Caractérisation des variations faux-sens à effet non-haploinsuffisant dans les maladies rares grâce à l'agrégation de données d'exome

François Lecoquierre

UNIVERSITÉ DE ROUEN
UFR DE MÉDECINE ET DE PHARMACIE

THÈSE
POUR LE DIPLÔME D'ÉTAT DE DOCTEUR EN MÉDECINE

Présentée et soutenue publiquement le 22 octobre 2018

par

François LECOQUIERRE

Né le 06 novembre 1989 à Ste Adresse (76)

Caractérisation des variations faux-sens à effet non-haploinsuffisant dans les maladies rares grâce à l'agrégation de données d'exome

Président du Jury	Monsieur le Professeur Thierry FREBOURG
Directeur de Thèse	Monsieur le Docteur Gaël NICOLAS
Membres du Jury	Madame le Professeur Christel THAUVIN-ROBINET
	Madame le Docteur Pascale SAUGIER-VEBER
	Monsieur le Docteur Dominique CAMPION

ANNÉE UNIVERSITAIRE 2017-2018

U.F.R. DE MÉDECINE ET DE PHARMACIE DE ROUEN

DOYEN : **Professeur Pierre FREGER**

ASSESSEURS : **Professeur Michel GUERBET**

Professeur Benoit VEBER

Professeur Guillaume SAVOYE

I - **MÉDECINE**

PROFESSEURS DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mr Frédéric **ANSELME**	HCN	Cardiologie
Mme Isabelle **AUQUIT AUCKBUR**	HCN	Chirurgie plastique
Mme Gisèle **APTER**	Havre	Pédopsychiatrie
Mr Fabrice **BAUER**	HCN	Cardiologie
Mme Soumeya **BEKRI**	HCN	Biochimie et biologie moléculaire
Mr Ygal **BENHAMOU**	HCN	Médecine interne
Mr Jacques **BENICHOU**	HCN	Bio statistiques et informatique médicale
Mme Bouchra **LAMIA**	Havre	Pneumologie
Mr Olivier **BOYER**	UFR	Immunologie
Mme Sophie **CANDON**	HCN	Immunologie
Mr François **CARON**	HCN	Maladies infectieuses et tropicales
Mr Philippe **CHASSAGNE** *(détachement)*	HCN	Médecine interne (gériatrie)
Mr Vincent **COMPERE**	HCN	Anesthésiologie et réanimation chirurgicale
Mr Jean-Nicolas **CORNU**	HCN	Urologie
Mr Antoine **CUVELIER**	HB	Pneumologie
Mr Pierre **CZERNICHOW** *(surnombre)*	HCH	Épidémiologie, économie de la santé
Mr Jean-Nicolas **DACHER**	HCN	Radiologie et imagerie médicale
Mr Stéfan **DARMONI**	HCN	Informatique médicale

Mr Pierre **DECHELOTTE**	HCN	Nutrition
Mr Stéphane **DERREY**	HCN	Neurochirurgie
Mr Frédéric **DI FIORE**	CB	Cancérologie
Mr Fabien **DOGUET**	HCN	Chirurgie Cardio Vasculaire
Mr Jean **DOUCET**	SJ	Thérapeutique - Médecine interne et gériatrie
Mr Bernard **DUBRAY**	CB	Radiothérapie
Mr Philippe **DUCROTTE**	HCN	Hépato-gastro-entérologie
Mr Frank **DUJARDIN**	HCN	Chirurgie orthopédique - Traumatologique
Mr Fabrice **DUPARC**	HCN	Anatomie - Chirurgie orthopédique
Mr Éric **DURAND**	HCN	Cardiologie
Mr Bertrand **DUREUIL**	HCN	Anesthésiologie et réanimation chirurgicale
Mme Hélène **ELTCHANINOFF**	HCN	Cardiologie
Mr Manuel **ETIENNE**	HCN	Maladies infectieuses et tropicales
Mr Thierry **FREBOURG**	UFR	Génétique
Mr Pierre **FREGER**	HCN	Anatomie - Neurochirurgie
Mr Jean François **GEHANNO**	HCN	Médecine et santé au travail
Mr Emmanuel **GERARDIN**	HCN	Imagerie médicale
Mme Priscille **GERARDIN**	HCN	Pédopsychiatrie
Mr Michel **GODIN** *(surnombre)*	HB	Néphrologie
M. Guillaume **GOURCEROL**	HCN	Physiologie
Mr Dominique **GUERROT**	HCN	Néphrologie
Mr Olivier **GUILLIN**	HCN	Psychiatrie Adultes
Mr Didier **HANNEQUIN**	HCN	Neurologie
Mr Fabrice **JARDIN**	CB	Hématologie
Mr Luc-Marie **JOLY**	HCN	Médecine d'urgence
Mr Pascal **JOLY**	HCN	Dermato - Vénéréologie
Mme Annie **LAQUERRIERE**	HCN	Anatomie et cytologie pathologiques
Mr Vincent **LAUDENBACH**	HCN	Anesthésie et réanimation chirurgicale
Mr Joël **LECHEVALLIER**	HCN	Chirurgie infantile
Mr Hervé **LEFEBVRE**	HB	Endocrinologie et maladies métaboliques
Mr Thierry **LEQUERRE**	HB	Rhumatologie
Mme Anne-Marie **LEROI**	HCN	Physiologie

Mr Hervé **LEVESQUE**	HB	Médecine interne
Mme Agnès **LIARD-ZMUDA**	HCN	Chirurgie Infantile
Mr Pierre Yves **LITZLER**	HCN	Chirurgie cardiaque
Mr Bertrand **MACE**	HCN	Histologie, embryologie, cytogénétique
M. David **MALTETE**	HCN	Neurologie
Mr Christophe **MARGUET**	HCN	Pédiatrie
Mme Isabelle **MARIE**	HB	Médecine interne
Mr Jean-Paul **MARIE**	HCN	Oto-rhino-laryngologie
Mr Loïc **MARPEAU**	HCN	Gynécologie - Obstétrique
Mr Stéphane **MARRET**	HCN	Pédiatrie
Mme Véronique **MERLE**	HCN	Épidémiologie
Mr Pierre **MICHEL**	HCN	Hépato-gastro-entérologie
M. Benoit **MISSET**	HCN	Réanimation Médicale
Mr Jean-François **MUIR** *(surnombre)*	HB	Pneumologie
Mr Marc **MURAINE**	HCN	Ophtalmologie
Mr Philippe **MUSETTE**	HCN	Dermatologie - Vénéréologie
Mr Christophe **PEILLON**	HCN	Chirurgie générale
Mr Christian **PFISTER**	HCN	Urologie
Mr Jean-Christophe **PLANTIER**	HCN	Bactériologie - Virologie
Mr Didier **PLISSONNIER**	HCN	Chirurgie vasculaire
Mr Gaëtan **PREVOST**	HCN	Endocrinologie
Mr Bernard **PROUST**	HCN	Médecine légale
Mr Jean-Christophe **RICHARD**	HCN	Réanimation médicale - Médecine d'urgence
Mr Vincent **RICHARD**	UFR	Pharmacologie
Mme Nathalie **RIVES**	HCN	Biologie de la reproduction
Mr Horace **ROMAN**	HCN	Gynécologie - Obstétrique
Mr Jean-Christophe **SABOURIN**	HCN	Anatomie - Pathologie
Mr Guillaume **SAVOYE**	HCN	Hépato-gastrologie
Mme Céline **SAVOYE–COLLET**	HCN	Imagerie médicale
Mme Pascale **SCHNEIDER**	HCN	Pédiatrie
Mr Michel **SCOTTE**	HCN	Chirurgie digestive
Mme Fabienne **TAMION**	HCN	Thérapeutique

Mr Luc **THIBERVILLE**	HCN	Pneumologie
Mr Christian **THUILLEZ** *(surnombre)*	HB	Pharmacologie
Mr Hervé **TILLY**	CB	Hématologie et transfusion
M. Gilles **TOURNEL**	HCN	Médecine Légale
Mr Olivier **TROST**	HCN	Chirurgie Maxillo-Faciale
Mr Jean-Jacques **TUECH**	HCN	Chirurgie digestive
Mr Jean-Pierre **VANNIER** *(surnombre)*	HCN	Pédiatrie génétique
Mr Benoît **VEBER**	HCN	Anesthésiologie - Réanimation chirurgicale
Mr Pierre **VERA**	CB	Biophysique et traitement de l'image
Mr Eric **VERIN**	HB	Service Santé Réadaptation
Mr Eric **VERSPYCK**	HCN	Gynécologie obstétrique
Mr Olivier **VITTECOQ**	HB	Rhumatologie
Mme Marie-Laure **WELTER**	HCN	Physiologie

MAITRES DE CONFERENCES DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mme Noëlle **BARBIER-FREBOURG**	HCN	Bactériologie – Virologie
Mme Carole **BRASSE LAGNEL**	HCN	Biochimie
Mme Valérie **BRIDOUX HUYBRECHTS**	HCN	Chirurgie Vasculaire
Mr Gérard **BUCHONNET**	HCN	Hématologie
Mme Mireille **CASTANET**	HCN	Pédiatrie
Mme Nathalie **CHASTAN**	HCN	Neurophysiologie
Mme Sophie **CLAEYSSENS**	HCN	Biochimie et biologie moléculaire
Mr Moïse **COEFFIER**	HCN	Nutrition
Mr Serge **JACQUOT**	UFR	Immunologie
Mr Joël **LADNER**	HCN	Épidémiologie, économie de la santé
Mr Jean-Baptiste **LATOUCHE**	UFR	Biologie cellulaire
Mr Thomas **MOUREZ**	HCN	Virologie
Mr Gaël **NICOLAS**	HCN	Génétique
Mme Muriel **QUILLARD**	HCN	Biochimie et biologie moléculaire
Mme Laëtitia **ROLLIN**	HCN	Médecine du Travail
Mr Mathieu **SALAUN**	HCN	Pneumologie

Mme Pascale **SAUGIER-VEBER**	HCN	Génétique
Mme Anne-Claire **TOBENAS-DUJARDIN**	HCN	Anatomie
Mr David **WALLON**	HCN	Neurologie

PROFESSEUR AGREGE OU CERTIFIE

Mme Mélanie **AUVRAY-HAMEL**	UFR	Anglais
Mr Thierry **WABLE**	UFR	Communication

I - PHARMACIE

PROFESSEURS

Mr Thierry **BESSON**	Chimie Thérapeutique
Mr Roland **CAPRON** (PU-PH)	Biophysique
Mr Jean **COSTENTIN** (Professeur émérite)	Pharmacologie
Mme Isabelle **DUBUS**	Biochimie
Mr François **ESTOUR**	Chimie Organique
Mr Loïc **FAVENNEC** (PU-PH)	Parasitologie
Mr Jean Pierre **GOULLE** (Professeur émérite)	Toxicologie
Mr Michel **GUERBET**	Toxicologie
Mme Isabelle **LEROUX - NICOLLET**	Physiologie
Mme Christelle **MONTEIL**	Toxicologie
Mme Martine **PESTEL-CARON** (PU-PH)	Microbiologie
Mme Élisabeth **SEGUIN**	Pharmacognosie
Mr Rémi **VARIN** (PU-PH)	Pharmacie clinique
Mr Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

MAITRES DE CONFERENCES

Mme Cécile **BARBOT**	Chimie Générale et Minérale
Mr Jérémy **BELLIEN** (MCU-PH)	Pharmacologie
Mr Frédéric **BOUNOURE**	Pharmacie Galénique
Mr Abdeslam **CHAGRAOUI**	Physiologie
Mme Camille **CHARBONNIER (LE CLEZIO)**	Statistiques
Mme Elizabeth **CHOSSON**	Botanique
Mme Marie Catherine **CONCE-CHEMTOB**	Législation pharmaceutique
Mme Cécile **CORBIERE**	Biochimie
Mr Éric **DITTMAR**	Biophysique
Mme Nathalie **DOURMAP**	Pharmacologie
Mme Isabelle **DUBUC**	Pharmacologie
Mme Dominique **DUTERTE- BOUCHER**	Pharmacologie

Mr Abdelhakim **ELOMRI**	Pharmacognosie
Mr François **ESTOUR**	Chimie Organique
Mr Gilles **GARGALA** (MCU-PH)	Parasitologie
Mme Nejla EL **GHARBI-HAMZA**	Chimie analytique
Mme Marie-Laure **GROULT**	Botanique
Mr Hervé **HUE**	Biophysique et mathématiques
Mme Laetitia **LE GOFF**	Parasitologie – Immunologie
Mme Hong **LU**	Biologie
Mme Marine **MALLETER**	Toxicologie
M. Jérémie **MARTINET** (MCU-PH)	Immunologie
Mme Sabine **MENAGER**	Chimie organique
Mme Tiphaine **ROGEZ-FLORENT**	Chimie analytique
Mr Mohamed **SKIBA**	Pharmacie galénique
Mme Malika **SKIBA**	Pharmacie galénique
Mme Christine **THARASSE**	Chimie thérapeutique
Mr Frédéric **ZIEGLER**	Biochimie

PROFESSEURS ASSOCIES

Mme Cécile **GUERARD-DETUNCQ**	Pharmacie officinale
Mr Jean-François **HOUIVET**	Pharmacie officinale

PROFESSEUR CERTIFIE

Mme Mathilde **GUERIN**	Anglais

ASSISTANT HOSPITALO-UNIVERSITAIRE

Mme Anaïs **SOAREZ**	Bactériologie

ATTACHES TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE

Mme Anne-Sophie **CHAMPY**	Pharmacognosie
M. Jonathan **HEDOUIN**	Chimie Organique
Mme Barbara **LAMY-PELLETER**	Pharmacie Galénique

LISTE DES RESPONSABLES DES DISCIPLINES PHARMACEUTIQUES

Mme Cécile **BARBOT**	Chimie Générale et minérale
Mr Thierry **BESSON**	Chimie thérapeutique
Mr Roland **CAPRON**	Biophysique
Mme Marie-Catherine **CONCE-CHEMTOB**	Législation et économie de la santé
Mme Élisabeth **CHOSSON**	Botanique
Mr Jean-Jacques **BONNET**	Pharmacodynamie
Mme Isabelle **DUBUS**	Biochimie
Mr Loïc **FAVENNEC**	Parasitologie
Mr Michel **GUERBET**	Toxicologie
Mr François **ESTOUR**	Chimie organique
Mme Isabelle **LEROUX-NICOLLET**	Physiologie
Mme Martine **PESTEL-CARON**	Microbiologie
Mme Élisabeth **SEGUIN**	Pharmacognosie
Mr Mohamed **SKIBA**	Pharmacie galénique
Mr Rémi **VARIN**	Pharmacie clinique
M. Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

III – MEDECINE GENERALE

PROFESSEUR

Mr Jean-Loup **HERMIL**	UFR	Médecine générale

MAITRE DE CONFERENCE MEDECINE GENERALE

Mr Matthieu **SCHUERS** (MCU-MG)	UFR	Médecine générale

PROFESSEURS ASSOCIES A MI-TEMPS

Mr Emmanuel **LEFEBVRE**	UFR	Médecine Générale
Mme Élisabeth **MAUVIARD**	UFR	Médecine générale
Mr Philippe **NGUYEN THANH**	UFR	Médecine générale
Mme Marie Thérèse **THUEUX**	UFR	Médecine générale

MAITRE DE CONFERENCES ASSOCIE A MI-TEMPS

Mr Pascal **BOULET**	UFR	Médecine générale
Mr Emmanuel **HAZARD**	UFR	Médecine Générale
Mme Marianne **LAINE**	UFR	Médecine Générale
Mme Lucile **PELLERIN**	UFR	Médecine générale
Mme Yveline **SEVRIN**	UFR	Médecine générale

MONO-APPARTENANTS

PROFESSEURS

Mr Serguei **FETISSOV** (med)	Physiologie (ADEN)
Mr Paul **MULDER** (phar)	Sciences du Médicament
Mme Su **RUAN** (med)	Génie Informatique

MAITRES DE CONFERENCES

Mr Sahil **ADRIOUCH** (med)	Biochimie et biologie moléculaire (Unité Inserm905)
Mme Gaëlle **BOUGEARD-DENOYELLE** (med)	Biochimie et biologie moléculaire (UMR 1079**)**
Mme Carine **CLEREN** (med)	Neurosciences (Néovasc)
M. Sylvain **FRAINEAU** (phar)	Physiologie (Inserm U 1096)
Mme Pascaline **GAILDRAT** (med)	Génétique moléculaire humaine (UMR 1079)
Mr Nicolas **GUEROUT** (med)	Chirurgie Expérimentale
Mme Rachel **LETELLIER** (med)	Physiologie
Mme Christine **RONDANINO** (med)	Physiologie de la reproduction
Mr Antoine **OUVRARD-PASCAUD** (med)	Physiologie (Unité Inserm 1076)
Mr Frédéric **PASQUET**	Sciences du langage, orthophonie
Mme Isabelle **TOURNIER** (med)	Biochimie (UMR 1079)

CHEF DES SERVICES ADMINISTRATIFS : Mme Véronique **DELAFONTAINE**

HCN - Hôpital Charles Nicolle

HB - Hôpital de BOIS GUILLAUME

CB - Centre Henri Becquerel

CHS - Centre Hospitalier Spécialisé du Rouvray

CRMPR - Centre Régional de Médecine Physique et de Réadaptation

SJ – Saint Julien Rouen

Par délibération en date du 3 mars 1967, la faculté a arrêté que les opinions émises dans les dissertations qui lui seront présentées doivent être considérées comme propres à leurs auteurs et qu'elle n'entend leur donner aucune approbation ni improbation.

REMERCIEMENTS

Aux membres du jury

À Monsieur le professeur Thierry Frébourg, pour l'enthousiasme de vos cours de PCEM1 (en 2007, avant même la *vision panoramique* apportée par le NGS), qui ont probablement joué un rôle décisif pour moi. Merci également pour votre soutien constant au cours de ces cinq années d'internat.

À Madame le professeur Christel Thauvin, pour accepter de juger ce travail. Merci pour votre accueil au sein de l'équipe Gad lors de cette excursion dijonnaise qui fut passionnante.

À Madame le Docteur Pascale Saugier-Veber, pour m'avoir enseigné la génétique moléculaire, et pour votre disponibilité au quotidien. Merci de ne pas tenir compte des quelques anglicismes qui persistent dans ce manuscrit, certains mots peuvent être *challenging* à traduire de manière convenable :)

À Monsieur le Docteur Dominique Campion, pour votre lecture de cette thèse et votre sympathie. Comme vous pourrez le remarquer, j'ai changé d'avis et je ne pense plus que le séquençage des régions codantes est *has been* :)

À Monsieur le Docteur Gaël Nicolas, mon directeur de thèse. Je me rappellerai toujours de notre rencontre dans le service de génétique de l'anneau central, j'avais été d'emblée très impressionné ! Merci infiniment pour ton aide durant ces années d'internat.

À l'équipe de génétique de Rouen

Je remercie toute l'équipe de génétique clinique qui m'a accueilli alors que je n'étais qu'un bébé. Merci **Alice** pour tes encouragements en fin d'externat, qui ont été très utiles pour moi. **Anne-Marie** et **Anne-Claire**, votre sympathie est précieuse. Merci **Géraldine** et **Pascal**, ainsi que la fabuleuse équipe des **techniciens de cytogénétique** pour m'avoir appris l'art de reconnaitre les chromosomes (et pour les stocks de chocolat dans la salle de pause).

De l'autre côté de la rue, j'ai été sympathiquement accueilli par les deux équipes de **techniciennes de neuro et d'onco**, que je remercie toutes *sans exception*. **Stéphanie**, merci pour l'agréable semestre d'été que j'ai passé en onco, où pour la première fois je me suis senti utile et où tu m'as permis d'approfondir ma connaissance du NGS. Merci aux **bioinformaticiens/biostatisticiens** et au reste de l'équipe diagnostique : **biologistes, secrétaires, ingénieurs et cadres**. **Pierre**, j'ai été ravi de partager ton bureau pendant 6 mois, surtout les jours ou tu ne mettais pas la clim à 17°C.

Je remercie vivement **l'équipe de recherche de l'unité U1245**, et en particulier **Anne**, qui m'a transmis sa passion pour le *wersten-blot* et pour le dosage des protéines. Mon seul regret restera de ne pas avoir élevé mes propres mouches avec **Magali** et l'équipe droso. Merci à **Marine** pour les balades nocturnes et au reste des thésards pour leur gentillesse. **Camille, Laetitia, Thomas, Sabrina, Asli et Marion**, l'année en votre compagnie était très plaisante, même si vous étiez là *tous les jours*.

Enfin, un immense merci à mes co-internes en or : **Maud**, qui a égayé une grande partie de mon internat et dont la sympathie n'a d'égale que sa fragilité osseuse, **Kévin**, qui n'a pas fait de *kévinade* depuis un peu trop longtemps à mon goût, **Gabriella**, **Vincent** et leur plaid violet, et les petits nouveaux **Ferdi** et **Juliette**, qui commencent à s'y connaître pas mal (en baby-foot).

À l'équipe dijonnaise

Je souhaite remercier chaleureusement l'équipe dijonnaise pour leur accueil et pour les connaissances que qu'ils m'ont apportées durant six mois. Merci aux biologistes et médecins, dont **Philippe**, **Fred** et **Arthur**, à **Yannis** pour son expertise bioinformatique, aux **techniciens et chercheurs** qui forment une équipe soudée et terriblement efficace. Merci à **Maud** et à **Benoit** pour avoir rendu ce bureau si plaisant, et aux **internes de biologie** dont la rencontre a été une chance. Merci à **Laurence Faivre**, dont la confiance m'honore. Je garde un souvenir nostalgique du soleil couchant sur la Bourgogne depuis le *rooftop* du laboratoire.

À mes amis

Avant de sortir de la génétique, je salue mes compatriotes français de la SIGF, en particulier **Sophie**, **Florence** et **Benjamin**, **Mathias** et **Afane** qui ont partagé mon bureau, et les autres, que j'ai beaucoup de plaisir à revoir régulièrement. Je remercie mes amis de l'externat et de l'internat, avec comme échantillon : **Édouard et Laura**, **Bruno et Caro**, **Thomas et Marine**, **Maud et Philippe**, **Boris et Amina**, **Jésus et Poca**, puis les rockstars **Julien**, **Aurélien**, **Benoit et Camille**.

Un petit message pour **les skaters de Rouen** : mon métier n'a rien à voir avec les dinosaures.

Merci aux chers *BDMs* (en particulier pour m'avoir supporté en Corse alors que j'avais le lancer de Cornetto un peu facile) : **Pierre et Margaux**, **Lucie**, **Pierre**, **Nikeu**, **Joachim**, **Magali et JB**, **Gabriel**, **Vincent et Sarah**, **Maxime**, **Romain**, **Rana**, **Sophie**, **Nico** et **Rozenn**, et **Marine**.

À ma famille que j'aime

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS

ACMG : The American College of Medical Genetics and Genomics
ACSG : Association for Clinical Genetic Science
ADN : Acide désoxyribonucléique
AMP : Association for Molecular Pathology
ANPGM : Association Nationale des Praticiens de Génétique Moléculaire
ARN : Acide ribonucléique
BAM : Binary alignment map
CADD : Combined Annotation Dependent Depletion
CIM : Classification internationale des maladies
CNV : Copy number variation
dbNSFP : Database for nonsynonymous SNPs' functional predictions
DDD : Deciphering Developmental Disorders
DDG2P : The Developmental Disorders Genotype-Phenotype Database
DM : Disease-causing mutation
DNM : De novo mutation
DNN : Deep neural network
DT : Decision tree
EBI : The European Bioinformatics Institute
ESP : Exome Sequencing Project
ExAC : The Exome Aggregation Consortium
FREX : The French Exome Project
GATK : Genome Analysis Tool Kit
GnomAD : Genome Aggregation Database
GO : Gene Ontology
GoNL : Genome of the Netherlands
GTEx : Genotype-Tissue Expression
HGMD : Human Gene Mutation Database
HGNC : HUGO Gene Nomenclature Committee
HI : Haploinsuffisant
HMM : Hidden Markov model
HPO : Human phenotype ontology
HTC : HaplotypeCaller
HUGO : Human Genome Organisation
Indel : Insertion/déletion
LGD : Likely gene disrupting
LOF : Loss-of-function
LOVD : Leiden Open Variation Database
MAF : Minor allele frequency
MKL : Multiple kernel learning
NBC : Naive Bayes Classifier
NCBI : National Center for Biotechnology Information
NGS : Next generation sequencing
NHI : Non haploinsuffisant
NHLBI : National Heart, Lung, and Blood Institute
NMD : Nonsense-mediated decay
NN : Neural network
nsSNV : Non-synonymouys single nucleotide variant
OMIM : Online Mendelian Inheritance in Man
OOA : Out of Africa
pLI : Probability of loss-of-function intolerance
PMID : PubMed ID
PSSM : Position-specific scoring matrix
PTC : Premature termination codon
PTV : Protein truncating variant
RF : Random forest
ROC : Receiver operating characteristic
SAM : Sequence Alignment Map
SIFT : Sorting Intolerant From Tolerant
SNV : Single nucleotide variant
SSC : The Simons Simplex Collection
SVM : Support vector machine
UCSC : University of California, Santa Cruz
UMD : Universal Mutation Database
VCF : Variant calling format
VEP : Variant Effect Predictor
WES : Whole exome sequencing
WGS : Whole genome sequencing

1 INTRODUCTION : LE DIAGNOSTIC GÉNÉTIQUE DES MALADIES RARES À L'ÈRE DU SÉQUENÇAGE À HAUT DÉBIT

L'apparition du séquençage à haut débit a été une étape importante dans l'histoire de la génétique. Grâce à cette nouvelle technologie, nous assistons actuellement à une transition de la génétique ou étude des gènes vers la génomique ou étude du génome. Dans le diagnostic génétique des maladies rares, le séquençage haut débit, ou NGS, pour « *next-generation sequencing* », a permis depuis le début des années 2010 des avancées considérables. Dans un premier temps utilisée majoritairement par la communauté scientifique, la puissance du NGS a permis l'identification d'innombrables nouvelles bases moléculaires à des maladies génétiques. Rapidement, la puissance du NGS a pu également être exploitée dans un cadre médical de diagnostic au bénéfice des patients, avec une augmentation des capacités diagnostiques et une diminution des délais. La détection des variations génétiques a été grandement facilitée, ce qui a permis de révéler l'étendue insoupçonnée du polymorphisme humain. La production de données massives de génomique a été à l'origine de problématiques nouvelles dans le traitement et l'interprétation des variations identifiées. L'abondance des données et leur traitement bio-informatique systématique ont propulsé la génomique au rang de *big data*, et de nombreux efforts pour organiser, analyser et partager ces données, ont été entrepris, dans le but d'en extraire des informations précieuses pour le diagnostic des patients atteints de maladies génétiques.

L'accès au NGS a donc révolutionné les capacités de détection de variations génomiques. En revanche, il a placé la communauté scientifique et médicale devant une difficulté bien plus grande encore : l'interprétation des variations génétiques. De manière naturelle, les efforts d'interprétation se sont focalisés sur les régions codantes du génome. Avec la montée en puissance actuelle du séquençage du génome complet individuel, la problématique complexe de l'interprétation des variations du génome non codant représente un défi grandissant. Cependant, les connaissances scientifiques actuelles semblent indiquer que la plupart des causes monogéniques restant à découvrir se situent néanmoins dans l'exome (ensemble des exons, ou portion codante de notre génome). Nous faisons l'hypothèse qu'en 2018, une grande proportion des gènes responsables de maladies monogéniques par un mécanisme de perte de fonction (ou mécanisme d'haploinsuffisance, correspondant à une diminution quantitative de la protéine mutée, le plus souvent par l'introduction d'un codon stop prématuré ou par une délétion totale ou partielle du gène) a déjà été décrite, et que les causes monogéniques restant à identifier reposent sur des mécanismes moléculaires plus subtils, plus diversifiés, et potentiellement plus rares. Certaines de ces variations particulières sur le plan fonctionnel, regroupées sous l'appellation de variations à effet non-haploinsuffisant, représentent le thème principal de ce travail. Identifier quelles variations, parmi les nombreuses variations changeant les séquences protéiques présentes chez tout un chacun, peuvent éventuellement être responsables d'une maladie monogénique reste encore à ce jour un défi quotidien dans les laboratoires de génétique moléculaire.

En introduction nous présenterons d'abord le concept de maladies rares monogéniques ou présumées comme telles et leurs caractéristiques générales. Nous introduirons ensuite la technique du séquençage haut débit et son intérêt dans le diagnostic des maladies monogéniques. Nous nous focaliserons ensuite sur les variations faux-sens, qui représentent un défi actuel dans l'interprétation des variations génétiques au sein des séquences codantes. Les multiples mécanismes d'effets biologiques des variations faux-sens, les différentes méthodes de prédiction *in silico* de pathogénicité, et leur distribution génomique en population générale et en pathologie seront abordés.

Dans l'idée que certaines variations faux-sens sont d'interprétation complexe, la deuxième partie de ce travail aura pour objectif de tenter d'identifier certaines propriétés de ces variations en utilisant les informations présentes dans des bases de données publiques.

Enfin, en troisième partie, nous proposerons une méthode simple d'identification de variations faux-sens pathogènes à effet non-haploinsuffisant, en se basant sur la récurrence de variations identiques entre plusieurs individus présentant un phénotype comparable.

1.1 Maladies rares et maladies génétiques

Dans cette partie introduisant le concept de maladies rares, nous décrirons les aspects épidémiologiques et les différentes classifications et bases de données, ce qui nous permettra d'évoquer ensuite la grande diversité des maladies génétiques rares.

1.1.1 Définition des maladies rares

Les maladies rares ont une définition variable en fonction des pays. En Europe, la définition admise est une maladie affectant moins d'un individu sur 2 000, alors qu'aux États-Unis par exemple, une maladie est définie comme rare lorsqu'elle concerne moins de 200 000 personnes. Si la grande majorité des maladies génétiques mendéliennes est constituée par les maladies rares, toutes les maladies rares ne sont pas génétiques. Aussi il est fréquemment mentionné sur des sources grand public que « 80 % des maladies rares sont d'origine génétique », mais cette affirmation ne semble pas avoir de fondement scientifique précis (http://www.thetgmi.org/genetics/how-many-rare-diseases-are-genetic/). Il n'en demeure pas moins que les affections génétiques mendéliennes représentent une cause majeure de maladies rares.

Le nombre de maladies rares est inconnu. Des chiffres de 5 000 à 8 000 maladies sont classiquement cités (notamment dans le rapport du 8 juin 2009 de la commission européenne se référant aux maladies rares). En réalité ce nombre dépend directement de la manière dont on définit une maladie, de la « résolution » de la définition. Par exemple on peut considérer le syndrome de Cornelia de Lange comme une entité à part entière (comme dans la base de données Orphanet (voir paragraphe 1.1.2.1) : entrée ORPHA199), ou bien être plus « résolutif » et différencier 5 maladies en fonction du gène causal, en argumentant que ces entités ont des subtilités cliniques, des modes de transmission différents, et peut-être un jour des traitements distincts (comme dans la base de données OMIM (voir paragraphe 1.1.2.1) : entrées #122470, #300590, #300882, #610759, #614701). Cet exemple souligne l'hétérogénéité de la classification des maladies rares, qui jadis

était majoritairement basée sur la distinction d'entités clinique, et qui se calque de plus en plus sur les causes moléculaires dans une classification d'entités « clinico-biologiques ».

1.1.2 Nosologie et ontologie phénotypique

Il existe plusieurs milliers de maladies génétiques avec bases moléculaires connues. Le nombre de gènes associés est du même ordre de grandeur, et le nombre de signes cliniques pouvant être identifiés est plus grand encore. Cette abondance de données est telle que leur accessibilité est sous-tendue par la nécessité d'une harmonisation, d'une organisation et d'une hiérarchisation des concepts. La classification des maladies, appelée « nosologie » doit dans le cadre des maladies rares être en interaction avec de nombreuses sources d'information, telles que les gènes et locus génétiques, des variations génétiques, des descriptions cliniques, etc. Nous présentons ici les deux grandes bases de données de maladies génétiques, Orphanet et OMIM, qui ont des modes de fonctionnement distincts et qui sont complémentaires. Ces deux outils serviront de base pour aborder d'une manière globale la diversité des présentations cliniques et familiales des maladies rares génétiques, ainsi que la complexité des relations génotype-phénotype.

1.1.2.1 Bases de données de phénotypes et maladies rares : Orphanet et OMIM

Il existe deux bases de données principales recensant les maladies rares : la plateforme européenne Orphanet[1] et le catalogue des maladies humaines mendéliennes OMIM (Online Mendelian Inheritance in Man), de l'université Josh Hopkins aux États-Unis[2]. De par les avancées rapides dans l'identification des maladies rares et de leurs causes moléculaires, ces bases de données sont mises à jour régulièrement. Cette réactivité, ainsi que les interactions qu'elles recensent entre les maladies et les gènes ou régions génomiques impliquées en font des outils infiniment plus adaptés que les terminologies médicales classiques telles que la classification internationale des maladies (CIM-10).

Orphanet : entrées cliniques

Orphanet a été fondée en France en 1997 dans le cadre de la Mission des médicaments orphelins du Ministère chargé de la santé. Cette plateforme a été confiée à l'Inserm, et a ainsi d'emblée été à l'intersection entre la santé publique et la recherche dans le champ des maladies rares[1]. Orphanet s'est vite transformée en une initiative européenne avec un rayonnement international. Actuellement Orphanet représente la principale classification des maladies rares. Les maladies y sont prioritairement définies et classées sur la base de la présentation clinique. Les causes génétiques sont également importantes dans la définition des maladies Orphanet, mais elles ne sont pas centrales comme dans la base OMIM. Orphanet est alimenté par une équipe dédiée qui effectue une veille scientifique basée sur la littérature et les avis d'experts et groupes de travail afin d'identifier de nouvelles entités clinico-biologiques. Chaque mise à jour est effectuée sur une base mensuelle. À chaque entrée correspond un identifiant de maladie unique, une position dans l'arbre nosologique, une description clinique, des liens vers d'autres classifications et bases de données, comme la nomenclature officielle des gènes HGNC, OMIM, la CIM-10, etc. Orphanet présente également des données d'épidémiologie, incluant une estimation de la prévalence des maladies. Ces données de prévalence sont élaborées et curées manuellement par l'équipe d'Orphanet à partir de plusieurs

sources d'information incluant des registres, des données issues d'instituts et agences sanitaires nationaux et internationaux, des requêtes automatiques ou non dans la littérature, ainsi que des avis d'experts.

