, Hongbin Zhong, Yuhuan Meng, et Hongli Du. 2019. « Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers ». *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-45835-3

Chen, Xiaoyu, Ole Schulz-Trieglaff, Richard Shaw, Bret Barnes, Felix Schlesinger, Morten Källberg, Anthony J. Cox, Semyon Kruglyak, et Christopher T. Saunders. 2016. « Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications ». *Bioinformatics* 32 (8): 1220-02. https://doi.org/10.1093/bioinformatics/btv710

Cibulskis, Kristian, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, et Gad Getz. 2013. « Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples ». *Nature Biotechnology* 31 (3): 213-09. https://doi.org/10.1038/nbt.2514

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, et Douglas M. Ruden. 2012. « A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff ». *Fly* 6 (2): 80-92. https://doi.org/10.4161/fly.19695

Clark, Michelle M., Zornitza Stark, Lauge Farnaes, Tiong Y. Tan, Susan M. White, David Dimmock, et Stephen F. Kingsmore. 2018. « Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases ». *npj Genomic Medicine* 3 (1). https://doi.org/10.1038/s41525-018-0053-8

Cleary, John G., Ross Braithwaite, Kurt Gaastra, Brian S Hilbush, Stuart Inglis, Sean A Irvine, Alan Jackson, et al. 2015. « Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines ». Cold Spring Harbor Laboratory. 2015. https://doi.org/10.1101/023754

COFRAC. 2019. « Guide technique d\'accréditation de la technologie de séquençage à haut débit »

D'Aurizio, Romina, Tommaso Pippucci, Lorenzo Tattini, Betti Giusti, Marco Pellegrini, et Alberto Magi. 2016. « Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2 ». *Nucleic Acids Research*, gkw695. https://doi.org/10.1093/nar/gkw695

Dai, Pei, Andrew Honda, Lisa Ewans, Julie McGaughran, Leslie Burnett, Matthew Law, et Tri Giang Phan. 2022. « Recommendations for next generation sequencing data reanalysis of unsolved cases with suspected Mendelian disorders: A systematic review and meta-analysis ». *Genetics in Medicine* 24 (8): 1618-29. https://doi.org/10.1016/j.gim.2022.04.021

Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. « Twelve years of SAMtools and BCFtools ». *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008

Devresse, Adrien, Fabien Delalondre, et Felix Schürmann. 2015. « Nix based fully automated workflows and ecosystem to guarantee scientific result reproducibility across software environments and systems ». In *SC15: The International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM. https://doi.org/10.1145/2830168.2830172

Dolstra, Eelco, Merijn de Jonge, et Eelco Visser. 2004. « Nix: A Safe and Policy-Free System for Software Deployment ». In . https://edolstra.github.io/pubs/nspfssd-lisa2004-final.pdf

Dolzhenko, Egor, Joke J.F.A. van Vugt, Richard J. Shaw, Mitchell A. Bekritsky, Marka van Blitterswijk, Giuseppe Narzisi, Subramanian S. Ajay, et al. 2017. « Detection of long repeat expansions from PCR-free whole-genome sequence data ». *Genome Research* 27 (11): 1895-903. https://doi.org/10.1101/gr.225672.117

Donato, Luigi, Concetta Scimone, Carmela Rinaldi, Rosalia D'Angelo, et Antonina Sidoti. 2021. « New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies ». *Neural Computing and Applications* 33 (22): 15669-92. https://doi.org/10.1007/s00521-021-06188-z

Duncavage, Eric J., Joshua F. Coleman, Monica E. de Baca, Sabah Kadri, Annette Leon, Mark Routbort, Somak Roy, Carlos J. Suarez, Chad Vanderbilt, et Justin M. Zook. 2023. « Recommendations for the Use of in Silico Approaches for Next-Generation Sequencing Bioinformatic Pipeline Validation ». *The Journal of Molecular Diagnostics* 25 (1): 3-16. https://doi.org/10.1016/j.jmoldx.2022.09.007

Eberle, Michael A., Epameinondas Fritzilas, Peter Krusche, Morten Källberg, Benjamin L. Moore, Mitchell A. Bekritsky, Zamin Iqbal, et al. 2017. « A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree ». *Genome Research* 27 (1): 157-64. https://doi.org/10.1101/gr.210500.116

Ewing, Adam D, None None, Kathleen E Houlahan, Yin Hu, Kyle Ellrott, Cristian Caloian, Takafumi N Yamaguchi, et al. 2015. « Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection ». *Nature Methods* 12 (7): 623-30. https://doi.org/10.1038/nmeth.3407

Firtina, Can, et Can Alkan. 2016. « On genomic repeats and reproducibility ». *Bioinformatics* 32 (15): 2243-07. https://doi.org/10.1093/bioinformatics/btw139

Garrison, Erik, et Gabor Marth. 2012. « Haplotype-based variant detection from short-read sequencing », juillet. http://arxiv.org/abs/1207.3907v2

Genomes Project Consortium, The 1000. 2015. « A global reference for human genetic variation ». *Nature* 526 (7571): 68-74. https://doi.org/10.1038/nature15393

Goodwin, Sara, John D. McPherson, et W. Richard McCombie. 2016. « Coming of age: ten years of next-generation sequencing technologies ». *Nature Reviews Genetics* 17 (6): 333-51. https://doi.org/10.1038/nrg.2016.49

Harpak, Arbel, Xun Lan, Ziyue Gao, et Jonathan K. Pritchard. 2017. « Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates ». *Proceedings of the National Academy of Sciences* 114 (48): 12779-84. https://doi.org/10.1073/pnas.1708151114

Hatem, Ayat, Doruk Bozdağ, Amanda E Toland, et Ümit V Çatalyürek. 2013. « Benchmarking short sequence mapping tools ». *BMC Bioinformatics* 14 (1). https://doi.org/10.1186/1471-2105-14-184

Head, Steven R., H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, et Phillip Ordoukhanian. 2014. « Library construction for next-generation sequencing: Overviews and challenges ». *BioTechniques* 56 (2): 61-77. https://doi.org/10.2144/000114133

Hu, Taishan, Nilesh Chitnis, Dimitri Monos, et Anh Dinh. 2021. « Next-generation sequencing technologies: An overview ». *Human Immunology* 82 (11): 801-11. https://doi.org/10.1016/j.humimm.2021.02.012

Hwang, Kyu-Baek, In-Hee Lee, Honglan Li, Dhong-Geon Won, Carles Hernandez-Ferrer, Jose Alberto Negron, et Sek Won Kong. 2019. « Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings ». *Scientific Reports* 9 (1). https://doi.org/10.1038/s41598-019-39108-2

Hwang, Sohyun, Eiru Kim, Insuk Lee, et Edward M. Marcotte. 2015. « Systematic comparison of variant calling pipelines using gold standard personal exome variants ». *Scientific Reports* 5 (1). https://doi.org/10.1038/srep17875

Jaganathan, Kishore, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, et al. 2019. « Predicting Splicing from Primary Sequence with Deep Learning ». *Cell* 176 (3): 535-48. https://doi.org/10.1016/j.cell.2018.12.015

Jarvis, Erich D., Giulio Formenti, Arang Rhie, Andrea Guarracino, Chentao Yang, Jonathan Wood, Alan Tracey, et al. 2022. « Semi-automated assembly of high-quality diploid human reference genomes ». *Nature* 611 (7936): 519-31. https://doi.org/10.1038/s41586-022-05325-5

Kim, Sangtae, Konrad Scheffler, Aaron L. Halpern, Mitchell A. Bekritsky, Eunho Noh, Morten Källberg, Xiaoyu Chen, et al. 2018. « Strelka2: fast and accurate calling of germline and somatic variants ». *Nature Methods* 15 (8): 591-04. https://doi.org/10.1038/s41592-018-0051-x

Koboldt, Daniel C., Qunyuan Zhang, David E. Larson, Dong Shen, Michael D. McLellan, Ling Lin, Christopher A. Miller, Elaine R. Mardis, Li Ding, et Richard K. Wilson. 2012. « VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing ». *Genome Research* 22 (3): 568-76. https://doi.org/10.1101/gr.129684.111

