UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ – ED 414

Laboratoire ICube – UMR7357

THÈSE présentée par :

Thomas WEBER

Soutenue le : **15 décembre 2021**

Pour obtenir le grade de : *Docteur de l'Université de Strasbourg*

Discipline : *Sciences de la vie et de la santé*

Spécialité : *Bioinformatique et biologie des systèmes*

Nouvelles méthodes d'évaluation des variations génétiques *via* une approche bioinformatique : application aux maladies humaines.

THÈSE dirigée par :

M POCH Olivier — *Directeur de recherche, CNRS*

RAPPORTEURS EXTERNES :

Mme BONNE Gisèle — *Directrice de recherche, INSERM*

M RAUSELL Antonio — *Maître de conférences, Université de Paris*

EXAMINATRICE INTERNE :

Mme BLOCH-ZUPAN Agnès — *Professeure des Universités, Université de Strasbourg*

Remerciements

Je souhaiterai tout d'abord exprimer ma profonde reconnaissance aux Pr Agnès Bloch-Zupan, Dr Gisèle Bonne et au Dr Antonio Rausell pour l'honneur qu'ils me font d'évaluer cette thèse de doctorat.

Même si je ne l'ai pas revu souvent depuis, merci à Valérie Cognat pour m'avoir aiguillé vers le laboratoire lors de mon tout premier stage de Master en bioinfo dans son équipe.

Je voudrai ensuite remercier d'abord l'équipe dans son ensemble, pour cette bonne humeur quotidienne, cette « positive attitude » inimitable, cette attention au bien être de chacun et cette bienveillance générale qui font que l'équipe CSTB est bien plus qu'une équipe. Chaque journée de ces quatre années a été une expérience humaine et professionnelle en soi, parsemée de rigolades et d'apprentissage permanent. Merci de manière générale pour tout ce que vous m'avez transmis autant humainement que scientifiquement.

Merci à Laetitia pour ses conseils sur la transcriptomique (mais également pour toutes les dégustations champagne auxquelles nous avons eu droit). Merci à Anne pour son aide administrative et humaine au quotidien, sa bienveillance à l'égard des jeunes étourneaux perdus dans le barouf administratif franco-français dans lequel nous vivons. Merci à Julie, pour m'avoir accueilli et encadré durant mes 2 stages de M2, pour ses nombreuses corrections dans la langue de Shakespear et pour son flegme british au quotidien, contrebalançant l'explosivité latine d'une autre personne qui se reconnaitra. Merci à Odile pour m'avoir fait confiance durant ces deux années de monitorat. J'espère ne pas avoir contribué à former une génération catastrophe mélangeant orthologie et paralogie. Merci à Claudine pour son écoute au quotidien, pour ses valeurs humaines, pour sa joie de vivre et sa bonne humeur, et surtout, pour les nombreuses fois ou elle m'a libéré des griffes de mon directeur de thèse à des heures tardives, après des journées de réunion. Merci à Jean-Sébastien, Christian, Pierre P. et Pierre C., Rabih, Nathalie, Anne J, pour votre partage dans vos domaines respectifs. Un grand merci à Raymond pour tous ses conseils et les nombreux échanges durant ces quatre années.

Merci à tous les anciens du labo : Julio, Yannis, Audrey, Gopal, Camille à travers qui des concepts tels que la trompette, le cricket ou le stéthoscope ont pris une toute autre dimension. Merci aux récemment nouveaux venus (presque deux ans avec les stages déjà pour certains) Christelle, Hiba, Lalla, Soumaya, Amani, Anna, Corentin pour votre engouement quotidien. La relève est entre vos mains pour faire autant de bêtises que nous ! Merci à Romain, mon camarade de thèse, pour m'avoir supporté pendant 2 ans dans notre bureau, pour ces nombreuses soirées partagées ensemble et pour ses nombreux conseils. Attention aux Lutti ! Merci à Célia pour sa gentillesse, son écoute et sa générosité. On compte sur toi pour la reprise des jeun's !!! Merci à Arnaud d'avoir « pimenter » son assistance informatique, mais surtout pour ton humour au quotidien, et tes compétences au combien indispensables à l'équipe. Prochaine étape la thèse !! Merci à Sarah pour avoir partagé ces moments de doute et de remise en question, mais aussi pour toutes ces soirées arrosées, je te souhaite le meilleur à Boston avec Maxime, vous allez déchirer !!! Merci à Bastien pour les soirées partagées (encore une fois) et pour ces bons moments passés ensemble.

Enfin j'en arrive au lieu central de cette thèse, le bureau 132, ou malheureusement la célèbre citation d'Einstein a dû se vérifier à maintes reprises. Merci à Luc, d'abord pour toute son aide et son écoute scientifique au quotidien, mais aussi d'avoir été mon interlocuteur sportif dans le laboratoire. D'avoir partagé la Ligue 1, la Ligue des Champions, les JO, l'Euro, Roland Garros, la Formule 1, le Tour de France et j'en oublie. De m'avoir appris qu'une bouteille de champagne ne devait jamais être refroidie à proximité d'un mur en hiver de risque que cela crée un mouvement brownien altérant ses propriétés gustatives. Merci pour ta modération au quotidien et ton humour sans limites. Et comme tu adores cette citation et que tu la représentes parfaitement : « On peut rire de tout, mais pas avec n'importe qui ».

Merci à Nico, mon camarade de classe, devenu compagnon de stage puis finalement compagnon de thèse, "collègue" et ami. Merci pour ces cinq années parsemées de rigolade, de mauvaise surprise puis de mémorables soirées pour les oublier (encore d'autres). Merci pour tous ces moments à se chamailler tels Statler et Waldorf où tout le labo a pris conscience que tu réalisais tes figures en insérant des valeurs manuellement. Pour m'avoir supporté durant ces quatre années dans la même pièce durant de nombreuses heures. Pour tes blagues, parfois au goût discutable. Merci pour ton investissement dans le labo, qui n'aurait certainement pas été le même, sans ta gestion des chapatis, jeux de rôle et autres boutades en interne. Sache qu'après ces trois ans de dur labeur, je suis là pour t'épauler et pour t'aider dans la gestion de tes multiples problèmes de santé latents (kératocône, poumons engorgés, calvitie, bedaine, collagène mou).

Merci à Kirsley, qui au-delà de son statut d'encadrant est devenu un ami. Merci de m'avoir supporté déjà durant mes stages, ce qui n'était pas une mince à faire vu mon niveau initial en biologie et en informatique. Merci de m'avoir également supporté pendant la première moitié de la thèse durant le développement chaotique de MISTIC. Malgré les multiples embuches tendues par des *reviewers* pas toujours bienveillants et recommandables, on s'est en sorti ! Merci pour tout ce que m'as transmis, ton sens de l'analyse, de la réflexion et du rebond permanent. Ta capacité de travail sans limites et tes compétences en biologie et en bioinfo. Merci d'avoir pu partager avec toi l'ensemble des sorties High-Tech, les soirées foot et les soirées tout court. Je te souhaite le meilleur pour la suite, je le pense sincèrement et tu le mérites amplement !

Il reste une personne du laboratoire à remercier. Olivier, que dire … Merci pour tout. Pour ton encadrement au quotidien durant ces quatre années et pour ne jamais m'avoir lâché, même lorsque nous nous sommes lancés dans cette véritable excursion dans les zones inexplorées du génome à travers des données scabreuses et pas toujours compréhensibles. Merci pour ta bonne humeur quotidienne et ton énergie inépuisable, pour les « Bonjour Monsieur », « À l'attaque », « Alors c'est fini ? » qui rythment le quotidien de l'équipe. Merci pour tout que tu m'as transmis scientifiquement, pour ta connaissance de tous les domaines de la biologie, ton sens de l'analyse et de la construction d'un raisonnement. Merci pour ces innombrables heures de réunion à essayer de faire émerger un message des dizaines de figures avec lesquelles je vous ai assommé, et pour ta capacité à détecter un signal (que dis-je, un pixel d'écart entre deux violin plots) que même une IA développée par DeepMind en 2050 ne saurait repérer. Je ne sais pas si cette thèse a été une réussite, mais cette partie de ma vie l'a été, et c'est entre autres grâce à ton soutien et à tout ce que tu m'as transmis. Mille mercis.

Je souhaiterais aussi remercier mes proches, mes parents, ma famille, mes amis de Strasbourg (Lise & Max, Suzi, Hélo & Thomas, Nakri & Ludo, Antoine K, Antoine L, Maxime M, Lucas M) ainsi que ceux qui sont plus éloignés, seulement par la distance, mais par pas la pensée (Shannen & Martin, Guillaume, Louis, Marc). Merci à la famille de Cécile, Pierrette, Alexandre et Anatole (sans oublier Ninja) de m'avoir soutenu et épaulé, de me considérer comme vous le faites. Merci pour votre générosité et toute l'aide que vous m'avez apportée, sans jamais rien calculer. Merci d'être ma deuxième famille. Enfin, il y en a une sans qui tout cela n'aurait aucun sens si elle n'avait pas été là. Merci à Cécile, ma moitié, mon pilier au quotidien, celle qui a toujours été là dans les bons comme dans les mauvais moments. Celle qui m'apprend à ne jamais baisser les bras et à toujours voir le positif. Merci pour ton soutien, ton amour, ta générosité, ta dévotion, ton investissement, sans oublier tous les petits plats que tu m'as concocté et qui m'ont redonné des forces au quotidien, et particulièrement durant la rédaction. Le meilleur reste à venir à tes côtés. Merci infiniment, pour le passé, le présent et le futur.

Sommaire

Résumé de la thèse

Accès aux variations génétiques

Le génome humain, récemment célébré au travers des 20 ans de la première version de celui-ci (*Human Genome Project*), a été le fruit d'un important investissement en ressources technologiques, financières et humaines, qui a marqué une nouvelle ère pour la compréhension de ce qui constitue notre patrimoine commun le plus primordial. À la suite de ce projet, l'avènement des technologies de séquençage à haut débit a permis une génération rapide, fiable et à moindre coût des séquences génomiques, au cœur de l'explosion de la quantité de données en biologie ou « *Big Data* » biologiques. L'exploitation de ces « *Big Data »* biologiques est aujourd'hui possible grâce à la bioinformatique, tirant profit des avancées récentes en informatiques, de leurs évolutions logicielles et matérielles et de leur optimisation. Ainsi, l'utilisation en bioinformatique de l'évolution des processeurs (CPU), cartes graphiques (GPU) et mémoire vive (RAM) ainsi que de l'intelligence artificielle (IA) ou du calcul distribué a entrainé une amélioration de l'assemblage des génomes (*e.g.* SOAP3, BarraCUDA), de la détection des variations génétiques (DeepVariant) ou bien encore de la gestion de « *Big Data* » hétérogènes issues des "omiques".

Emblématique de ces avancées bioinformatiques, le consortium gnomAD (*genome Aggregation Database*) vise à regrouper et exploiter de nombreux projets de séquençage à large échelle d'exomes (un exome étant l'ensemble des exons de tous les gènes codant pour les protéines) ou de génomes humains (impliquant notamment tous les types de gènes et leurs éléments de régulation). À ce jour, gnomAD répertorie, plus de 125 000 exomes et 76 000 génomes humains révélant environ 760 millions variations génétiques sur les quelques 3,055 milliards de nucléotides que compte le génome humain. Ces variations génétiques, qui matérialisent la variabilité génomique de l'espèce humaine, peuvent être de petite taille (*single nucleotide variation* : SNV ; petites insertions ou délétions : indels) ou de grande taille (*structural variation :* SV). La cartographie établie par gnomAD constitue une réelle avancée permettant entre autres de connaitre la fréquence allélique (*Minor Allele Frequency : MAF*) d'une variation dans la population ou d'identifier des « régions contraintes », présentant une absence de variations.

Classiquement, une variation génétique est étudiée au regard de sa localisation dans un gène. Cependant, un gène peut aboutir à différents types de transcrits ou isoformes et dont l'expression peut varier en fonction des différents stades développementaux ou tissus du corps humain. Cette variabilité est aujourd'hui accessible grâce à la transcriptomique, permettant

l'accès à l'ensemble quantitatif et qualitatif des isoformes présents dans un tissu donné et à un temps donné. À l'image de gnomAD, l'initiative GTEx (Genotype-Tissue Expression project) vise à séquencer et à analyser de manière intégrée le génome et le transcriptome de près de 1000 individus dans 54 tissus différents du corps humain. De ce fait, une variation n'est plus étudiée/interprétée seulement au niveau génomique, mais également au travers de son impact transcriptionnel, de manière trans-tissulaire et temporelle. Dès lors, les limitations actuelles ne résident plus en la génération/accumulation des données, mais en leur intégration, exploitation et interprétation multi-niveaux, qui nécessitent le développement de nouvelles méthodologies, outils ou métriques bioinformatiques.

Exploitation dans le cadre des maladies génétiques

Les innovations récentes en séquençage ont entrainé un changement des pratiques concernant l'analyse des maladies génétiques. Ainsi, les méthodes d'analyse génétique en clinique ont évolué de l'étude dédiée à un gène (*Targeted Sequencing* : TS), à celle d'un génome (*Whole-Genome Sequencing* : WGS) puis à un exome (*Whole-Exome Sequencing* : WES) par l'utilisation de kits de capture, ainsi que plus récemment à celui d'un transcriptome (RNA-seq). L'ensemble de ces approches recouvre le séquençage clinique ayant pour but d'identifier le dysfonctionnement génétique dont souffre un patient afin de poser un diagnostic efficace le plus précocement possible. Cependant, plus le champ à étudier s'étend (TS ► WES ► WGS ► RNA-seq), plus la quantité de données croît, augmentant également le nombre et la complexité des variations génétiques à analyser.

Lors d'un séquençage clinique, on priorise l'analyse des variants selon leur impact (en exploitant les critères de recommandation de l'ACMG ; *American College of Medical Genetics*). En effet, plus l'effet du variant recherché est délétère, plus le nombre de variants à analyser est réduit. Parmi les différentes classes de variations, la catégorie des variations faux-sens est une des plus délicates à étudier. Les variations faux-sens correspondent à la modification d'une paire de bases au niveau génomique qui entraine un changement d'acide aminé lors de l'étape de traduction et elles sont au nombre d'environ 12 000 par génome. Cependant, au regard du nombre des variations faux-sens et sur la seule base de leurs propriétés biologiques et physico-chimiques, il est très difficile d'identifier rapidement et avec certitude la variation responsable d'une maladie. C'est pourquoi aujourd'hui la proportion de variations faux-sens, dont la conséquence est inconnue (v*ariants of unknown significance* ; *VUS*), est logiquement la plus importante pour les variations faux-sens dans la base de données ClinVar (70% de faux-sens *VUS*).

MISTIC : prédicteur robuste de l'impact des variations faux-sens dans le génome

Dans ce contexte, j'ai pu développer MISTIC (*MIS*sense dele*T*eriousness pred*IC*tor), un nouveau prédicteur de variations faux-sens délétères, surclassant les outils existants en apportant différentes améliorations méthodologiques. MISTIC repose sur des techniques d'apprentissage automatique, représentant une des multiples avancées informatiques récentes. L'idée est (A) d'établir un modèle d'intelligence artificielle, à partir de (B) données étiquetées de très haute qualité (ici des variations connues comme étant délétères ou polymorphiques dans de nombreux ethnies) et de (C) paramètres pertinents les décrivant, afin de prédire de façon fiable le statut d'une nouvelle variation encore inconnue jusqu'alors. MISTIC apporte une amélioration majeure pour chacun des points précédemment cités, nécessaires au développement d'un modèle d'apprentissage automatique. Concernant les modèles d'intelligence artificielle (A), après de multiples étapes de tests intégrant douze méthodes différentes, un système de vote souple (combinaison pondérée des probabilités) intégrant deux algorithmes de familles différentes (forêt d'arbres décisionnels et régression logistique) a été retenu comme le plus performant dans différents scénarios d'analyses (variation avec / ou sans *MAF*, analyse d'exome…). Les descripteurs des variations génétiques (B) ont fait l'objet d'une sélection basée sur la pertinence, où 113 des 714 paramètres initiaux ont été conservés au vu de leur complémentarité et du poids important dans le calcul du score pour au moins un des deux algorithmes utilisés. Parmi ceux-ci, on peut noter : l'utilisation de *MAF* obtenue à partir de la dernière version de la banque gnomAD, permettant d'évaluer le degré de polymorphisme d'une variation, l'utilisation de mesures fonctionnelles telles que les 'régions contraintes' ou les propriétés physico-chimiques et biochimiques provenant de la base de données AAindex. Enfin, les variations utilisées lors de l'étape d'entrainement (C) ont été sélectionnées après de multiples étapes de filtrage afin de vérifier à la fois : la véracité du statut de celles-ci (délétères avec des critères de validation exigeants, bénins présentant une bonne couverture de séquençage), l'absence de recouvrement entre les jeux de variants délétères et bénins, l'absence de variations déjà utilisées par des outils intégrés dans MISTIC ou dont certains des paramètres seraient manquants. MISTIC a été testé sur différents jeux de données originaux, permettant d'estimer à la fois sa capacité à évaluer de nouveaux variants délétères jamais utilisés auparavant par les outils existants ou provenant de sources diverses (base de données de cancer DoCM). Afin d'évaluer les performances lors de la prédiction de variations avec et sans *MAF*, MISTIC a également été évalué en présence de variations dans gnomAD (avec *MAF*) à différents

niveaux de fréquences, et en présence de variations dites « spécifiques », encore absentes des bases de données actuelles. Dans l'ensemble des scénarios, MISTIC surclasse les outils récents (ClinPred, M-CAP, REVEL, CADD …) et permet d'identifier efficacement les variations délétères dans de réels exomes issues de cohortes de myopathes.

Cependant, celui-ci n'intègre pas pour le moment la gestion des différentes isoformes d'un gène au regard de leurs expressions tissulaires respectives.

Des variations génétiques à la prise en compte de leur expression

Dans ce contexte, je me suis intéressé aux variations présentes dans des gènes multi-isoformes ayant des expressions tissulaires singulières. Grâce aux avancées technologiques précédemment citées, de nombreux transcrits ont pu être identifiés, validés et quantifiés dans un certain nombre de tissus. Ainsi, l'expression ouvre la voie à une nouvelle manière de regarder un gène afin de savoir si une variation est présente ou non dans une isoforme s'exprimant dans un ou plusieurs tissus. Malgré la quantité importante de données disponibles concernant les gènes, transcrits et exons associés au génome humain, un faible nombre d'articles s'est intéressé à établir un bilan statistique de leurs propriétés intrinsèques (*e.g.* taille, longueur, distribution des exons/introns…) ou extrinsèques (*e.g.* fonction, nombre de paralogues, conservation, maladies génétiques…) au regard du statut d'un gène : SISOG (pour *Single transcript ISOform Gene* ou gène n'exprimant qu'une isoforme, quel que soit le tissu) ou MISOG (pour *Multiple transcript ISOforms Gene*). Cette analyse est essentielle afin de comprendre s'il existe des biais statistiques pouvant être expliqués biologiquement. Ainsi, j'ai dans un premier temps comparé les propriétés intrinsèques et extrinsèques des gènes SISOG *versus* MISOG. Les résultats de cette analyse ont notamment révélé des différences importantes en termes de taille, de nombre d'exons entre SISOG et MISOG ainsi qu'une relation entre la position ordinale des éléments géniques (exons et introns) et leur longueur. Par ailleurs, les gènes SISOG sont fortement enrichis en certaines familles fonctionnelles (entre autres les GPCRs) et environ deux tiers des gènes impliqués dans les maladies génétiques appartiennent au type MISOG.

À la suite de ce bilan, je me suis focalisé sur les variations présentes dans les exons dits « alternatifs », c'est-à-dire, des exons présents uniquement dans certaines isoformes d'un gène avec un profil singulier d'expression tissulaire. Pour cela, je me suis appuyé sur des travaux récemment publiés dans le cadre du projet GTEx, ayant pour but de normaliser l'expression d'un nucléotide au regard de son occurrence dans les différentes isoformes d'un gène (*proportion expressed across transcripts* ; pext). Afin d'évaluer si une variation est

différentiellement exprimée dans un tissu, j'ai développé une méthode orthogonale appelée duxt (*differential exon usage across tissues*), ayant pour objectif de normaliser l'expression d'un nucléotide au regard de son apparition dans les différents tissus afin d'identifier les nucléotides/exons spécifiques ou absents d'un tissu. En utilisant duxt, j'ai pu identifier, pour l'ensemble des gènes humains, ceux présentant des exons sur- ou sous-utilisés dans un nombre restreint de tissus. Cela a notamment permis de souligner que certains tissus ont de nombreux exons sur-utilisés quasi-spécifiques (*e.g.* le muscle squelettique) et qu'à l'inverse, certains tissus sont enrichis en exons quasi-absents (*e.g.* organes génitaux). Ces observations coïncident avec des résultats précédents de la littérature et obtenus dans le cadre d'autres approches à haut débit (entre autres la protéomique). Afin de corroborer cette nouvelle métrique, différentes applications ont été explorées. L'une d'entre elles est l'utilisation des sQTLs (*splicing Quantitative Trait Loci*), correspondant à des SNV statistiquement associées à une bascule de l'expression d'une sQTL isoforme vers une autre. En effet, en comparant les scores duxt et les sQTLs, nous avons constaté une forte corrélation entre exons à score duxt élevé (différentiellement sur- ou sous-utilisés) et à fort impact (plus l'impact est important, plus le changement d'isoformes est marqué).

Une seconde application a concerné l'identification de gènes associés à plusieurs maladies génétiques et présentant des disparités phénotypiques. En utilisant duxt, j'ai isolé des exons sur- ou sous-exprimés présentant des particularités intéressantes : une surexpression dans un tissu ou un groupe restreint de tissus et la présence sur l'exon d'un variant à caractère délétère. L'analyse détaillée des gènes et pathologies associées à ces exons a permis de révéler des différences phénotypiques marquées selon la présence/absence tissulaire de l'exon portant la variation délétère. Cette identification ouvre de nombreuses perspectives en termes d'analyse clinique, notamment la prise en compte de l'expression tissulaire d'une variation délétère en fonction de sa localisation exonique et pourrait aider à la compréhension des particularités phénotypiques observées chez certains patients atteints d'une même pathologie.

Conclusions et perspectives

Dans le cadre de ma thèse, j'ai développé MISTIC, qui se place comme l'un des outils les plus performants en matière de prédiction de variations faux-sens délétères. Dans le cadre de l'analyse d'un exome, celui-ci permet à la fois de réduire le nombre de variations à conséquence inconnue (*VUS*), tout en améliorant considérablement la priorisation de la variation délétère responsable. De plus, MISTIC a été invité à participer à la 6ème édition de la compétition *Critical Assessment for Genome Interpretation* (CAGI), visant à évaluer la

pertinence de ses scores contre d'autres prédicteurs l'ensemble des 74 278 013 variants faux-sens théoriques du génome afin d'intégrer d'être intégré à terme dans la ressource dbNSFP.