OMIM : entrées clinico-biologiques

OMIM est une initiative débutée il y a plus de 50 ans dans le but de lister les relations génotype-phénotype connues. Contrairement à Orphanet, il ne s'agit pas d'une nosologie mais plutôt d'un répertoire d'entités clinico-biologiques, qui ne sont pas hiérarchisées. Il existe néanmoins une ébauche de classification nosologique avec les « séries phénotypiques », qui sont des panels d'entités clinico-biologiques du même spectre. Par exemple la pathologie « Coffin-Siris syndrome 2 » (#614607), liée au gène *ARID1A* (*603024), fait partie de deux séries phénotypiques : « Coffin-Siris syndrome » et « Mental retardation, autosomal dominant ». Les séries phénotypiques sont générées manuellement par l'équipe d'OMIM. Contrairement à Orphanet où les entrées sont définies avant tout sur la présentation clinique, les entrées d'OMIM sont des entités clinico-biologiques, ce qui implique que la grande majorité des entrées OMIM sont associées à un gène unique. Cette propriété rend la base OMIM plus adaptée qu'Orphanet dans une utilisation génomique. Ainsi, le catalogue OMIM est utilisé de manière très vaste et représente une source de données incontournable dans l'étude des maladies rares. L'équipe d'OMIM effectue une veille de la littérature, et est également ouverte aux suggestions dans le but d'identifier de nouvelles entités et d'établir de nouvelles relations génotype-phénotype. Les mises à jour sont quotidiennes et la liste des nouvelles entrées et des mises à jour d'entrées existantes est disponible en ligne. L'intégralité des données est téléchargeable et une interface web (www.OMIM.org) est disponible. Les données disponibles sont identifiées par des entrées à 6 chiffres précédées d'un caractère précisant le type d'entrée. Les deux types d'entrées principales sont les « phénotypes » pour lesquels une base moléculaire est connue, identifiés par un symbole dièse, et les gènes, identifiés par un astérisque. Les « phénotypes » représentent des maladies génétiques mendéliennes, mais également des traits phénotypiques sans caractère pathologique (comme la couleur des yeux), des susceptibilités pharmacogénétiques, aux infections, au cancer, ou encore des syndromes de délétion/duplication récurrents. Pour chaque entrée de phénotype, des relations avec les entrées de gènes impliquées sont disponibles, et vice-versa. De nombreuses informations additionnelles sont disponibles dans chaque type d'entrée. Par exemple au sein des entrées de gènes il existe des « variants alléliques », qui sont des variants pathogènes d'intérêt. Dans les entrées de phénotype il existe, en plus d'une description en texte libre de la maladie, un « synopsis clinique », qui est une liste normée de symptômes observés dans la maladie, ainsi que des modificateurs apportant des précisions. Ces termes du synopsis clinique sont dans un vocabulaire contrôlé, utilisant en particulier la nomenclature HPO.

HPO : un vocabulaire standardisé

Le vocabulaire phénotypique le plus communément employé dans les maladies rares est la nomenclature HPO, pour « Human Phenotype Ontology »[3,4]. Le projet HPO a été initié en 2007 dans le but de pouvoir « capturer » la variabilité phénotypique des individus, et la rendre assimilable par des êtres humains et des ordinateurs. Initialement prévue pour les maladies rares, HPO intègre désormais également des descriptions phénotypiques impliquées dans les maladies

fréquentes. Cette base comprend cinq « sub-ontologies » nécessaires à la description des phénotypes : le sous-ensemble le plus vaste correspond aux « anomalies phénotypiques », auquel peuvent être ajoutés des modificateurs cliniques pour ces anomalies phénotypiques : la fréquence, le mode de transmission de la maladie, et l'âge de décès.

Les « anomalies phénotypiques » contenues dans HPO sont de plusieurs types : des anomalies morphologiques (exemple : arachnodactylie), des processus d'organe anormaux (exemple : épistaxis), des processus cellulaires anormaux (exemple : anomalie du métabolisme du cycle de Krebs), des anomalies biologiques (exemple : glycosurie), des anomalies électrophysiologiques, d'imagerie, ou des anomalies du comportement. Chaque entrée possède un nom, un numéro d'identification, ainsi qu'une liste de synonymes. En complément de cette liste de termes phénotypiques, il existe des relations logiques entre les différentes entrées, certaines entrées spécifiques incluses dans des entrées plus larges, par exemple « voix nasonnée » constitue, avec 10 autres entrées, les sous-classes du terme plus large « anomalies de la voix ».

La nomenclature HPO est utilisée dans un grand nombre d'applications, logiciels et bases de données. Leur utilisation systématique permet par exemple la priorisation automatisée des gènes lors des analyses pangénomiques au sein de cohortes hétérogènes, en proposant un *ranking* des variants en fonction de la similarité du phénotype du patient avec les connaissances scientifiques autour du gène.

Pour résumer, Orphanet et OMIM sont deux bases listant avec un objectif d'exhaustivité les maladies génétiques, ainsi que les relations génotype-phénotype connues. D'autres bases de données consacrées aux maladies rares génétiques existent, comme par exemple GeneReviews du NCBI, mais qui n'ont ni vocation d'exhaustivité, ni architecture modélisée pour des requêtes informatiques. Afin d'illustrer cette introduction avec des données générales sur les maladies rares, les données d'Orphanet et d'OMIM ont été téléchargées comme décrit dans le chapitre 2.1.

1.1.2.2 Diversité clinique des maladies génétiques rares

Les maladies génétiques rares sont extraordinairement variées, dans leurs symptômes, l'âge d'apparition, leur gravité, leur mode de transmission. Tous les organes peuvent être touchés, de manière isolée ou associée sous forme de syndromes. La base Orphanet propose une classification hiérarchisée des maladies, permettant d'observer de manière globale les grands champs cliniques. Ainsi on peut observer que parmi l'ensemble des maladies génétiques rares répertoriées dans cette base, les deux catégories les plus représentées sont « Anomalie rare du développement embryonnaire d'origine génétique » (n=3545 maladies) et « Maladie neurologique génétique rare » (n=3237 maladies) (Figure 1). Viennent ensuite par ordre de nombre de maladies répertoriées les maladies rares génétiques de l'œil (n=1235), des os (n=980), les erreurs innées du métabolisme (n=853), les maladies rares dermatologiques (n=699), endocriniennes (n=587), etc. Il est à noter que des maladies peuvent faire partie de plusieurs catégories, ce qui explique que les maladies rares du développement soient prépondérantes car elles englobent des anomalies malformatives pouvant se rapporter aussi aux catégories plus spécifiques d'organes. Aussi, le nombre de maladies dans chaque grand champ clinique n'est pas représentatif du nombre de patients atteints. Par exemple la catégorie « Syndrome familial avec prédisposition aux cancers » représente une part prépondérante

dans l'activité de génétique médicale de manière globale, avec en tête de file les prédispositions au cancer du sein et de l'ovaire, et les prédispositions aux tumeurs digestives, alors que ce champ ne regroupe *que* 114 affections répertoriées.

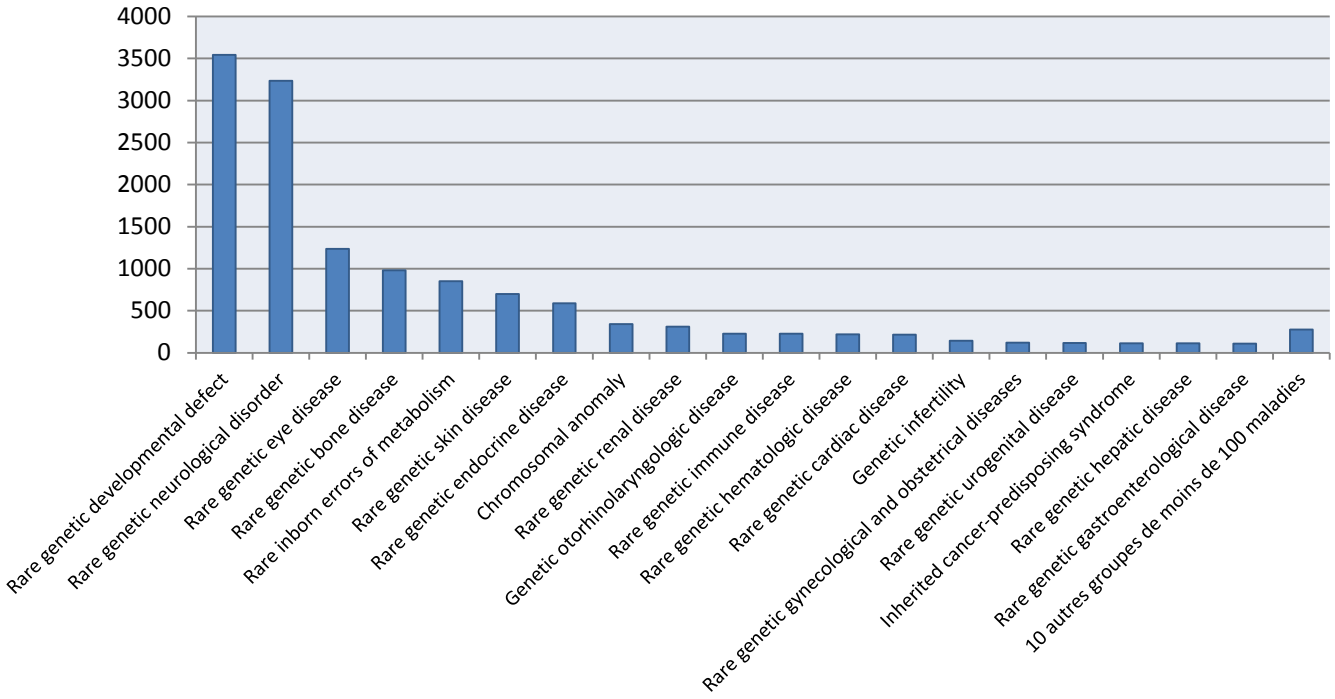


Figure 1. Grands champs cliniques des maladies génétiques rares selon la classification des maladies génétiques rares Orphanet
Dans un souci de lisibilité, les 10 catégories contenant le moins de maladies ont été regroupées, et le nom de la première catégorie a été tronqué, le nom complet étant « *Rare genetic developmental defect during embryogenesis* **».** Source : Orphadata (voir ressources web).

1.1.2.3 Variabilité des présentations familiales

En dehors de la grande diversité dans les formes cliniques, les maladies génétiques sont variées dans leur mode de transmission. Historiquement, les formes familiales ont été les plus évidentes et ont été les mieux documentées. Des études intra et inter familiales, basées notamment sur des études de liaison, ont permis d'identifier la cause génétique de nombreuses affections autosomiques dominantes, autosomiques récessives, et liées à l'X. Les maladies génétiques peuvent également être sporadiques et n'affecter qu'un individu dans une famille, par le biais d'une altération génétique dominante apparaissant chez l'individu atteint par exemple (variation de survenue *de novo*). Il est connu de longue date que l'immense majorité des maladies chromosomiques est due à des anomalies génétiques *de novo*. Plus récemment, l'impact des variations de séquence de survenue *de novo* a été démontré dans un grand nombre d'affections. Ainsi il est de plus en plus clair que les maladies génétiques ne sont pas synonymes de maladies familiales, et en particulier dans les maladies pédiatriques sévères, telles que la déficience intellectuelle, l'autisme, les épilepsies et les malformations cardiaques congénitales par exemple. En dehors des modes de transmission classiques et plus rares décrits dans les bases OMIM et Orphanet (Figure 2), il convient de mentionner certains modes de transmission liés à des anomalies génétiques particulières et qui peuvent être responsables de formes familiales inhabituelles. Ces anomalies particulières

comprennent entre autres les déséquilibres de translocations chromosomiques et autres anomalies chromosomiques, pouvant être responsables de récurrence dans les familles, les maladies à expansion de triplets, qui peuvent également se déstabiliser dans plusieurs branches familiales séparées par des individus sains, et les mosaïques germinales, à l'origine de possibles récurrences de mutations *de novo* au sein des fratries. Enfin, les gènes soumis à empreinte parentale, et les mécanismes digéniques et oligogéniques peuvent également donner lieu à des particularités dans les présentations familiales des maladies rares.

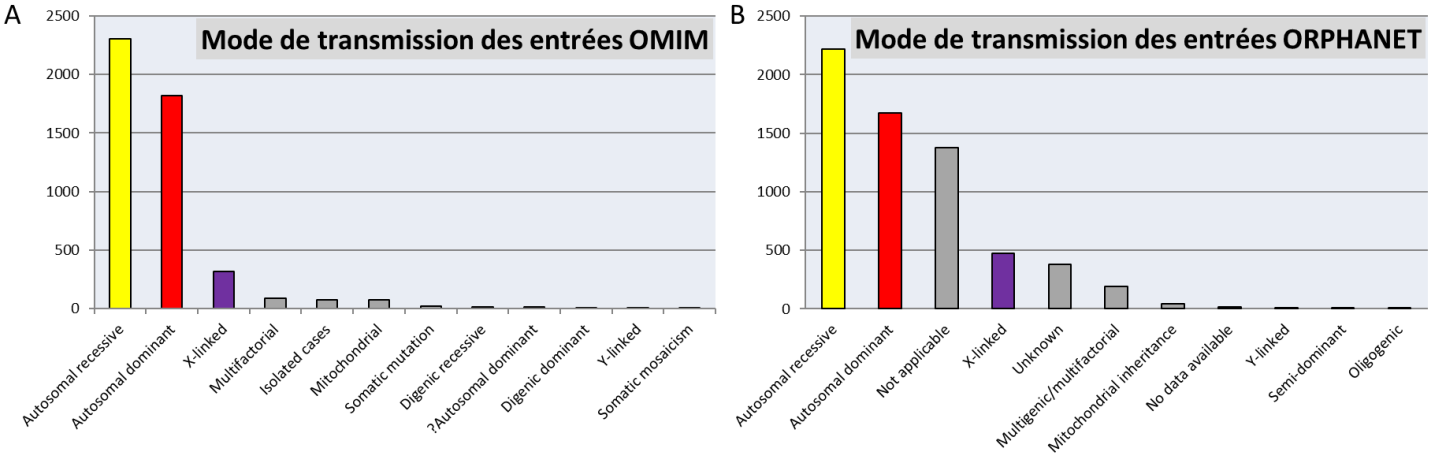


Figure 2. Mode de transmission des maladies génétiques au sein de deux bases de référence
A) Données issues de la base OMIM, selon la table genemap2 (une entrée par gène, plusieurs modes de transmission possibles par gène). B) Données issues de la base Orphanet (une entrée par maladie Orphanet, plusieurs modes de transmission possible par maladie).

1.1.2.4 Relations génotype-phénotype et hétérogénéité génétique

Une maladie, plusieurs gènes

Si certaines maladies sont dues à des mutations au sein d'un gène unique, de nombreuses maladies génétiques peuvent résulter d'une variation d'un gène parmi un ensemble de gènes plus ou moins vaste. Cette propriété des maladies génétiques est appelée *hétérogénéité génétique* et dépend comme déjà abordé du niveau de granularité de la maladie génétique considérée. La Figure 3A présente l'hétérogénéité génétique des maladies génétiques d'après la granularité Orphanet, basée avant tout sur des entités cliniques. Ces données indiquent que près de ¾ des maladies avec une cause génétique connue sont néanmoins associées à un gène unique.

Un gène, plusieurs maladies

Inversement, des mutations dans un gène donné peuvent causer diverses maladies (Figure 3B), par plusieurs moyens. D'une part il peut s'agir d'un mécanisme moléculaire différent (exemple : perte de fonction vs gain de fonction, comme par exemple dans le gène *SMAD4* pour lequel les variations perte de fonction entraînent une maladie associant polypes hamartomateux digestifs et télangiectasies hémorragiques, alors que certaines variations gain de fonction très spécifiques donnent lieu à une maladie multi viscérale avec déficience intellectuelle[5]). Un autre exemple est celui des variations perte de fonction germinales du gène *PDGFRB*, qui causent une maladie

neuropsychiatrique de transmission autosomique dominante (calcifications cérébrales primaires, OMIM #615007) alors que les mutations gain de fonction entraînent un phénotype d'hypercroissance (Syndrome de Kosaki, OMIM #616592), de vieillissement prématuré (Syndrome de Penttinen, OMIM #601812), ou une prédisposition héréditaire aux myofibromes (myofibromatose infantile OMIM #228550), toutes de transmission autosomique dominante. Deux maladies distinctes associées à un même gène peuvent également être dues à des variations partageant le même mécanisme mais avec un effet quantitativement différent. Par exemple de nombreuses maladies métaboliques ont des formes variables en fonction de la fonction résiduelle de l'enzyme déficiente. On peut citer également le gène *CFTR* à l'origine de la mucoviscidose, mais pouvant également causer un ensemble de symptômes du même spectre mais moins marqués regroupés sous le terme de *CFTR*-related disorders (CFTR-RD), en cas de variations d'effet atténué. Un troisième mécanisme à l'origine de l'observation de plusieurs phénotypes associés à un même gène peut passer par des modes de transmission différents. Un exemple classique est celui des atteintes mono-alléliques du gène *BRCA2*, à l'origine de prédisposition au cancer du sein et de l'ovaire, alors que les mêmes mutations à l'état bi-allélique sont à l'origine de la maladie très différente qu'est l'anémie de Fanconi. Enfin, certains variants peuvent être à l'origine de plusieurs phénotypes distincts, même au sein d'une même famille, soulevant l'hypothèse de facteurs modificateurs génétiques ou environnementaux. Tous ces mécanismes sont détaillés dans une revue centrée sur les maladies neuropsychiatriques[6].

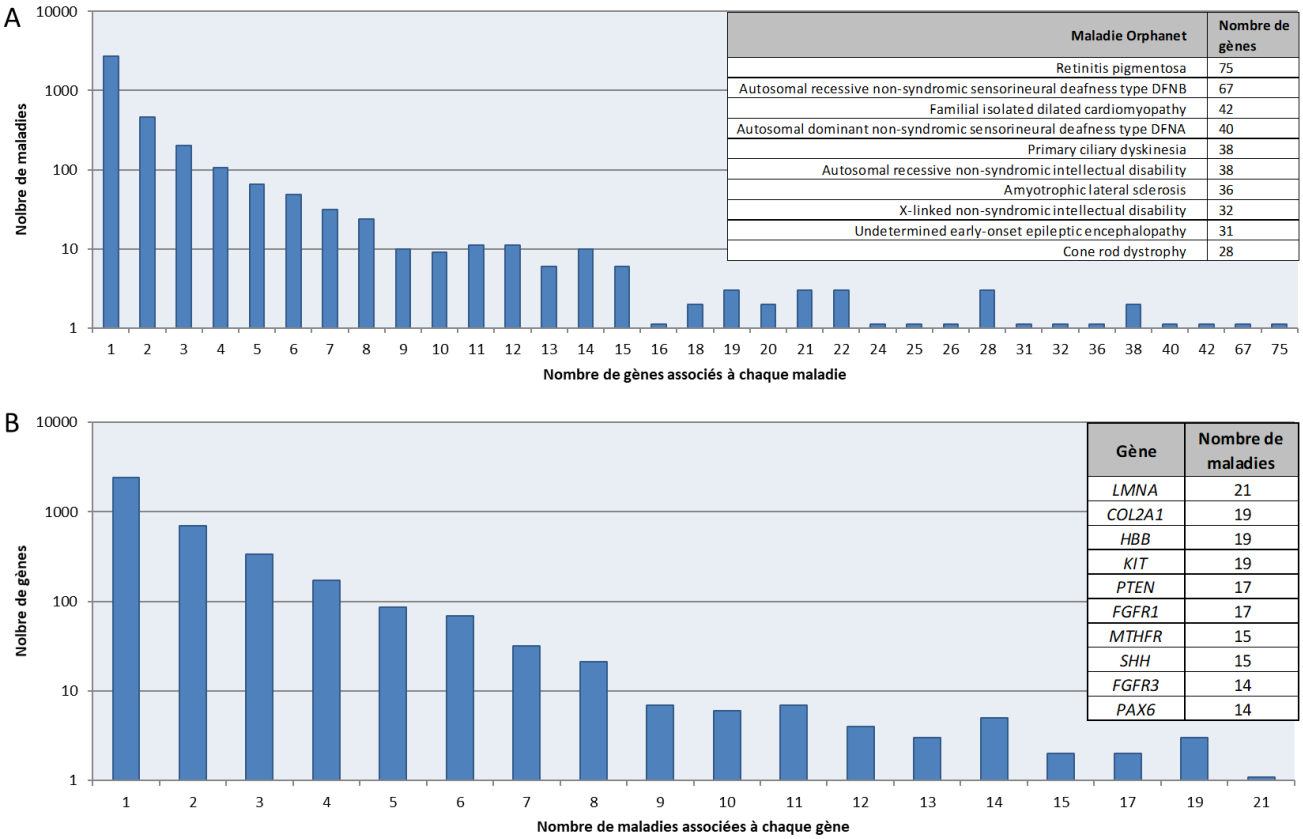


Figure 3. Hétérogénéité génétique dans les maladies génétiques humaines selon Orphanet
A) Nombre de gènes pour chaque maladie Orphanet associée à au moins un gène. En abscisse est représenté le nombre de gènes pour chaque affection et en ordonnée ne nombre de maladies (échelle logarithmique). Les dix maladies avec le plus grand nombre de gènes associés sont indiquées en encart.

Figure 3 (suite de la légende)
B) Nombre de maladies génétiques associées à chaque gène impliqué en pathologie humaine. En abscisse est représenté le nombre de maladies pour chaque gène et en ordonnée le nombre de gènes (échelle logarithmique). Les dix gènes avec le plus grand nombre de maladies associées sont indiqués en encart. Il est à noter que les maladies Orphanet incluent des maladies en mosaïque voir certains cancers. Source : Orphadata (http://www.orphadata.org/cgi-bin/inc/product6.inc.php).

1.2 Séquençage haut débit et analyses pangénomiques

Dans cette seconde partie introductive consacrée aux techniques de génétique moléculaire modernes basées sur le séquençage à haut débit, nous aborderons dans un premier temps les principes techniques du NGS, moléculaires et bio-informatiques. Nous évoquerons ensuite la variabilité et la mutabilité du génome humain dont l'évaluation précise a été possible grâce aux évolutions apportées par le NGS. Ces éléments nous permettront de détailler l'utilisation du NGS dans le contexte du diagnostic génétique des maladies rares.

1.2.1 Principes techniques du NGS

Le séquençage haut débit est une technologie apparue dans la deuxième moitié des années 2000 et qui s'est popularisée à partir de 2008 (Figure 4). Plusieurs domaines scientifiques ont exploité les possibilités offertes par le NGS. En dehors du séquençage de l'ADN, d'innombrables applications ont été développées, telles que des études de transcriptome, de méthylome, d'interaction du génome avec les protéines, de conformation 3D de l'ADN, etc. Néanmoins dans ce chapitre seront développées uniquement les applications de séquençage d'ADN génomique humain, avec comme modèle d'application le diagnostic des maladies monogéniques.

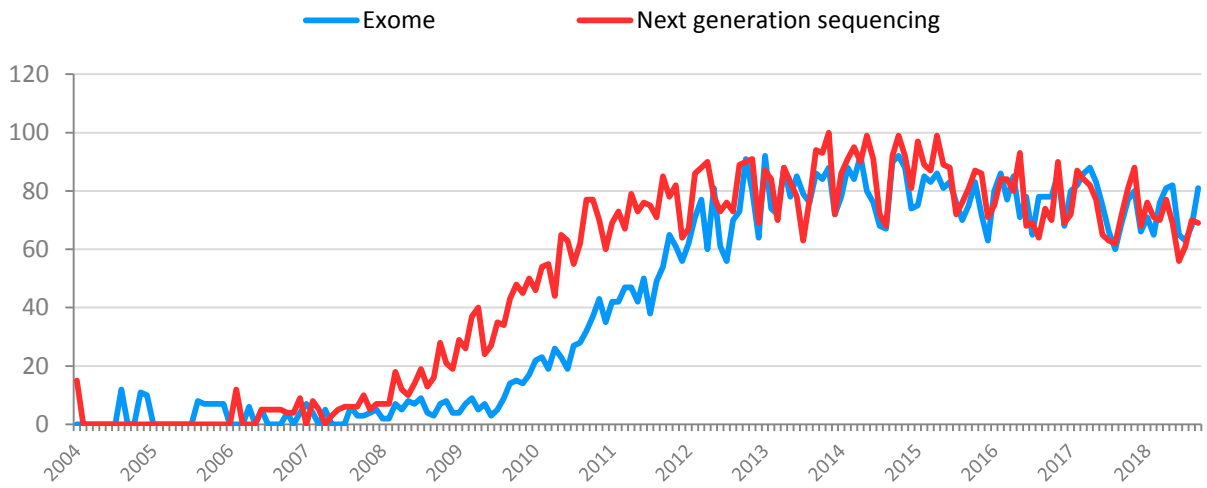


Figure 4. Évolution du nombre de requêtes Google pour les termes « exome » et « next generation sequencing », de 2004 à 2018
En ordonnée : pourcentage par rapport à la valeur la plus forte sur la période investiguée. Source : https://trends.google.fr/

Le NGS correspond au séquençage en parallèle d'un grand nombre de molécules d'ADN constituant l'échantillon testé. L'une des différences principales avec le séquençage Sanger est la

compartimentation des réactions. En effet lors du séquençage Sanger, la réaction se fait en phase liquide dans un tube, n'autorisant pas de multiplexage qui engendrerait un mélange des signaux. En NGS, les réactions se font également dans une phase liquide, mais très confinées spatialement grâce à des technologies supports solides recouverts de courtes chaînes d'acides nucléiques. Ainsi des millions de réactions sont possibles en parallèle et ne se mélangent pas. Ce processus mène à la génération par le séquenceur de fichiers bruts indiquant la séquence nucléotidique ayant été observée au sein de chaque réaction. Ces données sont par la suite traitées par une série d'algorithmes afin de détecter un signal d'intérêt au sein de cette immense masse de données, par exemple les variations génétiques codantes rares d'un individu. Ce traitement bio-informatique, adaptable en fonction des besoins, est une étape critique nécessitant une grande expertise afin d'obtenir des données de qualité contrôlée.

Chaque machine de séquençage à haut débit possède ses propres spécificités et protocoles. Nous citerons ici les processus généraux communs en prenant pour exemple le séquençage par synthèse, en short reads, d'ADN génomique lymphocytaire tel qu'il est pratiqué classiquement en génomique humaine en 2018. La Figure 5 représente l'ensemble de ce processus.

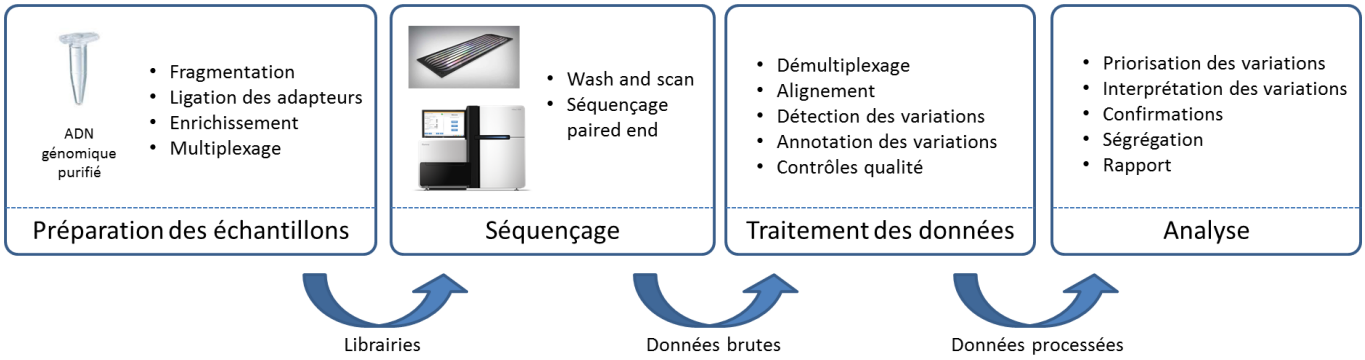


Figure 5. Éléments généraux du *workflow* de NGS, par exemple dans le contexte du séquençage d'exome

1.2.1.1 Préparation des échantillons

La première étape, préalable au séquençage et correspondant à la phase de *wet lab*, est appelée préparation des échantillons. L'objectif de cette étape est d'obtenir, à partir de l'ADN extrait des cellules, une « bibliothèque » (souvent improprement traduite « librairie » dans le langage courant français) de molécules d'ADN cible, fragmentées et modifiées de manière à répondre aux conditions requises par le séquenceur. Plusieurs protocoles sont possibles en fonction de l'objectif et de la chimie sous-jacente, mais les étapes les plus communes sont schématiquement la fragmentation, réalisée soit par un procédé mécanique (sonication) soit par action enzymatique, la ligation d'adapteurs, la sélection et l'enrichissement (réalisé classiquement en cas d'exome par capture à l'aide de sondes d'acides nucléiques biotinylées), et l'indexage, permettant de multiplexer les échantillons au sein d'un run unique de séquençage.

1.2.1.2 Séquençage

La seconde étape est celle du séquençage, où les librairies sont déposées sur un support (*flowcell* pour la technologie Illumina, puce semi-conductrice pour la technologie Ion Torrent par exemple)

qui est positionné dans le séquenceur. La séquence se fait de manière automatisée par cycles de *wash and scan* correspondant au séquençage d'une base à la fois, en utilisant la fluorescence (Illumina) ou le pH (Ion Torrent). Chaque extrémité du fragment d'ADN est séquencée (*paired-end sequencing*). Le séquenceur produit des fichiers convertibles en fichiers FASTQ, correspondants aux données brutes de séquençage.

1.2.1.3 Traitement bio-informatique des données

La troisième étape correspond à la gestion et la transformation bio-informatique des données primaires pour en extraire les informations d'intérêt, comme des données de qualité ou des variations génétiques. Ces pipelines bio-informatiques sont complexes et comprennent des problématiques propres à chaque situation (type de variations recherchées, méthodes de séquençage, etc.). Aussi il n'existe pas de processus consensuel de référence mais une panoplie d'outils indépendants dont le déploiement et le paramétrage nécessitent une expertise spécifique. Les principes de ces étapes, simplement évoqués dans ce chapitre, sont décrits plus en détail ailleurs : ref[7]. Pour chaque tache, des logiciels académiques gratuits sont disponibles, dont les plus communément employés seront cités.

Contrôle qualité des données brutes

Après le démultiplexage, les fichiers FASTQ comprennent l'information des bases nucléotidiques lues par le séquenceur, ainsi qu'une valeur de qualité associée. À cette étape, des contrôles de la qualité des données primaires sont faits, par exemple grâce à l'outil FastQC du Babraham institute.

Alignement

L'étape suivante correspond à l'alignement des séquences sur un génome de référence menant à la génération de fichiers SAM/BAM. Les logiciels classiquement utilisés incluent BWA[8], ou BowTie 2 de l'université Josh Hopkins. Des étapes techniques supplémentaires, ainsi que des contrôles sur la qualité de l'alignement, sont réalisés. Ces étapes permettent l'accès à des données importantes pour l'évaluation de la qualité du séquençage. En premier lieu, la profondeur correspond au nombre de molécules d'ADN séquencées à chaque position (exemple : 30x de profondeur moyenne pour un séquençage de génome complet = 30 molécules d'ADN séquencées pour chaque position). La couverture correspond au pourcentage des régions ciblées qui sont couvertes à Nx. En dehors de l'évaluation de la qualité du séquençage, le fichier BAM, correspondant au fichier d'alignement des séquences, sert de base pour l'étape suivante de détection des variations.

Variant calling

Plusieurs types de logiciels peuvent être utilisés en fonction du type de variations recherchées (mutations ponctuelles vs variations de structure). Les variations de séquence identifiées correspondent aux différences par rapport au génome de référence et sont exportées dans un fichier VCF. Un exemple commun dans la détection des variations de séquence est GATK-HTC par le Broad Institute.

Annotation

Ces variations sont ensuite annotées à l'aide de diverses sources de données, permettant d'associer des informations scientifiques et des prédictions d'intérêt à chaque variation. Des logiciels d'annotation classiques sont par exemple Variant Effect Predictor (VEP)[9], SnpEff[10] ou encore Annovar[11]. En plus de ces logiciels, des scripts maison peuvent également être implémentés afin d'annoter des informations additionnelles. Deux types d'annotations sont généralement réalisés : les annotations spécifiques du variant, et les annotations spécifiques du gène.

Les annotations spécifiques du variant incluent la prédiction de la fonction du variant, la fréquence du variant dans diverses bases de données de population, ou la prédiction de pathogénicité des variants faux-sens par exemple. On peut citer la ressource d'annotation automatisée dbNSFP[12], ou les auteurs ont pré-annoté de manière systématique toutes les substitutions nucléotidiques possibles des séquences codantes du génome avec 24 outils, incluant des scores de conservation interespèces, des scores de prédiction de pathogénicité des variations et des scores « ensemble » (voir chapitre 1.3.2.3). Cette base de données permet une annotation facilitée, en termes de ressources informatiques nécessaires et de risque d'erreur, au prix d'une certaine fixité dans les annotations, dbNSFP ne pouvant pas mettre à jour en temps réel les versions de chaque logiciel utilisé. En plus des logiciels de prédiction de pathogénicité, les bases de données de variations en population générale et les bases de données de variations en lien avec les maladies génétiques communément utilisées dans cette étape d'annotation seront détaillées dans des chapitres ultérieurs.

Les annotations spécifiques des gènes incluent par exemple le nom des maladies associées aux gènes (OMIM, ou Orphanet) ou la sensibilité du gène aux variations perte de fonction par exemple. Ces annotations peuvent être produites par des scripts simples (ou sur Excel en cas de données peu volumineuses).

1.2.1.4 Analyse des données

La quatrième étape correspond à la priorisation, l'interprétation et le rendu des résultats. Le faisceau d'éléments pris en compte dans ce processus sera revu plus en détail au chapitre 1.2.3 mais comprend l'effet biologique prédit de la variation sur l'ARN ou la protéine, la cohérence du phénotype et de la ségrégation familiale du variant, et la rareté de la variation en population non atteinte ou au contraire sa présence éventuelle en population atteinte.