Kowalewski, Markus, et Phillip Seeber. 2022. « Sustainable packaging of quantum chemistry software with the Nix package manager ». *International Journal of Quantum Chemistry* 122 (9). https://doi.org/10.1002/qua.26872

Krusche, Peter, None None, Len Trigg, Paul C. Boutros, Christopher E. Mason, Francisco M. De La Vega, Benjamin L. Moore, et al. 2019. « Best practices for benchmarking germline small-variant calls in human genomes ». *Nature Biotechnology* 37 (5): 555-60. https://doi.org/10.1038/s41587-019-0054-x

Kumaran, Manojkumar, Umadevi Subramanian, et Bharanidharan Devarajan. 2019. « Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data ». *BMC Bioinformatics* 20 (1). https://doi.org/10.1186/s12859-019-2928-9

Kweon, Jiyeon, An-Hee Jang, Ha Rim Shin, Ji-Eun See, Woochang Lee, Jong Won Lee, Suhwan Chang, Kyunggon Kim, et Yongsub Kim. 2020. « A CRISPR-based base-editing screen for the functional assessment of BRCA1 variants ». *Oncogene* 39 (1): 30-35. https://doi.org/10.1038/s41388-019-0968-2

Köster, Johannes, et Sven Rahmann. 2012. « Snakemake—a scalable bioinformatics workflow engine ». *Bioinformatics* 28 (19): 2520-02. https://doi.org/10.1093/bioinformatics/bts480

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2016. « ClinVar: public archive of interpretations of clinically relevant variants ». *Nucleic Acids Research* 44 (D1): D862–D868. https://doi.org/10.1093/nar/gkv1222

Leman, Raphaël, Béatrice Parfait, Dominique Vidaud, Emmanuelle Girodon, Laurence Pacot, Gérald Le Gac, Chandran Ka, et al. 2022. « SPiP: Splicing Prediction Pipeline, a machine learning tool for massive detection of exonic and intronic variant effects on mRNA splicing ». *Human Mutation* 43 (12): 2308-23. https://doi.org/10.1002/humu.24491

Li, Heng, Jonathan M. Bloom, Yossi Farjoun, Mark Fleharty, Laura Gauthier, Benjamin Neale, et Daniel MacArthur. 2018. « A synthetic-diploid benchmark for accurate variant-calling evaluation ». *Nature Methods* 15 (8): 595-07. https://doi.org/10.1038/s41592-018-0054-7

Li, Wentian, Jerome Freudenberg, et Jan Freudenberg. 2019. « Alignment-free approaches for predicting novel Nuclear Mitochondrial Segments (NUMTs) in the human genome ». *Gene* 691: 141-52. https://doi.org/10.1016/j.gene.2018.12.040

Li, Ziyang, Shuangsang Fang, Rui Zhang, Lijia Yu, Jiawei Zhang, Dechao Bu, Liang Sun, Yi Zhao, et Jinming Li. 2021. « VarBen: Generating in Silico Reference Data Sets for Clinical Next-Generation Sequencing Bioinformatics Pipeline Evaluation ». *The Journal of Molecular Diagnostics* 23 (3): 285-99. https://doi.org/10.1016/j.jmoldx.2020.11.010

Liao, Wen-Wei, Mobin Asri, Jana Ebler, Daniel Doerr, Marina Haukness, Glenn Hickey, Shuangjia Lu, et al. 2023. « A draft human pangenome reference ». *Nature* 617 (7960): 312-24. https://doi.org/10.1038/s41586-023-05896-x

Matthijs, Gert, Erika Souche, Mariëlle Alders, Anniek Corveleyn, Sebastian Eck, Ilse Feenstra, Valérie Race, et al. 2015. « Guidelines for diagnostic next-generation sequencing ». *European Journal of Human Genetics* 24 (1). https://doi.org/10.1038/ejhg.2015.226

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, et Fiona Cunningham. 2016. « The Ensembl Variant Effect Predictor ». *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0974-4

Miga, Karen H., Yulia Newton, Miten Jain, Nicolas Altemose, Huntington F. Willard, et W. James Kent. 2014. « Centromere reference models for human chromosomes X and Y satellite arrays ». *Genome Research* 24 (4): 697-07. https://doi.org/10.1101/gr.159624.113

Miller, Thiago L A, Fernanda Orpinelli Rego, José Leonel L Buzzo, et Pedro A F Galante. 2021. « sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies ». *Bioinformatics* 37 (3): 419-21. https://doi.org/10.1093/bioinformatics/btaa689

Musich, Ryan, Lance Cadle-Davidson, et Michael V. Osier. 2021. « Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider ». *Frontiers in Plant Science* 12. https://doi.org/10.3389/fpls.2021.657240

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2022. « The complete sequence of a human genome ». *Science* 376 (6588): 44-53. https://doi.org/10.1126/science.abj6987

Olson, Nathan D., Justin Wagner, Nathan Dwarshuis, Karen H. Miga, Fritz J. Sedlazeck, Marc Salit, et Justin M. Zook. 2023. « Variant calling and benchmarking in an era of complete human genome sequences ». *Nature Reviews Genetics*, n° 7 (avril). https://doi.org/10.1038/s41576-023-00590-0

Olson, Nathan D., Justin Wagner, Jennifer McDaniel, Sarah H. Stephens, Samuel T. Westreich, Anish G. Prasanna, Elaine Johanson, et al. 2020. « precisionFDA Truth Challenge V2: Calling variants from short- and long-reads in difficult-to-map regions ». Cold Spring Harbor Laboratory. 2020. https://doi.org/10.1101/2020.11.13.380741

Olson, Nathan D., Justin Wagner, Jennifer McDaniel, Sarah H. Stephens, Samuel T. Westreich, Anish G. Prasanna, Elaine Johanson, et al. 2022. « PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions ». *Cell Genomics* 2 (5): 100129. https://doi.org/10.1016/j.xgen.2022.100129

Plagnol, Vincent, James Curtis, Michael Epstein, Kin Y. Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W. Wood, et al. 2012. « A robust model for read count data in exome sequencing experiments and implications for copy number variant calling ». *Bioinformatics* 28 (21): 2747-54. https://doi.org/10.1093/bioinformatics/bts526

Poplin, Ryan, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, et al. 2018. « A universal SNP and small-indel variant caller using deep neural networks ». *Nature Biotechnology* 36 (10): 983-07. https://doi.org/10.1038/nbt.4235

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2017. « Scaling accurate genetic variant discovery to tens of thousands of samples ». Cold Spring Harbor Laboratory. 2017. https://doi.org/10.1101/201178

Rausch, Tobias, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, et Jan O. Korbel. 2012. « DELLY: structural variant discovery by integrated paired-end and split-read analysis ». *Bioinformatics* 28 (18): i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Regier, Allison A., Yossi Farjoun, David E. Larson, Olga Krasheninina, Hyun Min Kang, Daniel P. Howrigan, Bo-Juen Chen, et al. 2018. « Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects ». *Nature Communications* 9 (1). https://doi.org/10.1038/s41467-018-06159-4

Rhie, Arang, Sergey Nurk, Monika Cechova, Savannah J. Hoyt, Dylan J. Taylor, Nicolas Altemose, Paul W. Hook, et al. 2023. « The complete sequence of a human Y chromosome ». *Nature* 621 (7978): 344-54. https://doi.org/10.1038/s41586-023-06457-y

Rimmer, Andy, None None, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, Andrew O M Wilkie, Gil McVean, et Gerton Lunter. 2014. « Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications ». *Nature Genetics* 46 (8): 912-08. https://doi.org/10.1038/ng.3036

Roy, Somak, Christopher Coldren, Arivarasan Karunamurthy, Nefize S. Kip, Eric W. Klee, Stephen E. Lincoln, Annette Leon, et al. 2018. « Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines ». *The Journal of Molecular Diagnostics* 20 (1): 4-27. https://doi.org/10.1016/j.jmoldx.2017.11.003