Dans le but de perfectionner l'analyse des variations présentes sur des gènes à plusieurs isoformes, j'ai ensuite pu travailler au développement de duxt, une méthode d'identification des exons présentant une sur- ou sous-utilisation tissulaire. Ainsi, nous avons identifié des correspondances entre des gènes présentant des particularités phénotypiques intéressantes et des variations délétères présentes sur ces exons quasi-spécifiques ou quasi-absents de certains tissus. Les métriques développées durant ma thèse ont été pensées afin de tirer profit des mises à jour des ressources telles que gnomAD et GTEx. Ainsi, dans une logique évolutive, la prochaine version de MISTIC intégrera la nouvelle génération de MAFs composée d'une MAF globale calculée sur davantage d'individus ainsi que des MAFs associées à de nouveaux groupes ethniques. Quant à duxt, la métrique permettra l'identification de sur- ou de sous-utilisation à l'intérieur des différents types cellulaires d'un tissu grâce à l'exploitation du séquençage de noyaux uniques (single-nucleus RNA-seq).

L'exploitation des maladies génétiques en biologie permet de mettre en lumière les différents mécanismes moléculaires sous-jacents à un gène et son environnement. Bien que les méthodes d'étude clinique de ces dysfonctionnements aient évolué et reposent aujourd'hui sur le séquençage, une grande majorité des cas reste en errance de diagnostic. Cependant, dans le cadre des maladies génétiques, les méthodes d'analyse clinique actuelles résident en l'identification d'une variation unique comme responsable d'une pathologie. Une des perspectives d'évolutions est de basculer de cette relation de causalité directe à une analyse intégrative prenant en compte l'ensemble des variations d'un individu, chacune pouvant contribuer à son échelle à l'apparition d'un dysfonctionnement phénotypique visible. Demain, ces méthodes exploitant le paysage génétique des individus pourront aider à comprendre des susceptibilités ou des traits présentés par des individus non atteints de pathologie. Ces méthodes exploitant la combinatoire des variations tireront profit des différentes strates qui composent le « Big Data » biologique (le génome, l'expression temporelle et cellulaire de celui-ci, sa régulation …), mêlant ainsi intégration horizontale (les variations sur le génome) et verticale (les différents « omiques »).

Liste des figures

Liste des tableaux

Liste des équations

Abréviations

ACGS	*Association for Clinical Genomic Science*
ACMG	*American College of Medical Genetics*
ADN	Acide DésoxyriboNucléique
ADNc	ADN complémentaire
AF	*Allele Frequency*
AMP	*Association for Molecular Pathology*
API	*Application Programming Interface*
BSGM	*British Society for Genetic Medicine*
CAGI	*Critical Assessment for Genome Interpretation*
CCR	*Constrained-Coding Regions*
CDS	*Coding DNA Sequence*
CSV	*Comma-Separated Values*
DGV	*Database of Genomic Variants*
DNN	*Deep Neural Network*
EBI	*European Bioinformatics Institute*
EMBL	*European Molecular Biology Laboratory*
ENCODE	*Encyclopedia of DNA Elements*
ESP	*Exome Sequencing Project*
ExAC	*Exome Aggregation Consortium*
FTP	*File Transfert Protocol*
GFF	*General Feature Format*
gnomAD	*genome Aggregation Database*
GO	*Gene Ontology*
GPCR	*G Protein-Coupled Receptor*
GPU	*Graphics Processing Unit*
GRC	*Genome Reference Consortium*
GWAS	*Genome Wide Association Studies*
HGMD	*Human Gene Mutation Database*
HGNC	*HUGO Gene Nomenclature Committee*
HPCG	*Human Protein-Coding Genes*
HPO	*Human Phenotype Ontology*
IA	Intelligence Artificielle
LINE	*Long Interspersed Nuclear Elements*

LRS	*Long Read Sequencing*
MAF	*Minor Allele Frequency*
MISOG	*Multiple transcript ISOform Genes*
ML	*Machine Learning*
NCBI	*National Center for Biotechnology Information*
NGS	*Next Generation Sequencing*
OMIM	*Online Mendelian Inheritance in Man*
ONT	*Oxford Nanopore Technologies*
PCA	*Principal Component Analysis*
PCR	*Polymerase Chain Reaction*
QTL	*Quantitative Trait Loci*
ROC	*Receiver Operating Characteristic*
SISOG	*Single transcript ISOform Genes*
SMRT	*Single-Molecule Real Time*
SNP	*Single Nucleotide Polymorphism*
SNV	*Single Nucleotide Variation*
SO	*Sequence Ontology*
SQL	*Structured Query Language*
SV	*Structural Variation*
TER	*Translated Exonic Region*
TPM	*Transcripts Per Million*
TS	*Targeted Sequencing*
TSS	*Transcription Start Site*
TSV	*Tab-Separated Value*
TTS	*Transcription Termination Site*
UCSC	*University of California Santa Cruz*
UTR	*UnTranslated Region*
VCF	*Variant Call Format*
VUS	*Variant of Unknown Significance*
WES	*Whole Exome Sequencing*
WGS	*Whole Genome Sequencing*

Avant-propos

Dans le contexte du paysage des prédicteurs d'impact des variations génétiques en constante évolution, l'objectif général de ma thèse a été centré sur le développement de nouvelles méthodes permettant de mieux caractériser les variations causales des maladies génétiques rares. Dans cette optique, j'ai développé **MISTIC**, un nouvel outil de prédiction de l'impact des faux-sens, surclassant les outils disponibles, tout en réduisant le taux de faux-positifs. Dans le cadre d'une recherche exploratoire mettant à profit des données d'expression robustes disponibles pour 54 tissus humains, j'ai élaboré **duxt**, une méthode d'identification des exons alternatifs présentant une sur- ou sous-utilisation tissulaire. À terme, duxt a pour vocation de permettre une meilleure évaluation de l'impact des variations présentes dans les exons alternatifs des gènes à plusieurs isoformes de transcrits.

L'introduction de ce manuscrit comporte cinq chapitres (**Chapitre 1-5**) dont le but est de présenter à la fois les problématiques ayant donné lieu aux travaux de cette thèse, mais également les concepts sur lesquels ils s'appuient.

Le **Chapitre 1** a pour objectif de rappeler brièvement l'évolution et les fondements des méthodes de séquençage qui ont joué un rôle clé dans l'accès au génome humain et dans la détection des variations génétiques.

Le **Chapitre 2** est dédié aux variations génétiques et présente les différentes catégories de variations selon leur taille, leur localisation ou leurs mécanismes d'apparition. Enfin, quelques-unes des nombreuses applications dont elles font l'objet sont présentées.

Le **Chapitre 3** pose le contexte de l'étude des variations dans le cadre des maladies génétiques rares. Ce chapitre comprend un bref rappel du protocole d'exploitation des données de séquençage aboutissant à la détection des variations, suivi d'une section sur l'annotation et les différents critères utilisés afin d'évaluer une variation génétique dans le contexte clinique.

Le **Chapitre 4** se focalise sur les outils de prédiction de l'impact des variations, tant sur les différentes approches employées, que sur les catégories de descripteurs utilisés ou encore les données d'entrainement nécessaires au développement des prédicteurs.

Enfin, le **Chapitre 5** d'introduction se concentre sur l'exploitation des variations au regard des données massives d'expression génique (transcriptome). Les différentes ressources disponibles ainsi qu'un exemple d'utilisation de transcriptome dans le cadre clinique sont décrits.

La partie **Matériel et Méthodes** comporte deux chapitres (**Chapitre 6-7**) qui présentent dans le **Chapitre 6**, les ressources bio-informatiques (banques de référence, ressources

biomédicales, programme d'annotation, …), et dans le **Chapitre 7**, les outils statistiques et de programmation utilisés durant la thèse.

La partie **Contributions** se décline en deux chapitres majeurs (**Chapitre 8-9**).

Le **Chapitre 8** présente **MISTIC**, un nouvel outil de prédiction de l'impact des variations faux-sens, surclassant les outils actuels, et répondant aux problématiques du domaine par une augmentation de la spécificité et une diminution du nombre de variations à statut inconnu. Ce chapitre inclut la publication de MISTIC, des résultats complémentaires récents ainsi qu'une partie Conclusion et perspectives.

Le **Chapitre 9** est dédié aux résultats obtenus dans le cadre d'une recherche exploratoire visant à mieux caractériser les variations codantes présents dans des exons alternatifs, *i.e.* des exons présents dans un sous-ensemble des transcrits d'un gène. Dans ce cadre, nous avons réalisé une analyse comparant l'architecture des gènes à isoforme unique et aux isoformes multiple dont la publication dans le journal BMC Genome Biology, en cours d'évaluation, est fournie dans le chapitre. Puis, nous décrivons **duxt**, notre métrique pour identifier les différentiels d'utilisation tissulaire chez les exons alternatifs, à partir de données expérimentales d'expression. Ce chapitre inclus les méthodes utilisées pour élaborer duxt ainsi que des exemples d'utilisation dans le contexte biomédical.

Enfin, un court **Chapitre 10** de **Discussion ouverte et perspectives** considérera les futurs possibles des analyses et exploitations des variations génétiques en biomédecine génomique, notamment au travers de l'intégration graduelle des données de génomiques fonctionnelles et du rôle croissant des méthodes à bases d'intelligence artificielle.

INTRODUCTION

“Think of the human genome as the Book of Life. We are about to read the first chapter, as important an accomplishment as discovering the Earth goes round the sun or that we are descended from apes.”

John Sulston, 2001

2001

Cette date symbolique consacre l’entrée dans un nouveau siècle et dans un nouveau millénaire. Scientifiquement, 2001 coïncide avec la première mesure de l’atmosphère d’une exoplanète par le télescope Hubble, la première opération de téléchirurgie mondiale réalisée depuis New-York par le professeur Jacques Marescaux sur une patiente située à Strasbourg ou encore, la greffe du premier cœur artificiel autonome dans le Kentucky. C’est également le titre d’un film précurseur de Stanley Kubrick réalisé en 1968 et traitant de l’évolution humaine et de l’intelligence artificielle. Cependant, je retiendrai cette date comme l’année de l’accès à la séquence du génome humain, aboutissement du *Human Genome Project [HGP ;* (Lander et al. 2001)*]*. Bien que cette découverte ait sans doute eu un faible retentissement auprès du grand public, elle constitue un jalon décisif qui marquera à jamais notre histoire. On parle aujourd’hui d’ère post-génomique tant la séquence du génome humain a fait évoluer la science d’une vision centrée sur le gène vers une vision embrassant le génome dans sa totalité et sa complexité. Récemment célébré au travers de son 20ème anniversaire, le génome humain marque l’entrée dans un nouvel âge de la compréhension de notre patrimoine commun, le plus primordial, celui que John Sulston, ancien directeur de l’institut Sanger, appelait le « Livre de la Vie ».

Les étapes qui ont abouti à la découverte du génome et de ses constituants ont été nombreuses. Le concept de « génétique » a été introduit par Imre Festetics en 1819 (Poczai, Bell, et Hyvönen 2014), qui décrit pour la première fois quatre règles liées à l’hérédité, dont le principe de mutation. Ses travaux sont en partie poursuivis par Gregor Mendel qui définit trois règles liées à la génétique (1865) dont la deuxième est une démonstration mathématique de celle de Festetics (Mendel 1865). Puis en 1871, Friedrich Miescher identifie la présence de « nucléine », correspondant à l’ADN dans le noyau de globules blancs (Miescher 1871; Dahm 2005). Cinq nucléotides : adénine (A), cytosine (C), guanine (G), thymine (T) et uracile (U) sont identifiés au début du siècle dernier, en 1910, par Albrecht Kossel (prix Nobel 1910). Cependant, tandis que la communauté scientifique suspectait plutôt les protéines, c’est en

1952 que les expériences d'Alfred Hershey (prix Nobel 1969) et Martha Chase démontrent que l'ADN est le réel porteur de l'information génétique (Hershey et Chase 1952). Dans la foulée, en 1953, a lieu la découverte de la structure de la double hélice d'ADN (Watson et Crick 1953) par James Watson et Francis Crick (prix Nobel 1962), basée sur les travaux de Rosalind Franklin. Le langage de la vie devient « code de la vie » avec les travaux de Marshall Nirenberg identifiant la lecture en triplé (codon) en 1961 (Nirenberg et Matthaei 1961), puis viennent les avancées sur le fonctionnement du code génétique (Marshall 2014) et le séquençage du premier ARN de transfert en 1965 (Holley et al. 1965). Enfin, deux méthodes de séquençage direct de l'ADN voient le jour en 1977 (Sanger, Nicklen, et Coulson 1977; Maxam et Gilbert 1977), avancée majeure qui ouvrit la voie à l'obtention des séquences complètes de génomes d'êtres vivants.

L'ensemble de ces avancées fit prendre conscience au monde que l'information génétique est véhiculée par une chaîne chimique composée de quatre unités fondamentales, s'appariant sous la forme d'une double hélice. De cette prise de conscience vont découler d'innombrables découvertes scientifiques dans tous les domaines des sciences du vivant et notamment, en ce qui concerne notre espèce, dans la compréhension de ses maladies, de son évolution et de son hétérogénéité interindividuelle, reflet des variations génétiques accumulées au fil de notre histoire.

Chapitre 1. Les méthodes de séquençage et le génome humain

L'accès à la séquence des macromolécules biologiques (acides nucléiques et protéines) demeure un réel défi technologique, notamment du fait de leurs tailles, leurs nombres et leur extrême diversité.

En ce qui concerne les acides nucléiques, la détermination précise de leurs séquences n'a été rendue possible qu'à la fin des années 1970. Depuis, pas moins de trois générations biotechnologiques se sont succédé avant d'être en mesure d'aborder les multiples processus biologiques où la connaissance de la séquence constitue un enseignement décisif. Par-delà les prouesses technologiques, la stratégie générale demeure de recourir à un séquençage par fragment. Ainsi, dans le cas d'un génome, celui-ci se retrouve fractionné en millions de segments qui seront décryptés individuellement avant de reconstituer une séquence contiguë que l'on voudrait, dans l'idéal, identique au nucléotide près à celle d'origine. Cette exigence constitue encore un défi majeur pour les nombreux dispositifs technologiques, informatiques et bioinformatiques mis en œuvre pour atteindre cet objectif ambitieux.

Dans ce chapitre, je présenterai rapidement les trois générations de méthodes de séquençage et quelques-unes des applications exploitant le séquençage. Concernant le génome humain, au cœur de mes travaux, il est rapidement apparu que la séquence obtenue dans le cadre du *Human Genome Project* (Lander et al. 2001), ne constituait qu'une première étape et des travaux supplémentaires ont été nécessaires pour aboutir à des séquences de référence de très haute qualité à même de nous renseigner sur la totalité de chaque chromosome.

1.1 Des débuts révolutionnaires : le séquençage de 1ère génération

C'est la même année, en 1977, que les deux premières méthodes de séquençage à même de fournir des séquences à l'échelle d'un génome ont été publiées.

Toutes deux exploitent l'électrophorèse sur gels de polyacrylamide en plaque afin de séparer, au nucléotide près, des fragments d'ADN selon leur taille. Cependant, ces méthodes se différencient profondément par les techniques d'obtention et de marquage radioactif des fragments de séquençage. La méthode de MG (Maxam et Gilbert 1977) réalise une seule

cassure (statistique) par molécule d'ADN *via* un traitement chimique (méthode par dégradation de chaîne), puis un marquage radioactif de l'extrémité des fragments obtenus. La méthode de Sanger et Coulson (SC) obtient des fragments d'ADN radioactifs par une technique dite de terminaison de chaîne (Sanger, Nicklen, et Coulson 1977). Dans cette méthode, on réalise une copie de l'ADN à séquencer par une ADN Polymérase I modifiée (sans activité exonucléase 5' à 3') à l'aide d'une amorce complémentaire. Les nucléotides incorporés sont des désoxyribonucléotides standards (dNTPs) ainsi qu'une petite concentration de didésoxyribonucléotides dits « poisons » (ddNTPs, marqués radioactivement) qui entrainent un arrêt de la polymérisation et l'obtention de fragments radioactifs de longueurs différentes. Pour les deux méthodes, les fragments radioactifs étaient séparés sur des gels de polyacrylamide, déposés par la suite sur film radiographique exposé aux rayons X, permettant ainsi de lire la séquence et d'inférer l'ordre des bases.

L'amélioration de la technique de terminaison de chaîne de SC ainsi que la complexité pour automatiser la méthode de MG ont entrainé la disparition progressive de cette dernière. La méthode de SC a ensuite été automatisée (Figure 1) par les innovations de Smith, Hood et Applied Biosystems en 1986 (Smith et al. 1986). Cette automatisation a été possible entre autres, par l'utilisation de la fluorescence à la place de la radioactivité, qui fit passer le séquençage de quelques dizaines à plusieurs centaines de bases.

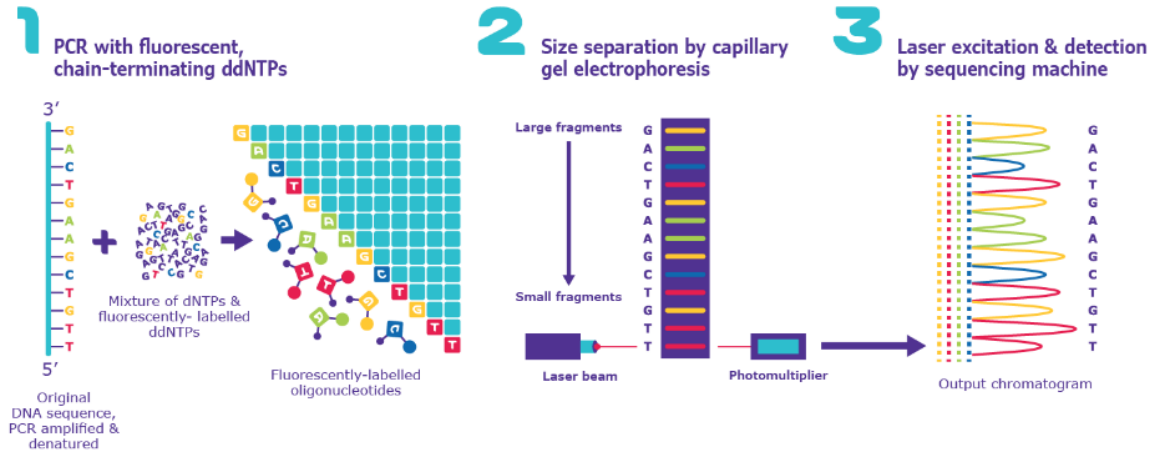


Figure 1 – Principe du séquençage Sanger après automatisation

Le séquençage Sanger automatisé peut se décomposer en trois étapes majeures. Durant la première étape (1), une séquence ADN d'intérêt est utilisée comme matrice pour un type spécial de PCR appelé PCR à terminaison de chaîne, étant donné l'ajout de désoxyribonucléotides (dNTP) modifiés appelés didésoxyribonucléotides (ddNTP). Ainsi, à chaque ajout d'un ddNTP marqué par fluorescence, la réaction se termine. Dans la deuxième étape (2), les oligonucléotides sont séparés selon leur taille par électrophorèse sur gel capillaire. La dernière étape (3) consiste à lire le gel pour déterminer la séquence de l'ADN d'entrée. Un ordinateur lit chaque bande du gel capillaire, en utilisant la fluorescence pour détecter chaque ddNTP en fin d'oligonucléotide. Chaque ddNTP étant marqué par une fluorescence différente, la lumière émise peut être directement liée à l'identité du ddNTP terminal. L'ensemble est visualisable sur un chromatogramme, qui montre le pic de fluorescence de chaque nucléotide sur toute la longueur de l'ADN séquencé.

Source : sigmaaldrich.com, consulté le 15/09/21

1.2 La force du nombre : le séquençage de 2nde génération

Les méthodes de séquençage dites de nouvelle génération (*Next-Generation Sequencing*; NGS) émergent à la fin des années 1990-début 2000 et sont toutes basées sur la méthode SC. Ces méthodes à haut débit permettent la génération de plusieurs millions, voir dizaines de millions de lectures par cycle d'exécution. Elles incluent le pyroséquençage (454), le séquençage par ligation (SOLiD), le séquençage basé sur la détection d'ions hydrogènes (Ion Torrent) et le séquençage par synthèse (Solexa, puis Illumina). Cette dernière méthode (Figure 2), qui devient rapidement prédominante pour sa rapidité, sa fiabilité et son coût (M. Meyer et Kircher 2010), repose sur trois étapes clés: fragmentation de l'ADN, amplification, séquençage et analyse des fragments de petite taille (≈ 150 pb).

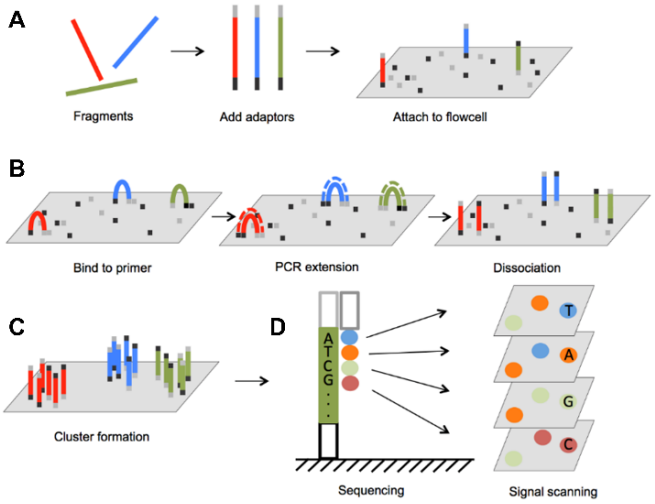


Figure 2 – Principe du séquençage Illumina

A – Fragmentation, ajout d'adaptateurs aux extrémités des fragments, puis fixation sur la « flowcell ».
B – Fixation de la partie opposée du fragment sur la plaque et formation d'un « pont ». Extension du brin complémentaire par PCR, puis dissociation.
C –Répétition de l'étape B et formation des « clusters ».
D – Identification des nucléotides incorporés par détection des fluorophores associés au ddNTPs.

Source : en.biomarker.com.cn, consulté le 16/09/21

Depuis la fin des années 2000, le séquençage Illumina a drastiquement modifié le paysage économique du NGS. Pour rappel, le *Human Genome Project*, initié en 1988 et achevé en 2003, permit le séquençage du premier génome humain pour environ 3 milliards de dollars. Avec le développement du NGS, notamment au travers de la méthode Illumina, le coût du séquençage d'un génome humain (Figure 3) est passé de 10 millions de dollars en 2007 à un coût de 1000 dollars en 2016 (facteur 10 000) (Wetterstrand 2020). Une des explications à cette chute des coûts est l'utilisation du multiplexage. Schématiquement, cette technique permet le séquençage simultané de plusieurs échantillons durant la même expérience *via* l'utilisation de marqueurs fonctionnant à la manière d'un code-barre. La génération massive et

le coût des séquences ne représentant plus des obstacles rédhibitoires, d'innombrables projets de séquençage voient le jour dont une partie est consacrée à notre espèce (Benson et al. 2018). Ces projets visent, pour l'essentiel, à établir le catalogue de la diversité génétique humaine afin de mieux comprendre l'évolution et l'histoire de l'humanité ou afin d'améliorer la santé humaine par l'étude des liens entre variations génétiques et traits phénotypiques. Cependant, plusieurs problèmes, inhérents aux limites des méthodes NGS et à la structure et complexité des génomes, ont nécessité de nouveaux développements.