Le chapitre suivant s'écarte de l'approche diagnostique décrite dans ce paragraphe pour montrer le rôle qu'a eu le NGS dans l'identification récente de l'incroyable polymorphisme génétique humain.

1.2.2 Variabilité du génome humain

1.2.2.1 L'apport des bases de données génomiques dans l'identification du polymorphisme génétique humain

La variabilité interindividuelle du génome humain, ou polymorphisme génétique, a pu être pleinement appréciée depuis l'agrégation de données de séquençage pangénomique. Des populations analysées par stratégies d'exome, comme dans la cohorte Exome Sequencing Project

(ESP) du NHLBI[13] agrégeant plus de 6000 individus, ou en génome, comme le projet 1000 Genomes avec 2504 individus[14] ont permis d'analyser quantitativement et qualitativement les variations identifiées à l'échelle d'une population. Plus récemment, des cohortes de plus grande ampleur ont été élaborées, avec le consortium ExAC (pour Exome Aggregation Consortium) incluant les données de 60 706 individus[15], puis GnomAD (Genome Aggregation Database) incluant les données de 123 136 exomes et 15 496 génomes. Ces cohortes ont permis de préciser les limites de la variabilité génétique humaine, faisant appel à des notions de génétique des populations, de contrainte fonctionnelle des variations, et de mutabilité du génome. Ces bases de données représentent actuellement des ressources primordiales pour l'interprétation des variations dans le contexte du diagnostic génétique des maladies rares.

ExAC : Exome Aggregation Consortium

Le consortium ExAC est une initiative américaine du Broad Institute visant à fournir à la communauté scientifique et médicale un jeu de données à grande échelle de variations génétiques humaines homogènes et de bonne qualité[15]. Les données d'exome de plusieurs études ont été mises en commun, et re-processées par un pipeline commun afin d'obtenir des données homogènes. Au total les données de 60 706 exomes ont été incluses dans ExAC (Figure 6), avec comme critères d'inclusion la nécessité de données de bonne qualité, l'absence d'apparentement entre les individus et l'absence de maladie pédiatrique sévère.

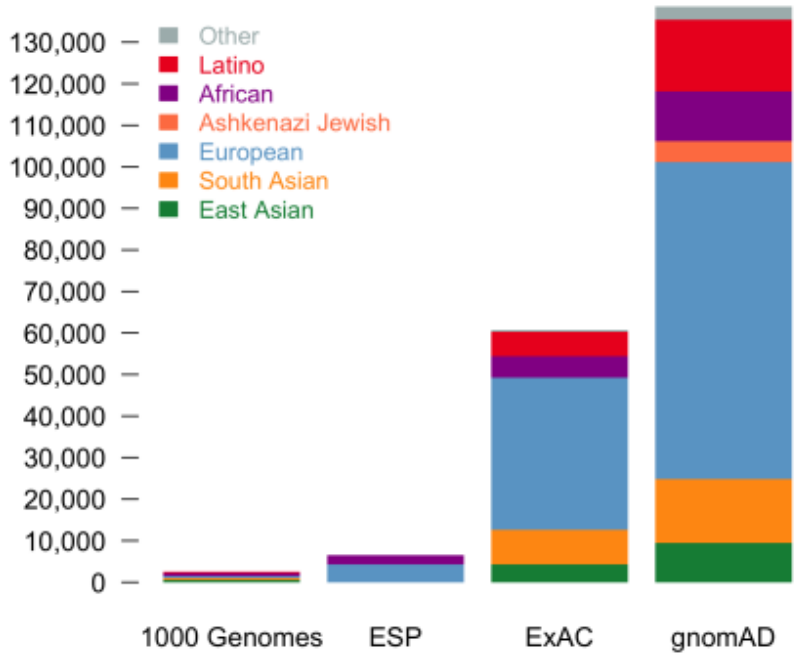


Figure 6. Comparaison de la taille et des populations constituant les bases de données actuelles de variations génétiques en population générale
Source : https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/

GnomAD : The Genome Aggregation Database

GnomAD constitue la mise à jour d'ExAC par la même équipe. Disponible pour sa version finale à partir de février 2017, les modifications par rapport à ExAC sont le doublement du nombre

d'échantillons, l'ajout de données de génome, ainsi qu'une optimisation dans les méthodes de détection des variants permettant d'augmenter la qualité des données. Les critères d'inclusion ont également été un peu modifiés par rapport à ExAC car les apparentés au premier degré d'individus avec maladie pédiatrique sévère sont désormais également exclus, en plus des cas index eux-mêmes, dans le but que GnomAD puisse servir de jeu de données encore plus pertinent pour l'étude de ces maladies (http://gnomad.broadinstitute.org/about). Au total GnomAD comprend des données issues de 123 136 exomes et de 15 496 génomes, provenant de 44 études indiquées en Figure 7, dont la plupart concernent des maladies fréquentes. Il est important de noter que si certaines de ces études ont permis de fournir des contrôles de population générale, il existe probablement un nombre significatif d'individus atteints de ces maladies fréquentes et leurs apparentés dans GnomAD, qui représente un biais potentiel dans l'utilisation de GnomAD comme cohorte contrôle pour ces maladies. En revanche GnomAD représente un outil extrêmement intéressant dans l'étude des maladies pédiatriques sévères.

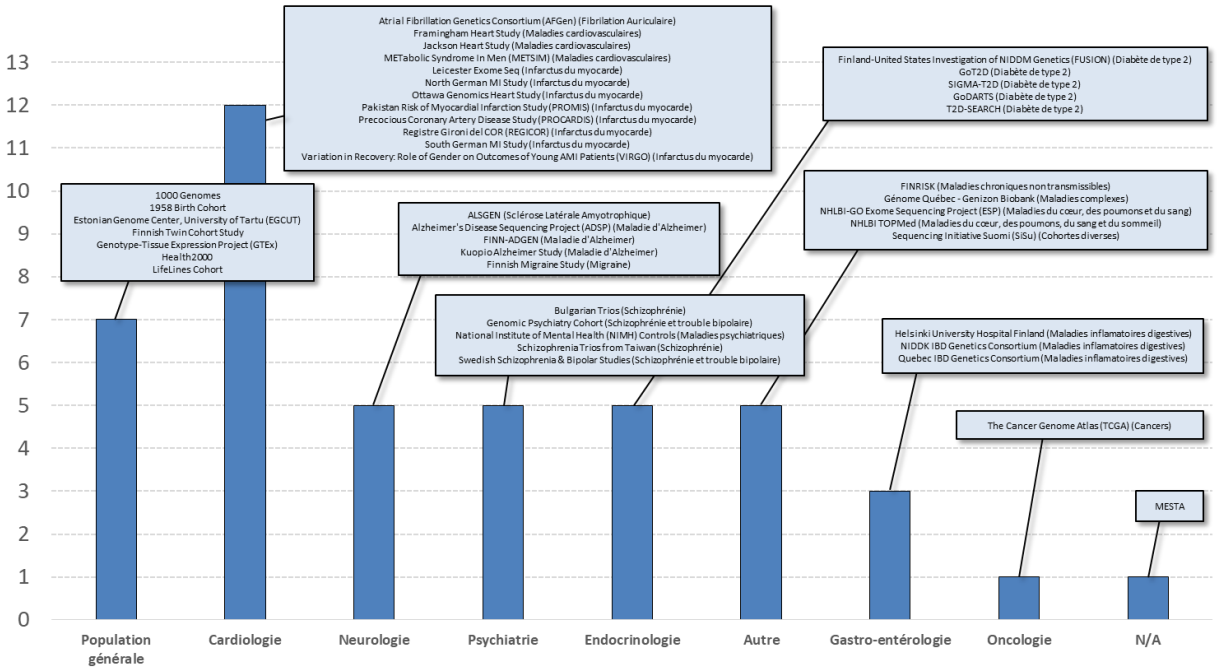


Figure 7. Cohortes constituant GnomAD
Les cohortes sont triées par groupes de maladies. En ordonnée : nombre d'études dans chaque catégorie. Source : http://gnomad.broadinstitute.org/about

Bases de données de populations ciblées

En dehors des bases de données globales de très grande taille, la création de bases locales concernant des populations ciblées peut s'avérer utile. On peut citer l'exemple du French Exome (FREX) Project, comprenant 573 individus nés de parents français analysés en exome, ayant pour objectif de cartographier l'architecture génétique française. Les populations françaises étant largement absentes de ExAC et GnomAD, FREX a permis de montrer certains *patterns* de variations distincts présents dans cette population. En particulier, 21% des variations identifiées dans FREX n'étaient pas présentes dans ExAC. Ainsi FREX représente une ressource pertinente de variabilité génétique en population générale française, ces exomes pouvant par exemple être utilisés

en tant que contrôles ajustés dans le cadre des études d'association basées sur l'exome[16]. D'autres bases de données de populations homogènes sur un plan ethnique incluent les par exemple les génomes du consortium *Genomes of the Netherlands* (GoNL)[17], les trios islandais du consortium deCODE[18], ou encore les génomes de l'*integrative Japanese genome variation database* (iJGVD)[19].

Variabilité génétique à l'échelle d'un individu

La cohorte ESP a permis de montrer la présence de 13 595 substitutions nucléotidiques dans un exome typique, comprenant 7 652 variations synonymes, 5 754 faux-sens, 12 variations canoniques d'épissage et 35 variations non-sens par individu[20]. Ces chiffres, assez anciens, sont sous-estimés par rapport aux données d'exome actuelles qui sont plutôt de l'ordre de 20 000 variations par exome. Aussi dans la cohorte relativement récente d'ExAC, le nombre de variations faux-sens et tronquantes est de 12 186 en moyenne par individu, soit environ deux fois plus que dans ESP (ref[15], calcul à partir des données supplémentaires). À l'échelle du génome, ce sont 4,1 à 5,0 millions de variations de séquence qui sont identifiées par individu[14], parmi lesquelles une grande majorité est fréquente (96 à 99 % des variations ayant une fréquence de l'allèle mineur (MAF), de plus de 0,5 %). Les variations de structure sont plus délicates à identifier, mais il a été estimé qu'un génome typique comprenait 2 100 à 2 500 évènements de type variations du nombre de copies (CNVs), inversions et insertions d'éléments mobiles du génome[14].

Variabilité génétique à l'échelle d'une population

L'agrégation des données d'une cohorte permet de montrer qu'une majorité des variants d'une population sont rares. Par exemple au sein de la population d'ExAC, plus de 99 % des variations identifiées ont une MAF de moins de 1 %[15]. En réalité cette tendance dépend uniquement de la taille de la cohorte : l'augmentation de la taille d'une cohorte ne permettant pas d'identifier plus de variations fréquentes mais uniquement des variations rares, la proportion de variations rares ne peut qu'augmenter avec la taille de la cohorte. Dans la base ExAC, il existe en moyenne au sein des séquences codantes une variation tous les 8 nucléotides. Ce consortium a montré que l'augmentation de la taille des cohortes permettait d'identifier une tendance à la saturation de certaines variations (transitions CpG, voir chapitre 1.2.2.2), indiquant que ces variations sont survenues de manière indépendante dans plusieurs populations. La grande taille de ces cohortes permet également de mettre en évidence des sites multi-alléliques, ou plusieurs allèles alternatifs sont identifiés.

L'évaluation du polymorphisme humain permet de retracer l'histoire évolutive d'homo sapiens

Il est apparu très vite que les populations africaines possédaient un polymorphisme plus important que les autres populations[20]. Ce phénomène a été mis en lien avec l'histoire évolutive des êtres humains, dans le modèle démographique *Out-of-Africa* (OOA), ou un petit nombre d'individus ayant quitté l'Afrique pour peupler les autres continents ont produit un « goulot d'étranglement » génétique réduisant le pool de variations polymorphiques par rapport aux populations Africaines. Par ailleurs, il a également été montré qu'en plus de l'aspect quantitatif, le type de variations n'était pas homogène entre les différentes populations. En effet, la proportion de variations rares

potentiellement délétères augmente dans les populations ayant migré hors de l'Afrique, et particulièrement dans les populations ayant subi des goulots d'étranglement récents (juifs ashkénazes, finlandais, québécois)[21]. Cette observation souligne l'effet moins prononcé de la sélection purificatrice dans les populations d'effectif efficace plus réduit avec goulot d'étranglement récent. De même, d'une manière générale, la datation de l'apparition des variations a permis de montrer que les variations les plus récentes étaient globalement plus délétères[13].

Intérêts des données de population générale dans l'étude des maladies génétiques

La fréquence des variations représente un élément clé de l'interprétation des variations dans le diagnostic des maladies génétiques. En effet, en dehors de rares situations telles que les effets fondateurs et la sélection balancée, les variations délétères sont rares en population générale. De fait, une fréquence trop élevée en population générale par rapport à la prévalence de la maladie est un élément fort à l'encontre de la pathogénicité du variant. Le consortium ExAC a montré que l'utilisation non pas de la fréquence sur la cohorte entière mais de la fréquence la plus élevée parmi les différentes populations géographiques incluses (popmax), permettait d'augmenter la capacité de filtrations des variations de fréquence supérieure à un seuil donné. La croissance des bases de données de population générale a permis de reclasser de nombreuses variations fréquentes considérées à tort comme pathogènes[15]. Inversement, les données de population générale peuvent également permettre d'évaluer la prévalence minimale de certaines maladies non pédiatriques sévères, comme par exemple dans les calcifications cérébrales primaires[22].

Dans le cas particulier des maladies sévères du développement, la filtration des variations en utilisant les données de population générale peut être assez drastique car les bases de données d'ExAC et GnomAD ne contiennent en théorie pas d'individus porteurs de maladies pédiatriques sévères. Dans une étude récente, les auteurs évaluaient la présence dans ExAC de génotypes pathogènes (d'après la base de données ClinVar, voir chapitre 1.2.3.3). Cette étude concluait que si de manière générale 2,8 % des individus d'ExAC étaient porteur d'un génotype pathogène dans une liste de 924 gènes associées à des maladies pédiatriques sévères, la très grande majorité des génotypes étaient en réalité compatibles avec une absence de phénotype pédiatrique sévère (plusieurs maladies associées au gène, présence de formes cliniques atténuées, facteurs de susceptibilité, etc..). Les auteurs ont identifié 18 individus d'ExAC avec des génotypes clairement pathogènes pour des maladies pédiatriques sévères, mais pour lesquels la réalité du génotype d'ExAC était douteuse. On peut noter l'exemple très particulier du gène *ASXL1* dont les variations tronquantes *de novo* sont à l'origine du syndrome de Bohring-Opitz, pédiatrique sévère et entièrement pénétrant, et pour lequel de nombreuses variations tronquantes sont présentes dans ExAC (n=342 individus porteurs). Une étude récente a montré que les variations présentes dans ExAC étaient en réalité des probables variations en mosaïque somatique, ces variations donnant un avantage sélectif aux lignées hématopoïétiques porteuses, expliquant la présence de mosaïques leucocytaires à fort taux chez des individus de population générale[23]. Une autre étude réalisée chez 589 306 individus visait à identifier des individus « résilients » vis-à-vis de maladies génétiques sévères[24]. Les auteurs ont identifié seulement 13 candidats, sans information sur la présentation clinique pour la plupart. Pour résumer, en dehors de cas particuliers très rares comme le cas du gène *ASXL1*, les génotypes pathogènes pour des maladies pédiatriques sévères pénétrantes sont

largement absents d'ExAC (et probablement par extension de GnomAD), ce qui permet une filtration forte grâce à ces jeux de données dans l'interprétation des variations.

Variation en population générale et contrainte fonctionnelle des gènes

L'étude des variations génétiques au sein d'une population permet d'identifier une contrainte, un appauvrissement statistique de certains variants fonctionnels, permettant de faire l'hypothèse que ces variations ont un effet biologique fonctionnel important. Le score RVIS[25], basé sur la proportion de variants fréquents parmi l'ensemble des variations non synonymes d'un gène (d'après les données de la cohorte d'exomes d'EVS), a vite été remplacé par les scores d'intolérance du consortium ExAC, avec en premier lieu le score pLI (*probability of loss-of-function intolerance*), où le nombre de variations rares tronquantes observées dans ExAC a été comparé au nombre de variations attendues par un modèle nul[26]. Indiqué pour chaque gène ayant une qualité satisfaisante dans ExAC, le score pLI varie de 0, ou absence de sensibilité aux variations tronquantes, à 1 ou le gène est très statistiquement déplété en variations tronquantes en population générale. Un seuil de 0,9 a été proposé pour définir 3 230 gènes intolérants à la perte de fonction. Cette mesure d'intolérance a été très largement employée par la communauté scientifique et médicale, en particulier pour identifier les gènes cibles de variations tronquantes dans les maladies génétiques pédiatriques sévères mono-alléliques.

1.2.2.2 La mutabilité du génome

Si la mutabilité du génome peut être approchée par l'étude du polymorphisme humain dans les bases de données de population, les stratégies contemporaines permettent de mesurer directement la mutabilité du génome à l'échelle de l'individu. Le séquençage de l'exome ou du génome d'un individu et de ses deux parents, définissant un séquençage en trio, permet par soustraction d'identifier un nombre restreint de variations *de novo*, absentes chez les parents et apparues chez l'individu. Par ces approches, le séquençage à haut débit a contribué à une augmentation considérable de la connaissance et de la compréhension de la mutabilité du génome[27]. Le nombre de variations *de novo* par individu a été identifié, et des facteurs influençant cette mutabilité ont été documentés, avec en premier lieu l'effet majeur et de découverte récente de l'âge paternel à la conception.

Nombre moyen de mutations *de novo* par individu

Le nombre de substitutions nucléotidiques *de novo* dans un génome typique est de l'ordre de 70,[18] mais ce chiffre peut varier du simple au double entre les individus (Tableau 1, Figure 8A). Concernant les variations au sein de la séquence codante, un nombre moyen de 1 à 2 variations *de novo* par individu a été identifié dans la grande majorité des cohortes (Figure 8C). Il a été observé que d'une manière générale, les études de génome permettaient d'identifier modérément plus que de DNM que les études basées sur l'exome (Figure 8B)[28]. On peut également observer une tendance globale à l'augmentation du nombre de DNM avec le temps, que l'on peut interpréter comme une probable augmentation de la sensibilité des techniques, que ce soit sur les étapes de séquençage ou de traitement bio-informatique des données.

Type de DNM	Taille	Nombre moyen de DNM par génome	Références (PMIDs)
Variation du nombre de copies (CNV)	>50pb	0,05-0,16	28965761, 27525107, 25883321
Insertion/déletion (Indels)	<50pb	2,6-9	28965761, 27525107, 25883321, 25597990
Substitution nucléotidique (SNVs)	1pb	45-89	28959963, 27322544, 24896178, 28965761, 27525107, 25597990, 28959963
Substitution en mosaïque	1pb	0,05-22,2	27525107, 25597990, 26054435, 27632392, 28867142
CNVs en mosaïque	>50pb	5.10-4-7,7.10-3	25634561, 28855261

Tableau 1. Nombre de variations *de novo* par individu
Données adaptées de Wilfert *et al.* : ref[28]. DNM : mutation *de novo*. Pb : paire de bases.

Le nombre de variations du nombre de copies (CNVs, délétions ou duplications de plus de 50pb) est moins précisément caractérisé et dépend fortement de la technique utilisée[28], mais il est bien plus rare, de l'ordre d'un évènement pour 10 individus. Dans la suite de ce chapitre, nous nous concentrerons sur les variations de séquence.

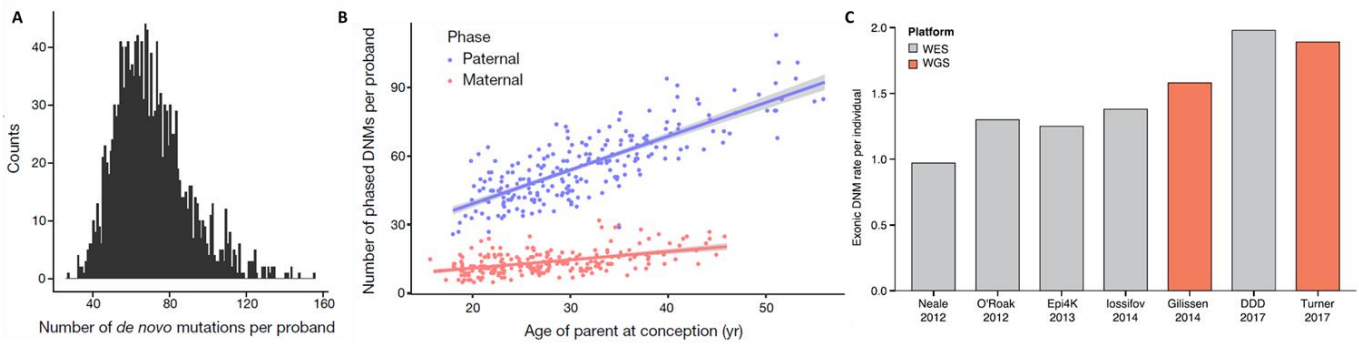


Figure 8. Variations *de novo* dans un génome typique et effet de l'âge parental
A) Distribution du nombre de variations *de novo* par individu au sein de 1 548 génomes. Source : ref[18].
B) Effet de l'age paternel et maternel sur le nombre de variations *de novo* chez 255 individus. Source : ref[18].
C) Nombre moyen de variations *de novo* par individu au sein de la séquence codante selon différentes études basées sur du séquençage d'exome et de génome. Source : ref[28].

Variations en mosaïque post-zygotique

Les évènements *de novo* survenus après la première division du zygote ou mosaïques post-zygotiques, sont bien plus rares que les évènements constitutionnels, mais leur nombre précis n'est pas bien connu, du fait des problématiques techniques spécifiques à leur mise en évidence. En particulier on peut concevoir qu'en diminuant le seuil de détectabilité, le nombre d'évènements détectés puisse augmenter, jusqu'au stade de cellule unique pour lequel on sait que le nombre d'évènements somatiques est de l'ordre de plusieurs milliers par cellule[29,30]. Ainsi la question du nombre de variations *de novo* en mosaïque post-zygotique par individu n'a pas de réponse unique. En revanche, concernant les variations détectables par les pipelines de séquençage classiques, certaines données sont disponibles. Afin d'identifier la proportion de variations somatiques, des auteurs ont évalué par WGS la concordance des mutations *de novo* au sein de 91 paires de jumeaux monozygotes[18]. 2,9 % des 6 034 DNM n'étaient identifiées que chez un des deux jumeaux, dont l'étude ciblée a montré que 57,9 % étaient réellement des variations présentes chez un seul des deux jumeaux. Ainsi, cette étude indique qu'environ 1,7 % des DNM identifiées par un pipeline classique seraient post-zygotiques. Cependant ce chiffre ne fait pas consensus dans la communauté, d'autres

études indiquant des chiffres allant jusqu'à 7 % des DNM[27]. Les variations *de novo* peuvent également être liées à une mosaïque post-zygotique chez un des parents. Ainsi, 3,8 % des DNM seraient identifiables dans au moins 1 % des cellules sanguines d'un des deux parents[31].

Causes et facteurs biologiques influençant le nombre de mutations *de novo*

Les mutations *de novo* résultent d'une lésion de l'ADN ayant échappé aux différents mécanismes de réparation. Les causes de mutations classiquement citées incluent en particulier différents mutagènes exogènes et endogènes, ou les erreurs de réplication. De nombreux facteurs associés à une variabilité du taux de DNM ont été identifiés, tels que le type de nucléotide et le contexte nucléotidique local, le timing de réplication, le taux de recombinaison, certaines contraintes fonctionnelles, et des facteurs épigénétiques[27]. Concernant l'impact de la séquence locale, on peut mentionner les transitions CpG (CpG → TpG), qui sont reconnues depuis très longtemps comme le type de substitution avec la mutabilité la plus élevée (de l'ordre de 10 à 18 fois plus élevée que les autres types de substitutions[32]), et dont l'explication biologique provient du mécanisme spécifique de déamination des cytosines méthylées en uracile.

Effet de l'âge paternel

80,4 % des mutations *de novo* surviennent sur l'allèle paternel[18]. Dès 2012, une relation positive a été observée entre l'âge paternel et le nombre de DNM dans la descendance[33]. Cet effet très important de l'âge paternel (Figure 8B), de l'ordre d'1,51 DNMs supplémentaires par an à la conception[18], explique une grande partie de la variabilité interindividuelle dans le nombre de DNM. Tous les types de substitutions nucléotidiques ont tendance à augmenter avec l'âge paternel, mais l'augmentation des transitions CpG est plus marquée[18]. Par ailleurs, une variabilité dans le timing de réplication a été observée en fonction de l'âge paternel. Pour expliquer cet effet majeur de l'âge paternel sur le nombre de DNM, l'hypothèse d'un stress réplicatif dû au grand nombre de mitoses subies par les spermatogonies âgées (23 mitoses additionnelles par an) a été proposée.

Effet de l'âge maternel

Plus récemment, un effet de l'âge maternel sur le nombre de DNM a également été démontré, bien que largement plus modeste que celui de l'âge paternel (de l'ordre de 0,37 DNMs supplémentaire par an à la conception[18]). L'effet de l'âge maternel a été montré comme étant associé à certaines spécificités moléculaires. En particulier, et à la différence des variations paternelles, l'étude du spectre mutationnel en fonction de l'âge maternel montre une augmentation forte du nombre de transversions C>G avec l'âge. Également, la répartition génomique des DNM est soumise à des biais avec l'âge, correspondant (i) aux clusters mutationnels et (ii) à des régions génomiques de mutabilité globalement augmentée. Les clusters mutationnels sont, chez un individu, des régions au sein desquelles plusieurs DNM surviennent séparées par une distance plus faible que ne le voudrait un modèle aléatoire (classiquement moins de 20kb). Les clusters mutationnels, présents sur les allèles paternels et maternels, augmentent de manière importante avec l'âge maternel. L'hypothèse de l'implication du système de réparation des cassures double brin de l'ADN a été soulevée pour expliquer cette clusterisation excessive. Aussi, certaines régions chromosomiques ont été identifiées comme très significativement enrichies en variations *de novo* d'origine maternelle. Des hypothèses

relatives à la dynamique chromosomique dans les ovocytes vieillissants ont été proposées pour expliquer ces disparités interchromosomiques marquées.

Pour résumer, entre 40 et 100 DNM, incluant en moyenne une à deux DNM au sein de la séquence codante, peuvent être identifiées dans un génome typique. L'âge paternel à la conception est le facteur de risque majeur augmentant le nombre de DNM. L'âge maternel est également mais plus faiblement corrélé au nombre de DNM. Des particularités biologiques de la gamétogenèse mâle et femelle peuvent être mises en avant pour expliquer les différences quantitatives et qualitatives ente les variations paternelles et maternelles. Par ailleurs on ne note à ce jour pas de facteur environnemental connu qui pourrait être associé à une augmentation du nombre de DNM.

1.2.3 Applications du NGS en génétique médicale

Depuis le début des années 2010, le NGS s'est imposé comme une technologie incontournable dans le diagnostic des maladies génétiques. Le screening des variations de séquence, et plus récemment des variations du nombre de copies, est aujourd'hui largement basé sur le NGS. Dans ce chapitre, nous évoquerons les différentes stratégies d'analyse à visée diagnostique, puis nous aborderons la question de l'interprétation des variations, désormais bien balisée par l'existence de recommandations précises. Enfin nous détaillerons certaines sources de données utiles dans ce processus d'analyses en NGS à visée diagnostique.

1.2.3.1 Stratégies ciblées vs pangénomiques

Dans le champ du diagnostic génétique médical, le NGS a été exploité d'une part par des approches ciblées de type panel de gènes, et d'autre part par des approches pangénomiques correspondant à l'exome, ou au génome.

Chaque méthode présente des avantages et défauts et répond à des problématiques différentes, représentées de manière schématique dans le Tableau 2. Les panels de gènes représentent une stratégie de choix lorsque la maladie étudiée est bien caractérisée sur le plan phénotypique avec une hétérogénéité génétique limitée et une connaissance des bases moléculaires stable dans le temps. Un bon exemple peut être le contexte du diagnostic des prédispositions aux tumeurs colorectales, pour lequel une approche par petit panel donne d'excellents résultats[34]. L'exome est une solution idéale en cas d'hétérogénéité génétique importante, d'évolution rapide des connaissances ou en cas d'exploitation en parallèle des données pour la recherche de nouvelles causes génétiques à la pathologie. L'exemple le plus commun est le diagnostic des maladies du développement et en particulier de la déficience intellectuelle[35]. Le séquençage du génome représente certainement le prolongement logique du séquençage d'exome, avec un bénéfice en termes de couverture et de capacité de détection de variations, notamment de structure, au prix d'une augmentation importante de la complexité de *management* des données et du coût.

	Panel de gènes	Exome 50-100x	Génome 25-50x
Rendement diagnostique	+	+++	+++
Détection des variations structurales	+	+	+++
Détection des mosaïques	+++	+	-
Prix des réactifs par échantillon	+/++	++	++++
Traitement et stockage des données	+	++	++++
Découvertes incidentales	+/-	+	+
Potentiel de ré-analyse ou de méta-analyse	+	+++	+++
Potentiel d'identification de nouvelles bases moléculaires	-	+++	++++

Tableau 2. Avantages et défauts des méthodes de NGS ciblé vs pangénomiques : éléments généraux

1.2.3.2 Recommandations d'interprétation des variations génétiques

L'interprétation des variations génétiques consiste à évaluer le lien de causalité entre les variations génétiques identifiées et le phénotype observé. Le grand nombre d'éléments à prendre en compte, comprenant une certaine part de subjectivité (concordance du phénotype par exemple), est à l'origine de discordances dans la classification des variations entre et au sein même des laboratoires de génétique moléculaire[36].

De manière à harmoniser les pratiques dans ce processus d'interprétation des variations génétiques constitutionnelles, l'*American College of Medical Genetics and Genomics* et l'*Association for Molecular Pathology* (ACMG-AMP) ont établi en 2015 des recommandations qui représentent aujourd'hui un cadre largement répandu dans la communauté génétique[37]. Des arguments standardisés en faveur et à l'encontre de l'implication des variants ont été identifiés et côtés, et un algorithme a été implémenté pour produire les conclusions à partir des arguments renseignés. Ces conclusions sont sous la forme de cinq classes : variation bénigne (classe 1), variation probablement bénigne (classe 2), variation de signification inconnue (classe 3), variation probablement pathogène (classe 4) et variation pathogène (classe 5). Le terme *probablement* a été défini comme « au moins 90 % de probabilité ». Les arguments pris en compte sont séparés en huit groupes : données de fréquence allélique, prédictions in silico, données fonctionnelles, données de ségrégation, données alléliques, classification dans les bases de données, présentation phénotypique. Ce *framework* d'interprétation des variations n'a bien sûr pas réglé l'intégralité des discordances[38], mais a permis aux équipes de s'accorder sur les points d'interprétations et leur poids. Devant le succès de ces règles, certaines équipes ont proposé des outils d'aide à la cotation ACMG, que ce soit par des formulaires en ligne[39] ou par des méthodes d'annotation des variations permettant dans une certaine mesure l'automatisation de cette classification[40,41]. Des ajustements de ces règles ont également été proposés[42].

Du fait du bien-fondé et de l'implantation importante de ces recommandations au sein des laboratoires de diagnostic, l'Association Nationale des Praticiens de Génétique Moléculaire (ANPGM), a proposé à l'été 2018 des recommandations françaises largement basées sur celles de l'ACMG (voir ressources web). Chaque point d'interprétation a été détaillé et des ressources recommandées y ont été associées. Certains points de détail ont par ailleurs été ajustés par rapport

aux recommandations américaines. De même, des recommandations anglaises, éditées par l'*Association for Clinical Genomic Science* (ACSG) ont également validé les recommandations ACMG comme base (voir ressources web).

1.2.3.3 Bases de données de variations en pathologie humaine

L'interprétation des données de NGS nécessite une confrontation des variations identifiées avec les connaissances scientifiques les plus à jour possibles. Ces connaissances sont organisées dans des bases de données exploitables de manière automatisée lors du processus d'annotation des variants, qui sont utilisés pour appliquer les recommandations ACMG-AMP. En dehors des bases de données de variations en population générale présentées précédemment, ce chapitre décrit dans un premier temps les bases de données de variations génétiques identifiées chez des patients les plus communément employées. Nous aborderons dans un second temps une ressource récente de variations *de novo*, appelée denovo-db, et qui sera utilisée dans la suite de ce travail. Enfin nous listerons d'autres ressources techniques utiles à l'interprétation des résultats génétiques. De nombreuses bases de régulation des gènes, de conformation tridimensionnelle de la chromatine, de méthylation, etc. sont également disponibles et d'une certaine utilité dans l'interprétation des résultats génétiques mais ne seront pas présentées en détail ici.

Bases de données de variations en pathologie humaine : ClinVar et HGMD

ClinVar et HGMD sont deux bases cataloguant les variations identifiées dans le cadre du diagnostic des maladies génétiques. Ces deux bases ont un fonctionnement très différent et sont complémentaires.

ClinVar est une initiative américaine du National Institute of Health agrégeant des variations génétiques humaines et leur signification clinique vis-à-vis des maladies génétiques[43]. Les données de ClinVar sont produites par les membres des communautés médicale et scientifique qui mettent à disposition de la collectivité leurs résultats dans un effort commun. Les laboratoires de diagnostic, les laboratoires de recherche et certains panels d'experts peuvent ainsi partager de manière proactive leurs interprétations et depuis quelque temps les éléments de preuve ayant permis de conclure. Cet aspect collaboratif est à la fois un élément extrêmement puissant de par la multiplicité des sources et la quantité de variations soumises, mais peut également être une limite car la pertinence des données n'est pas garantie. Afin de contourner cette critique potentielle, ClinVar a mis en place un système de classement des variations en fonction de leur degré de confiance, allant de 0 à 4 étoiles (Figure 9A). Ainsi une variation soumise sans indication des critères employés sera au niveau le plus bas (0 étoile). Une soumission unique indiquant la méthode de classification (« par exemple : application des recommandations ACMG ») ou plusieurs soumissions discordantes sont marquées d'une étoile. Plusieurs soumissions indépendantes concordantes sont identifiées par deux étoiles. Les variations validées par des groupes experts sont identifiées par 3 étoiles et celles incluses dans des recommandations de bonnes pratiques sont identifiées par 4 étoiles. La Figure 9B représente la signification clinique associée aux variations, montrant que les variations de signification inconnue sont prépondérantes dans ClinVar. On peut faire l'hypothèse qu'une part de ces variations de signification inconnue représente la classification par défaut par manque

d'information par certains utilisateurs soumettant à ClinVar de manière automatisée, venant « polluer » la base. Malgré la présence de nombreuses variations mal classifiées dans ClinVar[44], et son caractère très partiel car basé sur les soumissions volontaires, cette ressource reste néanmoins d'un intérêt majeur dans l'analyse des résultats de NGS.