Schneider, Valerie A., Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, et al. 2017. « Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly ». *Genome Research* 27 (5): 849-64. https://doi.org/10.1101/gr.213611.116

Sherry, S. T. 2001. « dbSNP: the NCBI database of genetic variation ». *Nucleic Acids Research* 29 (1): 308-11. https://doi.org/10.1093/nar/29.1.308

Souche, Erika, Sergi Beltran, Erwin Brosens, John W. Belmont, Magdalena Fossum, Olaf Riess, Christian Gilissen, et al. 2022. « Recommendations for whole genome sequencing in diagnostics for rare diseases ». *European Journal of Human Genetics* 30 (9). https://doi.org/10.1038/s41431-022-01113-x

Tan, Tiong Yang, Oliver James Dillon, Zornitza Stark, Deborah Schofield, Khurshid Alam, Rupendra Shrestha, Belinda Chong, et al. 2017. « Diagnostic Impact and Cost-effectiveness of Whole-Exome Sequencing for Ambulant Children With Suspected Monogenic Conditions ». *JAMA Pediatrics* 171 (9): 855. https://doi.org/10.1001/jamapediatrics.2017.1755

Thung, Djie Tjwan, Joep de Ligt, Lisenka EM Vissers, Marloes Steehouwer, Mark Kroon, Petra de Vries, Eline P Slagboom, Kai Ye, Joris A Veltman, et Jayne Y Hehir-Kwa. 2014. « Mobster: accurate detection of mobile element insertions in next generation sequencing data ». *Genome Biology* 15 (10). https://doi.org/10.1186/s13059-014-0488-x

Tommaso, Paolo Di, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, et Cedric Notredame. 2017. « Nextflow enables reproducible computational workflows ». *Nature Biotechnology* 35 (4): 316-09. https://doi.org/10.1038/nbt.3820

Tran Mau-Them, Frédéric, Alexis Overs, Ange-Line Bruel, Romain Duquet, Mylene Thareau, Anne-Sophie Denommé-Pichon, Antonio Vitobello, et al. 2023. « Combining globally search for a regular expression and print matching lines with bibliographic monitoring of genomic database improves diagnosis ». *Frontiers in Genetics* 14. https://doi.org/10.3389/fgene.2023.1122985

Vallet, Nicolas, David Michonneau, et Simon Tournier. 2022. « Toward practical transparent verifiable and long-term reproducible research using Guix ». *Scientific Data* 9 (1). https://doi.org/10.1038/s41597-022-01720-9

Wagner, Justin, Nathan D Olson, Lindsay Harris, Jennifer McDaniel, Ziad Khan, Jesse Farek, Medhat Mahmoud, et al. 2022. « Benchmarking challenging small variants with linked and long reads ». Cold Spring Harbor Laboratory. 2022. https://doi.org/10.1101/2020.07.24.212712

Wang, Jing, Leon Raskin, David C. Samuels, Yu Shyr, et Yan Guo. 2015. « Genome measures used for quality control are dependent on gene function and ancestry ». *Bioinformatics* 31 (3): 318-23. https://doi.org/10.1093/bioinformatics/btu668

Wang, Kai, Mingyao Li, et Hakon Hakonarson. 2010. « ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data ». *Nucleic Acids Research* 38 (16). https://doi.org/10.1093/nar/gkq603

Wetterstrand, Kris A. s. d. « DNA Sequencing Costs: Data ». https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Yang, Hui, et Kai Wang. 2015. « Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR ». *Nature Protocols* 10 (10). https://doi.org/10.1038/nprot.2015.105

Ye, Kai, Li Guo, Xiaofei Yang, Eric-Wubbo Lamijer, Keiran Raine, et Zemin Ning. 2018. « Split-Read Indel and Structural Variant Calling Using PINDEL ». In , 95-105. Springer New York. https://doi.org/10.1007/978-1-4939-8666-8_7

Yen, Jennifer L., Sarah Garcia, Aldrin Montana, Jason Harris, Stephen Chervitz, Massimo Morra, John West, Richard Chen, et Deanna M. Church. 2017. « A variant by any name: quantifying annotation discordance across tools and clinical databases ». *Genome Medicine* 9 (1). https://doi.org/10.1186/s13073-016-0396-7

Yu, Zhenhua, Fang Du, Rongjun Ban, et Yuanwei Zhang. 2020. « SimuSCoP: reliably simulate Illumina sequencing data based on position and context dependent profiles ». *BMC Bioinformatics* 21 (1). https://doi.org/10.1186/s12859-020-03665-5

Zeng, Tony, et Yang I Li. 2022. « Predicting RNA splicing from DNA sequence using Pangolin ». *Genome Biology* 23 (1). https://doi.org/10.1186/s13059-022-02664-4

Zhao, Sen, Oleg Agafonov, Abdulrahman Azab, Tomasz Stokowy, et Eivind Hovig. 2020. « Accuracy and efficiency of germline variant calling pipelines for hu-

man genome data ». *Scientific Reports* 10 (1). https://doi.org/10.1038/s41598-020-77218-4

Zook, Justin M., David Catoe, Jennifer McDaniel, Lindsay Vang, Noah Spies, Arend Sidow, Ziming Weng, et al. 2016. « Extensive sequencing of seven human genomes to characterize benchmark reference materials ». *Scientific Data* 3. https://doi.org/10.1038/sdata.2016.25

Zook, Justin M., Nancy F. Hansen, Nathan D. Olson, Lesley Chapman, James C. Mullikin, Chunlin Xiao, Stephen Sherry, et al. 2020. « A robust benchmark for detection of germline large deletions and insertions ». *Nature Biotechnology* 38 (11): 1347-55. https://doi.org/10.1038/s41587-020-0538-8

Zook, Justin M., Jennifer McDaniel, Nathan D. Olson, Justin Wagner, Hemang Parikh, Haynes Heaton, Sean A. Irvine, et al. 2019. « An open resource for accurately benchmarking small variant and reference calls ». *Nature Biotechnology* 37 (5): 561-06. https://doi.org/10.1038/s41587-019-0074-6

Zverinova, Stepanka, et Victor Guryev. 2022. « Variant calling: Considerations, practices, and developments ». *Human Mutation* 43 (8): 976-85. https://doi.org/10.1002/humu.24311

Annexes

Fig. 61. – Exemple d'un homopolymère, dinucléotide, trinucléotide et répétition en tandem (Olson et al. 2020)

Critères de validation selon l'Association for Molecular Pathology, College of American Pathologists

Traduit de (Roy et al. 2018)

1. Les laboratoires cliniques qui proposent des tests basés sur la NGS devraient procéder à leur propre validation du pipeline bioinformatique.
2. Un professionnel de la santé qualifié ayant reçu une formation appropriée en matière d'interprétation et de certification NGS doit superviser le processus de validation et y participer.
3. La validation ne doit être effectuée qu'après l'achèvement de la conception, du développement, de l'optimisation et de la familiarisation du pipeline bioinformatique et de ses composants.
4. la validation du pipeline bioinformatique doit reproduire fidèlement l'environnement réel du laboratoire dans lequel l'essai est effectué
5. la validation doit porter sur tous les composants individuels du pipeline bioinformatique utilisé dans l'analyse, et chaque composant doit être examiné et approuvé

par un professionnel de la biologie moléculaire médicale dûment qualifié et par le directeur du laboratoire