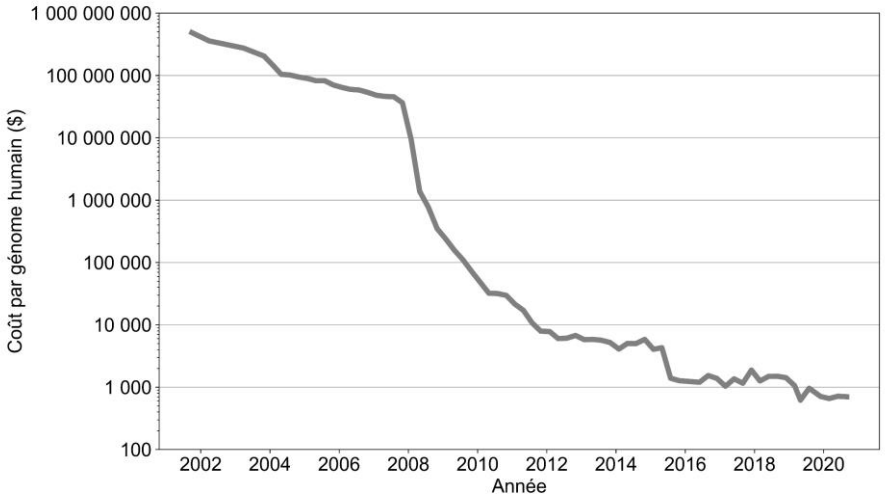


Figure 3 – Évolution du coût du séquençage d'un génome humain

Le coût d'un séquençage de génome humain est représenté en échelle logarithmique.
Source : Valeurs provenant du NCBI

1.3 Le séquençage de 3ème génération

Tout d'abord, les plateformes de séquençage NGS (incluant Illumina) impliquent toutes une amplification de l'ADN matrice pouvant entrainer des erreurs de copie et des biais. De plus, de nombreux organismes eucaryotes présentent des génomes complexes et riches en éléments répétés (54% chez l'homme, >90% chez le dipneuste géant australien). Chez l'homme, 630 Mpb sont composés de LINE (*Long Interspersed Nuclear Elements*) d'environ 7 kpb chacun (Nurk et al. 2021). L'utilisation de lectures courtes (150 pb) rend l'identification et l'agencement de ces éléments répétés difficiles dans le cadre d'un assemblage de génome. Deux nouvelles technologies aux méthodes orthogonales, mais avec un objectif commun, ont fait leur apparition : *Pacific Biosciences* (PacBio) et *Oxford Nanopore Technologies* (ONT) qui ont inauguré l'ère du séquençage par lectures longues (*Long Read Sequencing* ; LRS).

La première technologie repose (Eid et al. 2009) sur une observation optique en temps réel de la synthèse médiée par les polymérases sur une molécule unique d'ADN à partir de dNTPs fluorescents (Figure 4).

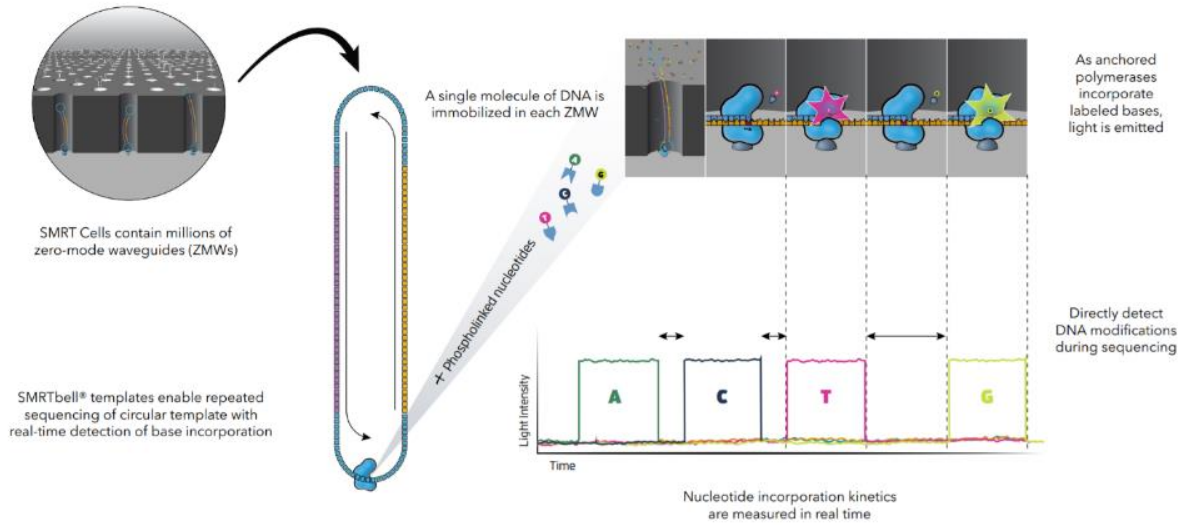


Figure 4 – Principe du séquençage PacBio

Après préparation de l'ADN en fragments d'environ 10 kpb, chacun d'entre eux est fixé un adaptateur afin de lui conférer une forme circulaire. Chaque puits appelé ZMW (Zero-Mode Waveguide) lit chaque fragment à plusieurs reprises grâce à la circularisation. Dans chaque ZMW, les nucléotides incorporés après polymérisation sont identifiés grâce à la fluorescence. Source : 2wordspm.com, consulté le 16/09/21

La seconde méthode (P. Chen et al. 2004) est basée sur un concept imaginé dans les années 1980, selon lequel une molécule d'ADN simple brin se déplaçant dans un canal étroit ou pore pourrait être analysée grâce à sa conductivité, révélant sa composition en acide nucléique. Au fil des améliorations technologiques (contrôle de la vitesse de passage de l'ADN dans le canal, ingénierie d'une protéine nanopore guidant l'ADN à l'intérieur d'une membrane résistante à l'électricité, circularisation…) les lectures ont évoluées d'une taille de 15 kpb jusqu'à 4 Mpb (Jain et al. 2018) et les problèmes de qualité de séquençage ont été résolus. Enfin, ces technologies permettent d'élaborer des séquenceurs de poche, se branchant en USB sur un ordinateur et demain, directement sur un smartphone (Figure 5) autorisant ainsi l'analyse d'un échantillon sur le terrain sans revenir en laboratoire.

Figure 5 – Exemples de séquenceurs Nanopore

À gauche – Séquenceur PromethION à grande capacité permettant d'utiliser 48 cellules en parallèle et générant jusqu'à 14 To de données par séquençage.
Au centre – Séquenceur MinION, se branchant sur le port USB d'un ordinateur, premier produit lancé en 2014.
À droite – Séquenceur SmidgION, futur produit se branchant directement sur un smartphone en exploitant l'évolution des téléphones et la connectivité 5G.

Source : Site Oxford Nanopore, consulté le 20/08/21

L'emploi des méthodes LRS permet depuis plusieurs années de réaliser des assemblages de très haute qualité pour des génomes complexes : le buffle d'eau (Low et al. 2019), le cobra (Suryamohan et al. 2020), le dipneuste géant australien (A. Meyer et al. 2021). Ces méthodes ont aussi contribué à l'amélioration de la détection des variants structuraux, dans le cadre de la génétique des populations (Beyter et al. 2021) mais également, dans le cadre clinique des maladies génétiques (Merker et al. 2018). Ainsi, de 2015 à 2018, le nombre de variants structuraux identifiés dans le génome d'un individu est passé de 2 200 à plus de 22 000 avec l'utilisation des méthodes LRS (Ho, Urban, et Mills 2020).

1.4 Applications des méthodes de séquençage

La révolution introduite par les méthodes de séquençage à très haut débit et à bas coût ne s'est pas limitée aux seules séquences des génomes. Tous les domaines de recherche où interviennent directement ou indirectement des acides nucléiques ont pu profiter des opportunités offertes depuis les études sur l'expression cellulaire, tissulaire et temporelle de chaque gène d'un organisme jusqu'aux travaux abordant l'organisation spatiale des chromosomes. Suite à de nombreuses adaptations des méthodes de séquençage ou des préparations des échantillons, le séquençage permet aujourd'hui d'aborder une multitude de processus dont je fournis ci-dessous une liste, certes non-exhaustive, mais que je voulais représentative des possibilités qui s'offrent désormais aux biologistes [(Soon, Hariharan, et Snyder 2013; Pachter 2013), Figure 6]. Ainsi, dans une présentation allant du génome aux produits des gènes, on peut aborder :

- Le génome complet, « cellulaire » ou « parental » par :
    - La détermination de la séquence dans un échantillon (DNA-Seq) ou dans une cellule unique (scDNA-Seq)
    - Le séquençage sélectif des brins d'ADN parentaux dans une cellule unique (Strand-Seq)
    - Le séquençage et le suivi des éléments transposables (Pool-Seq)
- La structure de la chromatine par :
    - Le séquençage des zones chromatiniennes en contact et la détermination de l'agencement tridimensionnelle des chromosomes (Hi-C-Seq)
    - Le séquençage des régions accessibles de l'euchromatine (DNAse-Seq, ATAC-Seq)
- L'épigénome par :
    - Le séquençage de l'ADN méthylé (MethylC-Seq, Bisulfite-Seq)
    - Le séquençage des régions 'protégées' par des protéines régulatrices (facteurs de transcription, histones modifiées…) (ChIP-Seq)

- Le séquençage évaluant l'activité des '*enhancers*' (STARR-Seq)

- La transcription par :
  - La quantification de tous types de transcrits (RNA non-codants, pré-messagers, messagers…) (RNA-Seq)
  - L'identification des variants d'épissage alternatifs (RNA-Seq)
  - Le séquençage des régions appariées offrant un accès à la structure des ARN (PARS-Seq)
- L'efficacité et la dynamique de la traduction (Ribo-Seq) par :
  - Le séquençage des transcrits présents au sein des ribosomes (RNA-Seq)

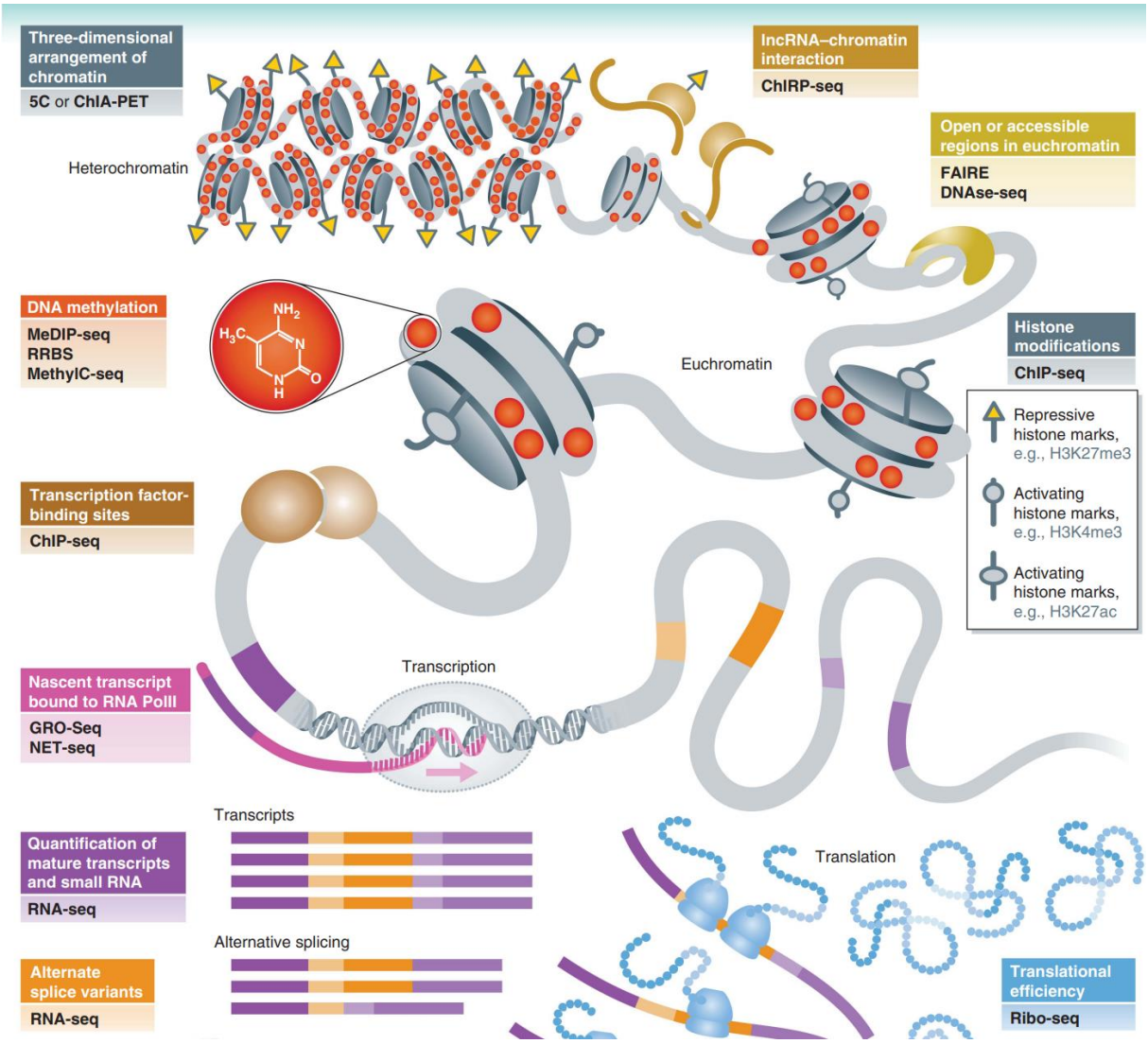


Figure 6 – Exemples d'exploitations de différentes méthodes de séquençage à haut débit

Diverses méthodes de NGS peuvent cartographier et quantifier avec précision les caractéristiques de la chromatine, les modifications de l'ADN et plusieurs étapes spécifiques de la cascade d'informations allant de la transcription à la traduction. Source : (Soon, Hariharan, et Snyder 2013)

1.5 Génomes humains de référence

1.5.1 Les assemblages de référence

Après la publication de la première version du génome humain en 2001 (Lander et al. 2001), différents assemblages successifs du génome ont été réalisés afin de compléter les régions non résolues et de corriger les biais de séquence existants. Ces assemblages sont gérés par le GRC (*Genome Reference Consortium*) cofinancé par différentes institutions, dont le NHGRI (*National Human Genome Reseach Institute*), le *Wellcome Sanger Institute* et l'EMBL-EBI (*European Molecular Biology Laboratory - European Bioinformatics Institute*).

Les deux assemblages utilisés couramment dans les projets de séquençage ou dans les bases de données sont le GRCh38.p13 (38 : 38ième version majeure ; p.13 pour patch ; 13 : 13ème version mineure) de Mars 2019 et le GRCh37.p13 de juin 2013. Les versions majeures présentent des changements de coordonnées chromosomiques tandis que les versions mineures contiennent des ajouts de séquences et des corrections, mais pas de changement de coordonnées. Ces assemblages sont composites. Ainsi pour GRCh38p.13, 70% de son contenu découle du génome d'un donneur anonyme masculin appelé RP11 et 23% de 10 librairies provenant des génomes d'individus distincts (Figure 7). Les 7% restants sont basés sur plus de 50 librairies composées d'un mélange d'hommes et de femmes anonymes ainsi que de chromosomes de diverses lignées cellulaires (Schneider et al. 2017).

On peut noter que sur le plan informatique, il est difficile de faire migrer la totalité des données, annotations et outils existants d'une version d'assemblage vers une autre. C'est pourquoi la majeure partie des sites et ressources sont aujourd'hui disponibles dans les deux versions. Lors de la manipulation des données, il est essentiel de contrôler la version d'assemblage utilisé afin d'éviter des comparaisons entre coordonnées chromosomiques différentes.

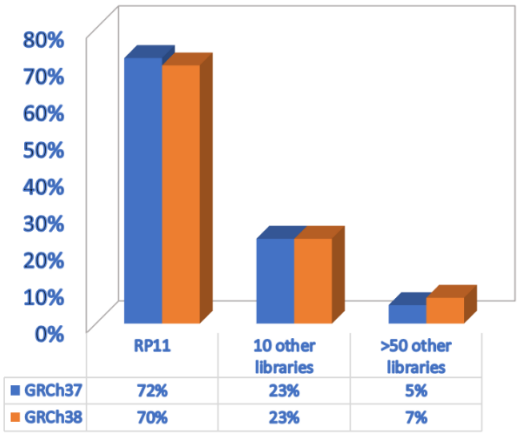


Figure 7 – Composition des assemblages GRCh37 et 38

Source : https://www.ncbi.nlm.nih.gov/grc/help/faq/ consulté le 15/08/21

1.5.2 Le nouveau génome humain

L'assemblageGRCh38, dernière version majeure d'assemblage humain publiée en 2013, est toujours incomplet avec près de 5% de « zones blanches » (nucléotides non caractérisés représentant 151 Mpb) et des régions complexes non séquencées. Début 2021, le consortium T2T, porté par K. H. Miga (UCSC) et A. Philippy (NIH), a obtenu le premier génome humain intégralement séquencé et assemblé, de télomère à télomère (Nurk et al. 2021). Cet assemblage, appelé T2T-CHM13, a introduit 200 Mpb jamais observées (Figure 8), 2 226 copies de gènes paralogues et 115 gènes codants pour des protéines. Ce nouveau jalon ouvre la voie à une réanalyse de l'ensemble des séquences et données associées et offre de nouvelles pistes pour comprendre les mécanismes cachés dans ces « zones blanches » du génome humain, très riches en éléments répétés.

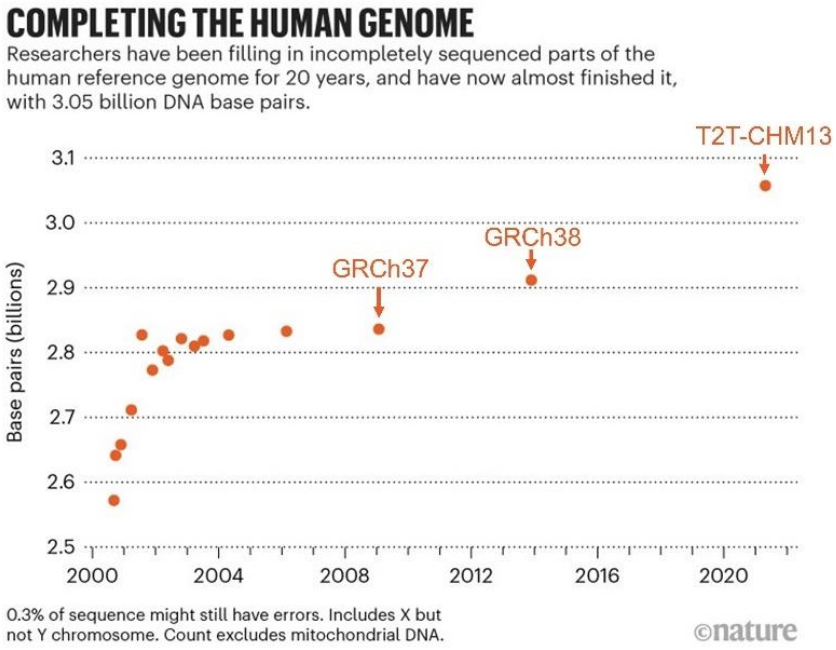


Figure 8 – Évolution de la taille du génome humain

Taille des génomes sans les nucléotides manquants , l'ADN mitochondrial et le chromosome Y
Source : (Reardon 2021)

Chapitre 2. Variations génétiques : reflet de la spécificité et de la diversité humaine

Les variations génétiques recouvrent l'ensemble des différences observées entre les génomes d'un individu ou d'un groupe d'individus et participent directement ou indirectement à la diversité phénotypique interindividuelle ou inter-éthnique. Les variations génétiques peuvent être abordées selon leur origine, leur taille, leur impact sur le génome et ses produits ou encore, leur fréquence dans la population. Cependant, toutes ces variations contribuent à l'hétérogénéité interindividuelle humaine au sein de laquelle se dissimule parfois la cause de dysfonctionnements moléculaires allant des maladies génétiques rares jusqu'aux maladies communes. Dans ce contexte, de nombreux efforts ont porté sur la caractérisation des fréquences alléliques des variations au sein de l'espèce humaine afin de pouvoir aborder l'étude de leur rôle dans son histoire et son évolution, dans l'émergence des maladies ou bien encore, dans la tolérance aux médicaments. La connaissance de ces fréquences est indissociable des multiples progrès réalisés non seulement, dans le nombre d'individus séquencés, mais également, dans l'organisation et la coordination des données de séquences humaines à l'échelle de consortiums planétaires.

2.1 Les variations génétiques chez l'être humain

2.1.1 Catégories de variations génétiques selon leur taille

On distingue deux catégories majeures de variations génétiques selon leur taille (Figure 9) :

- les variations ponctuelles d'une seule paire de bases appelées SNV (*Single Nucleotide Variation*) ainsi que les petites insertions et délétions (Indels) d'une taille comprise entre 2 et 50 nucléotides,
- les variations structurelles (*Structural Variation ;* SV) d'une taille supérieure à 50 nucléotides (Baker 2012).

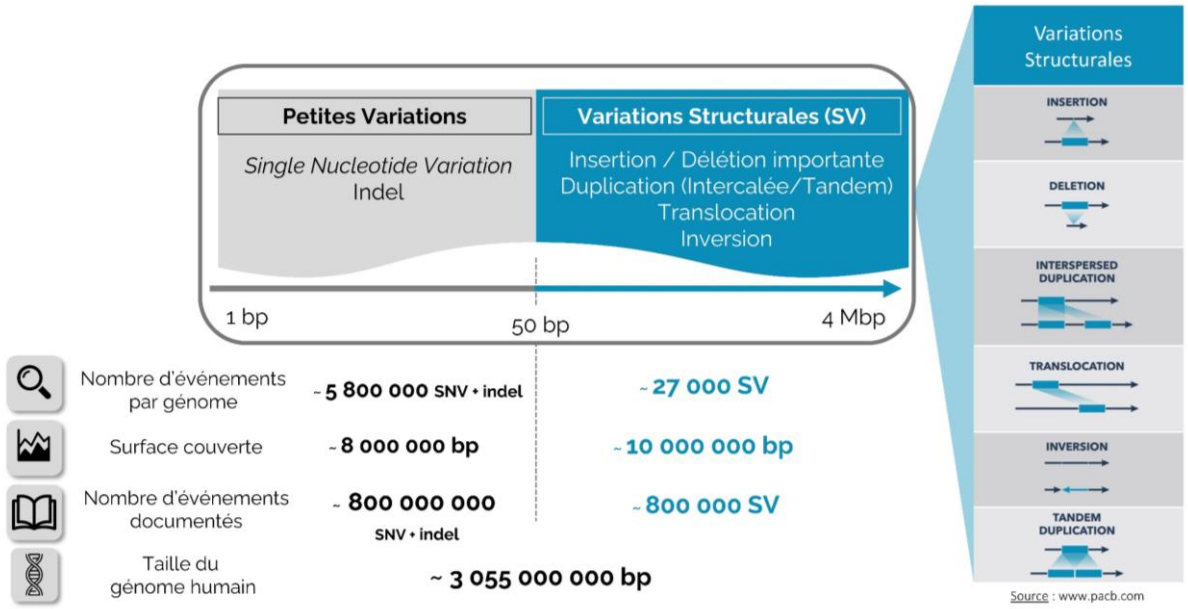


Figure 9 – Catégories et distribution des variations génétiques au sein du génome humain

Le nombre d'événements documentés est issu des bases gnomAD (section 6.4.2), dbSNP, dbVar et DGV. La taille du génome provient du dernier assemblage T2T (section 1.5.2).