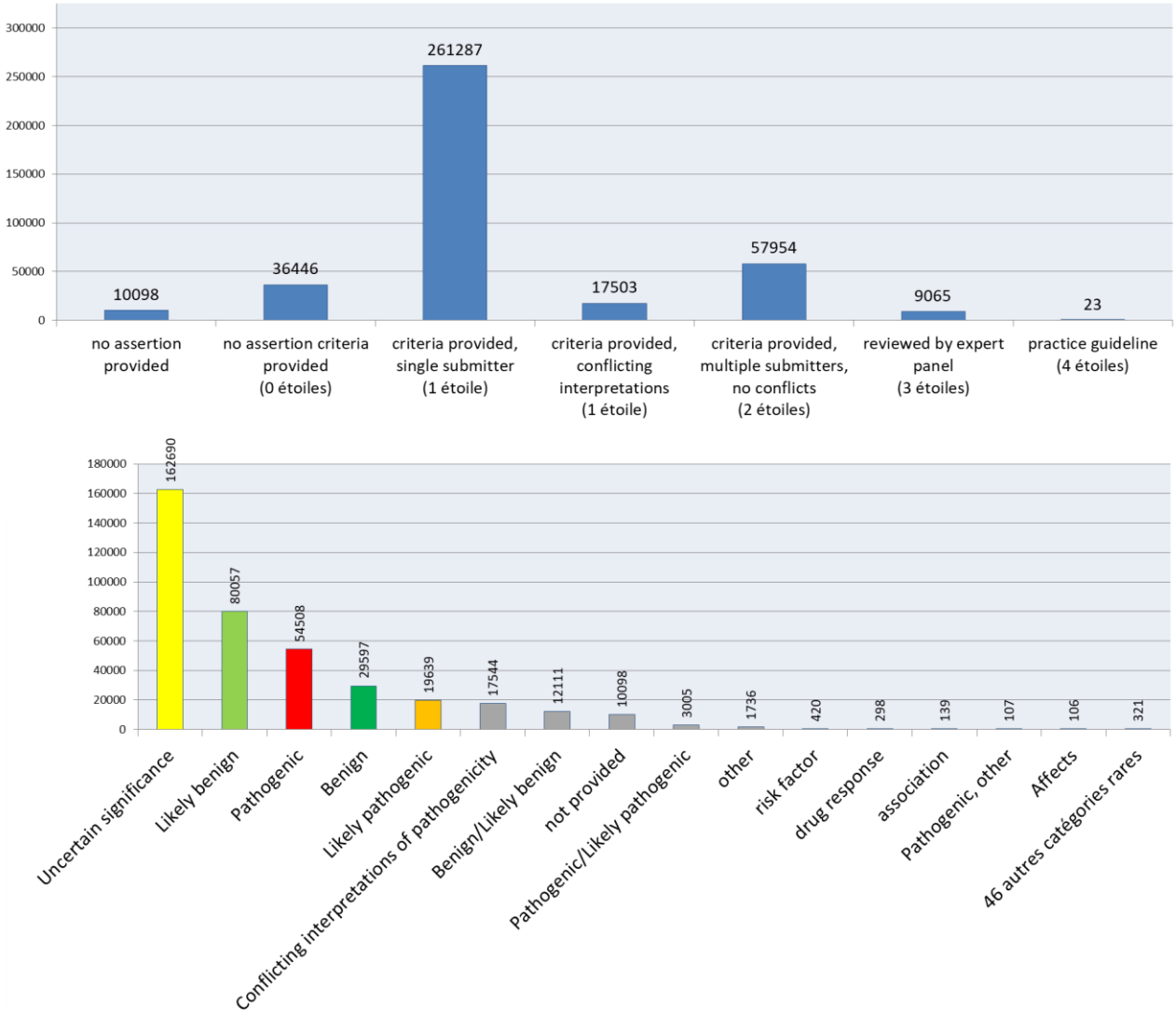


Figure 9. Caractéristiques des variations de la base ClinVar
A) Degré de confiance des variations. B) Signification clinique. De nombreuses classes de signification clinique très rares ont été agrégées dans un souci de visualisation des données.

HGMD, pour *Human Gene Mutation Database*, est une initiative de l'Université de Cardiff initiée en 1996 et financée par Qiagen. Contrairement à ClinVar, les variations ne sont pas incluses par la communauté mais par une équipe dédiée qui ajoute de manière systématique et prospective les variations publiées dans la littérature scientifique[45]. Dans sa version payante complète, HGMD comprend 224 642 variations extraites de la littérature, dont la grande majorité correspond à des variations considérées comme pathogènes (classe « DM », *disease-causing mutations* ; NB nous ne décrirons pas ici les différentes classifications des variations par HGMD qui ne sont pas celles utilisées par l'ACMG). Même si comme ClinVar, HGMD comprend des faux positifs, soit des variations neutres indiquées comme pathogènes, cette base constitue une ressource utile et

complémentaire à ClinVar dans le sens où elle permet d'identifier la présence de littérature scientifique associée aux variants, ce que ne permet pas directement ClinVar. La version gratuite de HGMD est quant à elle très limitée dans ses fonctionnalités et son contenu (données avec un délai de 3,5 ans).

Autres bases de données

En dehors de HGMD et ClinVar, il existe de nombreuses bases de données spécifiques des gènes et/ou des maladies. Concernant les bases gènes spécifiques, on peut citer les bases LOVD (pour *Leiden Open Variation Database*)[46], qui ont l'avantage d'être potentiellement mieux curées que ClinVar et HGMD, mais dont le caractère locus-spécifique limite leur utilisation dans une logique génomique. La base SwissVar[47] est spécifique des variations faux-sens. Cette base, correspondant à une des branches du réseau UniProtKB visant à centraliser les annotations fonctionnelles protéiques, est implémentée manuellement à partir de la littérature. La base MitoMap[48] est quant à elle spécifique des variations mitochondriale. Enfin, de nombreuses bases de données spécifiques de certaines maladies ont été produites par divers consortia et groupes experts, souvent de haute qualité et très exhaustives par rapport aux bases généralistes. En revanche, leur mode d'interrogation (interface web de format variable) et leurs classifications diverses de pathogénicité des variants les rendent non utilisables au stade de l'annotation automatisée. On peut citer comme exemples la base Alzforum/mutations qui répertorie les variations causales de formes autosomiques dominantes de maladie d'Alzheimer, la Rettbase qui liste les variants du gène *MECP2* à l'origine du syndrome de Rett, ou encore la *Cystic Fibrosis Mutation Database* concernant les variations associées à la mucoviscidose. On peut également citer l'initiative française *Universal Mutation Database* (UMD), proposant une architecture pour la création de bases de données locus-spécifiques[49], et comprenant à ce jour 37 bases.

Base de données de mutations *de novo* : denovo-db

Denovo-db est une base de données récente issue des grandes études de génomique, proposée par une équipe de l'Université de Washington à Seattle, et financée par plusieurs organismes[50]. L'objectif de denovo-db est de répertorier de manière systématique les variations *de novo* identifiées dans divers contextes cliniques et scientifiques, notamment par études d'exomes et génomes en trio. Pour cela, les auteurs de denovo-db ont recherché dans la littérature des études ayant identifié des variations *de novo* par technologie NGS. Par exemple sont inclus dans denovo-db les variations *de novo* identifiées chez des patients avec anomalies du développement au sein de la cohorte anglaise DDD[51], ou bien les variations *de novo* identifiées dans des grandes séries de patients avec autisme. Les auteurs ont extrait les informations essentielles pour chaque variation, incluant le nom de l'échantillon utilisé dans la publication originale, la position chromosomique du variant, l'allèle référence et l'allèle alternatif, ainsi que la validation potentielle en séquençage Sanger cher l'individu et ses parents. Une attention particulière a été apportée à la suppression des doublons, par exemple certains patients inclus dans la cohorte d'autistes SSC (Simon Simplex Collection). Dans sa dernière version 1.6, publiée en juillet 2018, denovo-db inclut 53 études concernant 21 phénotypes, incluant la SCC qui est traitée séparément. Au total, denovo-db 1.6 comprend 413 778 variants chez 15 668 individus. L'avantage de denovo-db est d'être une base

systématique, indépendante de toute interprétation. Dans la 3ème partie de cette thèse, cette base denovo-db sera au centre d'une stratégie permettant l'identification de nouvelles variations responsables de maladies du développement.

Autres ressources techniques utiles à l'interprétation des résultats génétiques

De nombreuses autres sources de données sont indispensables à prendre en compte afin d'évaluer une variation génétique. Ces ressources contiennent des informations techniques sur la complexité du génome, sur la complexité des gènes et transcrits ainsi que sur l'architecture des protéines. Concernant l'évaluation de la structure locale du génome, le navigateur du génome *UCSC genome browser*, par l'Université de Santa Cruz, Californie, est une solution permettant d'afficher de manière simultanée plusieurs types de régions complexes en un seul affichage[52]. Les petits éléments répétés et complexes sont regroupés dans la piste d'annotation RepeatMasker[53], et les répétitions de grande taille sont identifiables dans les pistes Segmental Dups[54] et Self Chain (UCSC). La prise en compte de l'organisation et de l'expression des transcrits peut donner des arguments lors de l'évaluation de variations affectant par exemple des exons alternatifs. Les bases de transcrits de RefSeq[55] et d'Ensembl[56] sont les plus utilisées, et sont également accessibles depuis UCSC. Les niveaux d'expression des transcrits dans les différents tissus sont accessibles depuis peu via la piste GTEx Transcript[57] basée sur les transcrits Ensembl. Concernant l'évaluation de l'architecture des protéines, notamment l'identification des domaines fonctionnels et structures, la base Uniprot[58] semble incontournable.

Suite à ce rapide tour d'horizon relatif au séquençage à haut débit dans le cadre de l'étude du génome humain, le chapitre suivant traite plus spécifiquement des variations des variations faux-sens, qui entrainent la substitution d'un acide aminé au sein des protéines, et dont l'interprétation peut être difficile.

1.3 Interprétation des variations faux-sens dans un cadre diagnostique

La première observation d'une variation faux-sens remonte à l'année 1957 où le Pr Vernon Ingram, qui étudiait la drépanocytose, tirait les conclusions suivantes : «*I have now found that out of nearly 300 amino-acids in the two proteins, only one is different ; one of the glutamic acid residues of normal haemoglobin is replaced by a valine residue in sickle cell anaemia haemoglobin*»[59]. La première mention du terme *missense* dans la base Pubmed remonte à 1964[60], en pleine période d'effort scientifique collaboratif visant à décrypter le code génétique[61], où les auteurs prêtaient à ces variations alors hypothétiques des rôles de régulation de la protéine. Depuis ces découvertes, les variations faux-sens se sont rapidement imposées comme le type de variations codantes les plus commun, à la fois en tant que polymorphisme et comme cause de maladies génétiques. Dans cette troisième partie introductive dédiée au défi d'interprétation des variations faux-sens, nous listerons les mécanismes par lesquels les variations faux-sens peuvent être à l'origine d'un effet délétère, nous aborderons les solutions *in silico* permettant de distinguer les variations faux-sens ayant un potentiel de pathogénicité, puis les approches génomiques ayant permis de caractériser leur distribution et leur impact, en population générale et en population atteinte.

1.3.1 Effets des variations faux-sens : de la biologie fonctionnelle à la génomique

Le mécanisme de pathogénicité des variations génétiques peut être évalué par deux approches complémentaires. D'une part, des tests fonctionnels réalisés sur des prélèvements des patients porteurs, ou bien dans des modèles cellulaires ou animaux dans lesquels la mutation a été introduite, peuvent permettre de mettre en évidence directement l'effet potentiel quantitatif ou qualitatif de la variation sur l'ARN, sur la protéine ou sur la voie biologique impliquée. Alternativement, l'agrégation de plusieurs patients porteurs de variations dans le même gène et ayant un phénotype commun permet d'avoir une vision globale du mécanisme mutationnel, permettant d'inférer l'effet fonctionnel associé à la maladie. Ainsi l'effet fonctionnel d'une variation, et *a fortiori* d'une variation faux-sens, peut être décrit d'un point de vue de biologie fonctionnelle ou d'un point de vue génomique intégratif, avec un vocabulaire distinct.

1.3.1.1 Effet fonctionnel des variations faux sens : point de vue de la biologie fonctionnelle

La substitution d'un acide aminé par un autre au sein d'une protéine peut avoir des effets extrêmement variés. On imagine aisément qu'une variation faux-sens équivalente n'affectera pas une protéine linéaire de la même manière qu'une protéine avec repliement tridimensionnel complexe. Les différents types d'effet moléculaire délétères des variations faux-sens ont été passés en revue par Stefl *et al*. en 2013[62] et sont repris dans le Tableau 3. Des exemples de maladies liées à chacun des effets sont disponibles dans le travail original.

Type d'altération moléculaire	Effet
Modification de la stabilité thermodynamique de la protéine	Déstabilisation
	Stabilisation
Modification des réseaux de liaison hydrogènes	Effet sur la structure de la protéine
	Effet sur l'agrégation / le repliement
	Effet sur la flexibilité
	Localisation subcellulaire
	Sensibilité aux modifications du pH
	Effet sur l'activité enzymatique
Modification de la dynamique conformationnelle	Effet sur la flexibilité conformationnelle
	Effet sur les régions désordonnées (*disordered regions*)
	Effet sur les transitions ordre-désordre (*order-disorder transitions*)
	Agrégation
	Mobilité locale

Tableau 3. Effet moléculaire des variations faux-sens selon Stefl *et al.*

Cette classification ne propose pas la classique opposition perte *versus* gain de fonction, probablement car il n'y a pas d'équivalence directe entre l'effet moléculaire (exemple : effet sur la flexibilité de la protéine), et l'effet fonctionnel global (exemple : perte de fonction), qui semblent être variables en fonction des protéines. Ces différents types d'effets fonctionnels avaient été proposés en 1932 par Muller (*Muller's morphs*)[63] et sont listés dans le Tableau 4. Ils comprennent les effets perte de fonction (« hypomorphe » et « amorphe »), et les effets gain de fonction

(« hypermorphe », « antimorphe » et « neomorphe »). Si certains termes ne sont aujourd'hui plus utilisés, les concepts de cette classification sont néanmoins toujours actuels.

Type d'altération moléculaire	Effet	Définition
Perte de fonction	Amorphe	Perte de fonction complète de l'allèle
	Hypomorphe	Perte de fonction partielle de l'allèle
Gain de fonction	Hypermorphe	Augmentation de la fonction normale du gène
	Antimorphe	Effet dominant négatif, fonction du gène opposée à la fonction normale
	Neomorphe	Apparition d'une nouvelle fonction de la protéine

Tableau 4. Effet fonctionnel des mutations sur les gènes selon Muller (1932)

Quelle que soit la terminologie utilisée, les généticiens moléculaires ont exceptionnellement accès à des études fonctionnelles. En revanche, des arguments génomiques peuvent aider à prédire l'effet fonctionnel d'une variation. Cependant, la nature des arguments utilisés ne permet pas de calquer la terminologie utilisée dans les études fonctionnelles.

1.3.1.2 Effet fonctionnel des variations faux-sens : point de vue génomique

L'observation globale des variations faux-sens par des approches génomiques peut également apporter des éléments permettant d'inférer des propriétés fonctionnelles aux variations associées à une maladie. Si par exemple au sein d'une cohorte de 10 patients atteints d'une maladie similaire, sept étaient porteurs d'une variation tronquante et trois d'une variation faux-sens, réparties tout le long du gène, le mécanisme de perte de fonction par haploinsuffisance (HI) ne ferait aucun doute. En revanche, si les dix patients étaient porteurs d'une variation faux-sens au sein d'un même domaine fonctionnel de la protéine, un mécanisme non-haploinsuffisant (NHI), de type gain de fonction ou effet dominant négatif par exemple, serait très probable. Un exemple de ces deux situations est présenté en Figure 10, comparant d'une part le gène *ARID1B*, responsable du syndrome de Coffin Siris et typique du mécanisme HI, et d'autre part le gène *CDK13*, responsable d'une maladie du développement de description récente et typique d'un mécanisme NHI.

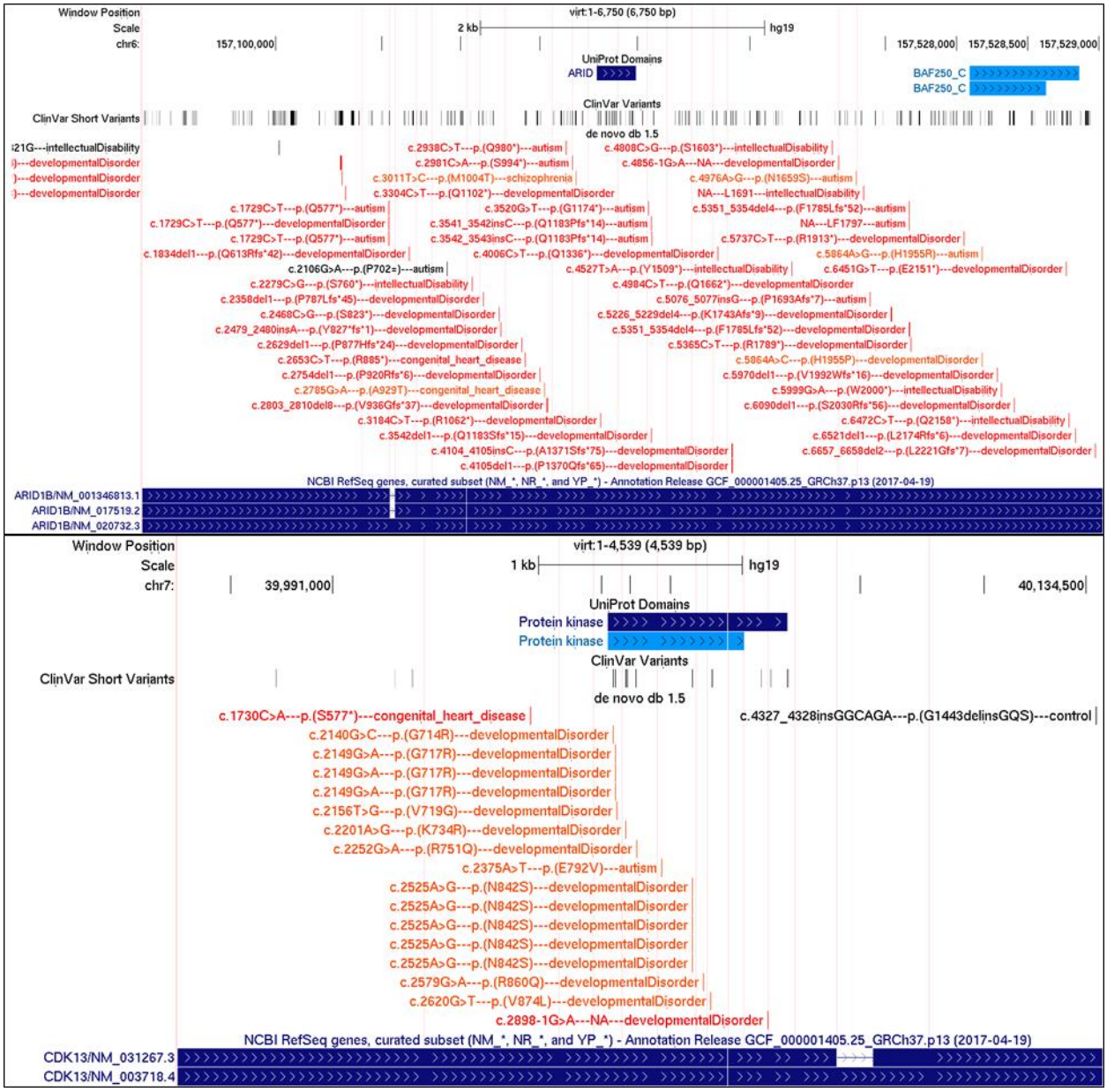


Figure 10. Exemple de *patterns* mutationnels typiques de mécanisme haploinsuffisant et non-haploinsuffisant

Vues du browser UCSC représentant les transcrits de chaque gène sans les introns, avec comme pistes d'annotation les domaines protéiques selon Uniprot et les variations répertoriées dans les bases ClinVar et denovo-db. Concernant la piste denovo-db, les variations faux-sens sont indiquées en orange, et les variations tronquantes apparaissent en rouge. A) Mécanisme d'haploinsuffisance : gène *ARID1B*, montrant des variations majoritairement tronquantes réparties de manière aléatoire et indépendante des domaines fonctionnels. B) Mécanisme non-haploinsuffisant : gène *CDK13*, montrant des variations majoritairement de type faux-sens, avec clusterisation forte dans un domaine fonctionnel de la protéine et récurrence mutationnelle.

Le consortium anglais DDD a estimé par une approche globale qu'environ 63 % des variations faux-sens *de novo* responsables de maladies du développement identifiées au sein de leur cohorte agissaient par un mécanisme non-haploinsuffisant (« *altered function* »), le reste agissant par haploinsuffisance[51]. De même, l'analyse de la répartition spatiale des variations au sein d'une cohorte, visant à identifier un biais de répartition de type clusterisation excessive, permet d'apporter des informations très précieuses dans le mécanisme de l'effet mutationnel (voir chapitre 1.3.3).

Ainsi, les tests fonctionnels, indispensables dans certains contextes, comme la caractérisation fine de l'effet des variations, mais aussi par la suite en tant que biomarqueur des variations à effet, peuvent être orientés de manière précise par des données génomiques.

Pour résumer, la Figure 11 reprend de manière schématique les principaux types d'effets fonctionnels des variations de séquence impliquées en pathologie, en fonction des types de mutation.

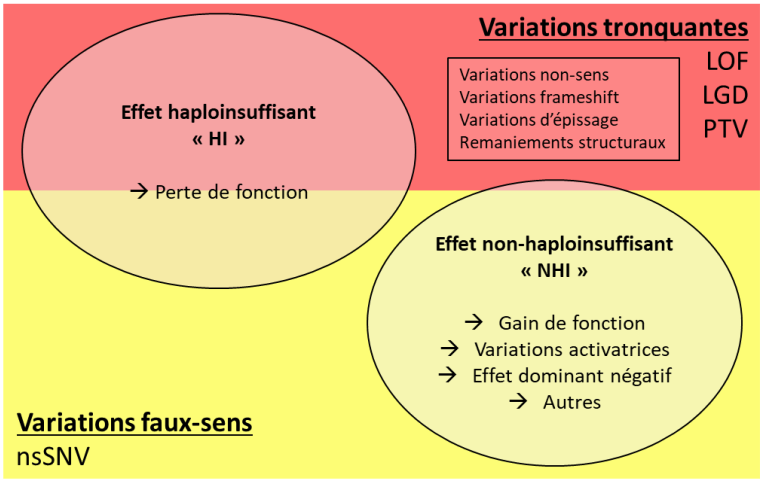


Figure 11. Lien entre les types de variations de séquence et les types d'effets fonctionnels
Les deux grands types d'effet, HI et NHI, sont représentés par rapport aux types de variations de séquence les plus communs : les variations tronquantes et variations faux-sens. Des termes synonymes de « variations tronquantes » et « variations faux-sens » communément employés dans la littérature sont indiqués. LOF : loss-of-function, LGD : likely gene disrupting, PTV : protein truncating variant, nsSNV : non-synonymous single nucleotide variant.

1.3.2 Outils de prédiction de pathogénicité des variations faux-sens

La difficulté de prioriser les variations faux-sens en fonction de leur probabilité d'effet biologique est une problématique très ancienne. Afin de distinguer les variations faux-sens à impact fort des variations neutres sur le plan biologique, de nombreux algorithmes ont été proposés, basés sur des modes de fonctionnement différents[64]. Ce chapitre passe en revue les trois types de scores disponibles, qui sont du plus simple au plus intégratif les scores de conservation, les scores de prédiction fonctionnelle et les scores d'ensemble. Les différents éléments pris en compte pour ces prédictions seront évoqués, avec en tête de files pour les scores de prédictions fonctionnelles la conservation de la séquence (homologie) et l'effet sur la structure tridimensionnelle de la protéine. Les méthodes permettant d'aboutir à des scores de prédiction à partir de ces données, désormais largement basées sur le *machine learning*, seront citées. Par la suite nous analyserons les stratégies d'évaluation et de comparaison de ces scores avant de conclure sur les intérêts et limites de ces différents scores dans le diagnostic génétique des maladies rares.

1.3.2.1 Scores de conservation

Les scores de conservation interespèces ne sont pas à proprement parler des logiciels de prédiction d'effet des variations mais plutôt des mesures relatives à la séquence étudiée (d'ADN ou de la

protéine), qui ne prennent pas en compte la possibilité d'un allèle alternatif. Ces scores sont basés sur des alignements des séquences nucléotidiques et protéiques de nombreuses espèces retraçant un arbre phylogénétique plus ou moins large. Certains scores sont déclinés en plusieurs versions en fonction du nombre d'homologues pris en compte, par exemple des homologues vertébrés ou uniquement des homologues mammifères. Ces scores diffèrent entre eux par le modèle permettant de définir la contrainte ou au contraire la divergence entre les espèces. Les principaux scores utilisés sont indiqués en Tableau 5.

Score	PMID	Année	Algorithme	Principe du fonctionnement
phastCons	16024819	2005	HMM	Basé sur un modèle statistique d'évolution de séquence
GERP	15965027	2005	-	Score basé sur la quantification de la sous-représentation de substitutions
SiPhy	19478016	2009	HMM	Score de conservation nucléotidique en fonction du spectre de variabilité interespèces
GERP++	21152010	2010	-	Évalue le maximum de vraisemblance du taux d'évolution au locus
phyloP	19858363	2010	HMM	Compare la variabilité interespèces avec un modèle nul de dérive génétique

Tableau 5. Scores de conservation
HMM : *Hidden Markov Model*

1.3.2.2 Scores de prédiction fonctionnelle des variations faux-sens

Les scores fonctionnels sont des algorithmes dont l'objectif final est d'obtenir la meilleure distinction entre les variations faux-sens neutres des variations délétères. Ces scores intègrent de nombreux éléments identifiés comme potentiellement associés à un caractère délétère, telles que la contrainte évolutive, la déstabilisation thermodynamique de la protéine, l'altération de sites fonctionnels d'importance, etc. De très nombreux scores ont été proposés (Tableau 6), basés sur des modèles de pathogénicité, des méthodes d'intégration et des jeux de données tests différents. Nous présentons ici les principales données employées par ces logiciels, correspondant à l'homologie de séquence, les modélisations structurales et des éléments additionnels, puis nous évoquerons ensuite les différents moyens d'intégration de ces données pour la création des scores de prédiction.

Homologie de séquence

Le principe de l'utilisation de l'homologie de séquence dans la prédiction fonctionnelle d'une variation faux-sens repose sur l'hypothèse qu'une forte conservation interespèces d'un acide aminé témoigne d'une contrainte biologique forte à cette position, et *de facto* d'un potentiel effet délétère de la substitution de l'acide aminé. Au contraire, une position peu conservée dans l'évolution oriente vers l'existence d'une tolérance à la variation à cette position, et donc une probabilité de pathogénicité plus faible. Par leur fonctionnement, les prédictions basées sur l'homologie présupposent que (i) tous les homologues ont une fonction et une contrainte fonctionnelle identique entre les espèces, et (ii) que chaque changement de résidu entre les espèces a un effet indépendant, sans interaction fonctionnelle avec le reste de la séquence protéique spécifique à l'espèce (épistasie). Ainsi ces deux points représentent des limites de ces approches. Aussi, la performance

des logiciels basés sur l'homologie est très dépendante de l'alignement interespèces et a tendance à diminuer dans les gènes ayant peu d'homologues[65].

Le logiciel basé sur l'homologie le plus utilisé est SIFT[66], qui recherche de manière directe la présence de la substitution nucléotidique parmi les séquences homologues. La génération suivante d'algorithmes a intégré certaines particularités biochimiques des résidus, telles que l'encombrement, la polarité, l'hydrophobie ou la charge, afin de déterminer un spectre de variabilité au sein des homologues, dans le but d'identifier si la variation étudiée était située en dehors de ce spectre, témoignant d'un potentiel effet délétère.

Récemment a été proposé l'algorithme PrimateAI, basé non pas sur la comparaison de l'homologie avec la séquence de référence dans les autres espèces, mais à partir des variations communes dans ces espèces[67]. Le rationnel vient du constat que les variations communes dans différentes espèces de grands singes sont très majoritairement bénignes chez l'homme. Du fait d'une grande variabilité génétique chez les grands singes (contrairement à l'homme qui a une population effective réduite du fait de *goulots d'étranglement* récents[68]), ces variations communes représentent un pool important de variations bénignes, qui associées à des données de structure, ont pu être utilisées pour entraîner l'algorithme de prédiction PrimateAI.

Modélisations structurales

De nombreux algorithmes de prédiction d'effet des variations faux-sens se basent sur la prédiction de modification des propriétés physico-chimiques de la protéine, et en particulier de la conformation tridimensionnelle. Le rationnel repose sur le fait qu'une grande proportion des variations faux-sens responsables des maladies mendéliennes est à l'origine d'une déstabilisation thermodynamique de la protéine[69]. Ces scores sont basés sur des modélisations empiriques ou statistiques d'efficacité énergétique, nécessitant habituellement comme base la modélisation tridimensionnelle de la région protéique, qu'elle soit expérimentale ou inférée par homologie d'après des structures connues. Les méthodes basées sur la modélisation de l'effet structural des variations permettent par exemple l'identification d'une perte ou d'un gain d'interactions disulfure, hydrophobes ou électrostatiques. Ainsi, la limite principale des algorithmes basés sur la structure repose sur la nécessité d'une structure protéique tridimensionnelle bien déterminée.

Autres éléments pris en compte

En dehors des informations d'homologie et des modélisations structurales, d'autres éléments sont classiquement pris en compte dans les scores de prédiction de la pathogénicité. Ces éléments incluent les propriétés physico-chimiques de la substitution, certaines annotations fonctionnelles de la protéine, ainsi que d'autres sources d'information plus rarement intégrées, telles que les propriétés d'agrégation, les régions intrinsèquement désordonnées, ou encore les réseaux d'interaction protéine-protéine[64].

Concernant l'analyse des propriétés physico-chimiques des substitutions, le score de Grantham est classiquement utilisé[70]. Pour chaque substitution possible, la matrice de distance de Grantham indique la similarité physico-chimique des deux acides aminés, basée sur la composition atomique du résidu, sa polarité et le volume moléculaire, et dont les valeurs varient entre 5 (Leucine ↔

Isoleucine, distance physico-chimique peu importante) et 215 (Tryptophane ↔ Cystéine, distance physico-chimique importante).

Intégration des données pour l'élaboration des outils de prédiction

La plupart des logiciels de prédictions prennent en compte à la fois des informations de structure, d'homologie et des éléments supplémentaires. La création de ces algorithmes de prédiction consiste généralement en un apprentissage automatique supervisé, basé sur des grands jeux de données avec pour objectif de pondérer le poids des différentes sources d'information de manière à obtenir la meilleure séparation possible entre les variations bénignes et les variations pathogènes. Une des limites de l'approche par *machine learning* vient du fait que par définition, ces algorithmes ne peuvent être performants qu'en présence de données similaires aux données utilisées pour l'entraînement. En particulier, ces stratégies peuvent être mises en défaut en cas de situation atypique sur le plan moléculaire, absentes du jeu de données d'entraînement. Les algorithmes de *machine learning* utilisés incluent par exemple les machines à vecteur de support (SVM), les forêts aléatoires (RF), ou plus récemment les réseaux neuronaux profonds (DNN). Le Tableau 6 reprend les principaux logiciels de prédictions fonctionnelles et leur mode de fonctionnement.