6. La conception et la mise en œuvre du pipeline bioinformatique doivent garantir la sécurité des informations identifiant les patients et être conformes à la loi nationale.
7. La validation du pipeline bioinformatique NGS doit être appropriée et applicable à l'utilisation clinique prévue, à l'échantillon et aux types de variants détectés dans le cadre du test NGS.
8. Les laboratoires doivent veiller à ce que la conception, la mise en œuvre et la validation du pipeline bioinformatique soient conformes aux normes et réglementations applicables en matière d'accréditation des laboratoires.
9. Le pipeline bioinformatique fait partie de la procédure d'essai, et ses composants et processus doivent être documentés conformément aux normes et réglementations d'accréditation des laboratoires.
10. L'identité de l'échantillon doit être préservée à chaque étape du pipeline bioinformatique NGS avec un minimum de quatre identifiants uniques, y compris un identifiant de localisation unique dans le contenu de chaque fichier de données lu et/ou généré par le pipeline.
11. Des paramètres spécifiques de contrôle et d'assurance de la qualité doivent être évalués au cours de la validation et utilisés pour déterminer les performances satisfaisantes du pipeline bioinformatique.
12. les méthodes utilisées pour modifier ou filtrer les séquences lues à tout moment dans le pipeline bioinformatique avant l'interprétation doivent être validées afin de garantir que les données présentées pour l'interprétation représentent de manière précise et reproductible la séquence dans l'échantillon, et une documentation complète de ces méthodes doit être conservée dans le cadre de la documentation d'essai conformément aux normes d'accréditation des laboratoires et aux réglementations en vigueur
13. Les laboratoires doivent prévoir des mesures spécifiques pour garantir que chaque fichier de données généré dans le pipeline bioinformatique conserve son intégrité et fournit des alertes ou empêche l'utilisation de fichiers de données qui ont été modifiés d'une manière non autorisée ou non voulue.
14. La validation in silico peut être utilisée pour compléter la validation du pipeline bioinformatique, mais ne doit pas remplacer la validation de bout en bout des pipelines bioinformatiques à l'aide d'échantillons humains.
15. La validation du pipeline bioinformatique doit inclure la confirmation d'un ensemble représentatif de variants avec des données indépendantes de haute qualité ; des mesures de validation appropriées par type de variant doivent être communiquées.

16. Les laboratoires cliniques doivent veiller à l'exactitude de la nomenclature et des annotations des variants HGVS générées par les logiciels et disposer d'un système d'alerte indiquant quand la nomenclature et les annotations générées par les logiciels doivent être revues et/ou corrigées manuellement, et la documentation relative à toute correction doit être conservée.
17. Une validation supplémentaire est nécessaire chaque fois qu'une modification importante est apportée à un élément du pipeline bioinformatique.

Catégorie	Métrique
Échantillon	Concentration ADN
	Taille des fragments d'ADN
	Quantification de la librairie
Exécution	Densité du cluster
	% bases >= score Phred minimal
	Réussite du démultiplexing
	% reads >= score Phred minimal
	Qualité de l'alignement
	Couverture moyenne
Alignement	% bases ciblées avec couverture >= minimum
	% bases alignées >= score Phred minimal
	% bases alignées >= score Phred minimal différentes de la référence
	% duplicats PCR
Par variant	profondeur
	Score de qualité
	Biais de brin
	Nombre de variants verticaux différents à cette position
	Nombre de variants horizontaux différents dans une fenêtre donnée
QC	Sexe observé correspond à celui prescrit

Tableau 16. – Métriques obligatoire de qualité selon (Roy et al. 2018)

Outils et version

Logiciels	Versions
awscli2	2.11.20
bcftools	1.17
bedtools	2.31.0
bwa	2022-09-23
nextflow	22.10.6
dos2unix	7.4.4
fastqc	0.11.9
gatk	4.4.0.0
hap-py	hap.py
htslib	1.17
mosdepth	0.3.3
multiqc	1.15
perl	5.36.0
picard-tools	3.0.0
python	3.10.12
R	4.2.3-wrapper
rtg-tools	3.12.1
samtools	1.17
SPiP	cae95fe0ee7a2602630b7a4eacbf7cfa0e1763f0
vcftools	0.1.16
vep	110

Tableau 17. – Versions des logiciels utilisés dans Bisonex

Pipeline

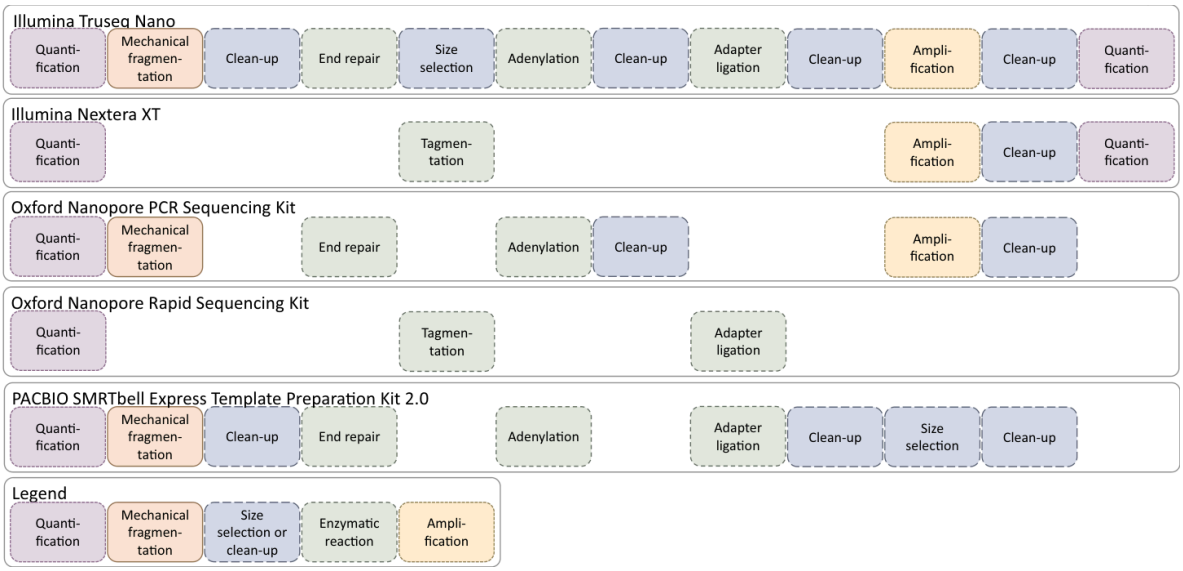


Fig. 62. – Présentations de protocoles commerciaux pour la préparation de librairies (Head et al. 2014).

Aligneurs

Aligner	publication	Indexing	Pairwise alignment	Max. read length (bp)
BWA	2009	BWT-FM	Semi-Global	125
Bowtie	2009	BWT-FM	HD	76
CloudBurst	2009	Hashing	Landau-Vishkin	36
GNUMAP	2009	Hashing	NW	36
Genome Mapper	2009	Hashing	NW	200
MOM	2009	Hashing	HD	40
PASS	2009	Hashing	NW	32
PerM	2009	Hashing	HD	47
RazerS	2009	Hashing	Myers Bit Vector	76
SHRiMP	2009	Hashing	SW	35
SOAP2	2009	BWT-FM	SW	44
Slider	2009	Hashing	HD	36

segemehl	2009	Suffix array	SW	35
BWA-SW	2010	BWT-FM	SW	10000
GASSST	2010	Hashing	Semi-Global	500
GSNAP	2010	Hashing	Non-DP Heuristic	100
SMALT	2010	Hashing	SW	150
SliderII (W)	2010	Hashing	HD	42
VMATCH (W)	2010	Suffix array	SW	N/A
mrsFAST	2010	Hashing	HD	100
LAST	2011	Suffix array	SW & NW	105
DynMap	2011	Hashing	NW	52
SHRiMP2	2011	Hashing	SW	75
SNAP	2011	Hashing	NW	10000
Stampy	2011	Hashing	NW	4500
TMAP	2011	BWT-FM	SW	N/A
X-Mate	2011	Hashing	Non-DP Heuristic	50
BLASR	2012	Suffix array	NW	8000
Batmis	2012	BWT-ST	HD	100
Bowtie2	2012	BWT-FM	SW & NW	400
GEM	2012	BWT-FM	SW & NW	150
RazerS3	2012	Hashing	Banded Myers Bit Vector	800
SeqAlto	2012	Hashing	NW	200
SplazerS	2012	Hashing	Banded Myers Bit Vector	150
WHAM	2012	Hashing	NW	74
YAHA	2012	Hashing	SW	10000
Subread	2013	Hashing	SW	202
BWA-MEM	2013	BWT-FM	SW & NW	650
Masai	2013	Suffix tree	Banded Myers Bit Vector	150
NextGenMap	2013	Hashing	SW & NW	250
SRmapper	2013	Hashing	HD	100