2.1.1.1 SNV et Indels

Les SNV représentent la classe la plus fréquente, la plus étudiée et la plus documentée des variations génétiques, car elles peuvent engendrer une modification des séquences protéiques. Elles correspondent à un changement de nucléotide dans une séquence ADN par rapport à une séquence considérée comme référence. On distingue (Figure 10) les transitions correspondant au remplacement d'une base de même catégorie (purine : A ⇆ G; pyrimidine : C ⇆ T) et les transversions correspondant au changement d'une catégorie par une autre (purine ⟶ pyrimidine : A ⇆ C/T / G ⇆ C/T ; pyrimidine ⟶ purine: C ⇆ A/G / T ⇆ A/G). On estime à 5 millions le nombre de SNV chez un être humain (Auton et al. 2015).

Les indels représentent les insertions ou délétions de 2 à 50 pb. Dans les régions codantes, celles-ci engendrent un décalage du cadre de lecture si leur longueur n'est pas un multiple de 3. On dénombre environ 800 000 pb associées aux indels par génome humain (Auton et al. 2015).

La maladie de Huntington a été la première maladie génétique rare associée à une variation génétique (Gusella et al. 1983). Bien que le gène de la huntingtine n'ait été isolé que 10 ans plus tard, celui-ci a pu être associé à la pathologie à l'aide d'analyse de liaison génétique et de marqueur exploitant le polymorphisme de la longueur des fragments de restriction.

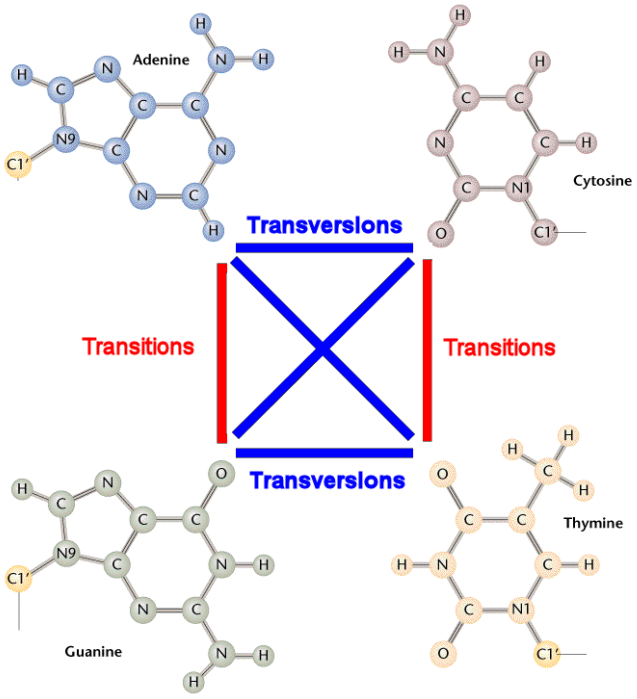


Figure 10 – Transitions et transversions aboutissant à un SNV

Source : (Carr 2014)

2.1.1.2 SV

Les variants structuraux concernent les variations génétiques de plus de 50 pb. Ceux-ci incluent les larges délétions et insertions dans le génome, les duplications en tandem ou intercalées, les inversions et les translocations (Figure 11). Les variations du nombre de copies (*Copy Number Variation* ; CNV) sont également souvent intégrées dans cette catégorie, qu'elles correspondent à des gains (*Copy Number Gain*) ou à des pertes (*Copy Number Loss*) de copies. Les SV représentent la classe de variations la plus récemment étudiée (Baker 2012). Grâce à l'utilisation des techniques LRS [section 1.3 ; (Jain et al. 2018)], le nombre de SV détectées par individu a récemment été revu à la hausse passant de 2 200 à 27 000 entre 2015 et 2018. On estime que l'ensemble des SV cumulés couvrent environ 10 Mpb sur le génome d'un individu, étendue supérieure à celle des SNV et Indels réunis (Chaisson et al. 2019). La première maladie génétique associée aux SV est la maladie de Charcot-Marie Tooth, identifiée par Lupski (Lupski et al. 1991) comme étant associée à une duplication en tandem de 1.5 Mpb sur le chromosome 17 qui entraine une triplication du gène PMP22.

Lors de ma thèse, je me suis focalisé sur l'étude des SNV et présenterait ci-après, leurs mécanismes d'apparition, leurs classifications selon leurs localisations ou impacts et leurs identifications.

2.1.2 Mécanismes d'apparition des SNV

Les SNV présents dans le génome humain sont dus au phénomène de mutagénèse, processus par lequel le contenu en nucléotides de l'ADN d'un organisme est modifié à une ou plusieurs positions données (Durland et Ahmadian-Moghadam 2021). La mutagénèse est un des moteurs de l'évolution en biologie, permettant aux individus d'une même population de développer des caractères phénotypiques avantageux ou désavantageux pouvant être soumis à une sélection. Bien que potentiellement utile d'un point de vue évolutif, la mutagénèse est strictement régulée et les cellules utilisent de nombreux mécanismes de réparation de l'ADN pour corriger les altérations nucléiques subies qu'elles soient endogènes (grande majorité des cas) ou exogènes (dues à un facteur environnemental) (Chatterjee et Walker 2017).

2.1.2.1 Mécanismes endogènes

Lors de la réplication cellulaire, les 3 milliards de nucléotides du génome humain sont copiés par les ADN polymérases (principalement δ, & et ϵ) qui possèdent un très faible taux d'erreurs estimé à 1 substitution de base sur 10^6 à 10^8 par cellule et par génération. Cette fidélité de la réplication est due à divers éléments : stabilité thermodynamique, sélection géométrique des dNTPs incorporés, suppression des dNTPs incorporés par erreur après relecture *via* une activité exonucléase… (Loeb et Monnat 2008). De plus, on trouve différents mécanismes de réparation de l'ADN, dont les défauts aboutissent à des mutations souvent associées à des cancers : réparation par excision des bases (*Base Excision Repair* ; BER) (Krokan et Bjørås 2013), réparation par excision des nucléotides (*Nucleotide Excision Repair* ; NER) (Schärer 2013), réparation des mésappariements (*DNA MistMatch Repair* ; MMR) (G.-M. Li 2008), recombinaison homologue (*Homologous Recombination* ; HR) (Sung et Klein 2006) et jonction des extrémités non homologues (*Non-Homologous End Joining* ; NHEJ) (B. Zhao et al. 2020). À noter que, parmi les mécanismes cellulaires endogènes, on trouve également des anomalies engendrées par les topoisomérases (Top I, II & III) (Pommier et al. 2006). L'ensemble de ces altérations constitue la principale source de mutagénèse spontanée.

D'autres phénomènes sont à l'origine de modifications telles que : la désamination spontanée de base entrainant une modification d'acide nucléique (Tomas Lindahl 1993), les sites abasiques (ou AP : apurinique / apyrimidique) ou sites « vacants » dus à des facteurs comme un pH extrême ou une température élevée (~10 000 sites par jour) et majoritairement corrigés par des endonucléases (T. Lindahl et Barnes 2000). Il existe également des phénomènes de lésions oxydatives de l'ADN dues à un excès de dérivés réactifs de l'oxygène (ROS), naturellement présents dans les cellules (Cadet et Wagner 2013).

2.1.2.2 Mécanismes exogènes

Les mécanismes exogènes conduisant à la mutagénèse se produisent lorsque des agents environnementaux rentrent en contact avec l'ADN, tels que :

- les rayonnements ionisants (Beta, Gamma, Neutron, rayons X) provenant de minerais, gaz ou dispositifs médicaux (Desouky, Ding, et Zhou 2015),
- les rayonnements UV du soleil [principalement UV-C (190 – 290 nm) étant donné la longueur d'onde d'absorption maximale de l'ADN (260 nm)] (Kiefer 2007).
- Les agents alkylants issus de l'alimentation, de la fumée de cigarette, de combustion, de traitements industriels ou de la chimiothérapie (Pegg 1990).
- Les amines aromatiques issues des pesticides, des automobiles, du charbon… (Skipper et al. 2010),
- ou bien encore, des hydrocarbones polycycliques aromatiques, des agents réactifs électrophiles ou des toxines (Chatterjee et Walker 2017).

Bien que la grande majorité des variations génétiques engendrées par la mutagénèse n'ont pas de conséquence délétère sur l'organisme, la localisation de certaines d'entre elles au sein, ou à proximité, d'un gène sont à l'origine de différents dysfonctionnements cellulaires pouvant engendrer des maladies génétiques.

2.1.3 Types de SNV

Les SNV sont habituellement classifiés selon leur impact sur le gène/transcrit. On distingue essentiellement les variations présentent dans les régions codantes des gènes, de celles situées dans les régions non-codantes ou les abords proches des gènes (Figure 11). Afin de normaliser l'impact des variations, une ontologie a été développée : la *Sequence Ontology* (SO) (Eilbeck et al. 2005). La nomenclature HGVS [*Human Genome Variation Society* ; (Dunnen et Antonarakis 2000)] est également communément utilisée, notamment dans le domaine biomédical.

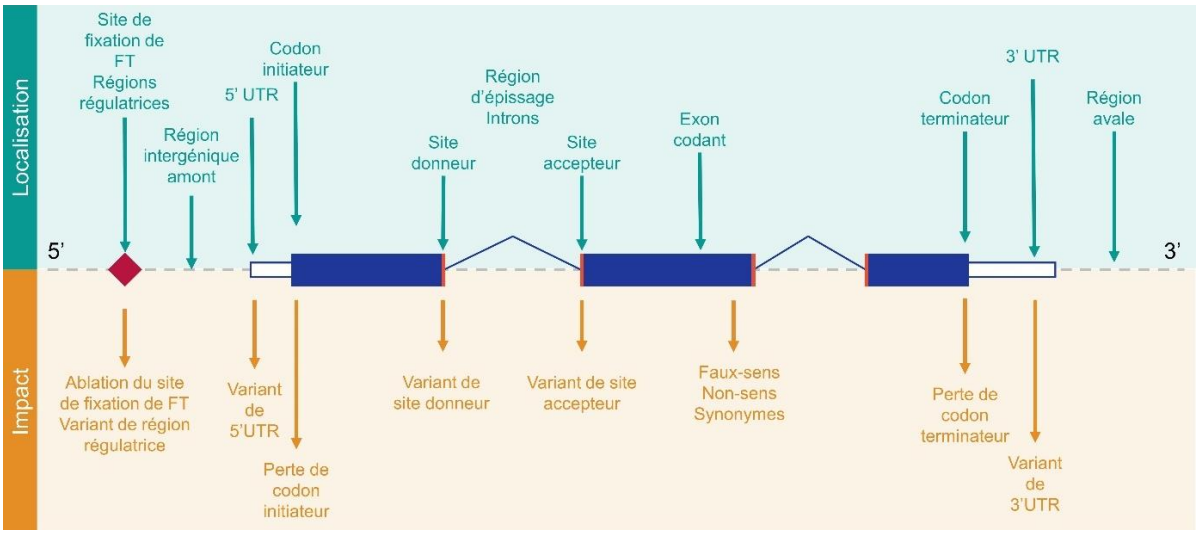


Figure 11 – Impact des SNV sur un gène selon leurs localisations

Adaptation de la représentation disponible dans la documentation de VEP (Variant Effect Predictor)
FT : Facteur de Transcription ; UTR : UnTranslated Regions

2.1.3.1 Variations codantes

Les variations codantes sont localisées dans les parties du gène (hors promoteur, *enhancers*, UTR…) impliquées dans la synthèse d'une protéine. Elles concernent environ 0,5% de l'ensemble des SNV (~25 000). On distingue les variations **synonymes ou silencieuses** (sSNV :~12 000 ; SO :0001819), sans impact au niveau de la séquence protéique en raison de la redondance du code génétique (Exemple : TAT ► TA**C** ; Tyr ► Tyr) des **variations non synonymes** (nsSNV : ~12 500) qui entrainent une modification de la séquence protéique et qui se répartissent en plusieurs catégories : les variations non-sens (~130 ; SO :0001587 ; Exemple : TAT ► TA**A ;** Tyr ► **STOP**) avec modification d'un acide aminé en codon terminateur, les variations entrainant une perte du codon initiateur (~120 ; SO :0002012 ; Exemple : ATG ► AT**A ;** Met ► **Ile**) ou du codon terminateur (~120 ; SO :0001578 ; Exemple : TGA ► T**C**A ; Stop ► **Ser**) et enfin, les variations faux-sens qui aboutissent au remplacement d'un acide aminé par un autre (~12 200 ; SO :0001583 ; Ex : TAT ► **C**AT ; Tyr ► **His**).

On distingue également les variations faux-sens conservatives, c'est-à-dire dont les propriétés physico-chimiques sont conservées, des variations faux-sens non-conservatives (Figure 12).

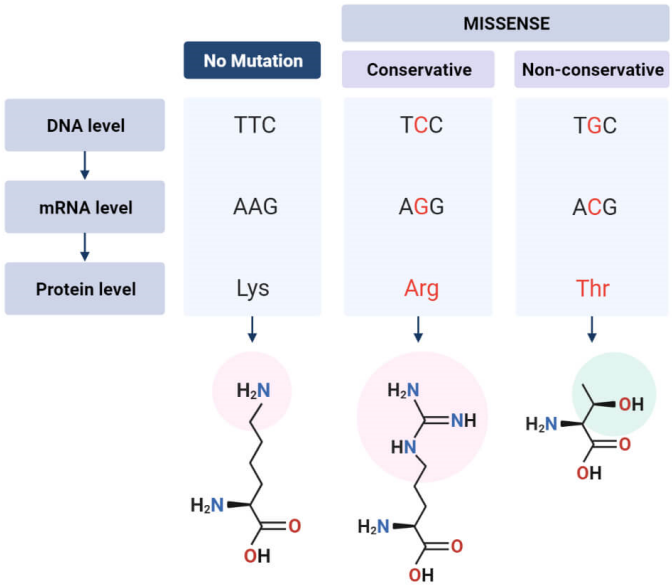


Figure 12 - Variations faux-sens conservatives et non-conservatives

Dans cet exemple, deux variations faux-sens présentes sur le même nucléotide conduisent dans un cas à une variation faux-sens conservative (Lys > Arg), étant donné la conservation de la charge positive, et dans l'autre cas à une variation non-conservative (Lys > Thr) ou la charge positive devient négative.
Source : https://microbenotes.com/missense-mutation/ consulté le 29/08/21

Les variations faux-sens sont celles que j'ai le plus étudiées durant ma thèse, notamment lors du développement du programme MISTIC.

2.1.3.2 Variations non-codantes

Les variations non-codantes représentent les variations hors des exons codants. Elles correspondent à la majorité des variations au sein du génome humain (~99.5% ; 5 000 000). En s'éloignant progressivement des exons, on identifie tout d'abord les variations touchant les sites donneur et accepteur d'épissage qui peuvent aboutir à une modification de la protéine produite (~500 ; SO :0001574 & SO :0001575), les variations introniques et les variations touchant les régions non traduites des gènes (*Untranslated Regions* ; UTR – SO :0001623 & SO :0001624). Enfin, on distingue les variations présentent dans les régions intergéniques en amont (SO :0001631) ou en aval du gène (SO :0001632) et celles affectant les régions régulatrices du génome (promoteurs, amplificateurs, inactivateur) (SO :0001566).

2.2 Exploitations des variations génétiques

2.2.1 Fréquence allélique dans la population

Une des caractéristiques essentielles des variations génétiques est leur fréquence dans la population. On parle de fréquence allélique pour définir la fréquence d'un allèle (*allele frequency* ; AF) et de fréquence allélique mineure (*Minor Allele Frequency* ; MAF) lorsqu'on

spécifie la fréquence d'une variation génétique. La fréquence allélique d'une variation est le fruit de l'évolution et du brassage entre individus dans les différentes ethnies composant la population mondiale. Ainsi, dès les premiers projets de séquençage à grande échelle, on distinguait la MAF calculée à partir de l'ensemble de la population (MAF globale) des MAF spécifiques à une ethnie (*e.g.* asiatiques de l'est du continent, afro-américains/africains, européens non-finlandais…).

L'utilisation massive du séquençage et l'essor du NGS ont entrainé une explosion des projets de séquençage humain, avec deux objectifs majeurs : identifier les sources de l'hétérogénéité interindividuelle ou inter-ethnique au sein de l'espèce humaine, et déterminer les causes des désordres phénotypiques liés aux maladies génétiques. Le premier objectif va se concrétiser par le développement de plusieurs consortiums successifs tirant profit des avancées technologiques du NGS afin d'agréger les données de projets de séquençage de plusieurs milliers [1000 genomes ; 1000G (Auton et al. 2015), *Exome Sequencing Project* ; ESP (Fu et al. 2013)], puis de dizaines de milliers d'individus [*Exome Aggregation Consortium* ; ExAC (Lek, Karczewski, et al. 2016), *genome Aggregation Database* ; gnomAD (Karczewski et al. 2020)], comme illustré sur la Figure 13. Le second objectif est principalement représenté par le développement de bases de données à vocation clinique visant à répertorier l'ensemble des variations observées chez des patients atteints de maladies génétiques [*Human Gene Mutation Database* (Stenson et al. 2014), ClinVar (Landrum et al. 2020)].

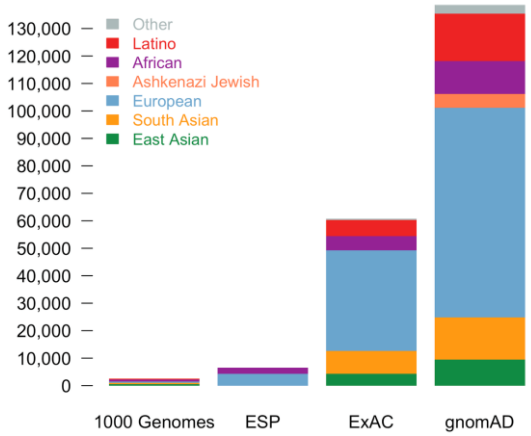


Figure 13 – Nombres d'individus séquencés par consortium

Source : (« gnomAD v2.1 | gnomAD news » 2018)

Avec l'évolution des consortiums (1000 genomes, ESP, ExAC, gnomAD) passant d'un ordre de grandeur de 1000 à 100 000 individus, la définition des ethnies devint plus précise. En effet, gnomAD distingue aujourd'hui les populations suédoises, bulgares, estoniennes, européennes du sud et du nord-ouest. Cependant, la composition de gnomAD demeure largement biaisée (Figure 14) par la prépondérance des populations européennes non-

finlandaises (plus de 50% de l'ensemble des individus étudiés) vis-à-vis des autres ethnies (*e.g.* population africaine/afro-américaine : moins de 10%).

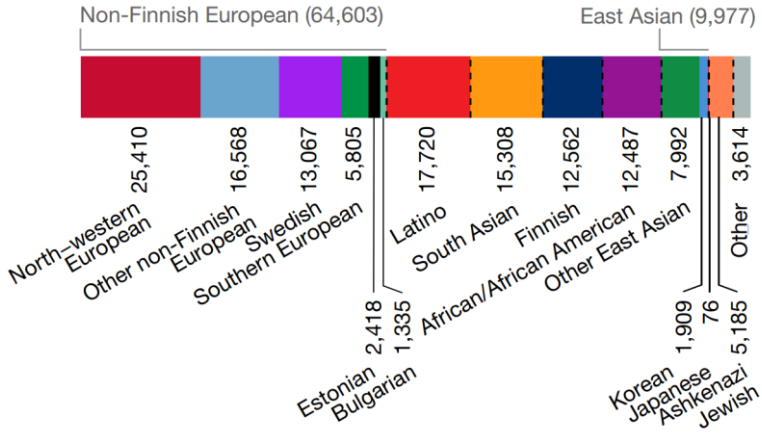


Figure 14 – Distribution des individus selon leur groupe ethnique d'origine dans gnomAD
Source : (Karczewski et al. 2020)

Lorsque la variation est fréquente dans la population (fréquence allélique > 1%), on parle alors de polymorphisme nucléotidique (*Single Nucleotide Polymorphism* ; SNP). La définition de variation commune/polymorphique (*i.e.* probablement non-délétère) dans la population générale a progressivement évolué, passant de 1-5% dans 1000 Genomes à 0.1% dans gnomAD. Dans la Figure 15, on note que parmi la totalité des variations répertoriées par le consortium ExAC (≈ 7.4M de variations à partir de 60 000 exomes), 99% ont une MAF inférieure à 1% dans la population et plus de la moitié sont des singletons, c'est-à-dire propre à un unique individu séquencé.

Aujourd'hui, environ 760 millions de variations sont répertoriées dans gnomAD (dont 16 millions dans les régions codantes). Cette notion de MAF est présente dans l'ensemble des sections suivantes notamment au travers de la compréhension de l'origine et de la diversité des ethnies, de la recherche de variations à MAF élevée en GWAS ou QTL (*Quantitive Trait Loci*) ou encore, de l'impact sur l'assimilation des médicaments. L'utilisation de ce type de données permet également d'exclure les SNV trop fréquents pour être des causes plausibles de maladies génétiques rares (Claussnitzer et al. 2020), présentées dans le Chapitre 3.