Nom	PMID	Année	Type	Algorithme	Résumé du fonctionnement
SDM	9051729	1997	Structure	-	Calcule la stabilité des protéines en utilisant un modèle statistique
SIFT	11337480	2001	Homologie	PSSM	Utilise la fréquence de substitution chez les homologues
Dmutant	12381853	2002	Structure	-	Modèle statistique pour estimer ddG
Polyphen	12202775	2002	Hybride	DT	Arbre de décision basé sur l'homologie, certaines annotations fonctionnelles et des caractéristiques physico-chimiques de la substitution
Panther	12952881	2003	Homologie	HMM	Utilise la conservation des résidus chez les homologues
LogR.E-value	14751981	2004	Hybride	-	Compare la capacité de la séquence wild-type et la séquence mutante à s'ajuster à la modélisation du domaine
LS-SNP	15827081	2005	Hybride	SVM	Utilise les propriétés structurelles des modèles protéiques, l'accessibilité aux solvants, et la conservation évolutive
MAPP	15965030	2005	Homologie	-	Compare les propriétés physico-chimiques d'une substitution avec les variations à une position donnée chez les homologues
MUpro	16372356	2005	Hybride	SVM	Calcule la stabilité des protéines, en utilisant des informations de séquence et de structure
nsSNPAnalyzer	15980516	2005	Hybride	RF	Utilise la fréquence de substitution dans les homologues et des paramètres structurels
pmut	15879453	2005	Hybride	NN	Utilise les propriétés des acides aminés, l'homologie, certaines annotations fonctionnelles, et des prédictions de structure
SNPeffect	15608254	2005	Hybride	-	Utilise des mesures du repliement, de la stabilité et d'agrégation des protéines, et des informations fonctionnelles (domaines, localisation subcellulaire)
A-GVGD	16014699	2006	Homologie	-	Compare la distance physicochimique de la substitution avec la variabilité au sein des orthologues
PhD-SNP	16895930	2006	Hybride	DT/SVM	Utilise les informations de séquence locale et des informations d'homologie
SNPs3D	16551372	2006	Hybride	SVM	Utilise des informations d'homologie et d'accessibilité stérique, électrostatique et du solvant
Imutant 3.0	18387208	2007	Hybride	SVM	Calcule ddG à l'aide de données thermodynamiques expérimentales
Parepro	18005451	2007	Hybride	SVM	Utilise les probabilités de substitution dans les homologues, les propriétés physico-chimiques de la substitution et les informations sur les acides aminés voisins
SAPRED	17384424	2007	Hybride	SVM	Utilise la conservation des résidus, les paramètres de structure, l'annotation fonctionnelle, les régions désordonnées et les propriétés d'agrégation
SNAP	17526529	2007	Hybride	NN	Utilise la fréquence de substitution chez les homologues, la structure secondaire prédite et l'accessibilité aux solvants, et des caractéristiques physicochimiques
LRT	19602639	2009	Homologie	-	Modèle basé sur la conservation de l'ADN
MutPred	19734154	2009	Hybride	RF	Modèle basé sur SIFT et sur 14 fonctions structurelles et fonctionnelles
PoPMuSiC	19654118	2009	Structure	NN	Utilise des potentiels statistiques pour calculer ddG
SNPs&GO	19514061	2009	Homologie	SVM	Utilise des informations de séquence protéique, de profil de séquence et de fonction
AUTO-MUTE	20573719	2010	Hybride	RF	Utilise la fréquence de substitution dans les homologues et des paramètres structurels
Mutation Taster	20676075	2010	Hybride	NBC	Combine la conservation évolutive, l'effet sur l'épissage, la perte de caractéristiques protéiques et l'effet sur l'expression de l'ARNm
Polyphen2	20354512	2010	Hybride	NBC	Utilise l'homologie, des paramètres structurels, des annotations de fonction et des caractéristiques physico-chimiques
MutationAssessor	21727090	2011	Homologie	-	Utilise le taux de conservation et de substitutions chez les homologues
FATHMM	23033316	2012	Hybride	HMM	Combine la conservation évolutive et la fréquence de variations neutres et pathogènes dans certaines régions protéiques
Provean	23056405	2012	Homologie	-	Score basé sur l'homologie
VEST	23819870	2012	Hybride	RF	Utilisation de 86 éléments
Evolutionary Action	-	2014	Homologie	-	Score basé sur l'homologie
SDS	24795746	2014	Structure	-	Intègre des mesures de la stabilité des protéines, de la flexibilité, du potentiel d'interaction protéine-protéine et de la liaison des petites molécules
SuSPect	24810707	2014	Hybride	SVM	Utilise des données d'interactions protéiques, d'homologie, certaines annotations fonctionnelles et l'accessibilité prédite aux solvants
VarMod	24906884	2014	Hybride	SVM	Utilise des informations de structure, d'homologie, et de prédictions fonctionnelles (sites de liaison des ligands et d'interactions protéine-protéine)
FATHMM-MKL	25583119	2015	Homologie	MKL	Score basé sur FATHMM, qui intègre des annotations encode pour fournir des prédictions sur les variations non codantes
PrimateAI	30038395	2018	Homologie	DNN	Utilisation de données d'homologie et les polymorphismes chez les grands singes
MutPred2	biorxiv134981	-	Hybride	NN	Utilisation de multiples sources d'information.

Tableau 6. Principaux logiciels de prédiction fonctionnelle des variations faux-sens (légende en page 33)

Tableau 6 (légende)
PSSM : position specific scoring matrix, DT : decision tree, HMM : Hidden Markov Model, SVM : Support Vector Machine, RF : Random Forest, NN : Neural Network, NBC: Naive Bayes Classifier, MKL : Multiple Kernel Learning, DNN : Deep Neural Network. ddG : variation d'enthalpie libre.

1.3.2.3 Scores d'ensemble

Plus récemment, des scores intégratifs ont été proposés. Ces scores « d'ensemble » intègrent les résultats de nombreux scores de pathogénicité, en plus d'informations très diverses, telles que la présence de maladies génétiques associées, la fréquence des variations, etc. Ces scores, dont le rationnel vient de l'observation que les différents scores sont souvent complémentaires et nécessitent une utilisation combinée, ont pour objectif de centraliser les multiples sources d'informations en un score unique plus performant que chaque algorithme pris séparément. Certains scores sont spécifiques des variations faux-sens alors que d'autres ont une vocation de pouvoir prédire l'impact de tous les types de variations. Eux-mêmes basés sur des principes de *machine learning* et donc de données d'entraînement et de test, leur apport par rapport à une approche combinée reste discutable.

Score	Année	PMID	Algorithme	Cible	Résumé du fonctionnement
Condel	2011	21457909	-	Faux-sens	Moyenne pondérée de 5 scores : Logre,MAPP, Mutation Assessor, Polyphen 2 et SIFT
KGGseq	2012	22241780	LR	Tous types	Priorisation des variations à 3 niveaux : génétique (mode de transmission), variants (fréquences) et intégration des connaissances sur les maladies
PON-P	2012	22505138	RF	Faux-sens	Intègre cinq scores : PhD-SNP, SIFT, PolyPhen-2, SNAP, I-mutant
CADD	2014	24487276	SVM	Tous types	Utilise 63 annotations extraites de Ensembl Variant Predictior, du projet ENCODE et de pistes UCSC Genome Browser
DANN	2014	25338716	DNN	Tous types	Données de CADD mais entrainement grâce à un réseau neuronal profond
MetaLR	2015	25552646	LR	Faux-sens	Intègre la MAF dans 1000 génomes et 9 scores : SIFT, PolyPhen-2, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy et PhyloP
MetaSVM	2015	25552646	SVM	Faux-sens	Intègre la MAF dans 1000 génomes et 9 scores : SIFT, PolyPhen-2, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy et PhyloP
Eigen	2016	26727659	-	Tous types	Apprentissage non supervisé basé sur des scores fonctionnels, de conservation, et la fréquence allélique dans 1000 génomes
REVEL	2016	27666373	RF	Faux-sens	Utilise 13 scores : MutPred, FATHMM, VEST, Poly-Phen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, et phastCons

Tableau 7. Principaux scores d'ensemble
LR : Logistic Regression, SVM : Support Vector Machine, DNN : Deep Neural Network, RF : Random Forest.

1.3.2.4 Comparaisons de performance des algorithmes

Les différents algorithmes de prédiction peuvent être comparés en fonction de leur performance sur des jeux de données dont caractère délétère ou non est connu. Les données classiquement employées pour ces comparaisons sont de deux ordres. D'une part des jeux de données basés sur des tests fonctionnels quantitatifs issus de *screenings* par mutagenèse dans des protéines modèles. Ces jeux de données ont l'avantage d'évaluer l'impact des variations de manière précise sur une échelle continue, mais sont de champ limité. Plus récemment, les jeux de données de variations associées en pathologie humaine ont été employés, tels que la base HGMD (voir chapitre 1.2.3.3) par exemple. Ces jeux de données ont l'avantage d'être spécifiques à l'homme, mais ont pour limites leur nature binaire (pathogène vs bénin) et la possibilité de certaines erreurs de classification. Il va de soi que les données utilisées pour comparer les différents scores doivent être indépendantes des données utilisées dans leur entrainement sous peine d'un biais. La performance des outils de prédiction est classiquement évaluée par l'aire sous la courbe des courbes ROC (pour

receiver operating characteristic : courbe du taux de vrais positifs en fonction du taux de faux positifs). Lors de la publication d'un nouvel algorithme, les auteurs proposent classiquement une comparaison avec les logiciels concurrents, montrant souvent une aire sous la courbe plus importante pour leur logiciel. Ces résultats suggèrent que ces algorithmes sont probablement testés dans les meilleures conditions possible lors de leur description initiale. Pour pallier ce biais, des groupes indépendants ont par le passé proposé des challenges sur des nouveaux jeux de données ou les différentes équipes pouvaient comparer leurs logiciels en aveugle (https://genomeinterpretation.org/content/5-challenge). Un autre moyen pour identifier les prédicteurs les plus pertinents est d'identifier leur popularité dans la communauté scientifique. Dans une étude, les auteurs ont représenté la popularité, indiquée par le nombre de citations, par rapport à l'année de parution des logiciels (Figure 12)[64].

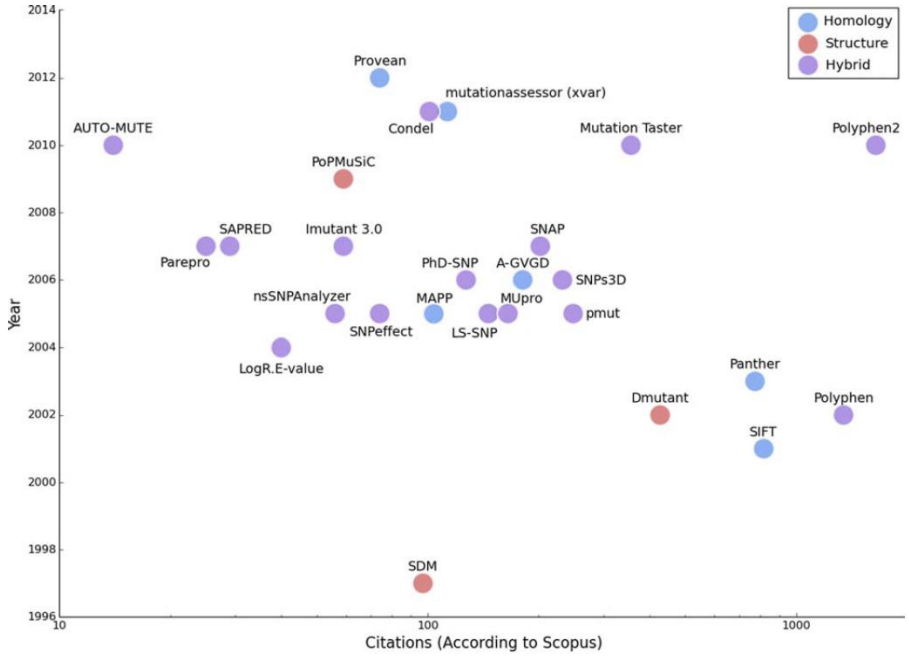


Figure 12. Popularité des logiciels de prédiction dans la communauté scientifique
Les algorithmes sont séparés en fonction du nombre de citations dans la littérature et de l'année de publication. Source : ref[64].

Des algorithmes de performances similaires peuvent dans de nombreux cas montrer des résultats discordants, ce qui indique une complémentarité des algorithmes, liée à la diversité dans leur mode de fonctionnement. Aucun logiciel ne surpasse tous les autres dans toutes les situations, ce qui suggère une utilisation combinée d'un jeu de plusieurs algorithmes.

1.3.2.5 Place des outils de prédiction en pratique médicale

Les scores de prédiction sont utiles à l'interprétation médicale des variations génétiques. En revanche il n'existe pas de logiciel ou d'ensemble de logiciels consensus validés et recommandés dans une utilisation diagnostique. Dans les recommandations ACMG précédemment évoquées, le poids apporté à ce logiciel est faible (*Pathogenic Supporting* 3, PP3), dans le cas d'une concordance de plusieurs scores « *Multiple lines of computational evidence support a deleterious effect on the gene or gene product* ». Ainsi, en l'état actuel des recommandations, l'utilisation de plusieurs

scores de prédiction fonctionnelle semble utile, contrairement à l'utilisation d'un score d'ensemble unique. Un autre élément qui s'oppose à l'utilisation des scores d'ensemble dans le contexte de l'application des recommandations ACMG provient du fait que ces scores d'ensemble peuvent prendre en compte des données cotées par ailleurs dans les éléments recommandés, telles que la fréquence allélique ou la ségrégation familiale par exemple, les rendant partiellement redondants.

Contrairement aux scores de prédiction, un rôle des scores d'ensemble qui semble se dégager est celui non pas dans l'interprétation des variations mais dans leur priorisation[71], dans le contexte de l'automatisation du processus d'analyse des données de génétique. Certains logiciels intégratifs récents (on peut citer the Exomiser[72], GenIO[73] ou SeqOne) utilisent d'une part les informations de pathogénicité des variations (par des scores d'ensemble), et d'autre part la priorisation des gènes en fonction de la présentation clinique du patient (revue des algorithmes disponibles : ref[74]), en lien avec des données de ségrégation, de qualité du génotype, etc., afin de proposer une priorisation globale et complètement automatisée des variations. Néanmoins, si l'utilisation de ces logiciels permet l'identification facilitée des variations les plus pertinentes, une expertise dans l'interprétation des variations est bien sûr toujours nécessaire.

L'utilisation des scores de prédiction fonctionnelle dans une démarche médicale semble dans ce contexte toujours pertinente en 2018 et pour les années à venir. Lors de l'utilisation d'un score, il semble indispensable d'identifier précisément les sources de données utilisées par le logiciel afin qu'il ne soit pas redondant avec les autres critères pris en compte. Afin de choisir les meilleurs algorithmes dans une utilisation donnée, le maintien d'une base de données locale de variations bénignes et pathogènes peut être utile et permettre d'effectuer une comparaison indépendante des différents logiciels (aire sous la courbe ROC) la plus adaptée au type de données étudiées.

1.3.3 Distribution des variations faux-sens en population générale et en pathologie

Les grands jeux de données de variations génétiques en population générale et en population atteinte par des maladies monogéniques permettent d'étudier à grande échelle les *patterns* de distribution des variations faux-sens. Contrairement à la très grande majorité des variations synonymes, les variations faux-sens peuvent entraîner un effet biologique pouvant biaiser leur distribution, que ce soit par un mécanisme de contrainte évolutive en population générale ou au contraire par un enrichissement en population malade du fait de la sélection des patients.

1.3.3.1 Distribution « bidimensionnelle » des variations faux-sens

L'étude de la distribution des variations par rapport à la séquence protéique, que l'on appellera ici distribution « bidimensionnelle » des variations, représente une méthode simple permettant d'identifier des biais distribution des variations faux-sens au sein des protéines. Dans une étude de 2015, les auteurs comparaient de manière globale la distribution des variations faux-sens issues 1000 Génomes considérées comme bénignes, et les variations pathogènes de HGMD[75]. Deux manières d'évaluer la clusterisation à partir de la distribution des variations le long de la séquence protéique (distribution « bidimensionnelle ») ont été réalisées : d'une part la mesure du *domaine occupancy score*, représentant la proportion des variations localisées dans un domaine fonctionnel de la protéine, et d'autre part l'utilisation d'un algorithme de détection de clusters « *de novo* »,

indépendamment de l'annotation des domaines protéiques. Ces deux approches ont montré une plus forte tendance des variations faux-sens pathogènes à clusteriser que les variations bénignes. Aussi la clusterisation et l'occupation des domaines étaient plus importantes pour les maladies dominantes que pour les maladies récessives. On peut faire l'hypothèse que cette clusterisation témoigne d'une sensibilité de certaines régions critiques de la protéine.

Plus spécifiquement, dans les maladies neurodéveloppementales, une équipe américaine a montré à partir de données d'exome une distribution biaisée de certaines variations, avec la présence de nombreux clusters de variations faux-sens *de novo* présents chez les cas mais pas dans des populations contrôles, ainsi que des « sites » de mutations *de novo*, correspondant à des acides aminés mutés de manière récurrente entre plusieurs individus[76]. En employant une méthode similaire d'étude de la distribution des variations faux-sens *de novo* dans les maladies du développement, une équipe néerlandaise a également identifié des clusters de variations faux-sens *de novo* significatifs à l'échelle du génome, dont certains n'étaient auparavant pas connus[77]. Cette étude, basée sur les données de denovo-db auxquelles ont été associées des données locales, a ainsi permis d'identifier les gènes *GABBR2*, *PACS2* et *ACTL6B* comme responsables de maladies du développement. Les deux premiers gènes ont été formellement confirmés depuis[78,79], et le troisième, *ACTL6B*, est en cours de validation, en particulier grâce à des données présentes dans ce travail (voir chapitre 3.3.3.1).

1.3.3.2 Distribution tridimensionnelle des variations faux-sens

En dehors de la répartition des variations en deux dimensions au long de la séquence de la protéine, il est possible pour certaines protéines pour lesquelles la structure tridimensionnelle est connue, d'analyser la répartition des variations en trois dimensions. Cette approche peut apporter des éléments complémentaires à une étude en deux dimensions basée sur la séquence. Par exemple dans une étude ayant déjà été évoquée[76], trois clusters de variations faux-sens *de novo* du gène *PTPN11*, responsable du syndrome de Noonan, ont été identifiés, et leur représentation en 3D a permis de montrer leur proximité tridimensionnelle au niveau du site actif de la protéine, correspondant en réalité à un unique cluster tridimensionnel.

Dans une étude récente, les auteurs analysaient de manière globale la répartition des variations faux-sens dans des modélisations tridimensionnelles des protéines[80]. La répartition des variations de GnomAD, considérées comme bénignes, et celle des variations pathogènes ClinVar étaient analysées. Pour chaque structure protéique disponible, un modèle statistique était appliqué pour identifier si les variations de chaque jeu de données étaient significativement clusterisées, de distribution aléatoire, ou au contraire significativement trop dispersées. Contrairement aux variations synonymes (issues de GnomAD), qui étaient largement réparties de manière aléatoire, les variations faux-sens bénignes et pathogènes avaient une distribution biaisée dans de nombreuses protéines. Les variations faux-sens bénignes avaient une forte tendance vers la dispersion. Les auteurs interprétaient ce résultat en évoquant la contrainte qui existe sur les résidus profonds dans les protéines, pouvant provoquer une déstabilisation thermodynamique de la protéine[81]. Ils proposaient l'idée que les variations bénignes étaient plus fréquemment localisées en surface de la protéine, dans un espace permettant leur dispersion tridimensionnelle. Au contraire, les variations faux-sens pathogènes montraient globalement une forte propension à la clusterisation. On peut

imaginer que cet effet soit lié à la fois aux variations faux-sens à effet perte de fonction (par clusterisation grossière, en particulier dans les résidus structuraux du cœur hydrophobe des protéines), et aux variations à effet gain de fonction (clusterisation fonctionnelle fine dans des régions critiques). Les auteurs mettaient à disposition un serveur web permettant de visualiser la répartition des variations faux-sens pour les structures disponibles. Basée sur cette ressource, la Figure 13 représente l'exemple de la protéine MSH6, dont les variations perte de fonction sont responsables du syndrome de Lynch, qui montre que les variations de la population générale ont une répartition aléatoire, alors que les variations pathogènes sont statistiquement clusterisées.

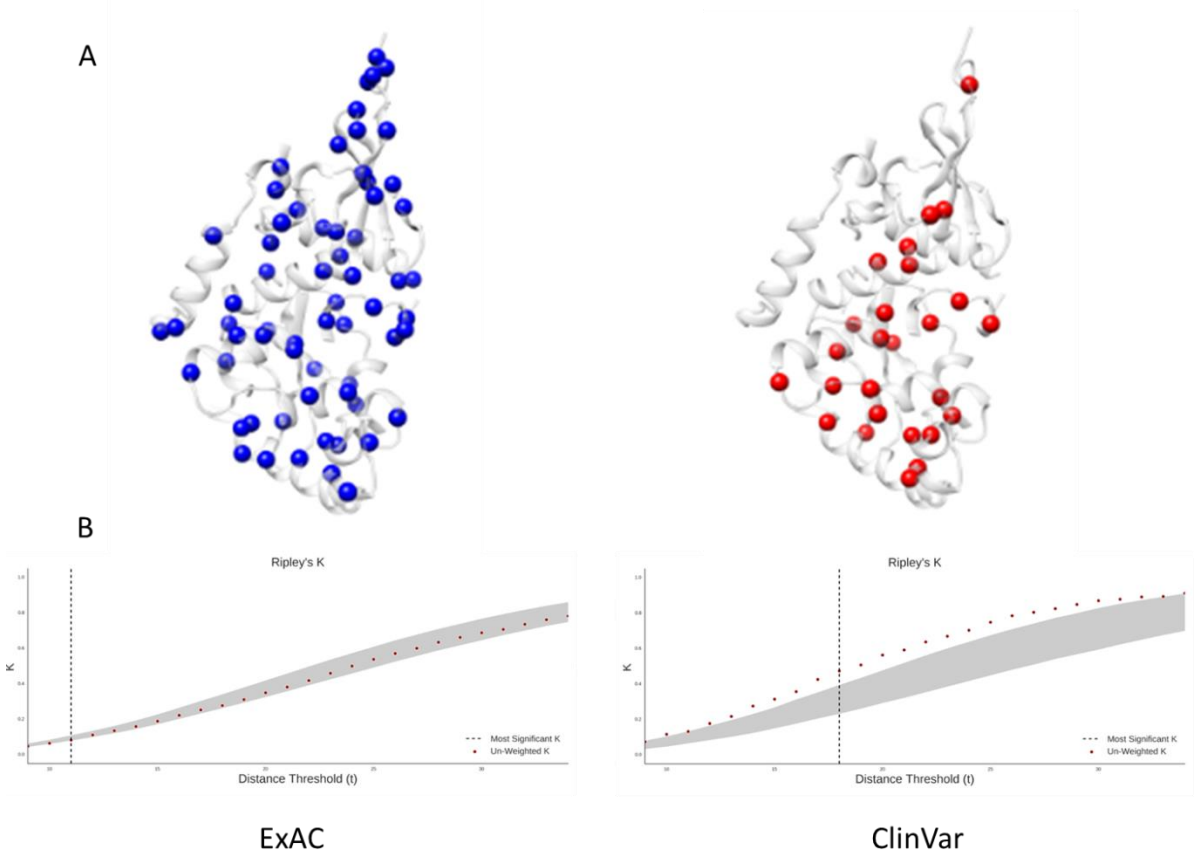


Figure 13. Comparaison de la distribution tridimensionnelle des variations faux-sens bénignes et pathogènes dans la protéine MLH1
Les variations considérées comme bénignes sont approximées par les variations de la base ExAC (en bleu). Les variations pathogènes (en rouge) proviennent de ClinVar. A) Visualisation des variants au sein de la structure protéique. B) Représentation statistique de la dispersion par rapport à un modèle aléatoire (zone grise). Des valeurs élevées de la mesure K indiquent une clusterisation. Source : http://astrid.icompbio.net/

Plus spécifiquement dans les maladies du développement, l'équipe néerlandaise montrait que d'une manière générale, les clusters identifiés d'après la séquence protéique étaient très majoritairement localisés en surface de la protéine[77]. L'évaluation des connaissances scientifiques concernant ces clusters montrait que l'effet de ces variations était très fortement associé à des mécanismes non-haploinsuffisants, de type gain de fonction ou dominant négatif.

Pour résumer sur la répartition des variations faux-sens en population générale et en pathologie, il apparaît que les variations situées dans les profondeurs de la protéine ont un potentiel de déstabilisation thermodynamique de la protéine (pouvant orienter vers un mécanisme de perte de

fonction), ce qui explique qu'elles apparaissent déplétées en population générale et au contraire enrichies en pathologie, probablement sous la forme d'une clusterisation grossière. À l'opposé, les variations situées en périphérie de la protéine sont globalement mieux tolérées, en dehors de certaines positions et résidus précis, souvent au sein de domaines fonctionnels, pouvant donner lieu à un effet de type gain de fonction ou dominant négatif par exemple. Cette clusterisation fine touche souvent un nombre restreint de résidus clés et est parfois identifiable sans prendre en compte la modélisation tridimensionnelle de la protéine.

On peut identifier deux limites principales des études globales basées sur ClinVar ou HGMD, diminuant leur potentiel de détection de clusters fins de variations à effet non-haploinsuffisant. D'une part le nombre d'occurrence des variations n'est pas pris en compte. Ainsi les maladies liées à une récurrence mutationnelle forte sur un seul résidu (par exemple le syndrome de Myhre lié à une position précise du gène *SMAD4*), ne peuvent pas être identifiées. Aussi, ces études sont basées sur la charge de variations pathogènes agrégées dans chaque gène, sans prendre en compte la maladie associée. Comme nous l'avons vu, il existe de nombreux gènes avec plusieurs maladies associées, liées à des mécanismes mutationnels différents. Pour reprendre l'exemple de *SMAD4*, la recherche de clusterisation à partir de ClinVar serait fortement polluée par les variations pathogènes faux-sens ayant un effet perte de fonction menant à la polypose juvénile ne permettant là encore pas d'identifier le cluster du syndrome de Myhre. Ainsi la méthode de choix pour identifier des clusters « fins » avec effet fonctionnel non-haploinsuffisant est d'analyser une cohorte de patients avec phénotype similaire en prenant en compte le nombre d'occurrences de chaque variation[76,77].

1.4 Formulation des hypothèses

En résumé, partant du postulat que :

1. Une grande proportion des relations génotype-phénotype restant à identifier dans les maladies du développement est capturée par l'exome

En comparant le nombre de variations *de novo* observées dans une cohorte de plus de 7 500 patients porteurs d'une maladie sévère du développement avec le nombre de variations attendues dans un modèle aléatoire[26], le consortium anglais DDD a évalué qu'une variation *de novo* jouait un rôle dans la pathologie de 42% des patients de cette cohorte[51]. Ce chiffre était bien plus élevé que la proportion de variations effectivement interprétées comme pathogènes avec les connaissances scientifiques de l'époque (23% des patients avec mutation *de novo* pathogène), suggérant qu'un grand nombre de corrélations phénotype-génotype bien « capturées » par le séquençage d'exome en trio n'est pas encore identifié.

2. Les variations à effet NHI sont plus complexes à mettre en évidence que les variations à effet HI

Dans cette même étude, les auteurs estimaient par deux approches complémentaires qu'environ la moitié des variations *de novo* agissaient par un mécanisme haploinsuffisant et que l'autre moitié agissait par un mécanisme non-haploinsuffisant. Il semble raisonnable de penser que le mécanisme pathogène de perte de fonction est globalement plus aisé à identifier et à interpréter que les mécanismes non-haploinsuffisants. En effet, plusieurs éléments rendent le mécanisme HI plus simple à identifier que l'effet NHI. Premièrement, dans les maladies du développement, les

variations *de novo* à effet HI ciblent de manière préférentielle une liste de gènes montrant une déplétion significative en variations tronquantes en population générale (ExAC pLI[15] > 0,9). Deuxièmement, la présence de variations tronquantes dans la charge mutationnelle associée à une maladie est fortement évocatrice d'une maladie causée par un effet HI. Troisièmement, les maladies à effet HI sont accessibles aux études d'enrichissement, les variations tronquantes étant modélisables selon un modèle nul[26,51,82], contrairement aux variations faux-sens. Au contraire, l'identification d'un effet non-haploinsuffisant, potentiellement causé par un set réduit de variations faux-sens spécifiques, semble plus complexe à identifier, et requérir des stratégies subtiles, telles que la recherche d'une clusterisation de faux-sens *de novo*.

Nous faisons l'hypothèse que le mécanisme NHI est plus commun parmi les maladies ultra-rares

Nous faisons l'hypothèse qu'une importante proportion des relations génotype-phénotype liées à des variations codantes qui restent à identifier en 2018 pourrait être liée à un mécanisme NHI. Aussi on peut imaginer que certaines maladies liées à des variations faux-sens restreintes uniquement à un ou quelques résidus critiques pourraient être extrêmement rares car la survenue de telles mutations serait très improbable d'un point de vue de mutabilité du génome. En conséquence nous faisons l'hypothèse que les maladies ultra-rares soient enrichies en variations à effet NHI, et donc en variations faux-sens.

En seconde partie de ce travail, nous tenterons d'apporter des éléments de réponse à cette hypothèse en analysant les informations de plusieurs jeux de données publiques déjà évoqués afin de préciser certaines caractéristiques des variations faux-sens observées au travers du prisme génomique. La distribution des variations faux-sens en fonction de la rareté des maladies, puis en fonction du mode de transmission associé, seront évaluées, puis certaines propriétés des variations *de novo* récurrentes dans les maladies du développement seront analysées.

Enfin, en troisième partie, nous exploiterons les résultats obtenus pour proposer une méthode visant à identifier rétrospectivement des variations faux-sens à effet NHI dans une cohorte de plus de 1 000 exomes pratiqués dans le contexte du diagnostic génétique de maladies du développement, à partir de bases de données publiques.

2 IDENTIFICATION DE PROPRIÉTÉS GÉNÉRALES DES VARIATIONS FAUX-SENS GRÂCE À DES BASES DE DONNÉES EN ACCÈS LIBRE

La quantité d'informations disponibles au sein des bases de données en accès libre est telle que leur interrogation peut permettre de répondre à certaines questions scientifiques. Dans cette deuxième partie, plusieurs questions sans réponse claire dans la littérature ont été posées, dans l'objectif de préciser certains aspects de l'impact des variations faux-sens en pathologie d'une manière globale. Nous proposons d'étudier dans un premier temps l'impact des variations faux-sens en fonction de la rareté des maladies. Ensuite nous évaluerons l'impact des variations faux-sens en fonction du mode de transmission des maladies génétiques. Enfin nous tenterons d'étudier certaines propriétés des variations récurrentes entre plusieurs individus, ce qui permettra de poser les bases pour la 3ème partie de cette thèse, dans laquelle la recherche de récurrence mutationnelle sera évaluée dans le but d'identifier des nouvelles variations pathogènes.

2.1 Méthodes : obtention des données

L'URL de chaque site internet utilisé est disponible dans la section ressources web.

Relations phénotype-génotype basées sur les gènes : OMIM

Le tableau genemap2 a été téléchargé en tant que base de données de relations gène-phénotype (disponible sur requête). Ce fichier comprend une ligne par gène avec différentes annotations incluant en particulier les symboles de gènes officiels selon la nomenclature HUGO, ainsi que les différentes maladies OMIM associées à chaque gène et leur mode de transmission.

Données épidémiologiques : Orphanet

Les données d'Orphanet ont également été exploitées, afin d'obtenir une deuxième source de maladies. Deux fichiers ont été téléchargés du serveur Orphadata : le fichier « Epidemiological data » contenant les codes Orpha des maladies et leur fréquence, disponible sur requête, et le fichier « *Disorders with their associated genes* », en accès libre. La prévalence des maladies Orphanet était annotée en 7 classes : 1-5/10 000 ; 6-9 /10 000 (3 entrées seulement); 1-9/100 000 ; 1-9/1 000 000 ; <1/1 000 000 ; et deux classes de prévalence inconnue ou non applicable. Une équivalence avec la ou les maladies OMIM était également disponible.

Mécanisme mutationnel des maladies : EBI Gene2Phenotype

Le fichier DDG2P.csv a été téléchargé du site de l'European Bioinformatic Institute, afin d'obtenir des annotations supplémentaires non disponibles dans OMIM et Orphanet. Ce fichier liste 2 331 maladies monogéniques du développement au sens large, en reprenant la nomenclature OMIM des maladies lorsque disponible. Un gène unique est indiqué pour chaque maladie, avec comme

informations pertinentes la colonne *allelic requirement*, que l'on peut rapprocher du mode de transmission de la maladie, mais surtout l'annotation *mutation consequence*, correspondant au mécanisme mutationnel connu de la maladie (e.g. *loss of function* ou bien *activating*). Une annotation de termes HPO associés à la maladie est également disponible.

Variations pathogènes : Clinvar

Les données de ClinVar ont été téléchargées (version du 01-07-2018), sous la forme d'un fichier VCF, avec pour annotations d'intérêt l'information CLNSIG, correspondant à la signification clinique de la variation, et CLNREVSTAT, correspondant au *reviews status*, à savoir le niveau de preuve de l'interprétation des variants (étoiles d'or). Le fichier VCF a été annoté par Ensembl Variant Effect Predictor afin d'obtenir des informations supplémentaires pour chaque variant, et en particulier l'effet fonctionnel prédit. D'une manière générale, les variations de classe 4 et 5, sans tenir compte de l'annotation CLNREVSTAT ont été incluses en tant que variations pathogènes.

Mesure d'intolérance des gènes à la perte de fonction : ExAC pLI

La pLI de chaque gène a été obtenue à partir de la table 13 dans les informations supplémentaires de l'article de Lek *et al.*[83]

Variations *de novo* chez les patients : Denovo-db

Les données de denovo-db 1.6 ont été téléchargées sous la forme de deux fichiers : « non-SSC Samples » et « SSC Samples », qui ont été fusionnés. Certains individus contrôles sont présents dans denovo-db, et ont été exclus des analyses. Ce fichier présente une ligne par mutation *de novo*, qui peuvent être dupliquées en cas de transcrits multiples. Ces duplicats ont été supprimés pour ne garder qu'une ligne par variation (suppression des lignes partageant le même #SampleID et la même variation).

Variations *de novo* en population générale : génomes des trios islandais

Les données de l'étude de Jónsson *et al.*[18] ont été téléchargées et ré-annotées par VEP.

2.2 Résultats

2.2.1 Maladies dominantes : impact de l'effet non-haploinsuffisant en fonction de la prévalence

Nous faisons l'hypothèse que l'impact du mécanisme non-haploinsuffisant dans le déterminisme des maladies génétiques dominantes pourrait augmenter avec la rareté des maladies. Ce modèle impliquerait (i) une plus grande proportion de variations faux-sens dans la charge mutationnelle responsable des maladies ultra-rares, et (ii) une plus faible sensibilité aux variations tronquantes (ExAC pLI) des gènes responsables de maladies ultra-rares, par rapport aux maladies moins rares. Deux approches basées sur ces propositions ont été appliquées, en plus d'une troisième approche plus directe concernant spécifiquement les maladies du développement au sens large pour lesquelles

la conséquence mutationnelle (par ex. perte de fonction) était disponible et a pu être mise en relation avec la fréquence des maladies.

Mécanisme moléculaire des maladies monoalléliques en fonction de la prévalence des maladies

Parmi les 2331 maladies du développement listées dans la base EBI Gene2Phenotype, 1 785 entrées comprenaient un identifiant OMIM unique, parmi lesquelles 767 maladies étaient mono-alléliques (mention *monoallelic*, *hemizygous* ou *x-linked dominant* dans *allelic requirement*). Ces entrées ont été annotées avec la fréquence des maladies Orphanet partageant le même identifiant OMIM. Ainsi des données de fréquence ont pu être établies pour 369 maladies du fichier EBI Gene2Phenotype. Du fait du petit nombre de maladies dans les trois groupes de prévalences les plus élevées (1-5/10 000 ; 1-9/100 000 et 1-9/1 000 000), ces trois groupes ont été regroupés en une catégorie >1/1 000 000 (n=117 maladies), dont le mécanisme moléculaire de la base EBI a été comparé avec le groupe de maladies ultra-rares <1/1 000 000 (n=252 maladies)(Figure 14).