mrFAST	2013	Hashing	Semi-Global	180
BWA-PSSM (W)	2014	BWT-FM	SW	100
CUSHAW3	2014	BWT-FM	SW & Semi-Global	100
Hobbes2	2014	Hashing	Banded Myers Bit Vector	
MOSAIK	2014	Hashing	SW	100
hpg-Aligner	2014	Suffix array	SW	5000
mrsFAST-Ultra	2014	Hashing	HD	100
ERNE2	2016	BWT-FM +hashing	HD	100
GraphMap	2016	Hashing	Semi-global	9000
NanoBLASTer	2016	Hashing	NW	7040
minimap	2016	Hashing	N/A	13000
rHAT	2016	Hashing	SW	8000
KART	2017	BWT-FM	NW	7118
LAMSA (W)	2017	BWT-FM + hashing	Sparse DP	1000
minimap2	2018	Hashing	NW	11628
DREAM-Yara (W)	2018	BWT-FM	Banded Myers Bit Vector	
MUMmer4 (W)	2018	Suffix array	SW	7821
NGMLR	2018	Hashing	SW	50000
lordFAST	2018	BWT-FM + hashing	SW & NW	35489
GraphMap2	2019	Hashing	Semi-global	9000
Magic-BLAST	2019	Hashing	Non-DP Heuristic	90000
BWA-MEM2	2019	BWT-FM	SW	650
HISAT2	2019	BWT-FM	Non-DP Heuristic	100
conLSH	2020	Hashing	Sparse DP	8000

Fig. 63. – Aligneur pour données génomique sur ADN après 2009. (W) = wrapper (utilise d'autres outils). GP = global position. Extrait de (Alser et al. 2021)

Appel de variants

Category	Name	Algorithm	Use	Paper
Small variants	GATK Haplotypecaller	Local reassembly of haplotypes	Germline, MNPs	(Poplin et al. 2018)
	BCFtools	Positional, pileups	Germline	(Danecek et al. 2021)
	FreeBayes	Haplotype-based, Bayesian model	Germline, MNPs	(Garrison et Marth 2012)
	GATK Mutect2	Local reassembly	Somatic	(Cibulskis et al. 2013)
	Strelka2	Tiered haplotype model	Germline, somatic	(Kim et al. 2018)
Structural variants	Delly2	RP, SR, RD	Germline SVs	(Rausch et al. 2012)
	Pindel	SR, RP	Germline SVs	(Ye et al. 2018)
	Manta	SR, RP, AS	Germline, somatic	(Chen et al. 2016)
	GRIDSS2	AS, SV Breakpoint	Somatic	(Cameron et al. 2021)
	Varscan2	RD, Circular Binary Segmentation	Exome, somatic, CNVs	(Koboldt et al. 2012)
	EXCAVATOR2	RD, In-,Off-target	Exome, CNVs	(D'Aurizio et al. 2016)
	ExomeDepth	RD, beta-binomial	Exome, CNVs	(Plagnol et al. 2012)
Other	Mobster	RP, clipped reads	MEIs	(Thung et al. 2014)
	ExpansionHunter	Reads spanning, flanking, in-repeat	Repeat expansion	(Dolzhenko et al. 2017)
	sideRETRO	SR, RP at insert	GRIP	(Miller et al. 2021)

Harpak et al. (2017)	HMM model	NAGC	(Harpak et al. 2017)
Li et al. (2019)	k-mer count, MDS	NUMT	(Li, Freudenberg, et Freudenberg 2019)

Tableau 18. – Principaux appels de variant selon (Zverinova et Guryev 2022) . Abbreviations: AS, assembly; CNV, copy-number variants; GATK, Genome Analysis ToolKit; MEI, mobile element insertions; RD, read depth; RP, read pairing; SR, split-read; SV, structural variants.

HiSeq2000	SRR1611178	SeqCap EZ Human Exome Lib v3.0	WES	79.93x
HiSeq2000	SRR1611179	SeqCap EZ Human Exome Lib v3.0	WES	79.84x
HiSeq2000	SRR292250	SeqCap EZ Exome SeqCap v2	WES	116.06x
HiSeq2000	SRR515199	SureSelect v4	WES	298.45x
HiSeq2000	SRR098401	SureSelect v2	WES	116.84x
HiSeq2500	SRR1611183	SeqCap EZ Human Exome Lib v3.0	WES	129.94x
HiSeq2500	SRR1611184	SeqCap EZ Human Exome Lib v3.0	WES	111.90x
HiSeq2000	ERR194147	UCSC Known gene	WGS	45.68x
HiSeq2000	SRX485062	UCSC Known gene	WGS	56.60x
HiSeq2500	SRX515284	UCSC Known gene	WGS	56.87x
HiSeq2500	SRX516752	UCSC Known gene	WGS	43.61x
IonProton	$NA12878_{combine}$	UCSC Known gene	WGS	9.87

Fig. 64. – Configuration pour (Hwang et al. 2015)

Sequencing Samples	Type	Bases (Gbp)	Read ($x10^6$)	Clean rare	>Q20	>Q30	GC content	Mean coverage
BGISEQ 500	WES	29.41	294.30	0.41%	96.72%	89.14%	49.75%	328
MGISEQ 2000	WES	16.34	163.55	0.25%	98.18%	92.08%	49.71%	129
HiSeq4000	WES	41.93	283.70	4.46%	97.36%	93.01%	50.63%	395
NovaSeq	WES	25.88	178.87	2.25%	95.33%	92.67%	49.73%	241
BGISEQ 500	WGS	126.86	1270.02	1.76%	93.73%	83.33%	41.76%	41
MGISEQ 2000	WGS	137.36	1374.87	0.21%	96.17%	88.19%	41.76%	45
HiSeq4000	WGS	191.00	1276.10	8.25%	95.90%	90.11%	41.69%	58
NovaSeq	WGS	98.30	657.45	1.28%	95.89%	93.86%	41.61%	29
HiSeq Xten	WGS	134.00	894.58	7.29%	94.50%	87.63%	40.71%	39

Tableau 19. – Méthodes pour (Chen et al. 2019).>Q20 = moins de 1% d'erreur. >Q30 = moins de 1‰

		exome	genome
BGI	bgiseq500	328X	41x
BGI	mgiseq2000	129	45x
illumina	hiseq	395	58
	novaseq	241	28.96 (*)

Fig. 65. – Méthodes pour (Chen et al. 2019)