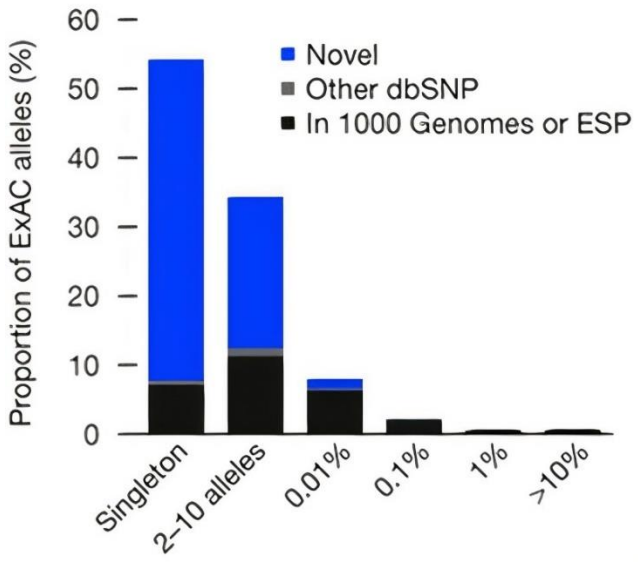


Figure 15 – Proportion de variants identifiés dans ExAC selon leur fréquence allélique
Source : (Lek, Fennell, et al. 2016)

2.2.2 Origines et ethnies

Une application de l'exploitation des SNV est l'étude des particularités génétiques présentent dans un groupe ethnique. En effet, selon sa localisation géographique et son passé, chaque population partage un patrimoine génétique propre. Un exemple est l'étude des populations finlandaises, nécessitant une attention particulière. En effet, les régions correspondant à la Finlande actuelle ont été inhabitées, puis peu peuplées entre la fin de la dernière glaciation (il y a 10 000 ans) et environ 1500 après JC. Ce n'est qu'à partir de cette période que les peuples occupant cette région ont commencé à migrer (Lamnidis et al. 2018). Ainsi, étant donné le manque de brassage génétique, certaines variations génétiques sont sur- ou sous-représentées au sein de la population finlandaise comparativement au reste de l'Europe. Lors de la construction des différents groupes ethniques étudiés, le consortium gnomAD a ainsi distingué la population finlandaise du reste de l'Europe afin de ne pas générer de biais statistique au niveau des fréquences alléliques calculées (Karczewski et al. 2020). Des divergences dans la structure génétique ont également été observées entre les populations vivant au Sud et au Nord de la Suède (Ameur et al. 2017).

Contrairement aux pays scandinaves, l'Afrique, berceau de l'humanité, présente une richesse en groupes ethniques qui corrèlent avec la diversité la plus importante en variations génétiques à travers le globe. Ainsi, dans une étude récente (Choudhury et al. 2020),des variations génétiques au sein d'une cohorte regroupant 426 individus de 50 groupes ethnolinguistiques africains, plus de 3 millions de nouvelles variations génétiques ont été répertoriées. En exploitant ces données, une hypothèse sur l'histoire migratoire des Bantous (locuteurs de 450 langues aux similitudes linguistiques) a pu être posée (Figure 16).

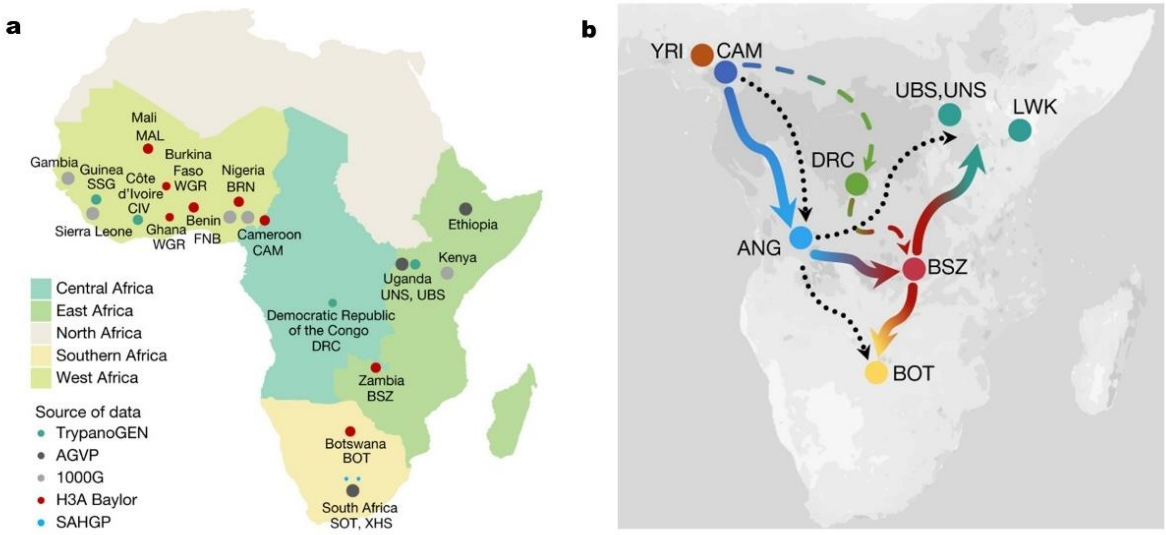


Figure 16 – Identification de mouvements de population à partir des variations génétiques

a – Expansions géographiques des populations étudiées par séquençage de génome complet (projet H3Africa) b – Mouvements supposés de populations durant la migration des groupes bantous à partir de leur histoire génétique. Les flèches pleines épaisses et colorées représentent les migrations inférées à l'aide des distances génétiques. Les flèches en pointillés noirs représentent l'ancien modèle utilisé. Source : (Choudhury et al. 2020)

2.2.3 Les associations pangénomiques

Pour l'analyse de traits (taille, poids, pression sanguine, certaines maladies communes ou cancers…), la variabilité génétique est utilisée *via* l'étude statistique de l'association entre ces traits et un continuum de variations génétiques polymorphiques (MAF > 0.5%). Dans ce cadre, on s'intéresse à l'ensemble des variations présentes dans le génome.

Dans le but d'identifier l'architecture génétique sous-jacente à ces traits, un nouveau modèle d'analyse s'est développé au début des années 2000 appelé études pangénomiques (*Genome Wide Association Studies* ; GWAS) (Dehghan 2018; Hirschhorn et Daly 2005; Uffelmann et al. 2021). Les GWAS se focalisent sur la recherche de variations déjà connues et relativement fréquentes dans la population. Schématiquement, ces études pangénomiques s'appuient sur un nombre important d'individus afin d'identifier les régions génétiques communes aux individus présentant un même trait phénotypique. Par la comparaison d'une cohorte de personnes cibles contre une cohorte témoin, on utilise des méthodes statistiques afin de faire ressortir les variations spécifiques au groupe cible (Figure 17). Plus le nombre d'individus dans le GWAS est important, plus la détection de variations sera précise. Plus de 5700 GWAS ont été réalisés sur 3300 traits différents en 20 ans (Uffelmann et al. 2021). En 2020, sur 107 000 SNPs statistiquement associés à un trait *via* GWAS, seuls 35% se situent dans les régions intergéniques tandis que les 65% restant se distribuent majoritairement dans

les introns des gènes (environ 62%) et seulement 3% dans les exons (Fauman 2020)[1]. Tous traits confondus, la surreprésentation des SNPs identifiées *via* GWAS dans des zones non-codantes illustre bien la difficulté à évaluer l'influence réelle de ces variations qui sont, majoritairement, toujours en attente d'une validation expérimentale démontrant un lien effectif aux traits associés.

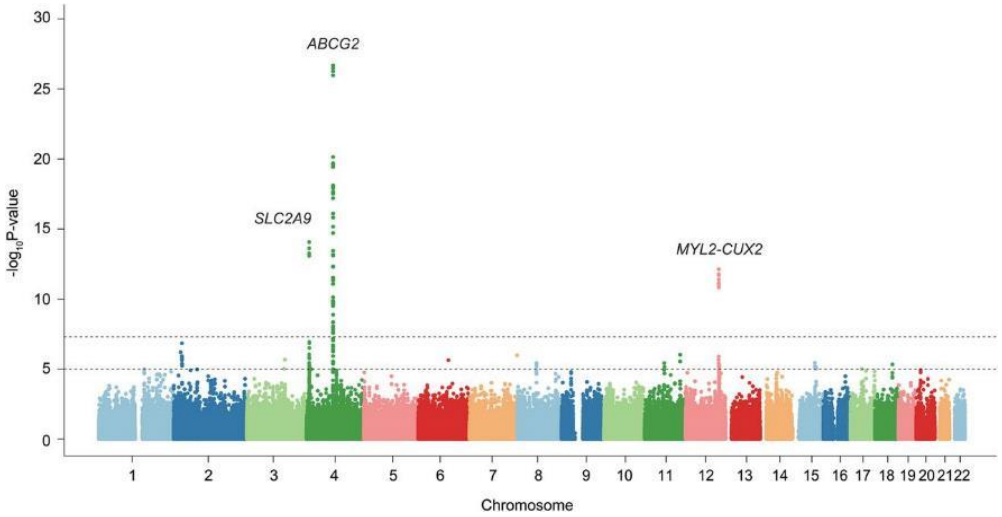


Figure 17 – Manhattan plot lors d'une étude GWAS portant sur la maladie chronique de la goutte

Les points passant les seuils de p-value (représentés par les lignes horizontales pointillées) correspondent aux associations statistiquement significatives. Source : (Matsuo et al. 2016)

2.2.4 Variations génétiques et traits quantitatifs

On désigne par trait quantitatif toute variation phénotypique **mesurable** (*e.g.* la taille, la couleur des yeux ou des cheveux) entre différents individus, que ce trait soit relié à l'environnement ou à la prédisposition génétique. On estime que la majorité des traits quantitatifs humains sont attachés à plusieurs loci génétiques, c'est pourquoi on parle de loci associés à des traits quantitatifs (*Quantitative Trait Loci* ; QTL).

Dans une logique complémentaire au GWAS qui s'intéresse formellement au lien discret (présence/absence) entre un nucléotide et un trait, l'analyse des QTL s'intéresse à évaluer le lien entre un locus génétique (correspondant à une région génomique plus ou moins grande) et un trait quantifiable de manière continue. Bien que ce lien soit étudié et utilisé depuis longtemps (Robertson 1967; Falconer 1996; Mackay 2001), le NGS et ses différentes applications (section 1.4) ont entrainé le développement de nouvelles méthodes évaluant les corrélations génome-épigénome/transcriptome/protéome *via* une approche intégrative multi-omiques. Cette révolution a porté l'analyse des QTL vers un changement de résolution, passant de l'identification d'une région génomique à un nucléotide spécifique permettant ainsi l'étude de divers types de QTL [(Shirai et Okada 2021) ;Tableau 1].

[1] Valeurs non publiées

Type de QTL	Description de l'acronyme
eQTL	*Expression*
sQTL	*Splicing*
mQTL	*Metabolite*
pQTL	*Protein*
meQTL	*Methylation*
rQTL	*Ribosome Profiling*
dsQTL	*DNAse*
hQTL	*Histone marks*
rdQTL	*RNA degradation*
mirQTL	*micro-RNA fixation*

Tableau 1 – Types de QTL rencontrés dans la littérature

Source : (Shirai et Okada 2021; Shi 2020)

Cependant, on étudie aujourd'hui principalement les QTL associés à un changement d'expression du gène (*expression* QTL ; eQTL) ou à une modification de la population de transcrits du gène (*splicing* QTL ; sQTL) à travers un couplage génome/transcriptome [(Consortium 2020), Figure 18]. L'analyse du génome fournit l'ensemble des variations génétiques tandis que l'étude du transcriptome permet la caractérisation des divers transcrits et leur niveau d'expression pour chaque individu.

On sait que les variations génétiques peuvent influer et moduler les machineries cellulaires liées à la transcription et à l'épissage. Ces mécanismes étant fortement couplés, un polymorphisme génétique peut engendrer une modification de l'expression d'un ou plusieurs gènes (eQTL) et/ou de l'épissage entrainant une modification des populations des transcrits (sQTL) d'un gène. Comme pour les études GWAS (Chapitre 2.2.3), une relation de puissance de détection est établie entre le nombre d'individus disponibles dans l'étude et la puissance de détection des QTL.

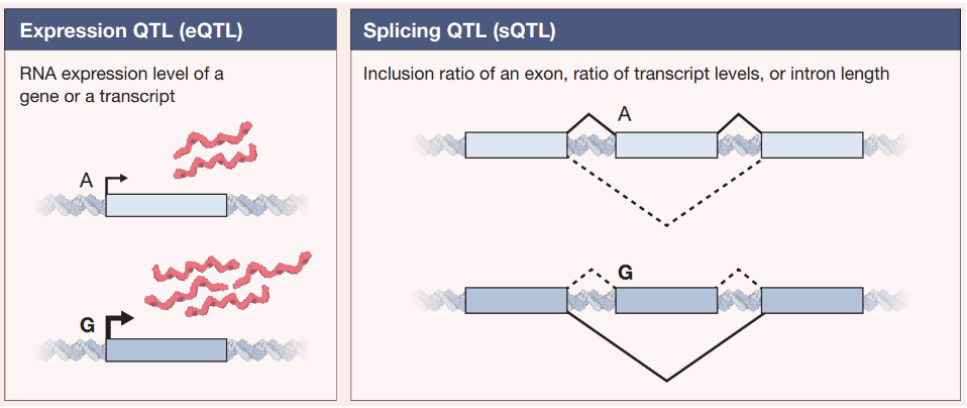


Figure 18 – expression et splicing Quantitative Trait Loci

Source : (Brandt et Lappalainen 2017)

2.2.5 Pharmacogénomique

La pharmacogénomique étudie l'influence des variations génétiques sur la réponse aux médicaments (Pinto et Dolan 2011). D'une part, cela permet de comprendre les liens entre variabilité génétique et métabolisation des composés pharmaceutiques, et d'autre part, de prédire la réponse d'un individu à un traitement pour l'adapter (dose, composition chimique). En effet, il est connu que l'hétérogénéité génétique interindividuelle participe de l'assimilation des médicaments, notamment par des différences de métabolisation, engendrant une inaction ou une toxicité du traitement. Cette différence de réponse, présentée en Figure 19, est abordée par la pharmacocinétique, étude du passage dans l'organisme d'une substance active (Absorption, Distribution, Métabolisation, Excrétion) et par la pharmacodynamique, effets qu'un principe actif produit sur l'organisme (Ahmed et al. 2016).

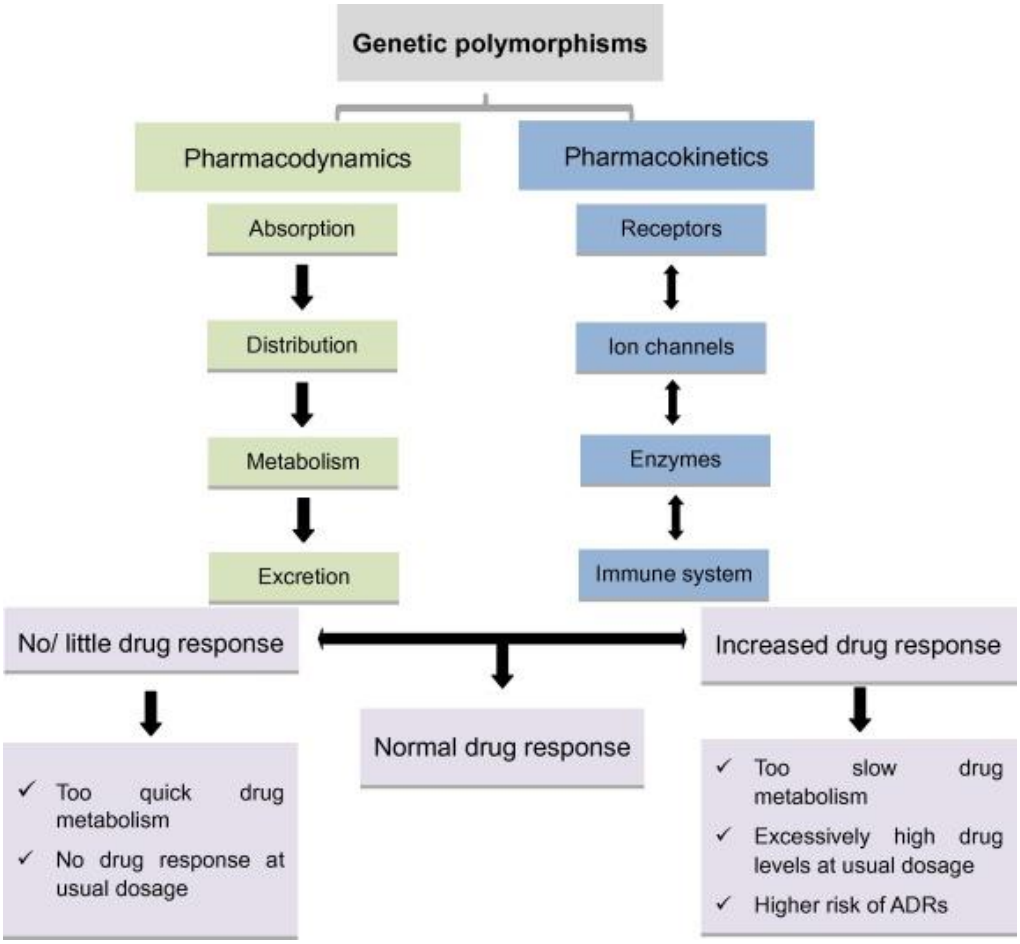


Figure 19 – Aperçu des conséquences des polymorphismes génétiques en pharmacogénomique

La pharmacocinétique et la pharmacodynamique sont les principaux déterminants des différences interindividuelles dans les réponses aux médicaments. Le polymorphisme génétique liés à ces processus peut entrainer des réponses hétérogènes aux médicaments. ADRs, adverse drug reactions. Source : (Ahmed et al. 2016)

Chapitre 3. Variations et maladies génétiques rares

Dans le Chapitre 2, nous avons abordé les différents types de variations génétiques et l'importance de leur fréquence dans la population au regard de l'origine des diversités ethniques, moléculaires ou phénotypiques. Ces variations peuvent être l'origine de désordres aboutissant à des maladies que l'on sépare classiquement en **maladies génétiques rares**, souvent à forte expressivité phénotypique et à déclenchement précoce et **maladies communes** dans la population. Toutefois, cette distinction est délicate à établir tant l'effet des variations génétiques s'inscrit plutôt dans un continuum que dans une hiérarchie stricte et la définition même de maladie génétique rare reste sujette à caution et varie selon les pays et les continents.

Pour l'étude des maladies génétiques rares, le séquençage constitue une méthode particulièrement adaptée, car il permet l'identification des variations causales, même si leur fréquence est extrêmement faible. Néanmoins, pour faire émerger les variations causales de la masse des variants bénins, délétères ou à signification inconnue identifiés par séquençage, plusieurs étapes de traitement des données sont nécessaires. Toutes ces étapes sont contraintes par un impératif de qualité lié aux besoins évidents d'un usage dans le cadre de la santé et de la pratique clinique. Face à cet impératif, ces étapes ont été codifiées par des recommandations et directives internationales afin de standardiser les multiples connaissances à mobiliser (fréquences alléliques, génotypes familiaux…) ainsi que les diverses normes et méthodes à appliquer (outils de prédiction *in silico*, validations expérimentales…).

Enfin, pour clore ce chapitre et illustrer un emploi clinique récent de séquençage combiné aux outils modernes de prédiction des variations causales décrits dans le Chapitre 4, je présenterai un exemple saisissant de diagnostic moléculaire ultra-rapide qui a sauvé la vie d'un nouveau-né atteint de maladie génétique rare.

3.1 Maladies génétiques rares et maladies communes

Une division, instaurée de longue date, distingue les maladies rares monogéniques et syndromiques des maladies communes. Les maladies rares, tel qu'illustré dans la Figure 20, sont classiquement associées à des variations à pénétrance élevée (proportion d'individus qui présenteront le phénotype en présence du génotype) et à MAF faible. À l'opposé, on trouve les maladies communes associées à des variations à faible pénétrance et à fréquence élevée. Cependant, d'autres situations ont été observées impliquant des variations rares à faible impact ou des variations communes à fort impact. Ceci illustre bien la difficulté à établir des relations strictes entre maladies et variations génétiques dont les effets s'inscrivent plutôt dans une sorte de continuum.

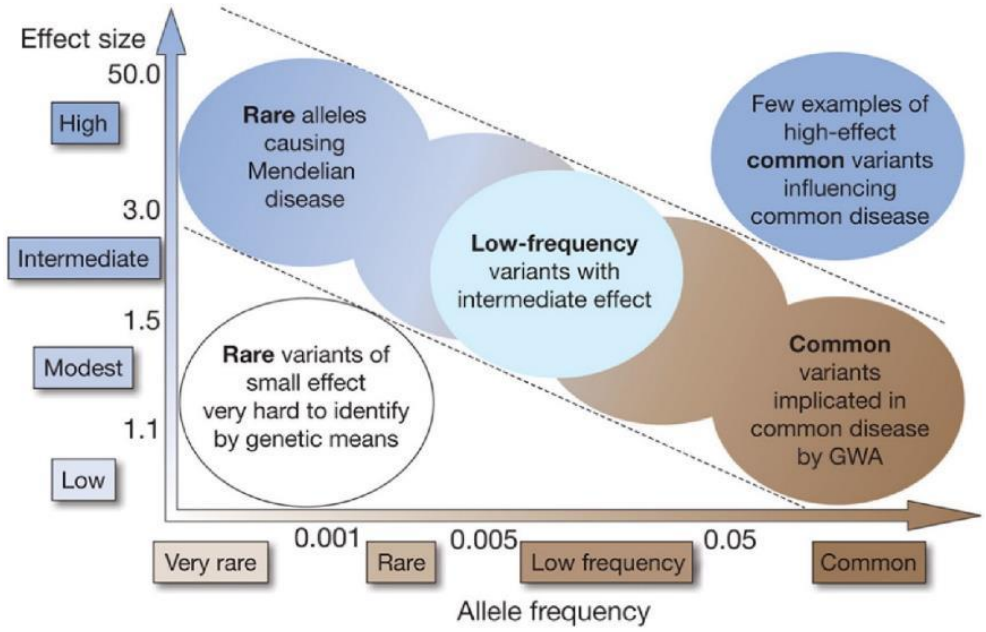


Figure 20 – Spectre des variations génétiques selon leur fréquence allélique et leur effet

L'effet (ordonné) correspond à un « Odds Ratio », risque relatif que la variation ait un impact au niveau génétique. Source : (Manolio et al. 2009)

On associe souvent les termes maladies génétiques rares et maladies mendéliennes étant donné la relation de causalité forte qui lie un locus génétique à la maladie (Rahit et Tarailo-Graovac 2020). En effet, un nombre élevé de maladies génétiques rares sont associées à un déclenchement précoce d'origine génétique et sont souvent monogéniques avec pour cause l'expression d'une variation délétère touchant directement ou indirectement un gène codant pour une protéine. À l'inverse, les maladies communes sont classiquement associées à des causes multiples au sein desquelles on trouve des facteurs environnementaux et des variations à faible impact et fréquence élevée, principalement localisées dans les régions intergéniques ou dans les régions non-codantes de plusieurs gènes (maladies polygéniques) (Claussnitzer et al. 2020).