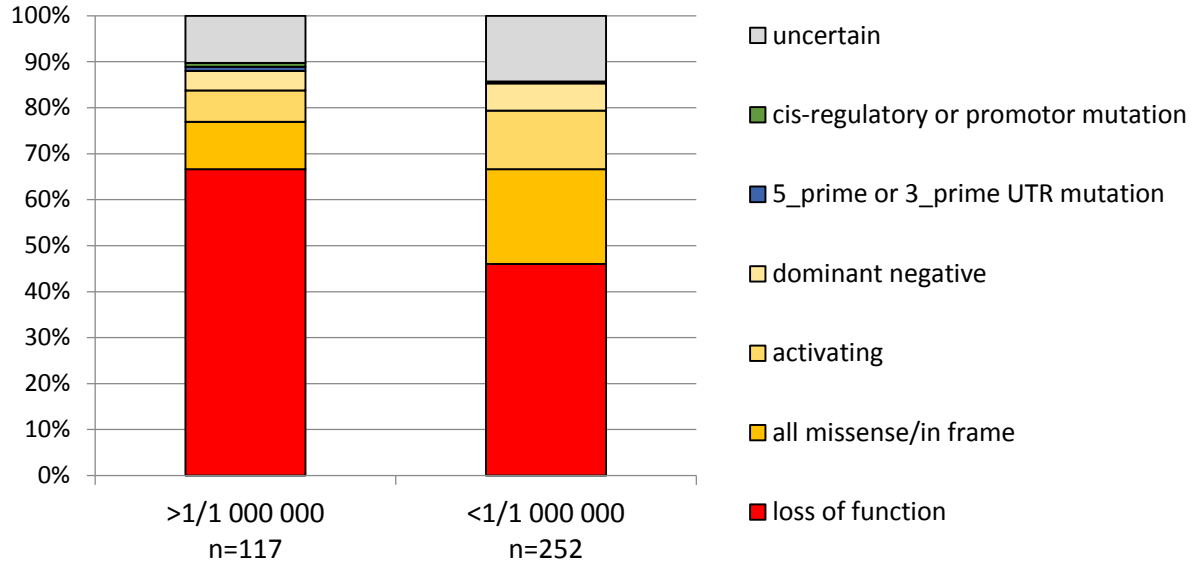


Figure 14. Effet fonctionnel moléculaire de 369 maladies génétiques monoalléliques du développement, en fonction de leur prévalence
Les maladies agissant par effet haploinsuffisant sont en rouge. Les maladies agissant par un mécanisme non-haploinsuffisant sont représentées avec un dégradé de jaune.

Ces données montrent une diminution relative du mécanisme haploinsuffisant (*loss of function*) dans le mécanisme des maladies ultra-rares, au profil du mécanisme non-haploinsuffisant, représenté par les 3 catégories *all missense/in frame*, *activating* et *dominant negative* ($p=2{,}1.10^{-4}$, test de Fisher comparant les groupes HI et NHI).

Proportion de variations faux-sens en fonction de la prévalence des maladies

Afin d'aborder cette question de l'impact de l'effet non-haploinsuffisant en fonction de la rareté des maladies sous un autre angle et d'après des données différentes, nous avons cherché à identifier une augmentation de la proportion des faux-sens avec la rareté de la maladie, pouvant témoigner d'une

augmentation d'un enrichissement en effet NHI. Dans ce contexte nous avons exploité les données de ClinVar pour l'identification des variations pathogènes et d'Orphanet pour l'identification de la rareté des maladies. ClinVar comprenait 76 858 variations pathogènes (classe 4 ou 5), dont 34 370 pour lesquelles les *submitters* avaient renseigné une (ou plusieurs) maladie Orphanet causée par la variation. 14 479 variations étaient associées à une maladie Orphanet pour laquelle le mode d'hérédité comprenait le terme « dominant » et ont été incluses. Les variations ont été regroupées au sein des 4 classes de prévalence, et leur effet fonctionnel, renseigné par Variant Effect Predictor, a été évalué (Figure 15). Cette analyse a montré que le spectre mutationnel variait en fonction de la rareté de la maladie ($p<10^{-16}$, test du khi-2 comparant les variations faux sens et les variations tronquantes). La proportion de faux-sens augmente avec la rareté des maladies, et au contraire les variations tronquantes (frameshift, non-sens et variations d'épissage) diminuent. Afin de s'assurer que cet effet n'était pas lié à un petit nombre de gènes/maladies fournissant un grand nombre de variations, nous avons refait cette analyse en ne prenant pas en compte les 23 maladies avec plus de 100 variations associées (jusqu'à 1623 variations pour le syndrome héréditaire de prédisposition au cancer du sein et de l'ovaire). Cette ré-analyse a de manière logique affecté uniquement les deux catégories les plus fréquentes (1-5/10 000 et 1-9/100 000), sans changer l'aspect global des spectres mutationnels (données non présentées).

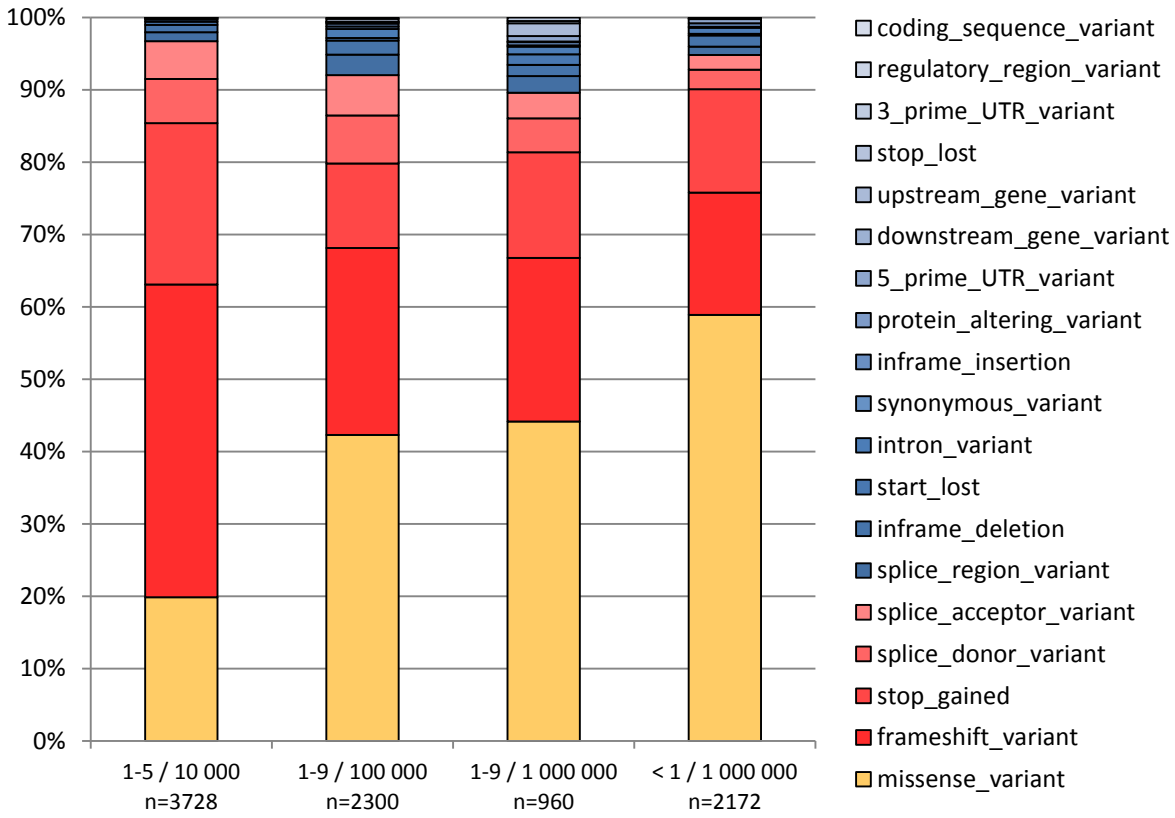


Figure 15. Spectre mutationnel des variations pathogènes ClinVar en fonction de la rareté des maladies
Les variations faux-sens sont indiquées en jaune. Les variations tronquantes sont indiquées en nuances de rouge. Les types variations plus rarement associées en pathologie sont indiqués en nuances de bleu.

Les gènes responsables de maladies ultra-rares dominantes sont-ils moins sensibles aux variations perte de fonction ?

On peut faire l'hypothèse que si les maladies ultra-rares étaient effectivement davantage liées au mécanisme NHI que les maladies moins rares, on pourrait observer une diminution de la sensibilité des gènes aux variations perte de fonction dans les maladies ultra-rares.

Dans un premier temps nous avons évalué si les gènes responsables de maladies dominantes par un mécanisme NHI étaient globalement moins sensibles aux variations perte de fonction que les gènes agissant par un mécanisme HI. La pLI des gènes identifiés comme dominants dans le listing EBI Gene2Phenotype a été comparée entre les mécanismes HI (annotation *loss of function*, n=368) et les maladies NHI (annotations *all missense/in frame*, *activating* et *dominant negative*, n=200). Comme attendu, ce test a montré une plus forte sensibilité aux mutations tronquantes des gènes agissant par effet HI que par effet NHI (pLI moyenne : 0,830 contre 0,639 respectivement, $p=6{,}36.10^{-8}$, test de Student).

Par la suite nous avons cherché le lien potentiel entre la prévalence de la maladie et la pLI. Les données d'Orphanet ont été utilisées. Cependant, contrairement à OMIM, les entrées de maladies Orphanet n'ont en général pas un gène unique associé (pour lesquels la pLI est disponible), mais il existe très fréquemment plusieurs gènes. Le nombre de maladies associées de manière isolée à un gène n'était pas suffisant pour tirer des conclusions. De fait, avec les données disponibles, la corrélation entre la fréquence de la maladie et la pLI n'a pas été possible.

Résumé et critique des résultats

Nous avons montré que les mécanismes mutationnels et le spectre mutationnel des maladies dominantes n'étaient pas homogènes en fonction de la rareté des maladies. La tendance qui se dégage est celle d'un plus grand rôle du mécanisme non-haploinsuffisant dans les maladies ultra-rares (<1/1 000 000), dans les maladies du développement. De plus, d'une manière globale, la proportion de variations faux-sens dans la charge mutationnelle responsable de maladies dominantes augmente avec la rareté des maladies. Par exemple, les variations faux-sens ne représentent que 20 % des variations responsables de maladies dont la prévalence est dans la gamme de prévalence 1-5/10 000, alors qu'elles représentent près de 60 % des variants impliqués dans les maladies les plus rares (<1/1 000 000). Une des critiques possibles de ces résultats provient du jeu de données utilisé pour évaluer la prévalence des maladies, qui comporte une importante part d'hétérogénéité génétique. En effet, l'analyse du nombre de gènes associé à chaque maladie Orphanet pour chaque classe de prévalence a montré un fort lien entre la rareté et le nombre de gènes associés (données non présentées). Par exemple pour la classe 1-9/10 000 (n=79 maladies Orphanet), 5,43 gènes étaient impliqués par maladie en moyenne, contre 1,28 pour la classe la plus rare (<1/1 000 000, n=1878 maladies). Ce lien entre le nombre de gènes ou de locus impliqués (*mutational target*) et la fréquence de la maladie a déjà été relevé mais sans données chiffrées[84]. On peut imaginer que ce nombre de gènes variable en fonction des classes de prévalence puisse être un biais dans les corrélations effectuées. L'utilisation des données d'OMIM, qui aurait été plus pertinente car les maladies OMIM sont bien plus souvent spécifiques d'un gène unique, n'a pas été possible du fait de l'absence de données épidémiologiques.

2.2.2 Impact des variations faux-sens en fonction du mode de transmission des maladies

Il est bien établi que les maladies récessives sont très majoritairement liées à un effet perte de fonction, contrairement aux maladies dominantes qui peuvent être dues à des mécanismes plus variés[75]. Cette constatation a été simple à vérifier dans les maladies du développement grâce à la base DDG2P qui indique pour chaque maladie le mécanisme de pathogénicité. La comparaison des spectres des maladies bialléliques et monoalléliques a bien montré des profils distincts, avec une diversité des mécanismes dans les maladies dominantes, s'opposant au mécanisme de perte de fonction prépondérant dans les maladies récessives (Figure 16).

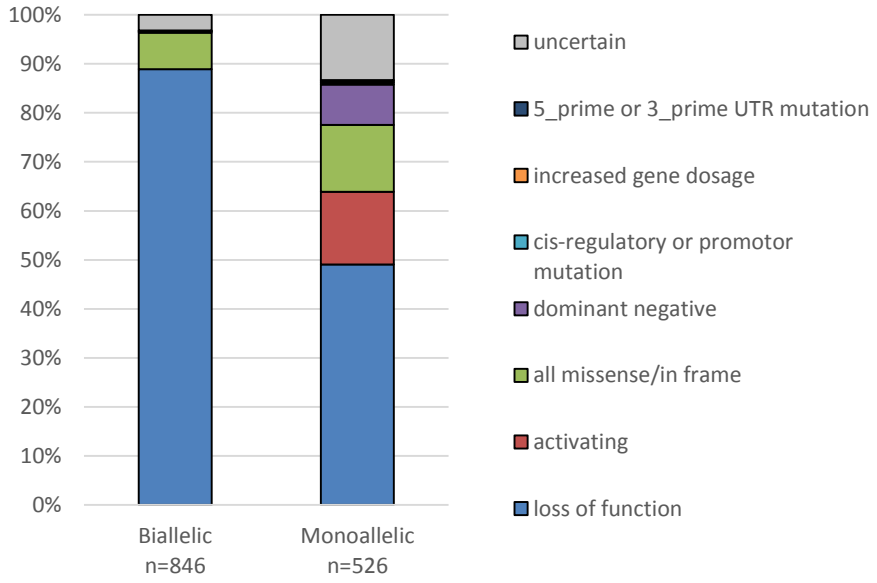


Figure 16. Mécanisme mutationnel en fonction du mode de transmission de la maladie
Données concernant les maladies du développement au sens large. Source : EBI Gene2Phenotype (voir ressources web).

Suite à ces observations, nous avons évalué l'impact des variations faux-sens en fonction du mode de transmission des maladies, dans l'hypothèse qu'il pourrait être plus fort dans les maladies dominantes qui reposent plus sur des mécanismes NHI. Pour répondre à cette problématique, nous avons étudié la proportion de variations faux sens par rapport aux variations tronquantes dans les maladies Orphanet, en se basant sur les données de ClinVar. Ce ratio était globalement constant en fonction du mode de transmission (autosomique dominant, autosomique récessif, dominant lié à l'X, récessif lié à l'X), et situé entre 0,6 et 0,7. Ces résultats indiquent que (i) l'impact des variations faux-sens est globalement le même dans tous les modes de transmission des maladies monogéniques, (ii) la proportion des faux-sens agissant par perte de fonction est probablement plus importante dans les maladies récessives que dans les maladies dominantes.

Pour résumer, les types d'altérations génétiques sont globalement homogènes en fonction des différents modes de transmission, mais leur mécanisme de pathogénicité est distinct : les maladies récessives sont quasi exclusivement liées à la perte de fonction, alors que les maladies dominantes sont souvent liées à des mécanismes différents. Dans cette optique, il a été montré que les variations faux-sens montraient une plus grande propension à la clusterisation dans les maladies dominantes

que dans les maladies récessives, et se localisaient plus fréquemment dans des domaines protéiques fonctionnels[75], cette tendance étant également observée en analysant leur distribution tridimensionnelle[80]. On peut faire l'hypothèse que cet effet est en grande partie médié par les variations faux-sens à effet non-haploinsuffisant.

2.2.3 Propriétés des variations *de novo* récurrentes dans les maladies du développement

Dans ce chapitre, nous proposons d'étudier certaines propriétés des variations survenant à l'état *de novo* chez plusieurs individus de manière indépendante, que nous appellerons variations récurrentes. Nous faisons l'hypothèse que l'identification d'une récurrence mutationnelle entre deux individus porteurs d'un phénotype du même spectre est un élément de grande valeur permettant l'implication de cette variation dans la pathologie. Nous évaluerons l'impact pathogène de ce groupe de variations récurrentes, leur spectre mutationnel, et tenterons d'évaluer des mécanismes possibles. Ces observations, basées sur les données de denovo-db, seront utiles pour la troisième partie de cette thèse, qui visera à exploiter cette récurrence afin d'identifier rétrospectivement de nouvelles variations pathogènes dans une cohorte de patients avec maladies du développement.

Pathogénicité des variations récurrentes chez les patients

Afin d'identifier un potentiel enrichissement en variations pathogènes dans le groupe des variations récurrentes, nous avons extrait de denovo-db 1.6 les variations de type substitutions, chez les cas (toutes maladies confondues), identifiées en exome. Les variations récurrentes (au moins deux occurrences dans denovo-db) et les singletons ont été séparés, et ces deux fichiers ont été ré-annotés grâce à VEP. De manière à enrichir ce jeu de données en variations pathogènes, les variants présents dans GnomAD ont été écartés. Suite à ces étapes, 11 290 singletons et 112 variations récurrentes ont été identifiés. La proportion de variations pathogènes connues dans ClinVar dans ces deux groupes a été comparée. Le groupe des variations récurrentes a montré 47,3 % de variations pathogènes (53/112), contre 2,5 % pour les singletons (284/11 290), soit une proportion 18,8 fois plus importante ($p=1,0.10^{-51}$, test de Fisher). On peut émettre comme critique la probable non-indépendance des données de denovo-db et de ClinVar. En effet il semble plausible que l'identification de plusieurs occurrences de la même variation, potentiellement par plusieurs équipes, puisse augmenter significativement la probabilité qu'elle soit incluse dans ClinVar. Afin de comparer la pathogénicité des variations récurrentes et singletons de manière indépendante d'une interprétation humaine, nous avons évalué la proportion des variations faux-sens avec un score CADD supérieur à 30 dans chacun des deux groupes. Cette analyse a montré un enrichissement significatif dans le groupe des variations récurrentes (16,7 % (n=161) contre 6,6 % (n= 13 989) dans le groupe des singletons, $p=1,62.10^{-6}$, test de Fisher). Au total, ces résultats montrent un fort enrichissement en variations pathogènes parmi les variations récurrentes entre plusieurs individus, dont la grande majorité présente une anomalie du développement.

Spectre mutationnel des variations récurrentes

Dans un travail déjà mentionné (au chapitre 1.3.3), une équipe néerlandaise a identifié 15 clusters de variations faux-sens *de novo* significatifs à l'échelle du génome. Le mécanisme de pathogénicité des 12 clusters déjà connus dans les maladies du développement a été analysé, ce qui a montré que le mécanisme était non-haploinsuffisant dans 8 cas, ce qui était statistiquement enrichi par rapport à la proportion de maladies liées à un effet non-haploinsuffisant de manière générale[77]. Comme les variations récurrentes représentent la forme extrême de clusterisation, nous avons évalué si le spectre mutationnel des variations récurrentes de denovo-db pouvait être biaisé vers les variations à effet NHI, et donc vers les faux-sens. Dans ce but, nous avons évalué le type mutationnel des variations *de novo* identifiées en exome chez les cas de la base denovo-db (Figure 17A). Cette analyse a montré que les variations non récurrentes avaient un spectre un peu différent de celui des variations non récurrentes. La proportion de variations affectant la séquence codante était plus importante dans le groupe des variations récurrentes, avec en particulier une augmentation nette des variations tronquantes aux dépens des variations synonymes ($p=4{,}7.10^{-3}$, est de Fisher comparant d'une part les variations synonymes et d'autre part la somme des variations faux-sens et tronquantes). Une autre manière de présenter ces données a été d'évaluer le pourcentage de récurrence parmi les différents types mutationnels (Figure 17B). L'analyse spécifique des variations synonymes, faux-sens et non-sens a montré que la proportion de variations non-sens récurrentes (2,22 %) était plus élevée que la proportion de faux-sens récurrents (0,99 %), elle-même plus élevée que la proportion de synonymes récurrents (0,63 %).

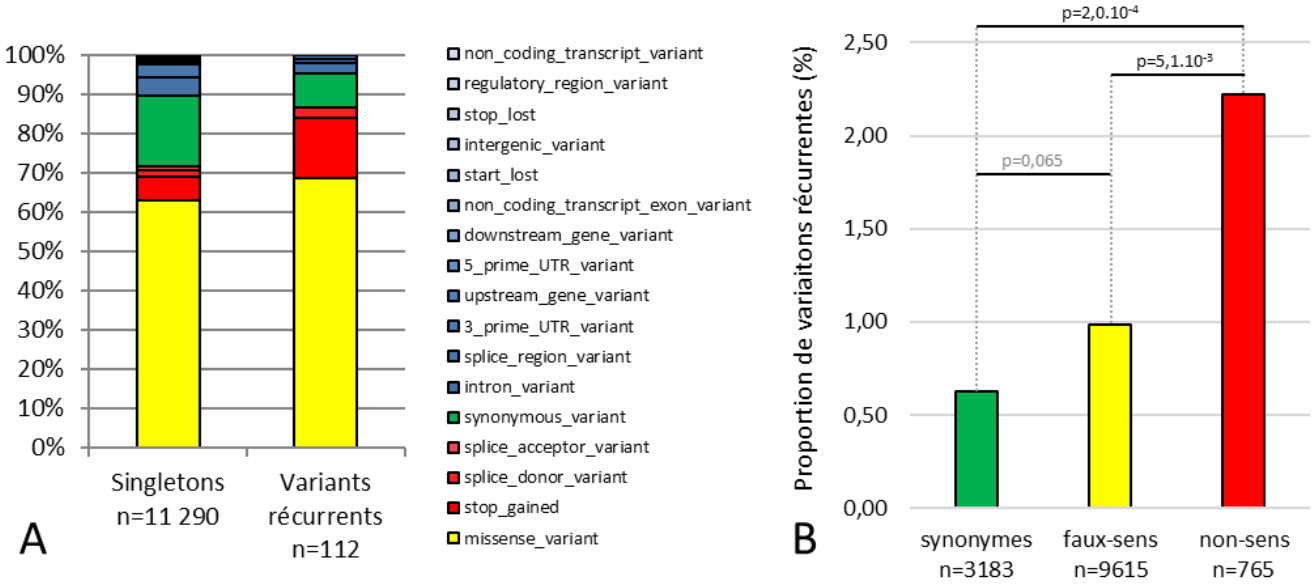


Figure 17. Spectre mutationnel des variations de denovo-db en fonction de leur statut de récurrence
Les variations évaluées sont les variations de type SNV provenant des cas dans denovo-db 1.6, identifiés en exome. A) Spectre mutationnel. Les variations faux-sens sont indiquées en jaune, les variations tronquantes en nuances de rouge, et les autres types mutationnels en nuances de bleu. B) Pourcentage de variations récurrentes pour plusieurs types mutationnels. Les tests statistiques sont des tests exacts de Fisher.

Pour résumer, la présence d'un grand nombre de variations tronquantes récurrentes n'oriente pas vers la prédominance d'un effet non-haploinsuffisant parmi les variations récurrentes. On observe

plutôt un enrichissement des variations récurrentes en variations pathogènes de manière globale, sans présager du mécanisme.

Mécanismes de récurrence : type de substitutions

On peut faire l'hypothèse que certains mécanismes moléculaires plus probables pourraient être à l'origine de « *hotspots* » mutationnels et être plus fréquents dans le groupe des variations récurrentes que dans celui des singletons. En particulier, les transitions CpG, qui sont de type CG > TG (ou CG > CA sur le brin antisens), et qui sont dues à la déamination des cytosines méthylées, sont le type de substitutions le plus commun[85] et ont déjà été observées comme à l'origine de récurrence mutationnelle au sein des populations[15]. Nous avons cherché à évaluer si les transitions CpG étaient plus communes au sein des variations récurrentes. Pour cela nous avons utilisé les données d'exome trio chez les cas de denovo-db et analysé le type de substitutions en fonction du statut de récurrence (Figure 18). Cette analyse a montré que le type de substitutions n'était pas homogène en fonction du statut de récurrence et que les transitions CpG étaient plus communes au sein des variations récurrentes (55,4 %) que chez les singletons (31,5 %, $p=7{,}8.10^{-10}$, test de Fisher). En conclusion, ces résultats montrent une association entre transitions CpG et récurrence mutationnelle au sein des régions exoniques.

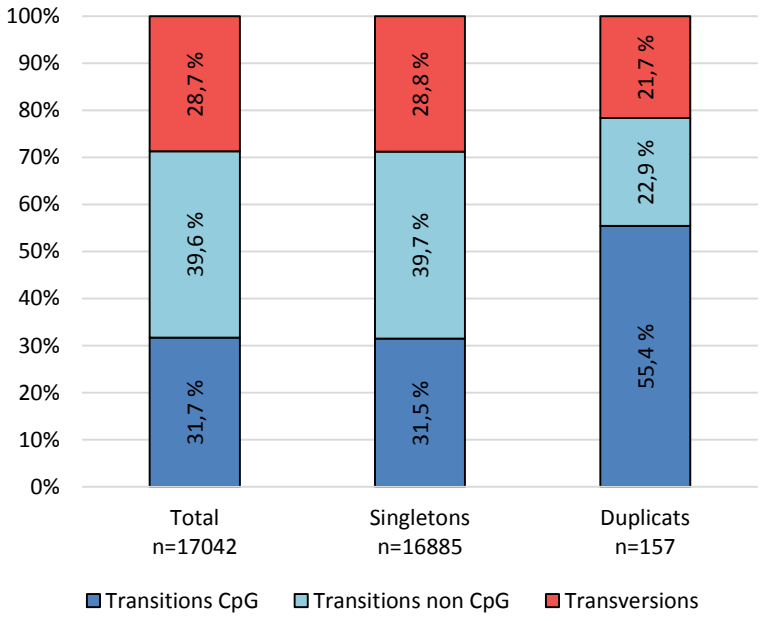
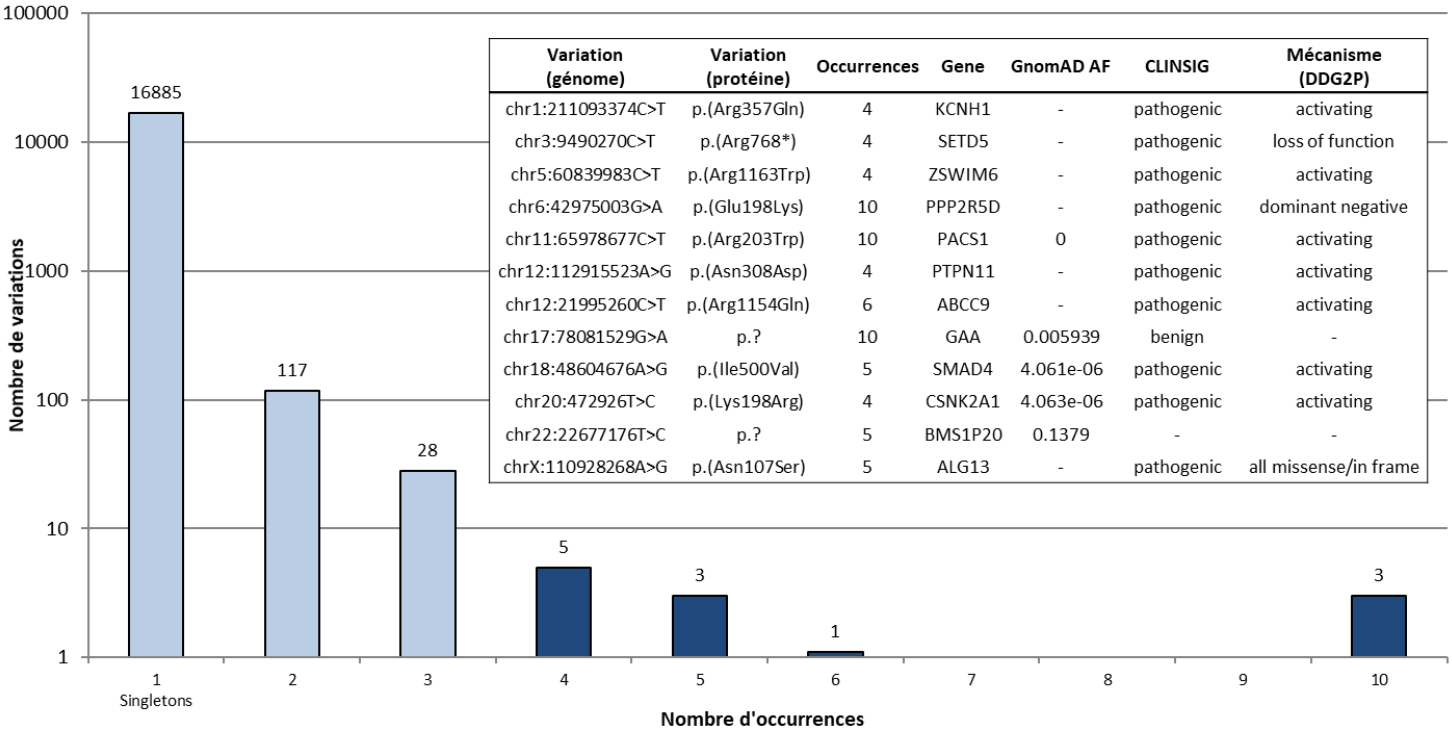


Figure 18. Récurrence mutationnelle et type de substitutions
Le spectre de substitutions *de novo* identifiées chez les cas de denovo-db analysés en exome est présenté en fonction de leur statut de récurrence ou non.

Mécanismes de récurrence : *hotspots* mutationnels

Si la plupart des variants récurrents dans denovo-db sont présents chez deux individus, certaines variations sont présentes chez 3 patients ou plus, jusqu'à 10 occurrences pour des variations codantes (Figure 19). Afin d'investiguer ces variations très récurrentes, nous avons étudié les

données d'exome de denovo-db chez les cas. Douze variations identifiées chez plus de trois individus ont été considérées comme très récurrentes (Figure 19, encart). En dehors de deux variations non codantes potentiellement artéfactuelles car identifiées par une seule équipe (*GAA* et *BMS1P20*), les 10 autres variations étaient pathogènes d'après la base ClinVar. Neuf de ces dix variations étaient faux-sens dans des gènes dont le mécanisme de pathogénicité était non-haploinsuffisant d'après la table GGD2P, avec pour sept d'entre elles le mécanisme « *activating* », qui ne rend compte que de 10,3 % des maladies monoalléliques (107/1031). Malgré les petits effectifs, et en prenant en compte les 12 variations non triées, l'enrichissement en variations dans des gènes avec mécanisme *activating* était statistiquement significatif dans le groupe des variations fortement récurrentes ($p=7,78.10^{-5}$, test de Fisher).



Variation (génome)	Variation (protéine)	Occurrences	Gene	GnomAD AF	CLINSIG	Mécanisme (DDG2P)
chr1:211093374C>T	p.(Arg357Gln)	4	KCNH1	-	pathogenic	activating
chr3:9490270C>T	p.(Arg768*)	4	SETD5	-	pathogenic	loss of function
chr5:60839983C>T	p.(Arg1163Trp)	4	ZSWIM6	-	pathogenic	activating
chr6:42975003G>A	p.(Glu198Lys)	10	PPP2R5D	-	pathogenic	dominant negative
chr11:65978677C>T	p.(Arg203Trp)	10	PACS1	0	pathogenic	activating
chr12:112915523A>G	p.(Asn308Asp)	4	PTPN11	-	pathogenic	activating
chr12:21995260C>T	p.(Arg1154Gln)	6	ABCC9	-	pathogenic	activating
chr17:78081529G>A	p.?	10	GAA	0.005939	benign	-
chr18:48604676A>G	p.(Ile500Val)	5	SMAD4	4.061e-06	pathogenic	activating
chr20:472926T>C	p.(Lys198Arg)	4	CSNK2A1	4.063e-06	pathogenic	activating
chr22:22677176T>C	p.?	5	BMS1P20	0.1379	-	-
chrX:110928268A>G	p.(Asn107Ser)	5	ALG13	-	pathogenic	all missense/in frame

Figure 19. Variations *de novo* très récurrentes
Les variations des cas de denovo-db identifiées en exome sont représentées en fonction de leur nombre d'occurrences, avec une échelle logarithmique. Les variations très récurrentes, avec plus de 3 occurrences, sont représentées en bleu foncé et détaillées en encart. La nomenclature protéique est définie sur le transcrit canonique selon RefSeq. Les autres annotations comprennent la fréquence allélique dans GnomAD, la pathogénicité selon ClinVar et le mécanisme de pathogénicité des gènes impliqués dans les maladies du développement selon la base EBI Gene2Phenotype.

Cette charge élevée en mutations activatrices parmi les variations très récurrentes permet de faire l'hypothèse d'un mécanisme prépondérant de *selfish spermatogonial selection*. Ce phénomène, de découverte récente et identifié par le biais d'un effet majeur de l'âge paternel sur l'apparition de certaines maladies, consiste en un effet activateur de certaines variations somatiques acquises, qui profèrent un avantage sélectif aux spermatogonies porteuses[86]. Ainsi, malgré une probabilité faible d'apparition, la prolifération des spermatogonies porteuses augmente considérablement le nombre de spermatozoïdes porteurs et donc l'incidence des maladies associées à ces variations. Aussi, concernant les 12 variations très récurrentes identifiées, le mécanisme de sélection positive des

spermatogonies a déjà été proposé pour certaines (*SMAD4*[87], *PACS1*[88], *PPP2R5D*[89]), ou bien clairement démontré (*PTPN11*[90]). Dans le but d'identifier des voies cellulaires affectées par ces variations, des analyses d'enrichissement (GO Panther, voir ressources web) ont été effectuées, mais n'ont pas été contributives devant le petit nombre de gènes évalués.

En conclusion, les variations très récurrentes sont très particulières sur le plan fonctionnel par rapport aux autres variations *de novo*, puisqu'elles sont très enrichies en variations faux-sens activatrices. La plupart de ces variations sont probablement responsables de clones cellulaires dans la lignée germinale paternelle expliquant leur forte récurrence. L'évaluation de l'âge paternel des patients porteurs de ces variations serait utile et montrerait probablement un effet marqué par rapport aux autres variations *de novo*.

2.3 Conclusions

En utilisant uniquement des données de génomique disponibles publiquement, nous avons pu identifier de nouvelles corrélations relatives à l'effet des variations faux-sens dans les maladies rares.

En premier lieu, et par deux approches complémentaires, nous avons montré que l'impact des variations faux-sens et la proportion d'effet non-haploinsuffisant augmentaient parmi les maladies ultra-rares (<1/1 000 000) mono-alléliques.

Ensuite, nous avons confirmé le fait bien établi que malgré une proportion similaire de variations faux-sens pathogène entre les maladies monogéniques autosomiques dominantes et autosomiques récessives, l'effet non-haploinsuffisant était spécifique des maladies mono-alléliques.