Génomes

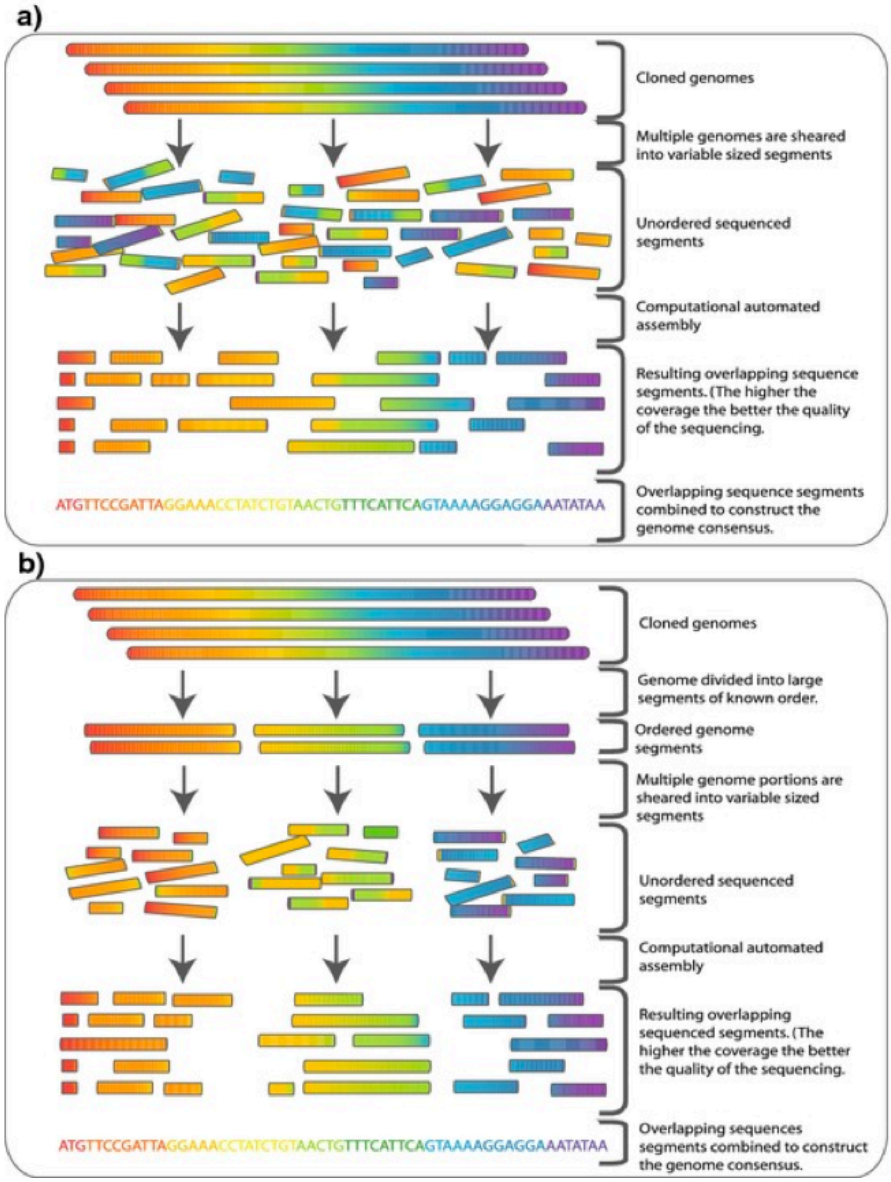


Fig. 66. – *Shotgun sequencing* : génome complet (A) et hiérarchique (B). Source: wikipédia

Bases de données

Records with assertion criteria	3310698
Total genes represented	92082
Unique variation records	2425225
Unique variation records with interpretations	2414095
Unique variation records with assertion criteria	2309591
Unique variation records with practice guidelines (4 stars)	663
Unique variation records from expert panels (3 stars)	15598
Unique variation records with assertion criteria, multiple submitters, and no conflicts (2 stars)	345874
Unique variation records with assertion criteria (1 star)	1841230
Unique variation records with assertion criteria and a conflict (1 star)	106226
Unique variation records with conflicting interpretations	106544
Genes with variants specific to one protein-coding gene	17210
Genes included in a variant spanning more than one gene	92411
Variants affecting overlapping genes	35523
Total submitters	2704

Tableau 20. – Statistiques clinvar au 17 décembre 2023

MIM Number Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
Gene description *	16,309	769	51	37	17,166
Gene and phenotype, combined	21	0	0	0	21
Phenotype description (molecular basis known)	6,344	383	5	34	6,766
Phenotype description or locus (molecular basis unknown)	1,390	111	4	0	1,505

Other (mainly phenotypes) (with suspected mendelian basis)	1,640	100	3	0	1,743
Totals	25,704	1,363	63	71	27,201

Tableau 21. – Statitisques OMIM au 19 décembre 2023

Annotation

Type	Détail
identification du gène affecté	identifiant Ensembl, nom « commun » du gène (ex HGNC)
identifiant du transcript	Ensemble, NCBI Refseq
identifiant pour CCDS	Consensus coding sequence
biotype selon GENCODE	codant pour protéine, pseudogène…)
coordonnées du variant	cDNA, processed coding sequence (CDS)
distance au transcrit	s'il est en dehors des bornes
conséquence sur transcrit	
nombre d'exons et introns touchés	
Transcript Support Level (TLS)	indique la fiabilité des modèles de transcrit
Annotation principle splice isoformes (APPRIS)	annotation de transcrit sur épissage alternatif (modèles informatique)

Tableau 22. – Annotation disponible dans VEP

Type	Detail
identifiant protéique	Ensembl, Refseq, UniProt (généré automatiquement, nettoyé la main ou combiné)
coordonées protéique	
codon	de référence et alternatif
acides aminés	de référence et alternatif
score	SIFT et PolyPhen2 prédictif de pathogénicitié
domaines protéiques	
notation HGVS	

Tableau 23. – Annotation protéique pour VEP

Type	Detail
Avertissement	
identifiant du gène	ENSEMBLE (principalement)
nom du gène	ENSEMBLE (principalement)
biotype	ENSEMBLE
Identifiant du transcript	ENSEMBLE (principalement)
Identifiant de l'exon	ENSEMBLE (principalement)
numéro de l'exon	sur le transcrit
impact du variant	
changement d'acide aminé	
changement de codon	
numéro du codon en CDS	
taille du CDS	
redondance du codon	

Tableau 24. – Annotation disponible dans SNPeff selon Cingolani et al. (2012)

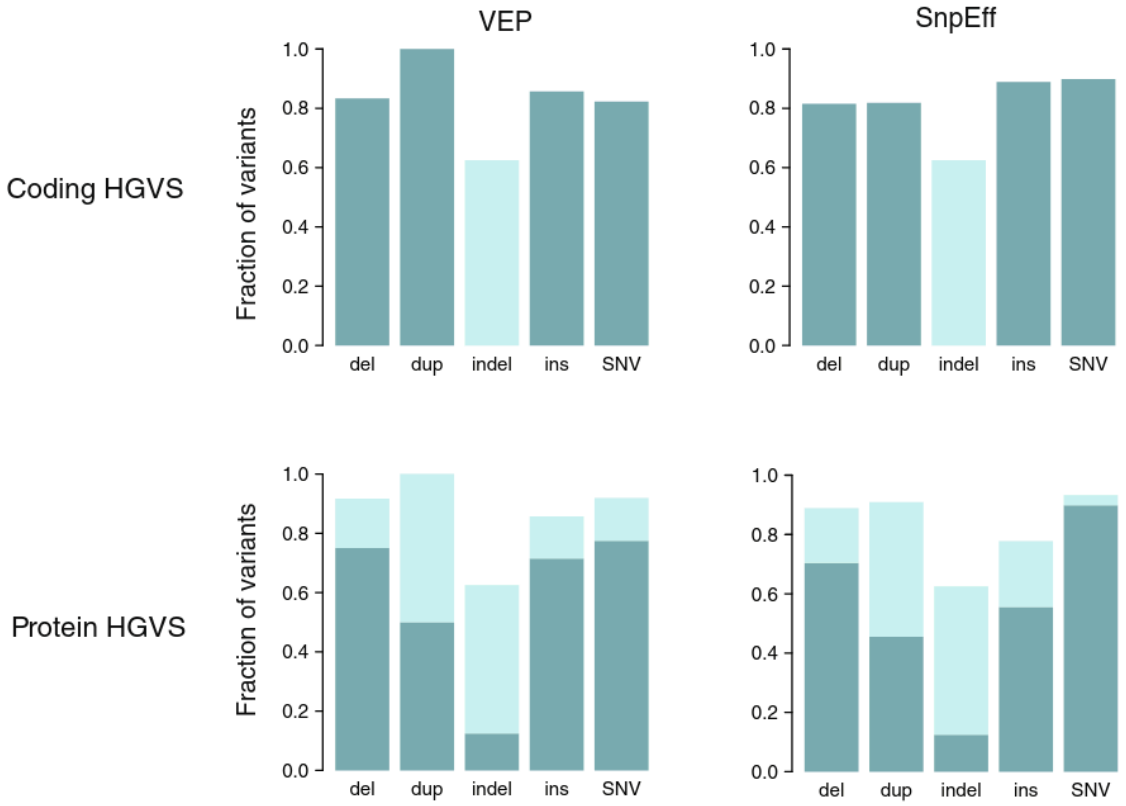


Fig. 67. – Comparaison de la précision sur 126 variants de référence selon de type de variants selon (Yen et al. 2017)