La définition des maladies rares ou communes varie selon les pays et continents (Figure 21A). Si une maladie rare est toujours associée à une pathologie touchant un nombre d'individus restreint, les seuils de prévalence varient entre l'Union Européenne (1 cas pour 2 000 habitants = 0.05%), les États-Unis (< 200 000 cas soit moins de 1 cas pour 1 500 habitants= 0.06%) ou le Japon (< 200 000 cas soit moins de 1 cas pour 2 500 habitants= 0.04%). On recense aujourd'hui environ 6 200 maladies génétiques rares selon la ressource Orphanet (Maiella et al. 2013) et l'on estime que l'ensemble de ces maladies touche environ 5% de la population mondiale (~400 millions d'individus), dont 3 millions en France et 25 millions aux USA.

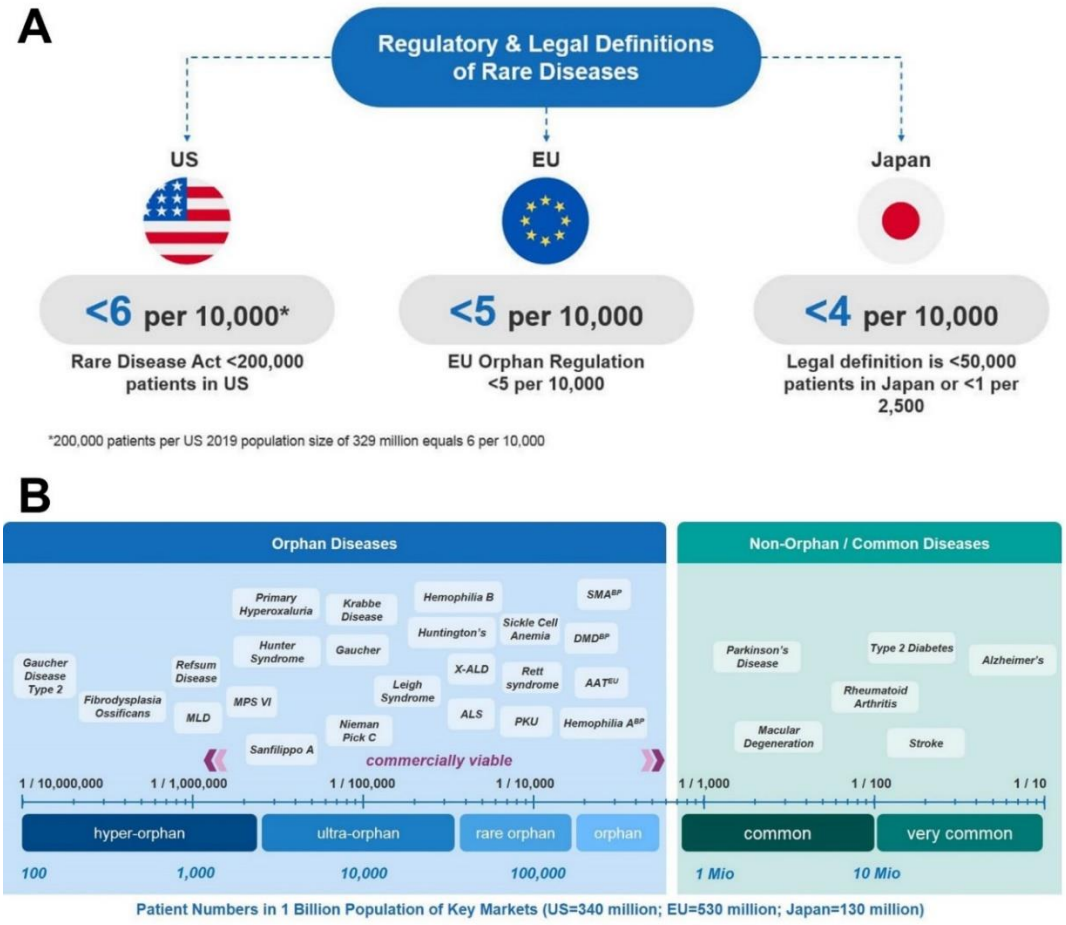


Figure 21 – Maladies rares et maladies communes

A – Définition des maladies rares aux États-Unis, dans l'Union Européenne et au Japon en fonction du nombre de cas pour 10 000 individus. B – Exemples de maladies rares orphelines et de maladies communes
Source : (Moritz 2020)

3.2 Séquençage clinique

3.2.1 Type de séquençage selon la surface du génome à couvrir

Différents types de séquençage NGS ont été progressivement intégrés en clinique afin d'identifier la cause d'une maladie génétique : le séquençage d'un panel de gènes cibles (*Targeted Sequencing* ; TS), connus comme étant reliés aux phénotypes du patient, le séquençage de l'ensemble des régions codantes pour les protéines (*Whole Exome Sequencing* ; WES) par le développement de kit de capture (Albert et al. 2007) et enfin récemment, le séquençage de l'intégralité du génome (*Whole Genome Sequencing* ; WGS).

Au regard du nombre limité de variations à analyser, l'utilisation du TS restreint les chances d'identifier un faux positif et facilite l'interprétation des données, cependant, comme on l'a déjà noté, l'analyse est limitée aux seuls gènes déjà associés aux maladies pressenties. En s'intéressant à l'ensemble des régions codant pour des protéines (2% du génome humain), le WES couvre les principales sources de maladies génétiques à savoir, en 2018, près de 5000 gènes (environ 1/4 des gènes codants pour les protéines) et plus de 85% des variations responsables des maladies génétiques connues (Caspar et al. 2018). Une des limites du WES est la définition des régions exoniques présente dans les kits de séquençage, pouvant aboutir à une absence de capture de certains exons ou à une mauvaise couverture de ceux-ci (Sastre 2014). Enfin, le WGS permet l'accès à l'ensemble des variations, codantes, non-codantes, aux variations structurales (SV), aux variants de nombre de copie (CNV) ainsi qu'aux variations affectant des éléments de régulation (*enhancers*, *silencers*, miRNA, snoRNA, lncRNA). Cependant, le volume important de variations et l'ampleur des impacts potentiels associés compliquent de façon exponentielle l'analyse pour l'homme et la machine (Shevchenko et Bale 2016).

3.2.2 Couverture de séquençage

L'emploi du séquençage dans le cadre clinique a exacerbé un problème important de toutes les méthodes de séquençage, le taux ou la profondeur de couverture. Classiquement, afin de pallier les biais liés aux différentes étapes d'amplification, séquençage, assemblage ou autres, chaque nucléotide est séquencé plusieurs fois, c'est ce que l'on appelle la profondeur de couverture (en X, correspondant au nombre de fois moyen qu'un nucléotide a été observé durant le séquençage) (Sims et al. 2014). Ce taux est souvent déterminant pour l'exploitation ultérieure des séquences et, à un coût équivalent, ce taux variera selon la taille des régions séquencées. Ainsi, un TS ou un WES vont aboutir à des taux de couverture supérieurs à 100X beaucoup plus importants que ceux d'un WGS (Tableau 2). Au-delà de la taille, l'objectif du séquençage influence le taux de couverture, ainsi les séquences de nouveaux génomes ou de transcriptomes (toutes espèces confondues) sont fréquemment publiées avec des couvertures allant de 5 à 10X. Cependant, dans le cadre d'un séquençage clinique de génome humain susceptible d'intervenir dans un diagnostic moléculaire, on exige toujours une couverture de séquençage beaucoup plus élevée comprise entre 30 et 60X afin d'être en mesure de différencier de façon non ambiguë, erreurs techniques et variations génétiques (Koboldt 2020).

Strategy	Panel	Exome	Genome
Size of target space (Mbp)	~ 0.5	~ 50	~ 3200
Average read depth	500–100×	100–150×	~ 30–60×
Relative cost	$	$$	$$$
SNV/indel detection	++	++	++
CNV detection	+	+	++
SV detection	–	–	+
Low VAF	++	+	+

Tableau 2 – Comparaison des approches TS (Panel), WES (Exome) et WGS (Génome) en clinique

Les symboles en dollars représentent les coûts relatifs approximatifs, bien qu'il faille noter que le coût du séquençage de TS dépend de la taille du panel de gènes à couvrir. La performance empirique de chaque stratégie pour la détection de variants de différentes classes est indiquée comme exceptionnelle (++), bonne (+) ou faible/absente (-).VAF = Variant Allele Frequency = MAF. Source : (Koboldt 2020)

3.2.3 Étapes d'analyse lors d'un séquençage

À ce jour, pour des raisons de coûts et d'accessibilité, le WES demeure la technique la plus communément utilisée en clinique. Son taux de résolution (pourcentage des cas où la variation délétère est identifiée sans ambiguïté) oscille entre 30 et 50% (Schwarze et al. 2018). Ce taux, qui peut paraître relativement faible, est lié à de multiples facteurs, entre autres, à la difficulté de caractériser les variations causales, parfois absentes des exons séquencés ou dans certains cas présentant des propriétés particulières, mais également aux nombreuses étapes nécessaires à l'identification de ces variations.

Comme pour tout processus de traitement des données, après l'obtention des données brutes d'un exome ou d'un génome, plusieurs étapes sont nécessaires pour identifier la variation délétère causale (Figure 22) : i) un prétraitement des lectures suivi de leur cartographie sur un génome de référence, ii) une identification des variations génétiques, puis iii) une analyse *via* l'annotation et l'interprétation des variants.

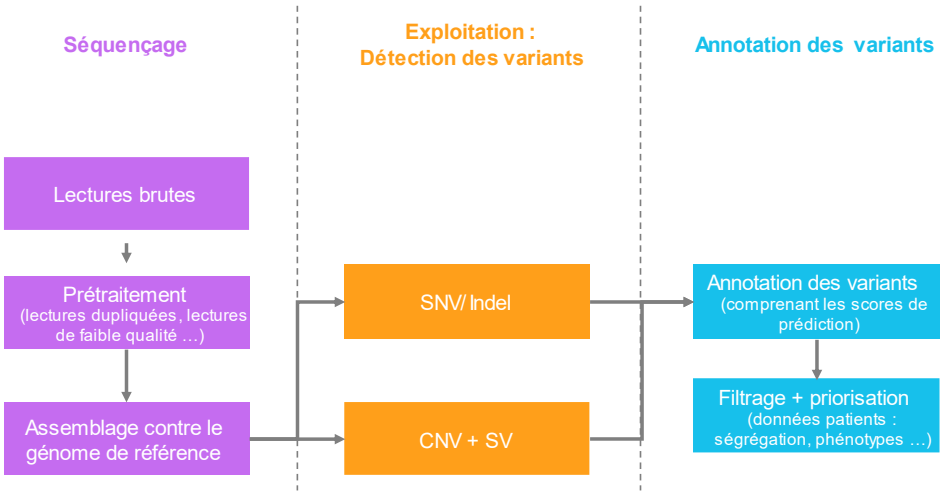


Figure 22 – Cascade d'analyse des variations génétiques en clinique

Source : inspiration de (Hiemenz et al. 2016)

3.2.4 Identification des variations de petite taille lors d'un séquençage

De multiples outils ont été développés afin d'identifier les variations génétiques de taille inférieure à 50 pb (SNV et indels, section 2.1.1.1). Cette détection est toujours réalisée à partir de la cartographie d'un exome/génome reconstitué sur un génome de référence (section 1.5). La plupart des outils dédiés aux variations de petite taille [*e.g. Samtools / BCFtools* (Danecek et al. 2021; H. Li et al. 2009) ou *FreeBayes* (Garrison et Marth 2012)] emploie des méthodes bayésiennes afin de définir de manière statistique le génotype le plus probable. D'autres outils [GATK *HaplotypeCaller* (Poplin, Ruano-Rubio, et al. 2018) et *Platypus* (Rimmer et al. 2014)] sont basés sur des outils de réalignements ou d'assemblage *de novo* des lectures de manière locale afin d'améliorer la détection des SNV/indels.

Enfin, plus récemment, des chercheurs de l'entreprise Verily (anciennement *Google Life Sciences*) ont développé *DeepVariant* (Poplin, Chang, et al. 2018), le premier outil de détection des SNV/indels basé sur des réseaux de neurones profonds (réseaux convolutifs) qui surpassent tous les outils existants en termes de fiabilité de détection. Pour l'ensemble de ces outils, plus la couverture de séquençage est importante, plus la fiabilité de la détection l'est également (Koboldt 2020). Les SNV/indels sont visualisables à l'aide d'outils spécialisés [*e.g.* IGV *genome browser ;* (Robinson et al. 2011)] permettant d'afficher un fichier d'assemblage des fragments séquencés et d'y faire apparaitre les variants détectés (Figure 23).

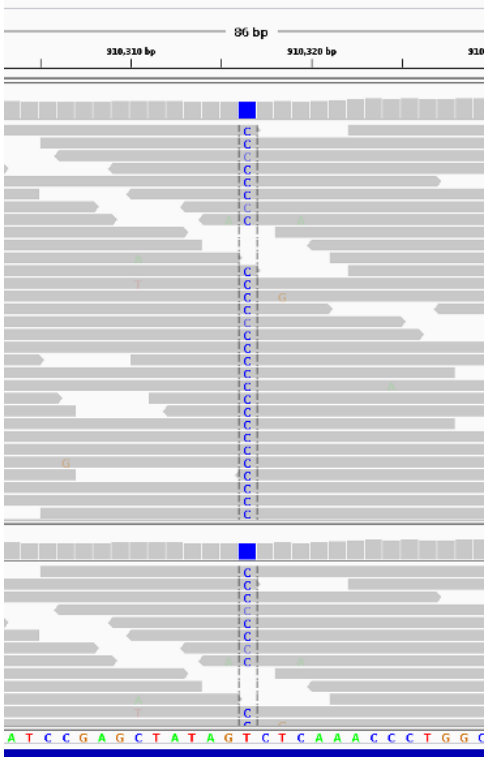


Figure 23 – Visualisation d'un SNV dans un assemblage de génomes dans le logiciel IGV
Source : wikis.utexas.edu, consulté le 25/09/21

3.3 Annotation et interprétation des variations

3.3.1 Annotation

Bien que les étapes de prétraitement, d'assemblage et d'identification des SNV/indels soient très semblables, quel que soit l'objectif du séquençage, l'étape d'exploitation et d'interprétation des SNV dans le cadre des maladies génétiques rares présente plusieurs spécificités. À la suite de l'identification des SNV/indels chez un individu, la première étape consiste à annoter ces variations afin de pouvoir les interpréter et diminuer le nombre de variations potentiellement causales. Cette étape nécessite l'emploi d'outils d'annotation [*Variant Effect Predictor* – VEP (McLaren et al. 2016); vcfanno (Pedersen, Layer, et Quinlan

2016); SnpEff (Cingolani et al. 2012)] et des bases de données [dbNSFP (Liu et al. 2020) ; dbscSNV (Jian, Boerwinkle, et Liu 2014)] permettant de transformer une variation en variation annotée avec de multiples informations (MAF et nombre d'homo/hétérozygote dans la population générale et dans les groupes ethniques, conservation, prédiction du caractère délétère, changement d'acide aminé au niveau protéique…). On distingue la « prédiction » de la conséquence d'une variation au niveau moléculaire (faux-sens, synonymes… ; section 2.1.3) de l'évaluation du caractère potentiellement délétère de cette variation (présenté ultérieurement dans le Chapitre 4).

3.3.2 « Bonnes pratiques » dans le cadre de l'analyse des variations en clinique

Plusieurs lignes directrices des « bonnes pratiques » en termes de manipulation et d'interprétation des SNV ont été publiées notamment, par le groupe de Daniel MacArthur (MacArthur et al. 2014), par l'ACMG-AMP (*American College of Medical Genetics - Association for Molecular Pathology*) (Richards et al. 2015) ou par l'ACGS-BSGM (*Association for Clinical Genomic Science - British Society for Genetic Medicine*) (Ellard 2020).

Afin d'évaluer si le statut attribué à une variation est délétère ou bénin, des catégories de critères à vérifier ont été suggérées (Tableau 3) :

- Données de populations pour cette variation : présence/absence dans la population ? (section 3.3.3)
    - Si présence, à quelle fréquence la variation est-elle présente dans la population générale / dans les groupes ethniques ?
- Données issues de prédiction ou de méthodes *in silico* (Chapitre 4)
    - Mécanisme de perte de fonction connu dans le gène (*Loss of Function* ; LoF) ou d'haploinsuffisance (HI) ? / même modification d'acide aminé à caractère pathogène connu ?
    - Consensus entre plusieurs outils de prédiction du caractère délétère ?
- Données fonctionnelles (section 3.3.4)
    - Études expérimentales démontrant une présence/absence du caractère délétère associé à la variation ?
- Données de ségrégation (section 3.3.5)
    - Présence de la variation chez des membres familiaux ?
    - *De novo* avec/sans validation du génotype parental ?
- Données alléliques
    - Présence en *cis* (dans le même gène) / *trans* (sur le gène situé sur l'autre copie du chromosome due à la diploïdie) de variant dominant/délétère ?

- Données provenant d'autres bases de données attestant du caractère délétère/bénin
- Autres données [ex : phénotype du patient spécifique d'un gène particulier tel que le *Gorlin syndrome* qui inclue un carcinome basocellulaire, des fosses palmoplantaires et des kératocystes odontogènes ; (Richards et al. 2015)]

	Benign		Pathogenic			
	Strong	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very strong**
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Tableau 3 – Tableau des « Standards & Guidelines » de l'ACMG

Ce tableau organise chacun des critères de l'ACMG selon le type de preuve ainsi que la force des critères pour une affirmation bénigne (côté gauche) ou pathogène (côté droit). BS, benign strong; BP, benign supporting; FH, family history; LOF, loss of function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong. Source : (Richards et al. 2015)

3.3.3 Emploi de la fréquence allélique mineure

Le premier filtre déterminant pour réduire le nombre de SNV/indels à étudier est celui de la MAF (section 2.2.1). Pour rappel, on estime que la grande majorité des variations impliquées dans les maladies génétiques rares présentent une MAF très faible, étant donné le niveau de pénétrance élevée de ces variations.

Dans le cadre de l'analyse de MAF calculées à partir des données du consortium gnomAD, il est recommandé d'étudier les variations présentant une fréquence inférieure à 0.1%, ce qui est en accord avec les niveaux de prévalence des maladies génétiques rares oscillant entre 0.04 et 0.07% selon les continents (section 3.1). Cependant, certaines maladies génétiques rares, comme la mucoviscidose, ne vérifient pas ces seuils et peuvent présenter des disparités importantes de fréquences des variations causales de la maladie. La plus fréquente dans le cadre de la mucoviscidose est la délétion de 3 pb Phe508del sur le gène CFTR (MAF globale dans gnomAD = 0.7%) qui est responsable de 83% des cas chez les individus de descendance européenne (MAF dans gnomAD chez les européens non Finlandais = 1.2%) (Boeck 2020), 30% des cas chez les individus de descendance africaine (MAF gnomAD chez les Africains/Afro-Américains = 0.26%) (Stewart et Pepper 2017) et aucun cas identifié en Chine (MAF gnomAD chez les Asiatiques de l'Est = 0%) (Zheng et Cao 2017). Dès lors, il est important de toujours garder à l'esprit, d'une part, que ces seuils sont arbitraires et ne s'appliquent pas à toutes les variations causales et d'autre part, que la MAF calculée sur l'ensemble de la population globale est insuffisante, et devra toujours être complétée par les MAF au niveau ethnique.

3.3.4 Études fonctionnelles

Les études fonctionnelles permettent de valider le caractère délétère des variations identifiées par NGS. De nombreuses techniques existent incluant : le « sauvetage » (*rescue*), des biomarqueurs, des cellules souches pluripotentes (*induced Pluripotent Stem Cell*), des modifications du gène cible (classiquement par *Knock Out*) dans des organismes modèles (souris, *zebrafish*) ou des dosages enzymatiques (Rodenburg 2018). Selon l'expérience réalisée, le niveau de fiabilité et de fidélité peut varier. Par exemple, un test enzymatique sur une biopsie issue d'un patient, voire d'un organisme modèle, démontre de manière plus fiable l'impact d'une variation que l'expression d'une protéine *in vitro*.

3.3.5 Utilisation des données de ségrégation familiale

Comme cela a été évoqué (section 3.1), les maladies génétiques rares sont aussi appelées maladies mendéliennes au regard de leur caractère souvent monogénique. Dès lors, ce type de pathologie répond aux lois de l'hérédité mendélienne. Etant diploïde, l'humain possède deux copies de chaque allèle. Ainsi, pour un même locus, on distingue les allèles identiques (**homozygotes**) des allèles distincts (**hétérozygotes**). Dans le cas d'une paire d'allèles hétérozygotes, chaque allèle n'a pas la même contribution et l'on distingue l'allèle s'exprimant (**dominant**) de l'allèle ne s'exprimant pas (**récessif**).

Il existe des cas rares de codominance où deux allèles hétérozygotes s'expriment. La dominance n'impacte pas les paires homozygotes.

En ce qui concerne les maladies génétiques rares, on séquence, si possible, l'exome/génome des membres de la famille afin d'identifier le mode d'hérédité associé à la variation causale. On distingue les modes d'hérédité dominant et récessif sur les autosomes (chromosomes 1 à 22) de ceux pouvant s'exprimer pour le chromosome X ou du mode dominant exclusif sur le chromosome Y (Figure 24). Il existe également des variations à caractère récessif dites hétérozygotes composites. Enfin, si la variation causale n'est pas retrouvée dans les génomes parentaux, celle-ci correspond à une mutation ***de novo*** à caractère dominant.