Enfin, nous avons analysé les variations *de novo* récurrentes chez les patients atteints de maladies du développement. Nous avons pu mettre en évidence le rôle des transitions CpG, ainsi que celui de la pression de sélection positive dans les spermatogonies, dans les mécanismes de récurrence. Enfin, nous avons montré que ces variations récurrentes étaient très fortement enrichies en variations pathogènes, ce qui nous a incités à utiliser cette stratégie sur des données de patients dans l'objectif d'identifier de nouvelles variations génétiques pathogènes potentiellement responsables d'un effet NHI.

3 MISE EN APPLICATION : UTILISATION DE DONNÉES PUBLIQUES POUR L'IDENTIFICATION DE VARIANTS FAUX-SENS PATHOGÈNES PAR RÉCURRENCE MUTATIONNELLE

3.1 Contexte

Les analyses en exome complet permettent d'identifier une cause monogénique dans une proportion importante de patients avec maladie du développement[91]. Cependant, un grand nombre de patients évocateurs de maladies monogéniques reste sans diagnostic après une analyse d'exome. Chez ces patients, des ré-analyses rétrospectives *a posteriori* ont montré leur intérêt[92], en permettant (i) l'identification de variations dans des gènes nouvellement impliqués en pathologie, (ii) la reclassification de variations grâce à des données supplémentaires, et (iii) l'identification de nouvelles variations par des approches additionnelles (e.g. recherche de variations de structure). Ainsi il est bien établi que les données d'exome contiennent une proportion d'informations de valeur potentiellement non exploitées pouvant être identifiées par des approches complémentaires.

Nous avons observé dans le chapitre précédent que les mutations *de novo* identifiées chez plusieurs patients non apparentés étaient très fortement enrichies en variations pathogènes. Nous avons fait l'hypothèse que cette récurrence pourrait être un argument de poids permettant d'identifier de nouvelles variations pathogènes rétrospectivement au sein d'une cohorte d'exomes chez des patients avec maladie du développement. Dans ce contexte, nous avons eu la chance de pouvoir accéder aux données d'exome de plus de 1200 patients avec anomalie du développement analysés au laboratoire de génétique du CHU de Dijon. Nous avons tenté d'identifier des variations « récurrentes » entre la série dijonnaise et denovo-db. Contrairement au chapitre précédent, cette approche comparait des variations *de novo* (présentes dans denovo-db) avec des variations dont la ségrégation parentale n'était pas forcément connue, les exomes dijonnais étant très majoritairement réalisés en solo. Nous avons fait le pari qu'un nombre conséquent de variations récurrentes seraient néanmoins également *de novo* dans la cohorte dijonnaise et que cette approche permettrait d'identifier des variations d'intérêt.

3.2 Méthodes

3.2.1 Description de la cohorte étudiée

Au moment de cette étude, la série dijonnaise comprenait 1271 entrées d'exomes chez des cas index avec anomalie de développement. La majorité avait été séquencée en solo (n=1036), mais des duos, trios et trio+ d'apparentés étaient également présents dans la base (n=235). La plupart des patients étaient inclus sous la mention générique "anomalie de développement" (n=717), alors que d'autres patients appartenaient à des sous-cohortes plus spécifiques mais toujours dans le champ des anomalies du développement, telles que « foetopathologie » (n=87), « déficience intellectuelle avec habitus marfanoïde » (n=79), « déficience intellectuelle », (n=66), etc. Parmi ces 1271 entrées, 376

analyses (29,6 %) étaient annotées comme positives, c'est-à-dire avec variation causale identifiée, 137 (10,8 %) comme non concluantes et 758 (59,6 %) comme négatives.

3.2.2 Établissement d'une liste de variations d'intérêt à partir de denovo-db

Nous avons téléchargé un fichier contenant 283,888 variations à partir de denovo-db version 1.5 depuis le site public (voir ressources web). Pour chaque variation, plusieurs classes d'annotations étaient disponibles, incluant des annotations standards comme la prédiction d'effet fonctionnel sur le transcrit et sur la protéine, la fréquence dans diverses bases de données, mais également la sous-cohorte de denovo-db de laquelle était issue cette variation *de novo* (ex : anomalie du développement, cardiopathie congénitale, etc., voir Figure 20), fournissant une information clinique minimale concernant le patient porteur de cette variation. Seules les séries cliniques en lien avec les présentations phénotypiques investiguées au laboratoire de génétique de Dijon ont été incluses : parmi les 18 cadres cliniques disponibles dans denovo-db, nous avons sélectionné les variations de six sous-cohortes en lien avec les maladies pédiatriques du développement : autisme, déficience intellectuelle, maladie du développement, épilepsie, anomalies du tube neural et dysplasie fronto-nasale acromélique.

Nous avons ensuite exclu les variations observées au moins une fois dans la base de données ExAC, sur les arguments suivants : (i) les individus séquencés dans ExAC sont exempts de maladie pédiatrique sévère et (ii) la plupart des SNVs impliqués dans les maladies du développement avec un mécanisme *de novo* sont de pénétrance forte, avec très peu d'individus résilients dans les bases de données d'individus contrôles (voir chapitre 1.2.2.1). Nous avons également filtré les variations de denovo-db sur la base de leur effet fonctionnel en ne conservant que les variations avec un effet codant prédit. Les variations intergéniques, introniques et synonymes ont été exclues car la plupart de ces variations n'étaient pas couvertes en exome, et car leur interprétation serait plus difficile, avec une proportion mineure de ces variations ayant un potentiel de pathogénicité. Enfin, nous avons exclu les indels, en ne conservant que les substitutions nucléotidiques, car les indels ont plus de probabilité d'être des faux positifs, que ce soit au sein des exomes réalisés et/ou analysés dans le laboratoire ou répertoriés dans la base denovo-db. Ce processus de filtration de denovo-db a conduit à sélectionner une liste de 8,435 évènements *de novo*, correspondant à 8,267 variations distinctes identifiées chez au moins un individu avec anomalie du développement.

3.2.3 Identification de récurrence entre les exomes dijonnais et denovo-db

Nous avons requêté dans la série dijonnaise la présence des 8,267 variations d'intérêt extraites de denovo-db, et également des variations avec changement nucléotidique distinct par rapport aux variations de denovo-db, mais à la même position génomique. Nous avons appliqué cette stratégie à l'ensemble de la série dijonnaise, sans prendre en compte, dans un premier temps, ni le phénotype investigué ni la présence éventuelle d'une variation causale préalablement identifiée par l'analyse d'exome. Les variations rares identifiées par cette stratégie ont été analysées plus en détail en commençant par une évaluation de la qualité des variations (visualisation des fichiers d'alignement sur le logiciel IGV si besoin), puis en prenant en compte la présentation clinique du patient, la cohorte clinique de l'individu de denovo-db et porteur de la même variation, la littérature disponible

sur le gène et la variation, et des logiciels d'analyse génomique en accès libre comme UCSC Genome Browser20 et VarSome (voir ressources web). Les confirmations et la ségrégation parentale des variations d'intérêt ont été réalisées par méthode de Sanger avec une procédure standard. La Figure 20 résume la stratégie globale de cette étude.

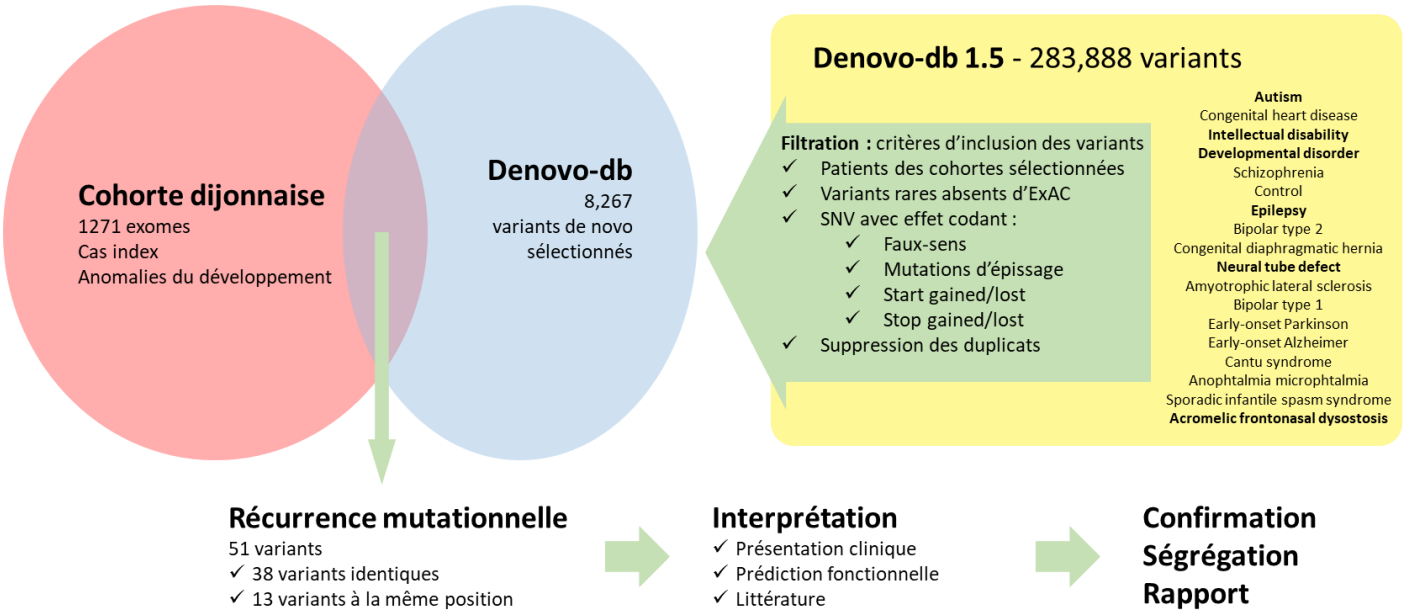


Figure 20. Schéma général de l'étude
Les sous-cohortes de denovo-db incluses sont indiquées en gras.

3.3 Résultats

3.3.1 Caractéristiques des variations identifiées

Notre stratégie a identifié 186 variants correspondant soit strictement à une variation identifiée chez un individu avec une maladie du développement et répertoriée dans denovo-db, soit à une variation distincte affectant le même nucléotide. Après exclusion des variants présents dans GnomAD et des variations de mauvaise qualité, 51 variations rares de bonne qualité ont été étudiées plus en détail (Figure 21).

La grande majorité de ces variations étaient des faux-sens (n=45), mais cinq variations non-sens et une variation synonyme ont également été identifiées. L'analyse du mécanisme moléculaire à l'origine de la récurrence a montré, en accord avec les résultats du chapitre précédent, qu'une forte proportion des variations avec récurrence mutationnelle stricte correspondait à des transitions CpG (21/38, 55 %). Le deuxième mécanisme de récurrence identifié dans le chapitre précédent correspondant à la sélection positive des spermatogonies, a également pu être mis en évidence dans cette étude, avec en particulier la présence au sein des exomes dijonnais de la variation extrêmement récurrente *PPP2R5D*(NM_006245.3): p.(Glu198Lys).

3.3.2 Validation de la pertinence de l'approche par identification de variations pathogènes connues

Comme anticipé, il est apparu très clairement que la liste des variations identifiées était fortement enrichie en variations pathogènes dans le champ des maladies du développement (Figure 21). En effet, une grande proportion des gènes identifiés était associée à des anomalies du développement causées par des mutations *de novo* (26/51, 51 %), selon OMIM. Par ailleurs, 15 variations (29,4 %) étaient annotées comme pathogène ou probablement pathogène par au moins un utilisateur dans la base ClinVar. Parmi les 51 variations identifiées, 23 variations (45 %) étaient déjà connues par l'équipe dijonnaise comme responsables de la maladie (ou fortement suspectées) chez 25 cas index (Tableau 8). Ces résultats ont validé notre approche basée sur la récurrence mutationnelle comme efficace pour identifier des variations pathogènes au sein de la série.

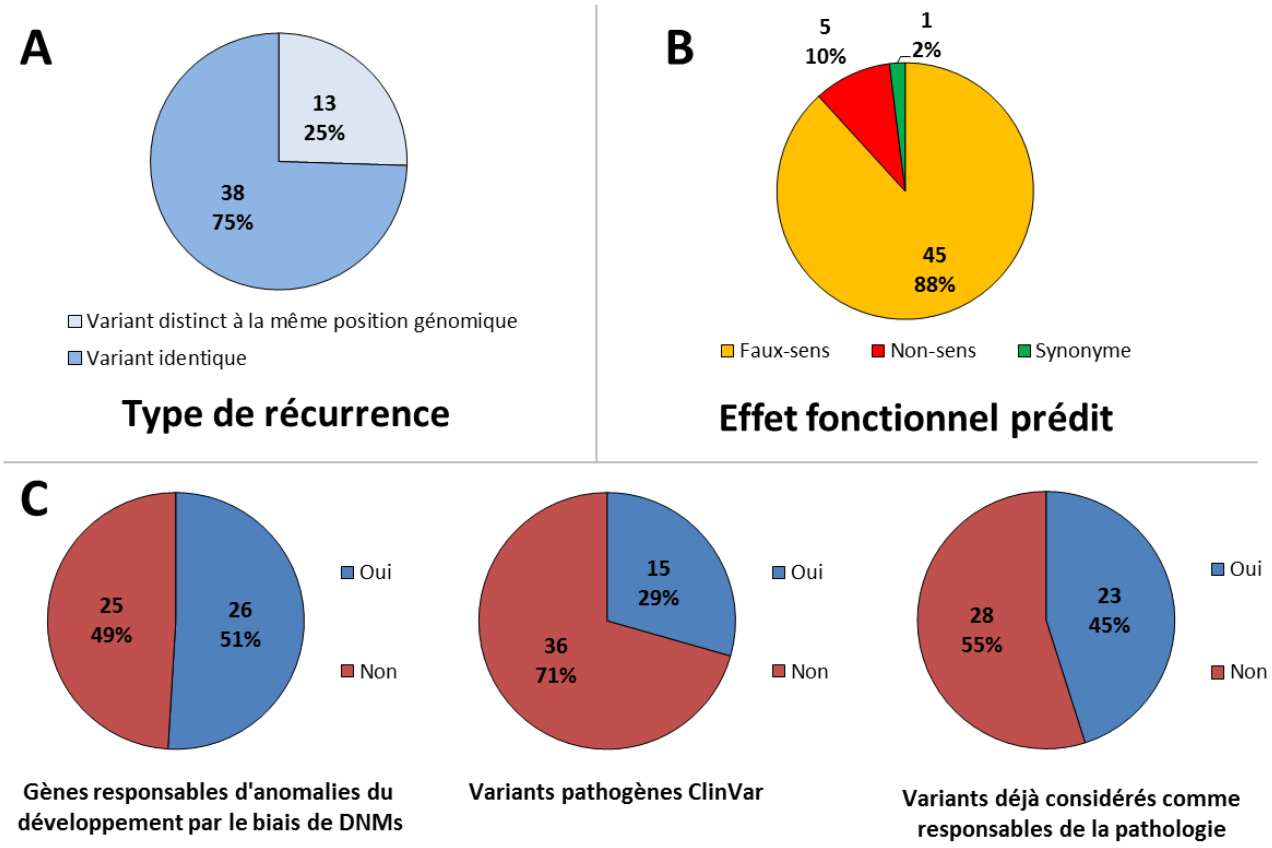


Figure 21. Caractéristiques des 51 variations ou positions génomiques récurrentes identifiées à la fois dans denovo-db et au sein de la série dijonnaise
A) Proportion de variations strictement récurrentes. B) Effet fonctionnel prédit. C) Enrichissement en variations d'intérêt clinique d'après les bases de données OMIM, ClinVar, et les interprétations clinico-biologiques dijonnaises préalables à cette étude.

.

Variation (hg19)	Variation (RefSeq)	Gène	# Individus denovo-db	# Individus cohorte locale	Clinvar	Ségrégation	#OMIM	Maladie OMIM compatible avec la présentation clinique	Date de création de l'entrée OMIM	Interprétation
chr1:g.210977491T>C	NM_172362.2:c.1480A>G, p.(Ile494Val)	*KCNH1*	1	1	Oui	*De novo*	#611816	Temple-Baraitser syndrome, Autosomal dominant	Février 2008	Pathogène
chr3:g.176750839A>G	NM_024665.5:c.1336T>C, p.(Tyr446His)	*TBL1XR1*	0	1	Non	*De novo*	#616944	Mental retardation, autosomal dominant 41	Mai 2016	Pathogène
chr6:g.33403367C>T	NM_006772.2:c.739C>T, p.(Gln247Ter)	*SYNGAP1*	1	1	Non	*De novo*	#612621	Mental retardation, autosomal dominant 5	Février 2009	Pathogène
chr6:g.42975003G>A	NM_006245.3:c.592G>A, p.(Glu198Lys)	*PPP2R5D*	10	1	Oui	*De novo*	#616355	Mental retardation, autosomal dominant 35	Mai 2015	Pathogène
chr7:g.40085606A>G	NM_003718.4:c.2525A>G, p.(Asn842Ser)	*CDK13*	3	1	Oui	*De novo*	#617360	Congenital heart defects, dysmorphic facial features, and intellectual developmental disorder, Autosomal dominant	Février 2017	Pathogène
chr7:g.44283125G>A	NM_001220.4:c.416C>T, p.(Pro139Leu)	*CAMK2B*	1	1	Non	*De novo*	#617799	Mental retardation, autosomal dominant 54	Décembre 2017	Pathogène
chr12:g.112926888G>A	NM_002834.4:c.1508G>A, p.(Gly503Glu)	*PTPN11*	1	1	Oui	Hérité de la mère symptomatique	#151100	LEOPARD syndrome 1, Autosomal dominant	Juin 1986	Pathogène
chr15:g.83348481C>T	NM_004644.4:c.1182G>A, p.(=)	*AP3B2*	0	1	Oui	Hérité de la mère, en mosaïque chez la mère	#617276	Epileptic encephalopathy, early infantile, 48, Autosomal recessive	Décembre 1997	Pathogène
chr16:g.56226265G>A	NM_138736.2:c.118G>A, p.(Gly40Arg)	*GNAO1*	0	1	Non	*De novo*	#615473	Epileptic encephalopathy, early infantile, 17, Autosomal dominant	Octobre 2013	Pathogène
chr19:g.13246913C>T	NM_052876.3:c.892C>T, p.(Arg298Trp)	*NACC1*	1	1	Oui	*De novo*	#617393	Neurodevelopmental disorder with epilepsy, cataracts, feeding difficulties, and delayed brain myelination, Autosomal dominant	Mars 2017	Pathogène
chr20:g.49509094G>GT	NM_181442.3:c.2156dupA, p.(Tyr719*)	*ADNP*	1	1	Non	*De novo*	#615873	Helsmoortel-van der Aa syndrome, Autosomal dominant	Juillet 2014	Pathogène
chr20:g.62073781G>A	NM_172107.3:c.794C>T, p.(Ala265Val)	*KCNQ2*	1	1	Oui	*De novo*	#613720	Epileptic encephalopathy, early infantile, 7, Autosomal dominant	Février 2011	Pathogène
chr20:g.62126409C>T	NM_001958.3:c.370G>A, p.(Glu124Lys)	*EEF1A2*	1	1	Oui	*De novo*	#616393	Mental retardation, autosomal dominant 38	Mai 2015	Pathogène
chrX:g.133559286C>T	NM_032458.2:c.1024C>T, p.(Arg342Ter)	*PHF6*	1	1	Oui	*De novo* hémizygote	#301900	Borjeson-Forssman-Lehmann syndrome, X-linked recessive	Juin 1986	Pathogène
chrX:g.153197526A>C	NM_003491.3:c.384T>G, p.(Phe128Leu)	*NAA10*	0	1	Oui	*De novo* hétérozygote	#300855	Ogden syndrome, X-linked dominant	Aout 2011	Pathogène
chrX:g.153296516G>A	NM_001110792.1:c.799C>T, p.(Arg267Ter)	*MECP2*	3	1	Oui	*De novo* hétérozygote	#312750	Rett syndrome, X-linked dominant	Juin 1986	Pathogène
chr19:g.13136154G>A	NM_001271043.2:c.371G>A, p.(Arg124Gln)	*NFIX*	1	1	Non	N/A	#602535 #614753	Marshall-Smith syndrome, Autosomal dominant Sotos syndrome 2, Autosomal dominant	Avril 1998, Aout 2012	VSI
chr22:g.41572254T>G	NM_001429.3:c.4783T>G, p.(Phe1595Val)	*EP300*	1	1	Oui	Hérité du père asymptomatique	#613684	Rubinstein Taybi syndrome 2, Autosomal dominant	Janvier 2011	VSI
chr8:g.133192493G>A	NM_004519.3:c.688C>T, p.(Arg230Cys)	*KCNQ3*	3	1	Oui	*De novo*	.	.	.	Variation recherche
chr9:g.2058457G>A	NM_003070.4:c.1514G>A, p.(Arg505Gln)	*SMARCA2*	1	1	Non	*De novo*	.	.	.	Variation recherche
chr10:g.78869939C>T	NM_001161352.1:c.1123G>A, p.(Gly375Arg)	*KCNMA1*	1	2	Non	*De novo / De novo* (deux cas index)	.	.	.	Variation recherche
chr14:g.105834449G>A	NM_001100913.2:c.625G>A, p.(Glu209Lys)	*PACS2*	3	2	Non	*De novo / De novo* (deux cas index)	.	.	.	Variation recherche
chr16:g.2226351G>A	NM_032271.2:c.1964G>A, p.(Arg655Gln)	*TRAF7*	1	1	Non	*De novo*	.	.	.	Variation recherche

Tableau 8 : Variations identifiées par récurrence mutationnelle déjà connues comme pathogènes ou candidates dans la cohorte dijonnaise
La mention 0 concernant le nombre d'individus porteurs dans denovo-db est due à une variation à la même position mais avec un changement nucléotidique distinct. La colonne ClinVar renseigne l'existence ou non d'au moins une soumission d'une variation comme pathogène ou probablement pathogène dans la base ClinVar. VSI : variation de signification indéterminée.

3.3.3 Identification de nouvelles variations pathogènes

3.3.3.1 Nouvelles variations dans des gènes récemment décrits

En dehors des 23 variations déjà interprétées comme pathogènes avant cette étude, notre approche a identifié 28 variations supplémentaires préalablement non retenues et ayant été identifiées à l'état *de novo* chez au moins un individu avec anomalie du développement dans une cohorte de denovo-db. Une évaluation de chacune de ces variations candidates a été effectuée, en se basant sur les recommandations ACMG[93]. Les critères pris en compte dans cette interprétation étaient la concordance entre le phénotype du patient et la sous-cohorte de denovo-db, la compatibilité du gène (fonction, expression et surtout maladie potentiellement associée dans OMIM et bibliographie), et la compatibilité de la variation (transcrit affecté, prédictions fonctionnelles, fréquence de la variation dans les bases de données de population générale et éventuellement présence de la variation dans les bases de données de variations pathogènes ClinVar et HGMD).

Dix-huit variations n'ont pas été retenues comme de bons candidats, du fait (i) qu'elles affectaient un gène non relevant (ex. exprimé uniquement dans le tissu cutané), (ii) qu'elles étaient également identifiées chez des contrôles ou parents sains, dans denovo-db ou dans la cohorte dijonnaise, ou (iii) qu'elles étaient identifiées chez un patient avec une autre variation définie comme pathogène avec contribution complète au phénotype.

Dix variations (20 %) ont été considérées comme des bonnes variations candidates et ont de fait été validées en séquençage Sanger chez le cas index et ses parents (et chez le frère atteint dans le cas du gène *ACTL6B*) (Tableau 9). Deux de ces 10 variations candidates ont été écartées car héritées d'un parent asymptomatique. Les 8 variations restantes étaient soit survenues *de novo* (n=6), soit montraient une ségrégation compatible avec un mode de transmission autosomique dominant (n=1, *ACTL6B*) ou récessif lié à l'X (n=1, *ZFX*). Cinq variations *de novo* affectaient des gènes avec une entrée OMIM compatible avec la présentation clinique du patient et ont été considérées comme pathogènes de rang diagnostique.

De manière à identifier la raison pour laquelle ces variations n'avaient pas été identifiées lors de la première analyse des exomes, nous avons relevé la date de création des entrées OMIM correspondantes, dans l'idée qu'elles pourraient être récentes. Effectivement, 3 entrées avaient été créées moins de 3 mois avant notre étude, et les deux entrées restantes dataient d'entre 1 et 2 ans, ce qui était globalement bien plus récent que les variations déjà identifiées comme pathogènes décrites dans le Tableau 8.

3.3.3.2 Identification de nouveaux gènes candidats

Trois de ces 10 variations sont restées candidates après la ségrégation : la variation p.(Arg764Trp) dans le gène *ZFX*, la variation p.(Gly343Arg) dans le gène *ACTL6B*, et la variation *de novo* p.(Arg126Gln) dans le gène *FEM1B*.

Des variations du gène *ZFX* n'ont jamais été associées aux maladies du développement. Nous avons identifié la variation p.(Arg764Trp) à l'état hémizygote chez un patient avec anomalie de développement syndromique de cause inconnue. Ce patient présentait une maladie diffuse avec

atteinte rénale malformative aggravée par une oxalurie avec néphrocalcinose, une atteinte endocrinienne, cardiaque, une angiomatose cutanée et hépatique, un retard de développement et une dysmorphie faciale. La variation identifiée dans *ZFX* était héritée de la mère asymptomatique et pour le moment aucune analyse de ségrégation plus étendue n'a été entreprise. Cette variation a également été identifiée dans denovo-db, également à l'état hémizygote, mais à l'état *de novo,* chez un individu avec anomalie de développement. Au total, ces résultats sont prometteurs et des investigations plus approfondies sont en cours afin de statuer sur cette variation d'intérêt.

Des éléments en faveur de l'implication des variations *de novo* du gène *ACTL6B* dans les maladies du développement sont venus d'une étude récente qui a montré l'existence d'une clusterisation significative de variations faux-sens *de novo* chez des patients avec maladies du développement[77]. Ce « cluster » était en réalité la même variation faux-sens à l'état *de novo* chez trois patients non apparentés. Le phénotype de ces patients n'a pas encore été décrit. Notre approche a identifié un changement de séquence différent mais menant au même effet faux-sens p.(Gly343Arg), chez deux enfants de la même fratrie présentant un retard global de développement avec hypotonie. Ce variant n'était pas hérité de la mère, et l'ADN du père, porteur d'une déficience intellectuelle, n'était pas disponible. De fait, cette variation était pertinente sur un mode de transmission autosomique dominant dans cette famille. Plusieurs patients porteurs de variations *ACTL6B* sont actuellement collectés afin de caractériser le phénotype associé plus en détail (manuscrit en cours par des collaborateurs).

FEM1B est un petit gène ubiquitaire qui n'a jamais été impliqué dans les anomalies du développement. La récurrence mutationnelle a mis en évidence la variation p.(Arg126Gln) chez une patiente avec retard global de développement syndromique. L'analyse de ségrégation a montré le caractère *de novo* de cette variation. L'individu porteur de la même variation *de novo* dans la base denovo-db était issu de l'étude britannique DDD. La collaboration avec cette équipe a permis d'identifier chez cette jeune patiente un phénotype très similaire de celui de la patiente française. Le site GeneMatcher, une plateforme de *data sharing* permettant d'établir des connexions entre des médecins, chercheurs et patients partageant un intérêt pour le même gène[94], a été utilisé, ce qui a permis d'identifier une troisième patiente porteuse de la même mutation *de novo,* détectée en exome trio. La précision du phénotype de cette patiente néerlandaise était également très concordante. Ainsi, nous avons identifié la même variation *de novo* chez trois individus porteurs d'un phénotype neurodéveloppemental très similaire, et assez spécifique, présenté en Tableau 10. Les particularités morphologiques de ces trois patientes, non présentées ici, ont été considérées comme ressemblantes par leurs trois généticiens cliniciens référents.

Variation (hg19)	Variation (RefSeq)	Gène	# Individus denovo-db	# Individus cohorte locale	Clinvar	Ségrégation	#OMIM	Maladie OMIM compatible avec la présentation clinique	Date de l'entrée OMIM	Interprétation
chr1:g.1737942A>G	NM_002074.4:c.239T>C p.(Ile80Thr)	*GNB1*	1	1	Oui	*De novo*	#616973	Mental retardation, autosomal dominant 42	Juin 2016	Pathogène
chr1:g.26784371G>A	NM_024887.3:c.632G>A p.(Arg211Gln)	*DHDDS*	1	1	Non	*De novo*	#617836	Developmental delay and seizures with or without movement abnormalities	Janvier 2018	Pathogène
chr5:g.160758065T>C	NM_021911.2:c.902A>G p.(Tyr301Cys)	*GABRB2*	1	1	Oui	*De novo*	#617829	Epileptic encephalopathy, infantile or early childhood, 2	Janvier 2018	Pathogène
chr17:g.57754422C>T	NM_004859.3:c.2669C>T p.(Pro890Leu)	*CLTC*	1	1	Non	*De novo*	#617854	Mental retardation, autosomal dominant 56	Janvier 2018	Pathogène
chr19:g.13342664C>T	NM_001127221.1:c.5263G>A p.(Gly1755Arg)	*CACNA1*	1	1	Non	*De novo*	#617106	Epileptic encephalopathy, early infantile, 42	Aout 2016	Pathogène
chr7:g.100244260C>G	NM_016188.4:c.1027G>C p.(Gly343Arg)	*ACTL6B*	0	1	Non	Validé chez le cas index et le frère symptomatique. Absent chez la mère. Père symptomatique non testé	-	-	-	Variation recherche
chr15:g.68582073G>A	NM_015322.4:c.377G>A p.(Arg126Gln)	*FEM1B*	1	1	Non	*De novo*	-	-	-	Variation recherche
chrX:g.24229365C>T	NM_001178084.1:c.2290C>T p.(Arg764Trp)	*ZFX*	1	1	Non	Hémizygote, hérité de la mère	-	-	-	Variation recherche / VSI
chr12:g.122252499C>G	NM_001353345.1:c.2378C>G p.(Pro793Arg)	*SETD1B*	1	1	Non	Hérité d'un parent asymptomatique	-	-	-	Variation recherche / Probablement bénin
chr16:g.2026948C>T	NM_006453.2:c.1426C>T p.(Arg476Cys)	*TBL3*	1	1	Non	Hérité d'un parent asymptomatique	-	-	-	Variation recherche / Probablement bénin

Tableau 9 : Variations d'intérêt identifiées rétrospectivement grâce à la récurrence mutationnelle
Ce tableau présente les neuf variations candidates ayant mené à une validation et ségrégation Sanger. La mention 0 concernant le nombre d'individus porteurs dans denovo-db est due à une variation à la même position mais avec un changement nucléotidique distinct. À noter que la variation identifiée dans la cohorte dijonnaise au sein du gène ACTL6B était une substitution différente de celle présente dans denovo-db, mais dont l'effet faux-sens était identique du fait de la redondance du code génétique. La colonne ClinVar renseigne l'existence ou non d'au moins une soumission de la variation comme pathogène ou probablement pathogène dans la base ClinVar. VSI : variation de signification indéterminée.

Patient	Méthode	Age	Indication	Malformations / déformations		Digestif	Développement neurologique				Comportement	Neurosensoriel	Dysmorphie faciale
				Cardiologie	Membres		Hypotonie	Retard global	Age de la position assise	Age de la marche			
Patiente 1 (France)	Exome solo	3,5 ans	Anomalie de développement	Foramen ovale perméable	Pieds bots varus équins	Sténose du pylore	+	+	Retardé : 13 mois	N/A	Auto agressivité	Surdité	+
Patiente 2 (Ecosse)	Exome trio	15 ans	Anomalie de développement	-	Pieds valgus	-	N/A	+	N/A	N/A Réticence à la marche	Anxiété Auto agressivité	N/A	+
Patiente 3 (Pays-Bas)	Exome trio	2 ans	Anomalie de développement	Communication interventriculaire	Camptodactylies des doigts	-	N/A	+	Retardé : 18 mois	N/A	Automutilations	Surdité	+

Tableau 10 : Résumé clinique des trois patientes porteuses de la variation *FEM1B*:p.(Arg126Gln) à l'état *de novo*
N/A : non disponible.

3.4 Discussion

Notre approche a rétrospectivement mis en évidence un petit nombre de variations au sein de la série de plus de 1200 exomes. Ce set de variations était fortement enrichi en variations pathogènes, aussi bien des variations déjà considérées comme pathogènes ou bien nouvellement identifiées. Une grande proportion de ces variations a été identifiée chez les patients dijonnais comme également de survenue *de novo*, indiquant la présence d'une récurrence du même évènement mutationnel chez deux individus avec maladie développementale. La récurrence mutationnelle n'est pas suffisante pour impliquer un variant avec certitude comme cause de la pathologie, mais semble être un élément de poids dans le processus de classification du variant. La récurrence mutationnelle peut survenir par hasard (comme le variant dans *TBL3*), dû à la taille importante des cohortes en jeu[15] et à la mutabilité du génome mais deux éléments nous ont permis de diminuer ce bruit de fond de manière intéressante. D'une part, nous avons seulement considéré les variations absentes de la population générale (MAF = 0 dans la population ExAC). De fait, de nombreuses positions génomiques avec haute mutabilité ont pu être exclues. Deuxièmement, l'intégration des données phénotypiques a été très utile, et la comparaison du phénotype des patients avec l'information clinique minimale représentée par la sous-cohorte de laquelle provenait l'individu de denovo-db était déjà très intéressante pour identifier des candidats sérieux.

La frontière entre variations de recherche et variations de rang diagnostique utilisables pour le conseil génétique des patients et de leur famille a été positionnée de manière pragmatique par l'équipe dijonnaise en fonction de la présence d'une description génotype-phénotype concordante dans le catalogue OMIM. Notre approche basée sur la récurrence mutationnelle a permis d'identifier aussi bien des variations candidates de recherche que des variations de rang diagnostique.