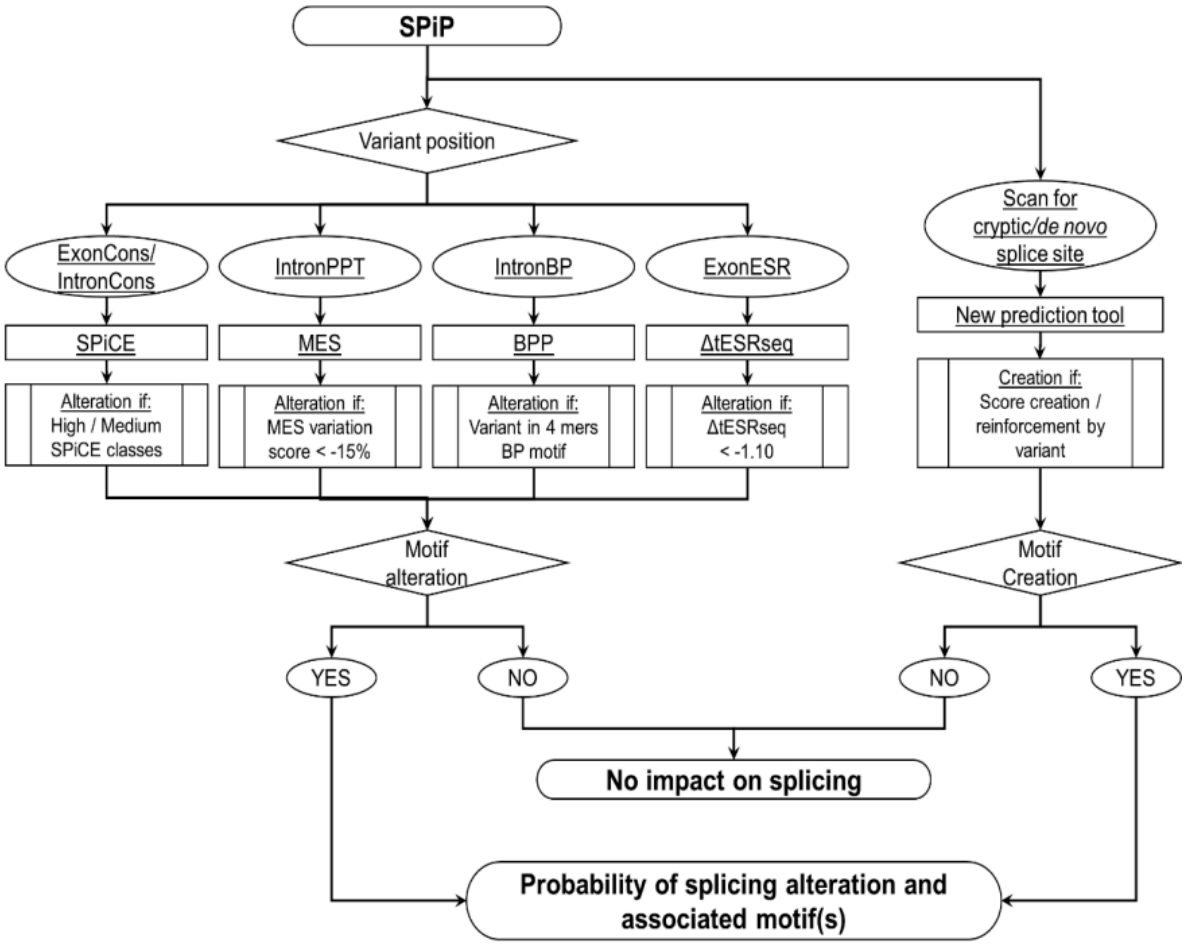


Fig. 68. – En cas de score SPiP supérieur à un seuil, le choix de motif est fait avec des scores pré-[existants ou un nouveau métascore développé par les auteurs (Leman et al. 2022)]

Reproductibilité, portabilité

Workflow	Nextflow	Galaxy	Toil	Snakemake	Bpipe
Platform	Groovy/JVM	Python	Python	Python	Groovy/JVM
Native task support	✓ (any)			✓ (BASH)	✓ (BASH)
Common workflow language		✓	✓		

Streaming processing	✓				
Dynamic branch evaluation	✓	?	✓	✓	Undocumented
Code sharing integration	✓				
Workflow modules		✓	✓	✓	✓
Workflow versioning	✓	✓			
Automatic error failover	✓		✓		
Graphical user interface		✓			
DAG rendering	✓	✓	✓	✓	✓
Container management					
Docker support	✓	✓	✓		
Singularity support	✓				
Multi-scale containers	✓	✓	✓		
Built-in batch schedulers					
Univa Grid Engine	✓	✓	✓	Partial	✓
PBS/Torque	✓	✓		Partial	✓
LSF	✓	✓		Partial	✓
SLURM	✓	✓	✓	Partial	
HTCondor	✓	✓		Partial	
Built-in distributed cluster					
Apache Ignite	✓				
Apache Spark			✓		
Kubernetes	✓				
Apache Mesos			✓		
Built-in cloud					

AWS (Amazon Web Services)	✓	✓	✓		

Tableau 25. – Comparaison de Nextflow par rapport à d'autres outils pour définir un workflow (Di Tommaso et al. 2017)

Puissance total (Tflops)	205.64 TFlops
Puissance CPU (Tflops)	63.55
Puissance GPU (TFlops)	142.09
Consommation électrique (kW)	15.20
Processseurs	Intel SandyBridge, Haswell, Cascade Lake
Carte graphique	K40, V100, A100
Réseau	40G/s (infinband), Réseau omnipath 100G/s (Omnipath)

Tableau 26. – Ressources du mésocentre de Franche-Compte

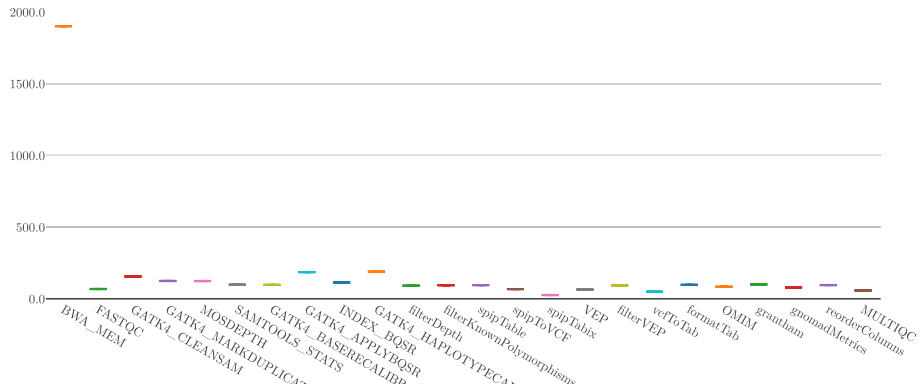


Fig. 69. – Temps de calcul en secondes processeur sur une exécution du pipeline pour chaque processus.

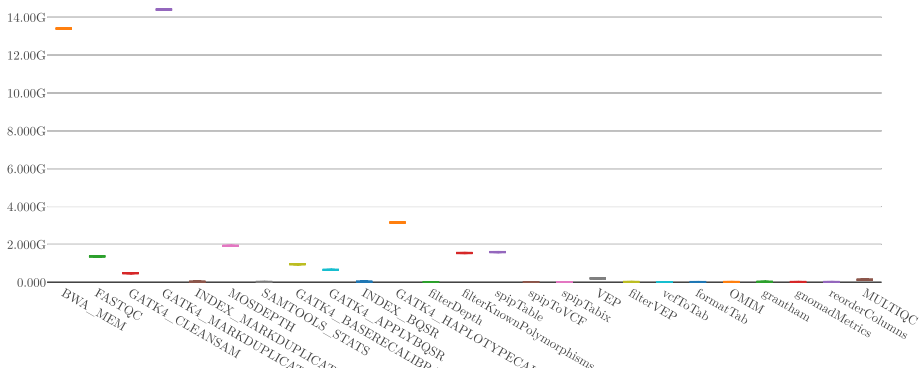


Fig. 70. – Coût en mémoire en gigaoctets d'une exécution du pipeline par processus