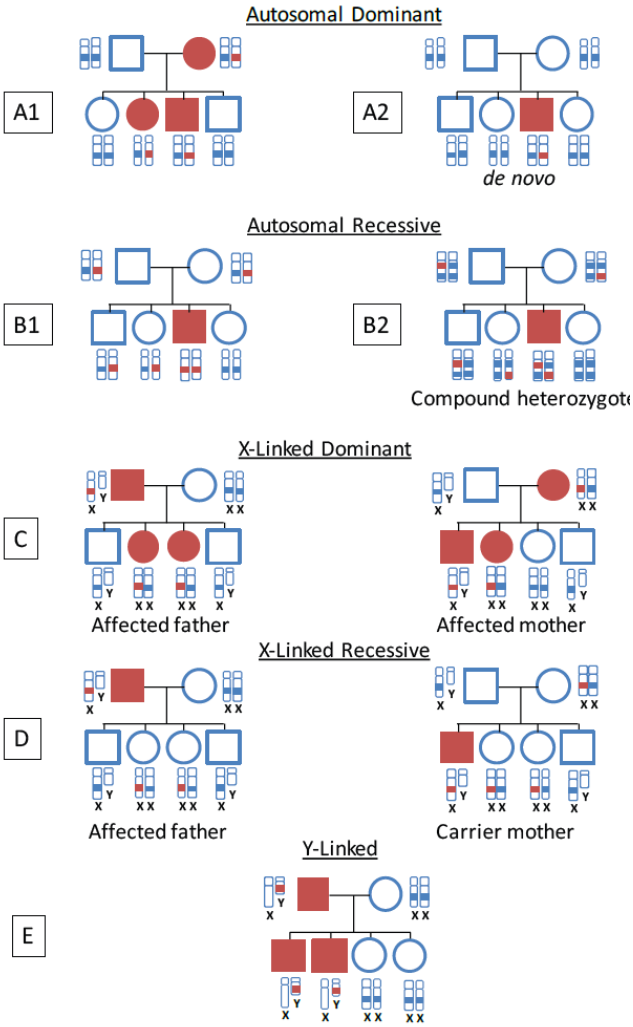


Figure 24 – Ensemble des différents modes d'hérédité

Carré : homme, rond : femme, couleur rouge : malade, couleur bleue : sain
Source : Thèse Kirsley Chennen

3.4 Répartition des variations génétiques dans ClinVar

La base de données ClinVar [décrite en 6.3.2, (Landrum et al. 2020)] fais partie des ressources d'études cliniques des variations génétiques offrant un accès à l'ensemble des informations décrites précédemment (fréquence allélique, validation expérimentale, pathologie associée…) et au statut clinique attribué (*pathogenic*, *benign*, *uncertain significance*) par l'auteur de l'analyse selon les critères fournis par l'ACMG. Le statut clinique attribué à une variation est une information majeure exploitée à la fois, par les experts dans le cadre de nouvelles analyses et par les outils de prédiction et bases de données associées pour construire les modèles de prédiction. Malgré les nombreux critères caractérisant une variation et les méthodes *in silico* (Chapitre 4) et *in vivo* (section 3.3.4) disponibles, près de la moitié du million de variations référencées dans ClinVar présente un statut clinique incertain (*Variants of Uncertain Significance* ; VUS ; Tableau 4). Le statut d'une variation n'est pas définitif et peut changer suite à une validation donnant lieu à des conclusions différentes ou suite au changement d'une fréquence allélique devenue trop fréquente (Shah et al. 2018). Tous ces éléments participent des conflits d'interprétation et de l'impossibilité de statuer.

La répartition des VUS n'est pas identique selon les conséquences moléculaires répertoriées dans ClinVar. Ainsi, environ 2/3 des VUS sont situés dans les régions codantes des gènes (Tableau 4). Parmi les classes impactant les protéines, la classe des faux-sens (*missense_variant*) est celle dont le taux de VUS est le plus important (70%) contre environ 10% pour les non-sens ou les variations de sites d'épissage donneur/accepteur. Enfin, en considérant la proportion de variations au statut délétère (*pathogenic*) sans conflit d'interprétation, les faux-sens représentent la classe la moins représentée (uniquement 10%) contre un taux de 85% pour les autres classes de variations à impact élevé.

Tous ces éléments indiquent à quel point, les variations faux-sens, objets d'études majeurs de mes travaux de thèse, constituent une population de variations difficile à cerner sans ambiguïté par les approches disponibles.

Région	Conséquence dans VEP	Nombre d'événements par individu sain (Eilbeck 2017)	Total dans Clinvar	"*Variants of Unkown Significance* (VUS)"	% du Total	"*Pathogenic without conflict*"	% du Total
Codante	*initiator_codon_variant*	117	1 633	622	38%	810	50%
	missense_variant	12 279	409 018	287 100	70%	41 228	10%
	nonsense	134	33 843	3 339	10%	29 009	86%
	stop_lost	120	509	274	54%	153	30%
	synonymous_variant	11 931	240 342	19 324	8%	701	0%
	Sous-total	24 581	685 345	310 659	/	71 901	/
Intronique	*intron_variant*	/	166 525	49 248	30%	8 750	5%
	splice_acceptor_variant	201	7 755	865	11%	6 461	83%
	splice_donor_variant	298	9 449	1 095	12%	7 858	83%
	Sous-total	499	183 729	51 208		23 069	/
UTR	*3_prime_UTR_variant*	/	55 974	34 552	62%	882	2%
	5_prime_UTR_variant	/	36 483	18 524	51%	2 800	8%
	Sous-total		92 457	53 076		3 682	/
Intergénique	*genic_downstream_transcript_variant*	/	66	35	53%	1	2%
	genic_upstream_transcript_variant	/	692	362	52%	10	1%
	Sous-total		758	397	/	11	/
Gènes non-codants	*non-coding_transcript_variant*	/	96 769	45 137	47%	9 127	9%
	Total	/	1 059 058	460 477		107 790	

Tableau 4 – Statistiques descriptives des variations génétiques répertoriées dans ClinVar

Version utilisée : clinvar_20210731.vcf.gz

3.5 Exemple de séquençage dans le cadre du diagnostic des maladies génétiques rares

Enfin, pour clore ce chapitre, je présenterai un exemple récent de séquençage clinique ultra-rapide (Owen et al. 2021) qui me tient à cœur, car il a permis l'obtention d'un diagnostic moléculaire en 11 heures (Figure 25). Cette prouesse a concerné un nourrisson de 5 semaines qui présentait de multiples phénotypes sévères (atteintes neurologiques et faciales) dont une encéphalopathie infantile associée à plus de 1500 maladies génétiques. La crainte du décès a nécessité la prise en charge rapide de l'enfant afin d'appliquer un traitement adapté dans les plus brefs délais. Après les tests neurologiques, un échantillon sanguin a été prélevé et préparé à 17h50 avant d'être séquencé à 19h23. Le lendemain matin, à 7h34, soit 11 heures plus tard, le séquençage du génome complet et son analyse bioinformatique étaient achevés. L'équipe a pu diagnostiquer, *via* l'utilisation d'une plateforme de traitement des variations exploitant une intelligence artificielle, la présence d'une duplication homozygote pathogène connue, entrainant un décalage de lecture sur le gène SLC19A3 (*thiamine transporter 2*), responsable d'un syndrome de dysfonctionnement du métabolisme de la thiamine. À 12h13, les premières doses de biotine et de thiamine ont été administrées et l'état du patient s'est amélioré de manière significative à 18h00. Cet exemple illustre bien l'intérêt du séquençage clinique dans le cadre de la génétique médicale personnalisée, qui peut permettre un traitement salvateur rapide.

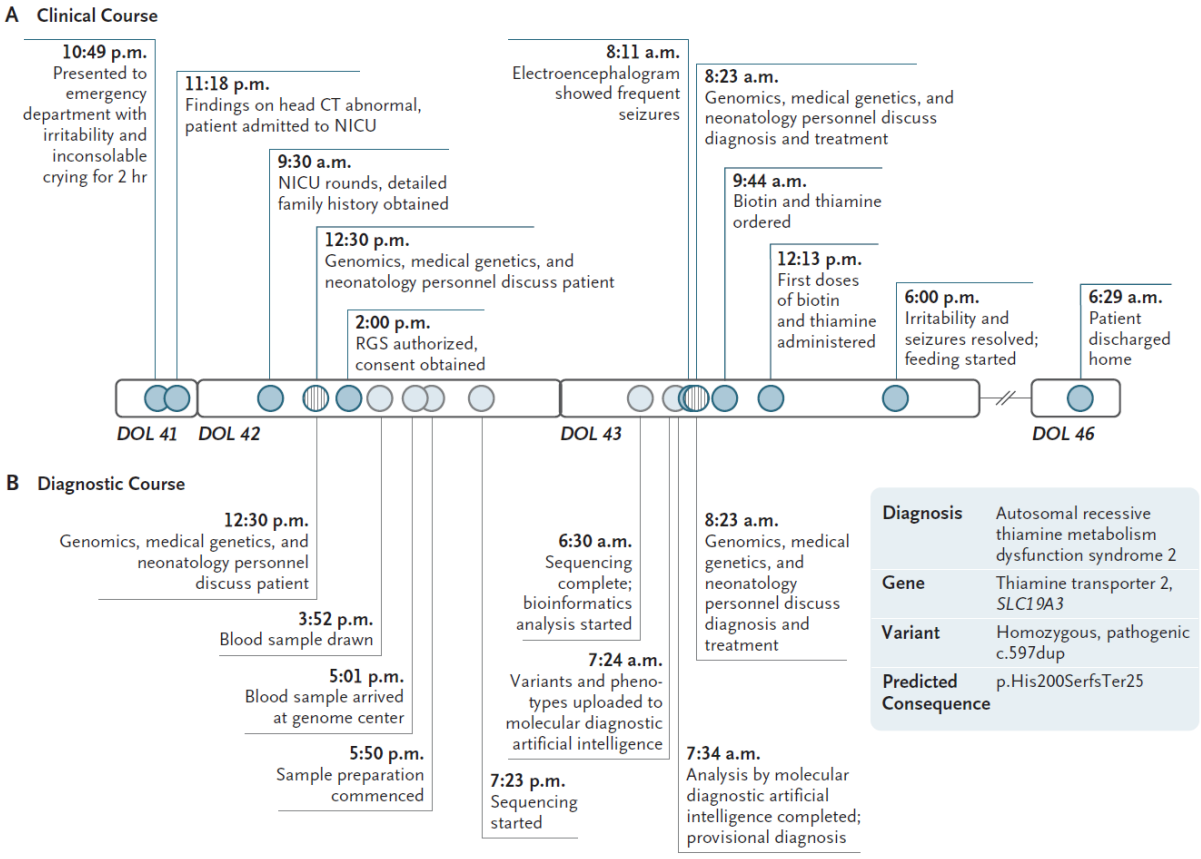


Figure 25 – Chronologie de la prise en charge clinique d'un nourrisson en situation critique

La variation homozygote du gène SLC19A3 du transporteur 2 de la thiamine détectée chez le patient avait déjà été signalée comme pathogène chez un enfant présentant un tableau similaire dans ClinVar. Les cercles le long de la ligne de temps indiquent les événements survenus au cours de l'évolution clinique (bleu foncé) et de l'évolution diagnostique (bleu clair). Les cercles avec des lignes verticales indiquent les points d'interaction entre le personnel de néonatologie, de génomique et de génétique médicale. CT : computed tomography, DOL : day of life, NICU : neonatal intensive care unit, et RGS : rapid genome sequencing. Source : (Owen et al. 2021)

Chapitre 4. Outils de prédiction de l'impact des SNV

Comme on l'a vu dans le chapitre précédent, la traque de la cause d'une maladie génétique rare est particulièrement ardue et implique de mobiliser de façon scrupuleuse tout un ensemble de connaissances et de moyens. Ceci est exacerbé dans le contexte d'une variation causale affectant un seul nucléotide (SNV). En effet, les génomes de deux êtres humains diffèrent d'environ vingt millions de nucléotides, dont cinq correspondent à des SNV (Auton et al. 2015). Globalement, cette diversité assure le brassage génétique permettant de pallier l'expression d'allèles récessifs hérités de manière ancestrale. Malheureusement, dans le cadre de la détermination des SNV délétères responsables de maladies génétiques rares, cette hétérogénéité naturelle complexifie le diagnostic (Eilbeck 2017). Dès lors, il a été nécessaire de faire évoluer les méthodes d'analyse manuelle vers une analyse informatisée et quasi-automatisée. Ce type d'analyse exploite la masse de données biologiques disponibles dont elles tirent de multiples catégories de **descripteurs** qui sont traités par de nouvelles approches algorithmiques, telle **l'intelligence artificielle**, afin d'améliorer sans cesse les performances des modèles de prédiction de l'impact des SNV.

Dans ce chapitre, je présenterai rapidement l'ensemble des outils de prédiction, leurs philosophies et leurs champs d'application (catégories de variations, exome/génome). J'aborderai également des aspects concernant les différentes familles de descripteurs employées dans le cadre des outils d'intelligence artificielle, l'importance des données d'entrainement ainsi que les méthodes d'évaluation des outils et les métriques associées. Ces notions ont été au cœur du développement de MISTIC (Chapitre 8), notre algorithme de prédiction de l'impact des variations les plus étudiées au sein des SNV non-synonymes (nsSNV) à savoir : les variations faux-sens.

Nom	Date	Faux-Sens	Synon.	Sites d'épissage	Non-codants	Nature du modèle	PMID
SIFT	2001	✓	✗	✗	✗	Prédicteur	11337480
PANTHER	2003	✓	✗	✗	✗	HMM	12952881
PANTHER-PSEC	2004	✓	✗	✗	✗	HMM	15492219
PMUT	2004	✓	✗	✗	✗	NN	15390262
MAPP	2005	✓	✗	✗	✗	Prédicteur	15965030
nsSNPAnalyzer	2005	✓	✗	✗	✗	RF^E	15980516
PhastCons	2005	✓	✓	✓	✓	Conservation	16024819
SNPs3D	2005	✓	✗	✗	✗	SVM	16169011
PhD-SNP	2006	✓	✗	✗	✗	SVM	16895930
MutationAssessor	2007	✓	✗	✗	✗	Prédicteur	17976239
SNAP	2007	✓	✗	✗	✗	NN	17526529
LRT	2009	✓	✗	✗	✗	Génomique comparative	19602639
MutPred2*	2009	✓	✗	✗	✗	RF^E	19734154
SiPhy	2009	✓	✓	✓	✓	Conservation	19478016
SNPs&GO	2009	✓	✗	✗	✗	SVM	19514061
GERP++*	2010	✓	✓	✓	✓	Conservation	21152010
MutationTaster2*	2010	✓	✗	✗	Intron	Naive Bayes	20676075
PhyloP	2010	✓	✓	✓	✓	Conservation	19858363
PolyPhen2*	2010	✓	✗	✗	✗	Classifier	20354512
Condel	2011	✓	✗	✗	✗	Système de $vote^E$	21457909
KGGSeq	2012	✓	✗	✗	✗	LR	22241780
PROVEAN	2012	✓	✗	✗	✗	Prédicteur	23056405
FATHMM	2013	✓	✗	✗	✗	HMM	23033316
VEST4*	2013	✓	✗	✗	✗	RF^E	23819870
CADD	2014	✓	✓	✓	✓	SVM puis LR	24487276
FunSeq2*	2014	✗	✗	✓	✓	*Weighted scoring scheme*	25273974
MetaLR	2014	✓	✗	✗	✗	LR	25552646
MetaSVM	2014	✓	✗	✗	✗	SVM	25552646
DANN	2015	✓	✓	✓	✓	DNN	25338716
DeepSEA	2015	✗	✗	✓	✓	CNN	26301843
deltaSVM	2015	✗	✗	✓	✓	SVM	26075791
FATHMM-MKL	2015	✓	✓	✓	✓	HMM + MKL	25583119
FitCons	2015	✓	✓	✓	✓	Evolution + Mesures Fonct.	25599402
GenoCanyon	2015	✗	✗	✓	✓	Non-supervisé	26015273
PON-P2*	2015	✓	✗	✗	✗	RF^E	25647319
SPANR/SPIDEX	2015	✗	✗	✓	✗	DNN	25525159
Eigen	2016	✓	✓	✓	✓	Non-supervisé	26727659
GenoSkyline	2016	✗	✗	✓	✓	Non-supervisé	27058395
GWAVA	2016	✗	✗	✓	✓	RF^E	24487584
M-CAP	2016	✓	✗	✗	✗	GB^E	27776117
PANTHER-PSEP	2016	✓	✗	✗	✗	Evolution	27193693
PredictSNP2*	2016	✓	✓	✓	✓	Consensus $classifier^E$	27224906
ReMM	2016	✗	✗	✓	✓	RF^E	27569544
REVEL	2016	✓	✗	✗	✗	RF^E	27666373
SIFT4G	2016	✓	✗	✗	✗	Prédicteur	26633127
DEOGEN2*	2017	✓	✗	✗	✗	RF^E	28498993
LINSIGHT	2017	✗	✗	✓	✓	Sigmoid linéaire	28288115
TraP	2017	✗	✓	✓	Intron	RF^E	28794409
CDTS	2018	✓	✓	✓	✓	*Metaprofile*	29483654
ClinPred	2018	✓	✗	✗	✗	$XGboost^E$ + RF^E	30220433
ExPecto	2018	✓	✓	✓	✓	CNN	30013180
FATHMM-XF	2018	✓	✓	✓	✓	HMM	28968714
FUN-LDA	2018	✗	✗	✓	✓	*Latent Dirichlet Allocation*	29727691
GenoNet	2018	✓	✓	✓	✓	Semi-supervisé / ElasticNet	30518757
PrimateAI	2018	✓	✓	✓	✓	CNN	30038395
eDIVA	2019	✓	✗	✗	✗	RF^E	31026367
MMSplice	2019	✗	✗	✓	✗	DNN	30823901
NCboost	2019	✗	✗	✓	✓	$XGboost^E$	30744685
S-CAP	2019	✗	✗	✓	✗	GB^E	30804562
SpliceAI	2019	✗	✗	✓	✗	CNN	30661751
LIST-S2*	2020	✓	✗	✗	✗	3 modules distincts	32352516
CADD-Splice	2021	✗	✗	✓	✗	DNN	33618777
MetaRNN	2021	✓	✗	✗	✗	RNN	*Preprint*
MTSplice	2021	✗	✗	✓	✗	DNN	33789710

Tableau 5 – Prédicteurs évaluant le caractère délétère des SNV identifiés dans la littérature
** : dernière version des prédicteurs. E : méthode d'ensemble ; RF : Random Forest ; SVM : Support Vector Machine ; HMM : Hidden Markov Model ; DNN : Deep Neural Network ; RNN : Recurrent Neural Network ; CNN : Convolutional Neural Network ; GB : Gradient Boosting ; MKL : Multiple Kernel Learning ; LR : Logistic Regression*

4.1 Prédiction et intelligence artificielle

Au regard du nombre de données à mobiliser, les prédicteurs et méta-prédicteurs actuels utilisent l'intelligence artificielle pour construire et entrainer leur modèle de prédiction (Greener et al. 2021; Libbrecht et Noble 2015). On distingue **l'apprentissage automatique** (*Machine Learning* ; ML) regroupant un nombre important de méthodes statistiques ayant pour objectif de réaliser une classification, de **l'apprentissage profond** (*Deep Learning* ; DL), une sous-catégorie du ML dédiée aux réseaux neuronaux multicouches. Dans le cadre du ML, les modèles entrainés apprennent à distinguer et prédire variations délétères et non-délétères en s'appuyant sur des jeux de données d'entrainement où ces deux catégories sont clairement **labélisées**. Différents développements récents ont permis la création de boites à outils dédié au ML telle que scikit-learn [section 7.1; (Pedregosa et al. 2011)] ou au DL telles que tensorflow/keras (Abadi et al. 2015) ou pytorch. Ces boites à outils simplifient l'utilisation non seulement, des méthodes d'intelligence artificielle, mais également des méthodes d'optimisation des hyperparamètres (paramètre dont la valeur est utilisée pour contrôler le processus d'apprentissage) utilisés par les algorithmes lors de l'apprentissage. Dans le cadre d'une classification binaire (variation délétère/non-délétère), l'amélioration des performances se concentre principalement sur la construction de jeux de données originaux et sur les développements de nouveaux descripteurs ou de nouveaux systèmes combinant des familles algorithmiques distinctes. Sans entrer dans la description détaillée des algorithmes communément employés en ML, leur usage dans de nombreux domaines a permis d'identifier certaines forces et faiblesses (illustrées pour cinq des méthodes parmi les plus employées dans le Tableau 6) que l'on retrouve dans le cadre de la prédiction de l'impact des variations délétères/non-délétères. Enfin, on peut noter que dans le cas spécifique du DL, la construction de réseaux multicouches et l'optimisation des différents paramètres demeure une étape complexe qui limite aussi bien leur usage généralisé que leurs performances.

Une autre différence concerne le niveau de supervision employé lors du développement des prédicteurs basés sur de l'intelligence artificielle. Sur le plan algorithmique, la grande majorité des méthodes de ML utilise des méthodes **supervisées** d'apprentissage reposant sur des algorithmes à base d'arbres (*random forest :* PON-P2, GWAVA, DEOGEN2*, gradient boosting :* M-CAP, S-CAP), de régression logistique (MetaLR, CADD) ou de machine à support de vecteurs (*Support Vector Machines :* SNPs&GO, MetaSVM) (Tableau 5). Cependant, on note l'émergence de prédicteurs tirant parti des avancées en apprentissage profond (*deep learning :* SPANR/SPIDEX, MMSplice) et des réseaux neuronaux à convolutions (*convolutional neural network*), spécialement dans l'analyse des régions non-codantes. Enfin, on peut noter l'apparition de méthodes basées sur un apprentissage **semi-supervisé**

(GenoNet) combinant un modèle supervisé (quand la labélisation des variations est disponible) et une classification non-supervisée ainsi que des méthodes entièrement **non-supervisées** (Eigen, GenoCanyon…) qui ne nécessitent pas des données préalablement labélisées.

Pour conclure, on distingue les modèles algorithmiques de classification dits « classiques/simples » comme la régression logistique et les machines à support de vecteurs, des méthodes de classification par « ensemble » (*Ensemble learning*) comme la forêt aléatoire, le *gradient boosting* ou la classification par vote. Les méthodes de classification par ensemble combinent plusieurs algorithmes (plusieurs arbres de décisions dans le cas de la forêt aléatoire par exemple) afin d'améliorer la robustesse et la « généralisabilité », c'est-à-dire la capacité de généraliser des propriétés à partir d'un nombre important d'observations, dans un modèle unique. Ces méthodes représentent une part importante des différents modèles utilisés dans le domaine de la prédiction de l'impact des variations aujourd'hui (Tableau 5), étant donné leur performance élevée et leur grande polyvalence.

Algorithme	Forces ✔	Faiblesses ✘
***Support Vector Machine* (SVM)**	- Bonnes performances en présence de beaucoup de dimensions - Forte importance des hyperparamètres - Faible impact des *outliers*	- Temps de traitement long et besoin de mémoire important - Faible performance si recouvrement entre classes (variants bénins mal classifiés)
***Naive Bayes* (NB)**	- Temps de traitement très rapide - Applicable à de grands jeux de données - Faible impact des *outliers*	- Requiert une indépendance totale des descripteurs - Jeu d'entrainement nécessite de coller parfaitement à la population décrite
***Logistic Regression* (LR)**	- Performance élevée dans beaucoup de situations - Mise à l'échelle des valeurs non nécessaire - Interprétabilité (poids des descripteurs) - Applicable à de grands jeux de données	- Faible performance sur données non linéaires (image) - Pas assez flexible pour réaliser des relations complexes entre descripteurs
***Random Forest* (RF)**	- Performance élevée dans beaucoup de situations - Bonne gestion des valeurs manquantes - Capable de modéliser des données non linéaires - Interprétabilité (poids des descripteurs) - Faible impact des *outliers*	- Besoin de descripteurs avec une valeur informative importante pour la prédiction
***Neural Networks* (NN)**	- Capacité à réaliser des relations complexes entre descripteurs - Applicable à de grands jeux de données - Bonne gestion des valeurs manquantes	- Difficulté à trouver l'architecture adéquate - Résultats difficiles à interpréter

Tableau 6 – Forces et faiblesses des cinq algorithmes les plus utilisés en intelligence artificielle

Données issues de elitedatascience.com ; kaggle.com ; aquero.com

4.2 SNV et génome : intrication entre localisation et classification

Comme on l'a vu, la grande majorité des SNV présentes chez un individu est éparpillée au sein de régions non-codantes du génome (régions intergéniques, introns, …) tandis qu'une faible fraction se situe dans les régions codantes traduites en protéine (2.1.3). Ces localisations vont jouer un rôle déterminant sur l'impact moléculaire potentiel d'un SNV, et depuis 20 ans, plus de 60 prédicteurs distincts ont cherché à modéliser le rôle de chaque nucléotide du génome humain pour évaluer et prédire l'impact des SNV (Tableau 5).