3.4.1 Identification de variations pathogènes dans des gènes OMIM

La plupart des variations identifiées par récurrence mutationnelle dans des relations génotype-phénotype connues avaient été identifiées préalablement par la première lecture de l'exome (20/25 patients). Cinq nouveaux diagnostics ont pu être faits, dans le cadre de relations génotype-phénotype de description récente, qui étaient inconnues lors de l'analyse première des exomes. La plupart des variations identifiées étaient impliquées dans des maladies autosomiques dominantes ou liées à l'X causées majoritairement par des mutations *de novo*. Le gène *AP3B2* a fait exception puisqu'il s'agit d'un gène responsable d'encéphalopathie épileptique autosomique récessive. Le patient identifié était porteur de deux variations hétérozygotes composites héritées. Néanmoins, le processus mutationnel a quand même pu être observé puisque la variation identifiée par récurrence mutationnelle était présente en mosaïque dans le sang maternel, comme décrit précédemment par l'équipe dijonnaise[95].

Un autre élément relatif aux variations de rang diagnostique implique un faux-sens dans le gène *EP300* identifié comme *de novo* chez un patient de la cohorte DDD. Ce patient présentait des caractéristiques évocatrices de syndrome de Rubinstein-Taybi mais dans une

forme potentiellement plus modérée[96]. De fait ce variant avait été considéré comme pathogène et indiqué comme tel dans la base ClinVar. Nous avons observé ce même variant chez un cas index avec maladie neurodéveloppementale non résolue, mais ne présentant pas les caractéristiques morphologiques associées au syndrome de Rubinstein-Taybi. De plus, ce variant était hérité du père strictement asymptomatique, sans arguments pour une éventuelle mosaïque. Ainsi, ce variant peut être un variant à pénétrance incomplète, un facteur de risque de maladie neurodéveloppementale non associé à un phénotype spécifique, ou un variant bénin impliqué de manière erronée par son caractère *de novo*. Il n'a pas été rendu comme pathogène et a été considéré comme étant de signification inconnue.

3.4.2 Identification de nouvelles relations génotype-phénotype

Notre approche a mis en évidence des variations d'intérêt dans des gènes sans phénotype neurodéveloppemental concordant décrit. Cinq de ces variations avaient déjà été identifiées par l'équipe dijonnaise, incluant un variant de *PACS2* de description récente[78] à l'état *de novo* chez deux individus non apparentés, un variant de *TRAF7* (collaboration en cours), et trois variations dans des gènes avec un nouveau phénotype possible. Ces variations, dans les gènes *KCNMA1*, *KCNQ3* et *SMARCA2*, ont été identifiées chez des patients présentant un phénotype non compatible avec les maladies associées : la dyskinésie paroxystique non kinésigénique avec ou sans épilepsie généralisée (OMIM #609446), l'épilepsie néonatale bénigne (OMIM #121201) et le syndrome de Nicolaides-Baraitser (OMIM #601358), respectivement. L'identification d'une récurrence mutationnelle conforte le choix de ces variations en tant que bons candidats dans les maladies neurodéveloppementales. La description clinique de ces patients, au sein de cohortes plus larges, est actuellement en cours par l'équipe dijonnaise.

La récurrence mutationnelle a également permis d'identifier de nouvelles variations candidates, dans les gènes *ACTL6B*, *FEM1B* et *ZFX*. La variation du gène *FEM1B* montre la puissance de la stratégie basée sur la récurrence mutationnelle pour identifier des nouveaux gènes responsables de maladies du développement liées à des variations très spécifiques et clusterisées. Le mécanisme de pathogénicité de cette variation n'a pas pu être précisé sur le plan fonctionnel. Les données de contrainte en population générale ne montrent pas d'intolérance du gène *FEM1B* aux variations perte de fonction (pLI = 0,54), mais la fiabilité de cette mesure est probablement mise en cause par (i) la petite taille du gène et (ii) le fait que le gène *FEM1B* n'est constitué que de deux exons codants, les variations tronquantes pouvant potentiellement être de conséquence atypique (échappement au *nonsense mediated decay*, NMD). La protéine FEM1B, tout comme la protéine VHL bien connue en oncogénétique, correspond à une sous-unité de reconnaissance du substrat de certains complexes ubiquitine ligase CUL2-RING E3[98]. La variation faux-sens que nous avons identifiée se localise au sein d'un des domaines ankyrine, correspondant à des domaines d'interaction protéine-protéine. Ainsi on peut imaginer que cette variation puisse modifier l'interaction de ce complexe d'ubiquitinylation avec son substrat, et ainsi perturber la spécificité du ciblage des protéines modifiées par ce complexe. Un autre élément en faveur d'un effet potentiellement non-haploinsuffisant de cette variation est que le modèle murin *knock-out* homozygote ne présente

pas de problème développemental mais un phénotype très différent d'intolérance au glucose[99]. Au total, ces éléments semblent indiquer que la maladie du développement associée à *FEM1B* puisse résulter d'un mécanisme non-haploinsuffisant de type gain de fonction.

3.4.3 Variations récurrentes et effet non-haploinsuffisant

Comme nous avons vu au chapitre 1.1.2.4, certains gènes peuvent être responsables de plusieurs maladies génétiques différentes, potentiellement par le biais de mécanismes distincts. En conséquence, dans le chapitre 2.4, il n'a pas été possible d'analyser de manière fine le mécanisme d'effet des variations récurrentes dans denovo-db, qui n'a été évalué que pour le petit groupe des variations très récurrentes. L'accès aux données cliniques précises des patients dijonnais, avec une maladie bien déterminée pour chaque patient (Tableau 8, Tableau 9), nous a permis d'évaluer le mécanisme d'effet associé à ces variations récurrentes d'après la base DDG2P. Ainsi, parmi les 21 patients de la cohorte dijonnaise avec une mutation récurrente pathogène dans une maladie mono allélique, 11 patients étaient porteurs d'une maladie annotée comme à effet NHI (*all missense/in frame*, *activating* et *dominant negative*) dans DDG2P (52,4 %, les autres variations étant celles à effet perte de fonction). En comparaison, la proportion globale de l'effet NHI dans DDG2P (maladies monoalléliques et liées à l'X) était plus faible (39,6 %; 360/910), mais cette différence n'était pas statistiquement pas significative du fait du nombre réduit d'effectifs (p=0,26 ; test de Fisher). En conclusion, la récurrence mutationnelle permet d'avoir accès aux variations à effet NHI, mais de manière non spécifique, permettant également d'identifier de nombreuses variations à effet HI.

3.4.4 Récurrence mutationnelle et denovo-db en routine diagnostique

Notre approche s'est basée sur la puissance qu'apporte l'information de la récurrence mutationnelle dans les maladies ultra-rares. L'originalité de ce travail était d'identifier des variations candidates d'une manière non biaisée, initialement indépendante de tout processus d'interprétation, ou de variations ou gènes candidats. L'identification d'une récurrence phénotypique via des stratégies classiques de partage de données, incluant GeneMatcher, a montré son efficacité pour la délinéation de nouvelles relations phénotype génotype.

Cette approche était simple à mettre en place car les données de denovo-db sont accessibles et faciles à manipuler. Par exemple, les données peuvent être facilement visualisées dans le navigateur du génome UCSC, qui s'est avéré être un outil utile en interprétation de NGS de routine, présentant une vision globale de la mutabilité des gènes, ainsi que le potentiel clustering des variations et le type de variations retrouvées à l'état *de novo* dans les présentations cliniques diverses incluses dans denovo-db.

En dehors d'une ré-analyse rétrospective occasionnelle, nous proposons que l'annotation prospective avec denovo-db puisse aider à mettre en évidence des variations d'intérêt d'une manière non biaisée, avec un chevauchement limité avec ClinVar et HGMD. Du fait de la croissance actuelle des cohortes de trio, on peut estimer que de plus en plus de récurrence

mutationnelle pourra être observée et de fait notre approche a le potentiel de gagner en efficacité dans le futur.

L'équipe dijonnaise a montré les bénéfices d'une ré-analyse annuelle des données d'exome chez les patients sans diagnostic, avec jusqu'à 15,4% de nouveaux diagnostics chez ces cas index[97], liés notamment à la rapidité des découvertes dans le champ des maladies du développement. Cependant, la ré-analyse complète des données de NGS est longue et chère, ce qui pose la question de la faisabilité à moyen et long terme. L'approche basée sur la récurrence mutationnelle apporte un nombre limité de variations à considérer, compatible avec une ré-analyse occasionnelle de grande ampleur.

3.4.5 Partage de données à grande échelle et récurrence mutationnelle

Nous faisons l'hypothèse que la récurrence mutationnelle sera un élément clé à considérer pour l'identification future de maladies ultra-rares, en particulier celles associées à des variations faux-sens spécifiques avec un mécanisme distinct de la perte de fonction. Aussi elle représente une approche non biaisée permettant d'identifier des nouvelles relations génotype-phénotype dans des gènes déjà impliqués en pathologie humaine, pour lesquels l'interprétation classique serait mise en échec du fait d'une incompatibilité phénotypique apparente. L'agrégation systématique des variations à la manière de denovo-db est une approche particulièrement pertinente. Différemment, la mise en commun de cas non résolus au sein de grandes cohortes pourrait mener à l'augmentation du signal de récurrence mutationnelle associée à la maladie. Dans cette optique, l'initiative européenne Solve-RD agrégera les données de plus de 18 000 patients avec une maladie monogénique présumée non identifiée, dans l'objectif d'augmenter notre connaissance des maladies ultra-rares.

4 DISCUSSION ET CONCLUSION

L'analyse de données de génomique déjà produites (*data-mining*), que ce soit à partir de données publiques uniquement (partie 2) ou en introduisant des données locales (partie 3) a permis l'identification de certaines variations et propriétés génomiques d'intérêt. Pour ce travail, nous avons étudié plus spécifiquement le rôle des variations faux-sens à effet non-haploinsuffisant, qui nous semblent représenter une source importante de nouvelles relations génotype-phénotype restant à identifier.

Les variations faux sens à effet NHI sont rares et complexes à identifier

Nous avons observé que le mécanisme NHI, ainsi que les variations faux-sens en général, étaient enrichis au sein des maladies ultra-rares (par rapport au mécanisme HI et aux variations tronquantes, respectivement). Nous interprétons cet effet par la diminution de la « cible mutationnelle » (*mutational target*[84]) dans les maladies par mécanisme NHI par rapport au mécanisme HI. En effet, il est facilement concevable que les maladies liées à des modifications très spécifiques de la protéine, que ce soit par la perturbation d'un domaine protéique, voire d'un résidu unique, puissent être moins fréquentes, car moins probables, que des maladies liées à des variations tronquantes ou faux-sens déstabilisant l'architecture de la protéine potentiellement réparties sur toute la longueur de la protéine. Dans ce travail, la recherche de récurrence mutationnelle a été réalisée de manière simple à l'échelle des nucléotides. De manière à identifier des possibles substitutions faux-sens différentes, nous avons élargi l'analyse à la recherche de variations distinctes du même nucléotide, ce qui a permis l'identification de variations d'intérêt, en particulier la variation récurrente au sein de *ACTL6B*, déjà identifiée par plusieurs études[67,77]. On peut faire l'hypothèse que cette stratégie puisse gagner en puissance en se basant sur une récurrence non pas à l'échelle du nucléotide, mais de l'acide aminé. En effet, des variations de certains acides aminés cibles peuvent être provoquées par des substitutions nucléotidiques de deux voire des trois nucléotides d'un même codon, avec de nombreux exemples en génétique somatique (par exemple dans la protéine KRAS, pour laquelle n'importe quelle substitution nucléotidique des deux premiers nucléotides du codon 12 est *driver* dans divers cancers), mais également dans des maladies du développement (par exemple dans le gène *SMAD4* déjà évoqué, pour lequel des variations des 3 nucléotides du codon 500 peuvent provoquer le syndrome de Myhre). Des variations plus complexes de type delins ou délétions / insertions en phase peuvent également être rencontrées. Par ailleurs, la répartition très précise des variations à effet NHI au sein des gènes les rend difficiles à détecter dans des études d'enrichissement. En effet, ces études sont le plus souvent réalisées soit à l'échelle du variant, avec un cruel manque de puissance pour les variations très rares, soit à l'échelle du gène, mises au point pour pallier la rareté des variants en intégrant tous les variants rares d'une certaine catégorie (par exemple, faux sens prédits délétères) d'un gène. Pour identifier ces variations, des stratégies plus fines permettant d'identifier un enrichissement non pas global, mais dans une portion de la protéine, semblent indispensables. L'analyse par domaines fonctionnels, la recherche de clusterisation au niveau de la séquence protéique ou de la modélisation tridimensionnelle, ou encore la recherche de

récurrence mutationnelle, sont autant d'approches pouvant permettre d'identifier ces variations, gènes et maladies génétiques. Ainsi, cette approche pourrait permettre d'étendre le spectre des gènes de malades monogéniques très rares restant à identifier mais également de détecter de nouveaux signaux d'association (facteurs de risque) dans les maladies rares mais aussi les maladies communes. De telles approches sont par exemple en cours de développement dans l'étude de la maladie d'Alzheimer jeune dans le laboratoire.

Identification d'un nouveau gène associé à une anomalie du développement par probable effet NHI

Dans ce contexte, nous avons cherché des variations *de novo* récurrentes entre plusieurs individus porteurs d'une anomalie du développement. L'analyse de ces variations récurrentes nous a permis de monter qu'elles étaient très fortement enrichies en variations pathogènes, ce qui nous a menés à appliquer cette stratégie sur une cohorte de plus de 1 200 patients avec maladie de développement analysés en exome par l'équipe dijonnaise. Outre l'identification de plusieurs variations *de novo* pathogènes dans des gènes récemment impliqués en pathologie humaine, la récurrence mutationnelle a mis en lumière trois variations dans des gènes candidats. En particulier, une même variation faux-sens *de novo* dans le gène *FEM1B* a été observée chez trois patientes porteuses d'un phénotype neurodéveloppemental sévère et assez spécifique sans cause connue, ce qui nous a incités à considérer *FEM1B* comme un probable nouveau gène associé aux maladies du développement. Plusieurs arguments nous ont orientés vers un potentiel effet non-haploinsuffisant associé à cette variation très spécifique. Cette hypothèse pourrait faire l'objet de tests fonctionnels, pouvant passer dans un premier temps par la quantification de la protéine mutante, puis par des tests d'affinité avec les différents substrats, et des mesures de l'ubiquitinylation des protéines cibles.

Variations tronquantes à effet NHI

En contraste avec les données présentées ici, d'autres types mutationnels que les variations faux-sens peuvent aboutir à un effet non-haploinsuffisant. En particulier, certaines variations « tronquantes » avec codon stop prématuré (PTC) peuvent en cas d'échappement au NMD (NMD-) mener à la production de transcrits stables, dont l'expression protéique peut donner lieu à une maladie par un effet non-haploinsuffisant. On peut citer l'exemple marquant du gène *ROR*, pour lequel les variations tronquantes NMD+ hétérozygotes ne produisent pas de phénotype (statut de porteur dans le syndrome de Robinow), alors que les variants tronquants NMD- sont à l'origine de la brachydactylie autosomique dominante de type B1[100]. Dans ce contexte, une équipe a récemment proposé un algorithme permettant de prédire si les variations tronquantes donnaient lieu à un mécanisme de NMD[101]. Dans ce travail, les auteurs ont appliqué cette méthode pour identifier à partir de larges jeux de données de génomique des gènes sensibles à l'échappement au NMD, fournissant une liste de gènes candidats dans lesquels les codons stop prématurés NMD- pourraient produire un effet pathogène passant par un mécanisme NHI. Ces résultats indiquent que ce type de variations particulières est probablement plus commun qu'anticipé et fournissent un modèle afin de les identifier.

Impact fonctionnel des variations et transition génomique

Il semble que les connaissances sur l'effet fonctionnel des variations génétiques soient, bien que très importantes dans la littérature, assez difficilement accessibles en 2018. Au sein de ce travail il est apparu que les types d'effets des variations sur la fonction de l'ARN, des protéines ou des voies cellulaires (par exemple : activation constitutionnelle du récepteur, altération de la localisation cellulaire de la protéine, ou encore saut d'exon), n'étaient pas classés et structurés de manière consensuelle. On peut émettre le besoin d'une organisation des types d'effets fonctionnels des variations au sein d'une ontologie similaire à la *gene ontology*, relative à la fonction des gènes[102]. Cette classification permettrait d'associer un effet fonctionnel (obtenu expérimentalement, à partir de la littérature) aux données de génomique, qui en sont actuellement largement exemptes. Ces annotations complémentaires aux annotations communément disponibles pourraient bénéficier au diagnostic des maladies génétiques, mais également à la recherche, autorisant des approches statistiques globales visant à identifier des patterns fonctionnels des variations identifiées au sein des cohortes. Ces annotations d'effet fonctionnel pourraient permettre grâce à des techniques de *machine learning* de préciser les connaissances relatives à chaque type d'effet, voire d'inférer un effet fonctionnel possible pour les nouvelles variations identifiées.

Conclusion

Depuis une dizaine d'années, la génétique a franchi un cap majeur en passant dans l'ère de la génomique. Les variations du génome entier sont maintenant d'accessibilité grandement facilitée et le défi du généticien du 21ème siècle n'est plus la détection des variations génétiques, mais leur interprétation. Parmi les variations du génome, les variations non codantes restent certes les plus complexes à caractériser. Néanmoins, nous rappelons ici qu'un grand nombre de variations codantes restent d'interprétation difficile, et en particulier les variations faux-sens.

Dans ce travail, nous avons étudié le mécanisme non-haploinsuffisant par diverses approches dans une logique génomique, ce qui a mené à l'observation que (i) les variations faux-sens pathogènes et le mécanisme NHI sont plus communs parmi les maladies génétiques ultra-rares, et (ii) la recherche de mutations *de novo* récurrentes entre plusieurs individus porteurs d'un phénotype similaire est une méthode permettant l'identification de variations pathogènes de mécanisme NHI. Cette méthode a mené dans un second temps à l'identification de nouvelles variations pathogènes au sein d'une cohorte locale de patients avec maladie du développement.

Nous proposons que l'utilisation intégrée de bases de données aide à l'interprétation des variations génomiques. Bien que des outils d'aide à la priorisation et de visualisation sont disponibles et d'une grande utilité, une annotation la plus exhaustive possible intégrant les données génomiques, fonctionnelles et phénotypiques est aujourd'hui difficile à appréhender par un généticien moléculaire du fait de la multiplicité des informations. L'intégration de l'ensemble des données et de l'expertise humaine ouvre la voie vers une aide encore plus importante des outils informatiques pour l'interprétation des données génomiques.

RESSOURCES WEB

ASTRID	http://astrid.icompbio.net/
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/
EBI Gene 2 Phenotype	https://www.ebi.ac.uk/gene2phenotype/downloads
Go Panther	http://www.pantherdb.org/
MuPIT	http://hg19.cravat.us/MuPIT_Interactive/
OMIM	https://www.omim.org/
Orphanet	https://www.orpha.net/consor/cgi-bin/Disease.php?lng=FR
OrrphaData	http://www.orphadata.org/cgi-bin/index.php/
Recommandations ACGS	http://www.acgs.uk.com/media/1059647/uk_practice_guidelines_for_variant_classification_2017.pdf
Recommandations ANPGM	http://www.anpgm.fr/images/BP-NGSDiag_001_Interpretation_Variants_V1.pdf
SwissVar	https://swissvar.expasy.org/
UCSC genome browser	http://genome-euro.ucsc.edu/cgi-bin/hgTracks
Variant Effect Predictor	http://grch37.ensembl.org/Homo_sapiens/Tools/VEP
VarSome	https://varsome.com/

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Maiella, S., Rath, A., Angin, C., Mousson, F. & Kremp, O. [Orphanet and its consortium: where to find expert-validated information on rare diseases]. *Rev. Neurol. (Paris)* **169 Suppl 1,** S3-8 (2013).

2. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43,** D789-798 (2015).

3. Köhler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42,** D966-974 (2014).

4. Köhler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* **45,** D865–D876 (2017).

5. Lin, A. E. *et al.* Gain-of-function mutations in SMAD4 cause a distinctive repertoire of cardiovascular phenotypes in patients with Myhre syndrome. *Am. J. Med. Genet. A.* **170,** 2617–2631 (2016).

6. Zhu, X., Need, A. C., Petrovski, S. & Goldstein, D. B. One gene, many neuropsychiatric disorders: lessons from Mendelian diseases. *Nat. Neurosci.* **17,** 773–781 (2014).

7. Davis-Turak, J. *et al.* Genomics pipelines and data integration: challenges and opportunities in the research setting. *Expert Rev. Mol. Diagn.* **17,** 225–237 (2017).

8. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma. Oxf. Engl.* **25,** 1754–1760 (2009).

9. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).

10. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).

11. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

12. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum. Mutat.* **37,** 235–241 (2016).

13. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220 (2013).

14. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

15. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

16. Nicolas, G. *et al.* SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. *Mol. Psychiatry* **21,** 831–836 (2016).

17. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46,** 818–825 (2014).

18. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549,** 519–522 (2017).

19. Yamaguchi-Kabata, Y. *et al.* iJGVD: an integrative Japanese genome variation database based on whole-genome sequencing. *Hum. Genome Var.* **2,** 15050 (2015).

20. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).

21. Ganna, A. *et al.* Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *Am. J. Hum. Genet.* **102,** 1204–1211 (2018).

22. Nicolas, G., Charbonnier, C., Campion, D. & Veltman, J. A. Estimation of minimal disease prevalence from population genomic data: Application to primary familial brain calcification. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **177,** 68–74 (2018).

23. Carlston, C. M. *et al.* Pathogenic ASXL1 somatic variants in reference databases complicate germline variant interpretation for Bohring-Opitz Syndrome. *Hum. Mutat.* **38,** 517–523 (2017).

24. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* **34,** 531–538 (2016).

25. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9,** e1003709 (2013).

26. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat. Genet.* **46,** 944–950 (2014).

27. Acuna-Hidalgo, R., Veltman, J. A. & Hoischen, A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.* **17,** 241 (2016).

28. Wilfert, A. B., Sulovari, A., Turner, T. N., Coe, B. P. & Eichler, E. E. Recurrent de novo mutations in neurodevelopmental disorders: properties and clinical implications. *Genome Med.* **9,** 101 (2017).

29. Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350,** 94–98 (2015).

30. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348,** 880–886 (2015).

31. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48,** 126–133 (2016).

32. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet. TIG* **29,** 575–584 (2013).

33. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–475 (2012).

34. Baert-Desurmont, S. *et al.* Optimization of the diagnosis of inherited colorectal cancer using NGS and capture of exonic and intronic sequences of panel genes. *Eur. J. Hum. Genet. EJHG* (2018). doi:10.1038/s41431-018-0207-2

35. Gieldon, L. *et al.* Diagnostic value of partial exome sequencing in developmental disorders. *PloS One* **13,** e0201041 (2018).

36. Brownstein, C. A. *et al.* An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome Biol.* **15,** R53 (2014).

37. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17,** 405–424 (2015).

38. Amendola, L. M. *et al.* Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am. J. Hum. Genet.* **99,** 247 (2016).

39. Kleinberger, J., Maloney, K. A., Pollin, T. I. & Jeng, L. J. B. An openly available online tool for implementing the ACMG/AMP standards and guidelines for the interpretation of sequence variants. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18,** 1165 (2016).

40. Li, Q. & Wang, K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am. J. Hum. Genet.* **100,** 267–280 (2017).

41. Patel, R. Y. *et al.* ClinGen Pathogenicity Calculator: a configurable system for assessing pathogenicity of genetic variants. *Genome Med.* **9,** 3 (2017).

42. Nykamp, K. *et al.* Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **19,** 1105–1117 (2017).

43. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46,** D1062–D1067 (2018).

44. Shah, N. *et al.* Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *Am. J. Hum. Genet.* **102,** 609–619 (2018).

45. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136,** 665–677 (2017).

46. Fokkema, I. F. A. C. *et al.* LOVD v.2.0: the next generation in gene variant databases. *Hum. Mutat.* **32,** 557–563 (2011).

47. Mottaz, A., David, F. P. A., Veuthey, A.-L. & Yip, Y. L. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinforma. Oxf. Engl.* **26,** 851–852 (2010).

48. Lott, M. T. *et al.* mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinforma.* **44,** 1.23.1-26 (2013).

49. Béroud, C. *et al.* UMD (Universal Mutation Database): 2005 update. *Hum. Mutat.* **26,** 184–191 (2005).

50. Turner, T. N. *et al.* denovo-db: a compendium of human de novo variants. *Nucleic Acids Res.* **45,** D804–D811 (2017).

51. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542,** 433–438 (2017).

52. Karolchik, D., Hinrichs, A. S. & Kent, W. J. The UCSC Genome Browser. *Curr. Protoc. Bioinforma.* **Chapter 1,** Unit1.4 (2012).

53. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet. TIG* **16,** 418–420 (2000).

54. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297,** 1003–1007 (2002).

55. Pruitt, K. D. *et al.* RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42,** D756-763 (2014).

56. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662-669 (2015).

57. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–585 (2013).

58. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40,** D71-75 (2012).

59. Ingram, V. M. Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature* **180,** 326–328 (1957).

60. Nirenberg, M. & Leder, P. RNA CODEWORDS AND PROTEIN SYNTHESIS. THE EFFECT OF TRINUCLEOTIDES UPON THE BINDING OF SRNA TO RIBOSOMES. *Science* **145,** 1399–1407 (1964).

61. Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192,** 1227–1232 (1961).

62. Stefl, S., Nishi, H., Petukh, M., Panchenko, A. R. & Alexov, E. Molecular mechanisms of disease-causing missense mutations. *J. Mol. Biol.* **425,** 3919–3936 (2013).

63. Muller, H. J. Further studies on the nature and causes of gene mutations. *Proceedings of the 6th International Congress of Genetics* (1932).

64. Katsonis, P. *et al.* Single nucleotide variations: biological impact and theoretical interpretation. *Protein Sci. Publ. Protein Soc.* **23,** 1650–1666 (2014).

65. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15,** 978–986 (2005).

66. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11,** 863–874 (2001).

67. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **50,** 1161–1170 (2018).

68. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538,** 201–206 (2016).

69. Yue, W. W., Froese, D. S. & Brennan, P. E. The role of protein structural analysis in the next generation sequencing era. *Top. Curr. Chem.* **336,** 67–98 (2014).

70. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185,** 862–864 (1974).

71. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.* **18,** 599–612 (2017).

72. Smedley, D. *et al.* Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10,** 2004–2015 (2015).

73. Koile, D., Cordoba, M., de Sousa Serro, M., Kauffman, M. A. & Yankilevich, P. GenIO: a phenotype-genotype analysis web server for clinical genomics of rare diseases. *BMC Bioinformatics* **19,** 25 (2018).

74. Smedley, D. & Robinson, P. N. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med.* **7,** 81 (2015).

75. Turner, T. N. *et al.* Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. *Hum. Mol. Genet.* **24,** 5995–6002 (2015).

76. Geisheker, M. R. *et al.* Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat. Neurosci.* **20,** 1043–1051 (2017).

77. Lelieveld, S. H. *et al.* Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. *Am. J. Hum. Genet.* **101,** 478–484 (2017).

78. Olson, H. E. *et al.* A Recurrent De Novo PACS2 Heterozygous Missense Variant Causes Neonatal-Onset Developmental Epileptic Encephalopathy, Facial Dysmorphism, and Cerebellar Dysgenesis. *Am. J. Hum. Genet.* **102,** 995–1007 (2018).

79. Yoo, Y. *et al.* GABBR2 mutations determine phenotype in rett syndrome and epileptic encephalopathy. *Ann. Neurol.* **82,** 466–478 (2017).

80. Sivley, R. M., Dou, X., Meiler, J., Bush, W. S. & Capra, J. A. Comprehensive Analysis of Constraint on the Spatial Distribution of Missense Variants in Human Protein Structures. *Am. J. Hum. Genet.* **102,** 415–426 (2018).

81. Zhang, Z., Miteva, M. A., Wang, L. & Alexov, E. Analyzing effects of naturally occurring missense mutations. *Comput. Math. Methods Med.* **2012,** 805827 (2012).

82. Lelieveld, S. H. *et al.* Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nat. Neurosci.* **19,** 1194–1196 (2016).

83. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

84. Veltman, J. A. & Brunner, H. G. De novo mutations in human genetic disease. *Nat. Rev. Genet.* **13,** 565–575 (2012).

85. Ségurel, L., Wyman, M. J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu. Rev. Genomics Hum. Genet.* **15,** 47–70 (2014).

86. Goriely, A. & Wilkie, A. O. M. Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease. *Am. J. Hum. Genet.* **90,** 175–200 (2012).

87. Michot, C. *et al.* Myhre and LAPS syndromes: clinical and molecular review of 32 patients. *Eur. J. Hum. Genet. EJHG* **22,** 1272–1277 (2014).

88. Schuurs-Hoeijmakers, J. H. M. *et al.* Clinical delineation of the PACS1-related syndrome--Report on 19 patients. *Am. J. Med. Genet. A.* **170,** 670–675 (2016).

89. Houge, G. *et al.* B56δ-related protein phosphatase 2A dysfunction identified in patients with intellectual disability. *J. Clin. Invest.* **125,** 3051–3062 (2015).

90. Maher, G. J. *et al.* Visualizing the origins of selfish de novo mutations in individual seminiferous tubules of human testes. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 2454–2459 (2016).

91. Vissers, L. E. L. M., Gilissen, C. & Veltman, J. A. Genetic studies in intellectual disability and related disorders. *Nat. Rev. Genet.* **17,** 9–18 (2016).

92. Nambot, S. *et al.* Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genet. Med. Off. J. Am. Coll. Med. Genet.* (2017). doi:10.1038/gim.2017.162

93. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17,** 405–424 (2015).

94. Sobreira, N., Schiettecatte, F., Valle, D. & Hamosh, A. GeneMatcher: a matching tool for connecting investigators with an interest in the same gene. *Hum. Mutat.* **36,** 928–930 (2015).

95. Assoum, M. *et al.* Autosomal-Recessive Mutations in AP3B2, Adaptor-Related Protein Complex 3 Beta 2 Subunit, Cause an Early-Onset Epileptic Encephalopathy with Optic Atrophy. *Am. J. Hum. Genet.* **99,** 1368–1376 (2016).

96. Hamilton, M. J. *et al.* Rubinstein-Taybi syndrome type 2: report of nine new cases that extend the phenotypic and genotypic spectrum. *Clin. Dysmorphol.* **25,** 135–145 (2016).

97. Nambot, S. *et al.* Clinical whole-exome sequencing for the diagnosis of rare disorders with congenital anomalies and/or intellectual disability: substantial interest of prospective annual reanalysis. *Genet. Med. Off. J. Am. Coll. Med. Genet.* (2017). doi:10.1038/gim.2017.162

98. Dankert, J. F., Pagan, J. K., Starostina, N. G., Kipreos, E. T. & Pagano, M. FEM1 proteins are ancient regulators of SLBP degradation. *Cell Cycle Georget. Tex* **16,** 556–564 (2017).

99. Lu, D. *et al.* Abnormal glucose homeostasis and pancreatic islet function in mice with inactivation of the Fem1b gene. *Mol. Cell. Biol.* **25,** 6570–6577 (2005).

100. Ben-Shachar, S. *et al.* Dominant versus recessive traits conveyed by allelic mutations - to what extent is nonsense-mediated decay involved? *Clin. Genet.* **75,** 394–400 (2009).

101. Coban-Akdemir, Z. *et al.* Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *Am. J. Hum. Genet.* **103,** 171–187 (2018).

102. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43,** D1049-1056 (2015).

RÉSUMÉ

Depuis la fin des années 2000, l'apparition du séquençage à haut débit, ou *next generation sequencing* (NGS) a métamorphosé le paysage de la génomique humaine. Le séquençage des régions codantes de tous les gènes en parallèle, l'exome, voire le séquençage du génome complet, sont maintenant possibles à l'échelle individuelle, permettant d'identifier pratiquement toutes les variations génétiques portées par un individu. Ces technologies ont révélé l'existence d'un polymorphisme insoupçonné, dont le décryptage dans le but d'en extraire des informations utiles pour le diagnostic génétique des maladies humaines représente toujours une préoccupation majeure des généticiens. Ainsi, la détection des variations génétiques n'est désormais plus limitante, mais leur interprétation, correspondant à l'établissement d'un lien de causalité entre une variation génétique et la maladie présentée par le patient, constitue un véritable défi.

Les variations de séquence les plus fréquemment associées à des maladies génétiques sont globalement dichotomisées en deux grands ensembles : (i) d'une part les variations tronquantes, qui altèrent la fonction du gène de manière quantitative, et (ii) les variations faux-sens, dont l'effet biologique est moins facilement prédictible et présente une grande diversité mécanistique. Certaines variations faux-sens particulières sur le plan fonctionnel, appelées variations à effet non-haploinsuffisant (NHI), engendrent la production de protéines qualitativement altérées avec une fonction anormale. Nous faisons l'hypothèse que les maladies causées par des variations à effet NHI sont complexes à mettre en évidence et qu'elles représentent une part importante des maladies génétiques qui restent à identifier en 2018.

Dans ce contexte, nous avons exploité des jeux de données disponibles publiquement dans l'objectif d'identifier certaines propriétés des variations à effet NHI dans une logique génomique globale. Les analyses réalisées ont permis de montrer une association entre l'effet NHI et les maladies ultra-rares (prévalence < 1/1 000 000). Nous interprétons ce résultat par le fait que les maladies associées au mécanisme NHI sont potentiellement liées à des variations faux-sens très spécifiques et donc d'apparition moins probable que des variations à effet haploinsuffisant pouvant être réparties dans toute la séquence codante du gène. Ensuite, dans l'objectif de pouvoir identifier de nouvelles variations pathogènes à effet NHI dans les maladies du développement, nous avons évalué les propriétés des variations *de novo* identifiées en parallèle chez plusieurs patients non apparentés porteurs de phénotypes similaires. En se basant sur la base denovo-db, nous avons pu identifier des facteurs expliquant en partie la récurrence d'une même mutation chez plusieurs individus, puis nous avons montré que les variations récurrentes étaient particulièrement enrichies en variations pathogènes.

Ces résultats ont été mis en application par la recherche de récurrence mutationnelle entre les données publiques de variations *de novo* issues de la base denovo-db et les variations identifiées chez plus de 1200 patients avec anomalie du développement analysés par séquençage d'exome dans le laboratoire de génétique du CHU de Dijon. En plus de l'identification de plusieurs nouvelles variations pathogènes dans des gènes connus, cette approche a permis l'association du gène *FEM1B* à un phénotype neurodéveloppemental par le biais d'un mécanisme potentiellement NHI.

Au total, ce travail contribue à la caractérisation des variations faux-sens à effet NHI et s'inscrit dans un effort général complexe d'identification des mécanismes mutationnels menant aux maladies génétiques humaines.