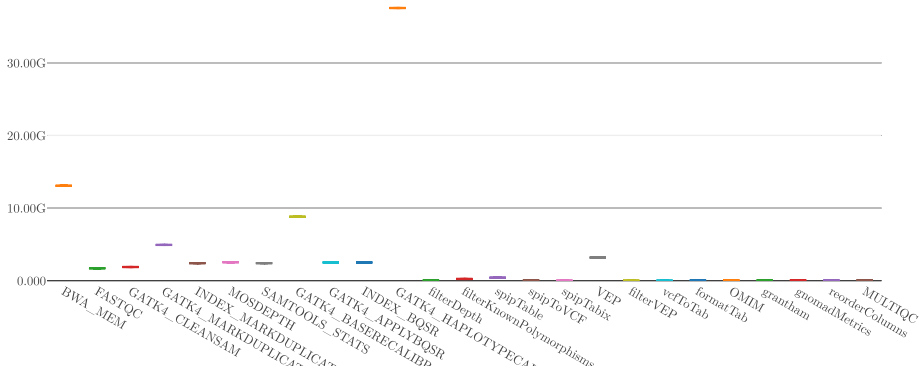


Fig. 71. – Nombre d'octets écrits pour une exécution du pipeline par processus

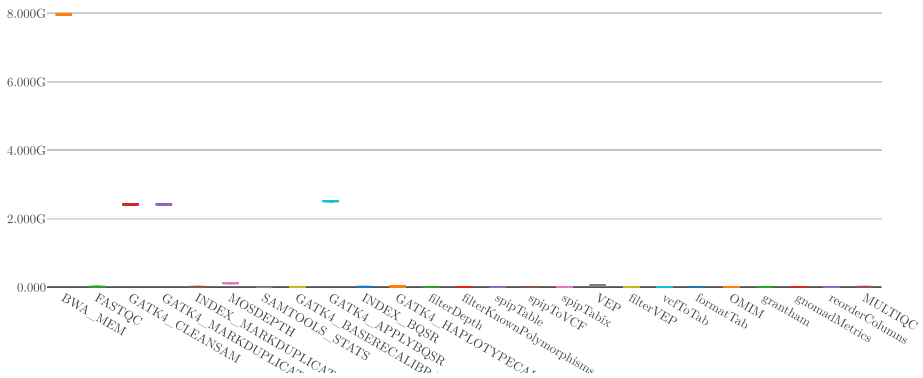


Fig. 72. – Nombre d'octets lus pour une exécution du pipeline par processus

Recrutement

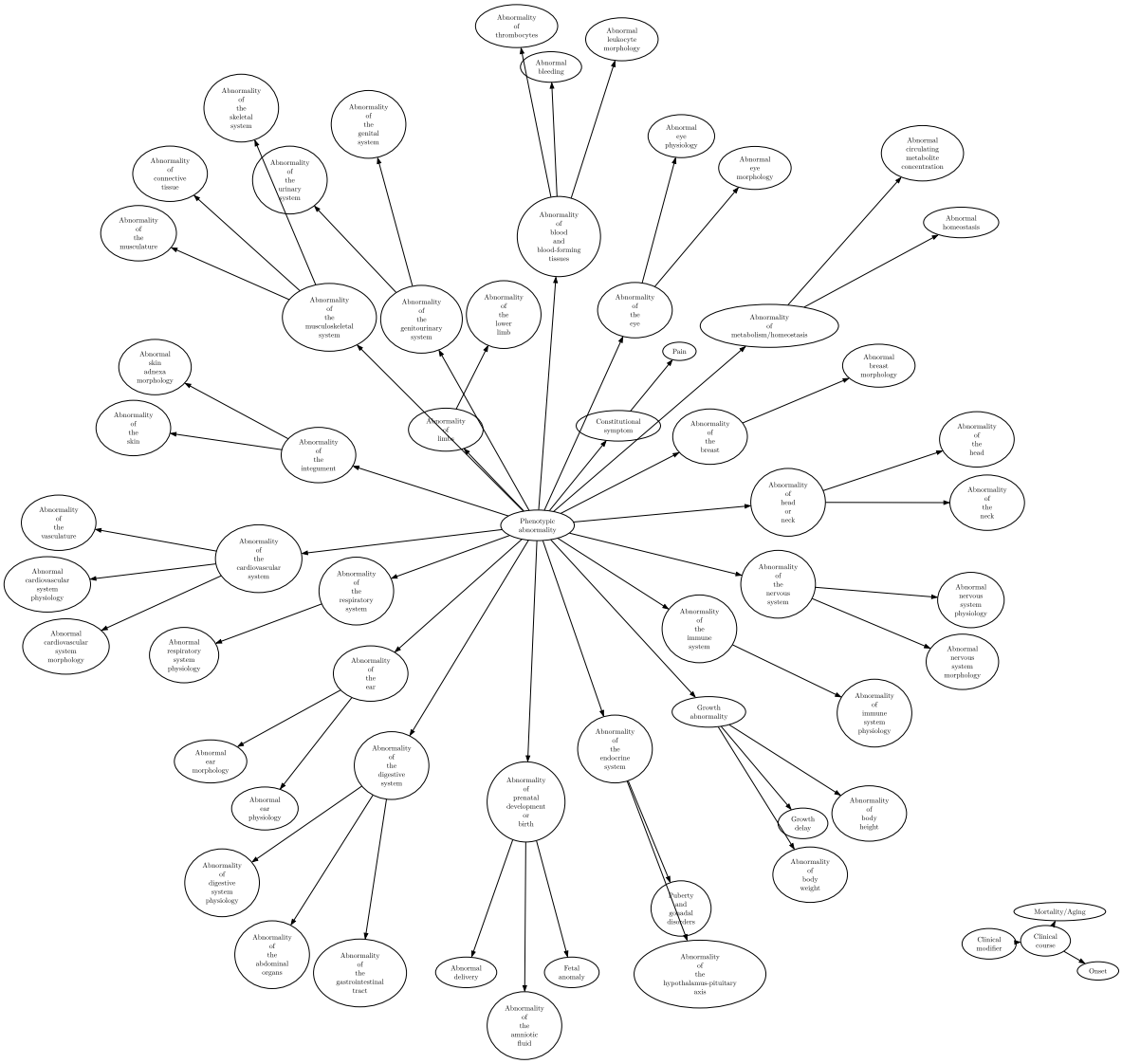


Fig. 73. – Catégories selon la nomenclature HPO les plus fréquentes dans la clinique extraite des compte-rendus pour les patients considérés.

Table des matières

UFR SCIENCES DE LA SANTÉ
UNIVERSITÉ DE FRANCHE-COMTÉ

RÉSUMÉ

_____

Nom – Prénom : PRAGA Alexis

Thèse soutenue le : 18/04/2024

Titre de la thèse : Un pipeline bioinformatique de ré-interprétation d'analyses constitutionnelles d'exome

Résumé :

Au sein du service de consultation de génétique constitutionnelle de Besançon, l'exome est un examen couramment prescrit afin de déterminer une cause génétique aux symptômes des patients. Cependant, près de 50% des résultats sont négatifs. S'agissant d'un examen sous-traité, il existe donc un besoin pour ré-analyser les données brutes dans un contexte diagnostique et de recherche. Cette thèse propose donc un pipeline bio-informatique pour répondre à cette problématique.

Les apports de cette thèse consistent, sur le plan informatique, à proposer un pipeline reproductible à 100% en exploitant l'outil Nix. Il est également facilement portable sous différentes architectures grâce à l'outil Nextflow et ses performances ont été ainsi testées sur le supercalculateur de Franche-Comté. Sur le plan biologique, nous avons procédé à une validation à l'aide d'un échantillon de référence, de données informatiques de référence, et de deux approches purement informatiques en utilisant un patient « de synthèse » et des données générées in silico. Sur le plan médical, nous avons examiné sa non-infériorité par rapport au laboratoire sous-traitant et des résultats préliminaires de la ré-interprétation sont présentés.

Mots clés : génétique,pipeline,réinterprétation,constitutionnel