La première distinction séparant ces prédicteurs concerne leur champ d'application : l'ensemble des SNV d'un génome ou un type particulier (faux-sens, sites d'épissage, SNV dans les régions non-codantes). Dans la première catégorie, on trouve des outils tels que PhastCons, CADD ou Eigen (Tableau 5). En ce qui concerne les outils ciblés, on trouve des outils comme PolyPhen-2, MutationTaster2 ou M-CAP dédiés aux faux-sens ou des outils tels que S-CAP, CADD-Splice ou SpliceAI dédiés aux variations touchant les sites d'épissage.

Les variations faux-sens entrainant un changement de la séquence protéique représentent la catégorie la plus étudiée dans le domaine de la prédiction de l'impact des SNV (Eilbeck 2017). Cet intérêt est lié d'une part, à leur impact potentiel sur la protéine et d'autre part, au fait que les variations faux-sens représentent la classe de variations non-synonymes la plus fréquente dans le génome : 10-15 000 par individu au regard des autres variations non-synonymes (120 pour les gains de STOP et 120 pour les pertes de STOP) ou des variations touchant les sites d'épissage (environ 500) (Eilbeck 2017). On peut rappeler que les variations non-codantes ont également fait l'objet de développements, mais ce n'est qu'après la publication de CADD en 2014 (*Combined Annotation Dependent Depletion*), que de nouvelles approches ont été explorées pour évaluer les variations non-codantes sur d'autres bases que la conservation au cours de l'évolution. Depuis, sont apparus des outils susceptibles d'évaluer l'ensemble des SNV à l'échelle du génome voire, d'outils se consacrant exclusivement aux régions non-codantes (*e.g.* GWAVA). Enfin, comme on le verra dans les sections suivantes, selon le champ d'application (génome complet ou type de variations), on observe une grande hétérogénéité des philosophies, descripteurs, données ou algorithmes composant un prédicteur.

4.3 Philosophies/approches de prédiction

Schématiquement, les approches de prédiction des variations génétiques exploitent essentiellement l'évolution et les propriétés physico-chimiques et structurales des protéines auxquelles peuvent s'ajouter des informations provenant de différents niveaux de connaissances biologiques.

4.3.1 Utilisation de la conservation et de l'évolution

SIFT (Tableau 5) a été le premier logiciel évaluant l'impact d'une variation faux-sens à partir du degré de conservation de l'acide aminé modifié et de sa localisation au sein de la protéine. La conservation est un des **descripteurs** essentiels d'une variation car elle informe sur la tolérance au changement d'une position qui découle de l'évolution humaine [(Lindblad-Toh et al. 2011) ; Figure 26]. Ainsi, de multiples mesures de conservation pour prédire l'impact des variations ont été développées, souvent couplées à la phylogénie (PhastCons, PhyloP, GERP puis GERP++). Cependant, à la différence de SIFT, dédié à la prédiction de l'impact des variations faux-sens au sein des protéines, les autres méthodes citées sont applicables à de nombreuses positions d'un génome *via* l'alignement de régions génomiques d'espèces plus ou moins proches et l'analyse des conservations. Ces méthodes constituent les premiers outils de prédiction du caractère délétère des variations sur l'ensemble du génome, incluant les régions non-codantes.

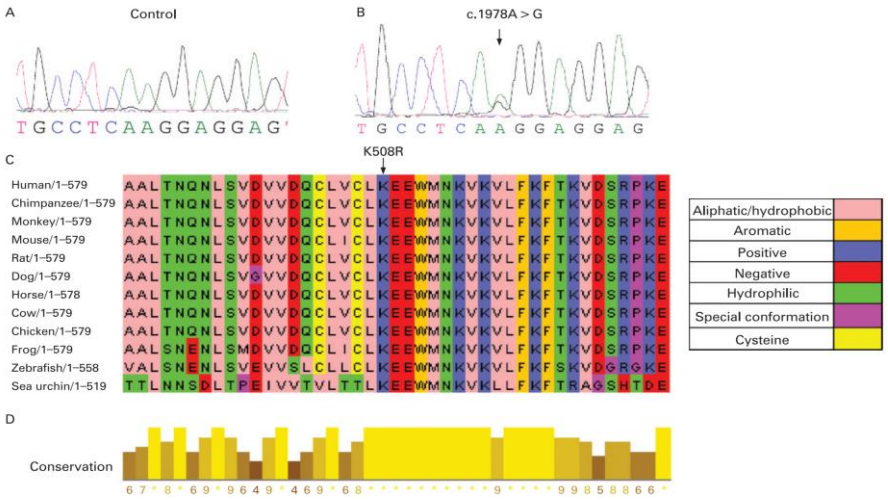


Figure 26 – Illustration de l'utilisation de la conservation en prédiction

A, B – Fragments de chromatogrammes issus du séquençage du gène BHD sauvage (A) et muté (B, variation faux-sens c.1978A>G). C – Alignement multiple des séquences BHD de différentes espèces : la variation faux-sens induit le changement de la Lysine 508, invariante dans toutes les espèces, en Arginine (K508R) D – niveau de conservation (échelle 1-9) et positions invariantes indiquées par un astérisque. Source : (Toro et al. 2008)

4.3.2 Exploitation des propriétés physico-chimiques et structurales des protéines

Une propriété intrinsèque des variations faux-sens est leur impact sur la séquence d'une protéine et sur sa structure tridimensionnelle liée aux contraintes spatiales issues des propriétés physico-chimiques des acides aminés (Ittisoponpisan et al. 2019). Ainsi, pour évaluer l'impact des faux-sens au niveau protéique et structural, des outils se sont développés combinant : prise en compte de la conservation ; calcul du degré de divergence physico-chimique entre l'acide aminé original et l'acide aminé mutant ; évaluation de l'impact structural dans l'environnement 3D du résidu modifié (PolyPhen2, MutationTaster, nsSNPAnalyzer, MAPP, PMUT, SAPRED, SNAP, SNPs3D, MISTIC) (Figure 27).

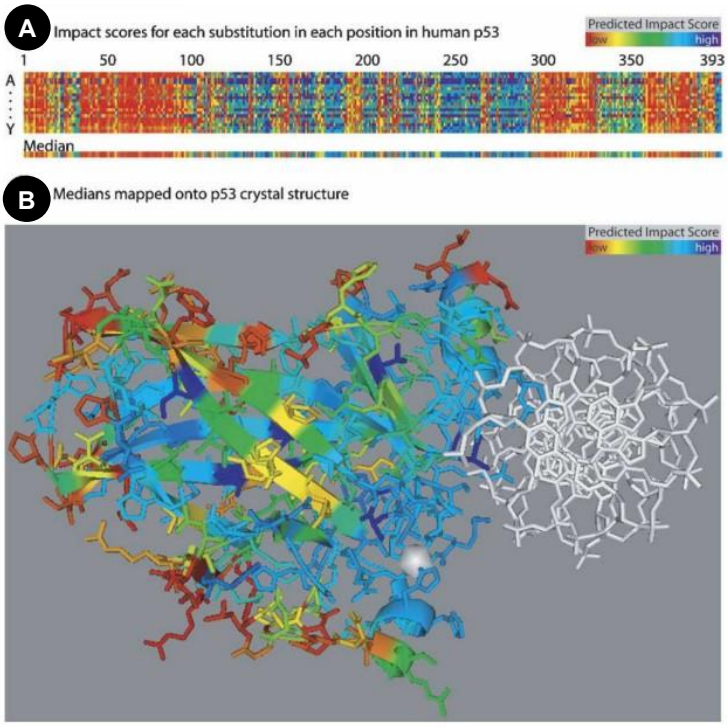


Figure 27 – Scores d'impact de substitutions faux-sens prédits par l'outil MAPP

A – scores d'impact des faux-sens pour chaque position de p53 (gène TP53) et pour les 19 substitutions possibles prédites sur la base des conservations, divergences physico-chimiques et localisations spatiales. On observe la présence de zones (en bleu) où les substitutions à fort impact délétère potentiel sont concentrées
B– visualisation sur la structure 3D de p53. Structures secondaires colorées en fonction des scores d'impact prédits. On retrouve plusieurs zones (en bleu) au cœur de la protéine où les substitutions à fort impact délétère potentiel sont concentrées. Source : (Stone et Sidow 2005)

4.3.3 Agrégation et méta-prédiction

Les technologies NGS ont permis de disposer d'informations sur des centaines de millions de variations génétiques et sur leurs fréquences d'occurrence dans la population générale ou dans les groupes ethniques. Au travers de consortium tel qu'ENCODE (Moore et al. 2020), les NGS ont aussi fourni d'autres types d'informations ayant trait, entre autres, à l'épigénétique, à la structure de la chromatine, aux sites reconnus par des facteurs de transcription ou hypersensibles à la DNAse I, à l'état de modification des histones… (section 1.4).

L'accès à ces informations, utilisées comme des **descripteurs** complétant ceux décrits précédemment, a révolutionné le domaine de la prédiction de l'impact des variations. Ceci a permis le développement d'outils améliorant les performances non seulement au sein des régions codantes (MutationTaster2, REVEL, M-CAP, ClinPred, MetaSVM, MetaLR), mais également dans les régions non codantes (CADD, DANN, GWAVA, GenoNet, Eigen, PrimateAI, DeepSEA, LINSIGHT, ExPecto, NCBoost). Dans ce contexte, on distingue les **prédicteurs** qui agrègent un nombre important de descripteurs des **méta-prédicteurs** qui vont intégrer des descripteurs et des prédicteurs déjà établis devenant ainsi eux-mêmes des descripteurs (Figure 28)

La construction de ces méta-prédicteurs, s'inscrit dans la logique d'amélioration des performances. Cependant, bien que communément utilisée depuis plusieurs années, la juxtaposition de descripteurs (conservation, MAF, outils de prédiction comme SIFT) et d'outils déjà entrainés sur les mêmes descripteurs (PolyPhen-2, CADD) suscite de plus en plus d'interrogations. Il devient très délicat de comprendre la structuration de l'information au sein de ces méta-prédicteurs et d'identifier les problèmes de surapprentissage et de circularité, évoqués dans la section 4.6.3 (Heijl et al. 2020).

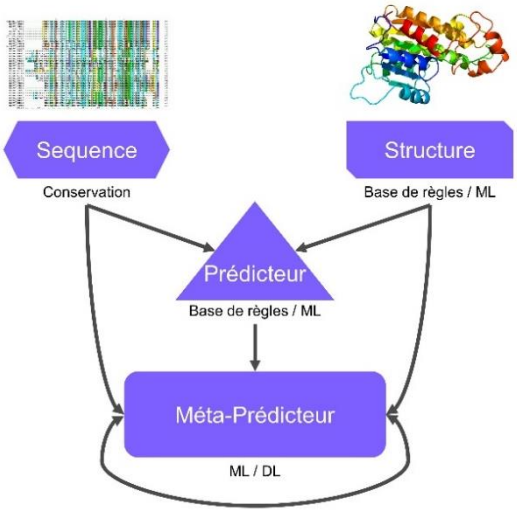


Figure 28 – Les différentes catégories d'outils de prédiction des variations génétiques

L'orientation des flèches symbolisent l'intégration des outils en tant que descripteurs au sein des prédicteurs et méta-prédicteurs

4.4 Catégories de descripteurs employées

Comme nous l'avons introduit précédemment, les descripteurs (le terme anglais original étant *features*) constituent une part essentielle de tout prédicteur. Sans détailler les descripteurs déjà cités, la classification des descripteurs selon leur niveau d'application (gène ou région génique/exon, nucléotide) permet de mieux cibler leur influence potentielle dans les résultats des prédictions fournies.

4.4.1 Descripteurs au niveau du gène et des régions géniques

Au niveau génique, on peut noter l'utilisation de valeurs d'**haploinsuffisance** basées sur un modèle prédictif (Huang et al. 2010) qui reflète la tolérance d'un gène à n'exprimer qu'une copie sauvage suite à la présence d'une variation hétérozygote délétère dans l'autre copie. Il existe également de multiples métriques reflétant la **contrainte** d'un gène ou d'une région selon la densité en variations observées. La contrainte reflète la pression évolutive s'appliquant sur un gène afin de limiter l'apparition des variations. Au niveau du gène, on note

le développement de plusieurs métriques calculant la contrainte [*Residual Variation Intolerance Score* ; RVIS (Petrovski et al. 2013), *probability of being Loss-of-function Intolerant* ; pLI (Lek, Karczewski, et al. 2016), *z-score* (Karczewski et al. 2020)] à partir du nombre observé de variations sur le nombre attendu de variations (selon un modèle statistique intégrant notamment la taille du gène). Il existe également des méthodes basées sur les contraintes locales [*Constrained Coding Regions* ; CCRs (Havrilla et al. 2019) ; section 6.2.3.1], signes d'une pression évolutive liée à des régions d'importance fonctionnelle différente au sein d'une protéine (*e.g.* domaines, sites catalytiques, signaux de localisation cellulaire…). Une autre façon d'intégrer cette notion de contrainte locale est de moyenner la MAF des variations observées au sein d'une fenêtre glissante sur le génome.

4.4.2 Descripteurs au niveau de la variation

Comme on l'a vu précédemment, au niveau de la variation, les principaux descripteurs employés actuellement sont les mesures de conservation inter-espèces et les propriétés physico-chimiques et structurales. De plus, un paramètre essentiel, déjà cité à de nombreuses reprises, est la MAF caractérisant les fréquences de variations observées. Souvent utilisée en tant que filtre univoque pour les variants polymorphiques s'appuyant sur une valeur seuil de fréquence (section 3.3.3), ou retravaillée sous la forme d'une moyenne calculée dans une fenêtre glissante, la MAF a également été employée récemment en tant que descripteur brut non transformé afin de décrire une variation (Alirezaie et al. 2018). Enfin, concernant spécifiquement les variations faux-sens, on peut citer la base de données AAindex, qui répertorie les différences de propriétés physico-chimiques (polarité, volume, point isoélectrique) pour tous couples d'acides aminés, ainsi que les matrices de similarité et cartes de contacts potentiels au sein des protéines (Kawashima et Kanehisa 2000).

Au niveau des régions non-codantes, on peut noter l'utilisation de descripteurs basés sur des données : d'épigénétique, de contacts chromatiniens, sur les sites reconnus par les facteurs de transcription ou d'hypersensibilité à la DNAse I ou encore, sur l'état de modification des histones, accessibles *via* ENCODE (Moore et al. 2020). Enfin, certains méta-prédicteurs dédiés aux sites d'épissage (donneur et accepteur) intègrent les scores d'autres outils ayant pour objectif d'identifier les motifs impliqués dans l'épissage de l'ARN [MaxEntScan (G. Yeo et Burge 2004)] ou récemment, des données d'expression (section 5.2.2).

Comme on le voit, la multiplication des données fait qu'aujourd'hui un grand nombre de paramètres sont utilisés pour caractériser une variation. L'accès à ces données a été simplifié par le développement de bases de données et d'API [dbNSFP (Liu et al. 2020), myvariant.info (Xin et al. 2016), VariantDB (Vandeweyer et al. 2014)] qui agrègent l'ensemble des informations dans un même répertoire (Figure 29). En permettant l'accès aux descripteurs des variations et aux outils qui les intègrent, ces ressources ont facilité leur évaluation et leur comparaison (section 4.6).

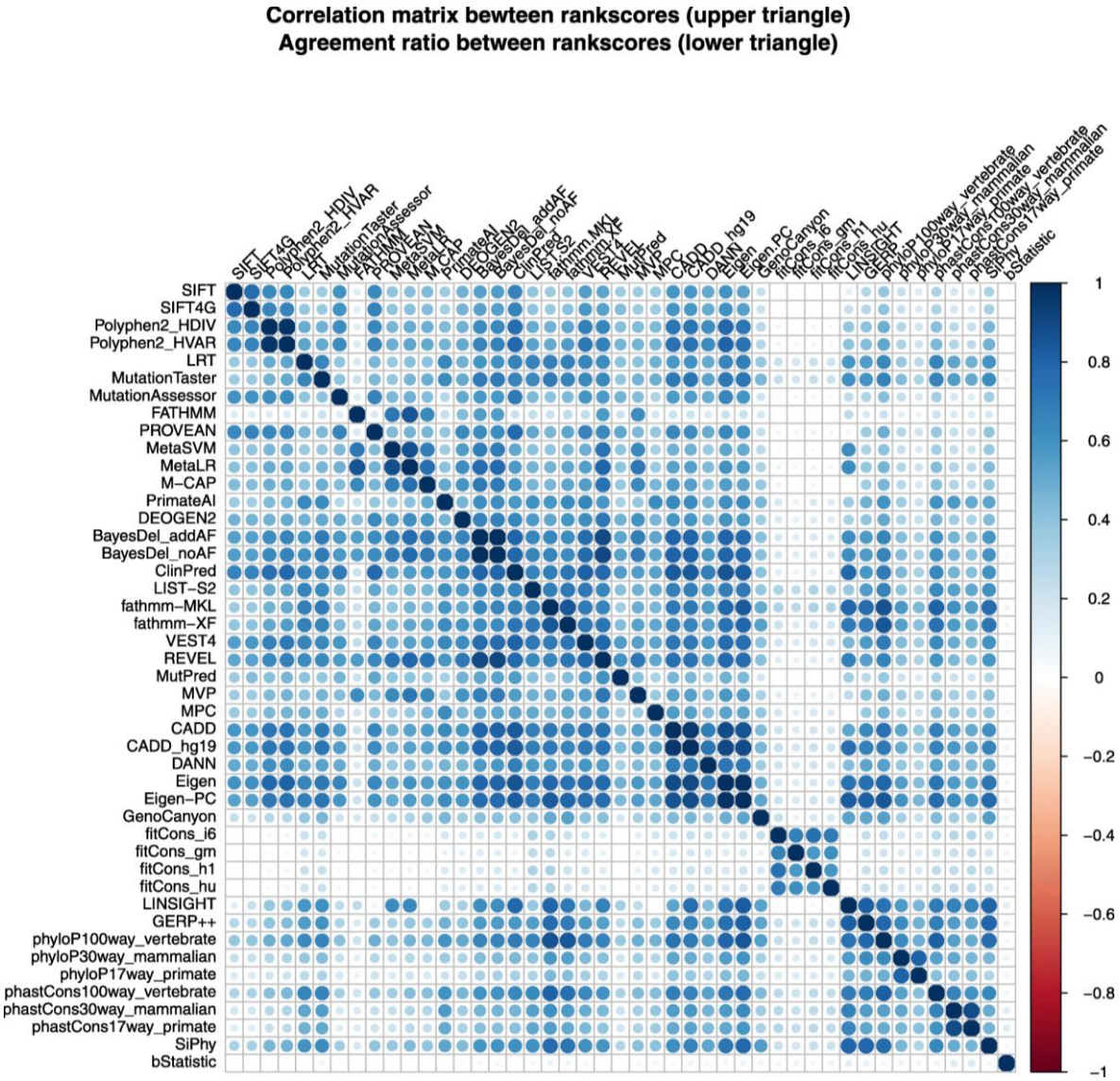


Figure 29 – Matrice de corrélation de descripteurs présents dans la base de données dbNSFP

Selon les auteurs, la partie supérieure de la matrice (triangle supérieur droit) représente la corrélation entre les différents scores fournis par les descripteurs et la partie inférieure (triangle inférieur gauche) l'accord entre les différents scores, c.-à-d. la capacité à fournir la même conclusion quant au statut de la variation (délétère, toléré). Source : (Liu et al. 2020)

4.5 Importance des données d'entrainement

4.5.1 Construction classique d'un jeu d'entrainement

La stratégie la plus commune vise à opposer, un jeu **positif** de variations délétères validées, à un jeu **négatif** de variations issues de la population saine ou validées comme non responsables de maladies (Libbrecht et Noble 2015). Les principales bases de données exploitables dans le cadre du jeu positif sont ClinVar (Landrum et al. 2020) et HGMD (Cooper et Krawczak 1996). Quant au jeu négatif, les prédicteurs intègrent dans certains cas les variations dites « bénignes » de ClinVar ou les variations de population de 1000 génomes, ExAC ou gnomAD. Cependant, une première distinction peut être réalisée entre variations codantes (principalement les faux-sens) et variations non-codantes. En effet, 65% des variations validées comme délétères et associées aux maladies génétiques dans ClinVar sont codantes (38% faux-sens et 27% de non-sens ; Tableau 4). Les 35% restants se distribuent principalement dans les régions introniques et plus particulièrement, au niveau des sites d'épissage. Comme la qualité des modèles prédictifs est liée à la quantité et à la qualité des données utilisées lors de l'entrainement, les prédicteurs dédiés aux régions codantes présentent un niveau de fiabilité nettement plus important.

4.5.2 Équilibrage des jeux d'entrainement

Un aspect important à considérer est la proportion des jeux positifs et négatifs. Certains prédicteurs s'entrainent sur un jeu de données dit « balancé » et contenant une quantité identique de variants délétères et bénins tandis que certains optent pour un entrainement sur un jeu « débalancé » [REVEL ; (Ioannidis et al. 2016)] afin d'augmenter la taille du jeu négatif. Certaines études (Nair et Vihinen 2013) ont pointé l'importance d'équilibrer la proportion de variations en considérant les gènes. En effet, certains gènes, tels que TP53 ou BRCA1 et 2, peuvent se trouver surreprésentés au vu du nombre important de variations délétères associées.

4.5.3 Stratégie alternative de construction d'un jeu d'entrainement pour les régions non-codantes

Comme énoncé en 4.5.1, les variations validées comme ayant un impact délétère incluent peu de variations situées dans les régions non-codantes. Pour contourner ce manque de données, une stratégie alternative développée par le prédicteur CADD (Tableau 5) a consisté à générer des variants *de novo* sur l'ensemble du génome. Ceci a été réalisé à partir : (1) d'un modèle empirique d'évolution de séquences, (2) d'un taux spécifique de dinucléotides CpG, et (3) de taux de mutation mis à l'échelle localement dans des fenêtres de l'ordre de la mégabase. En utilisant ce modèle, 14,7 millions de variations ont pu être simulées, aboutissant

à une grande proportion de variations sans effet délétère et une petite fraction de variations ayant un impact. Ces variations simulées constituent souvent le jeu positif d'entrainement. Le jeu négatif, quant à lui, a été construit sur la base d'une extraction des positions divergentes entre le génome humain de référence et un génome ancestral inféré d'une comparaison génome humain - génome de chimpanzé. Après une étape d'exclusion des variations fréquentes dans la population (MAF > 5% dans 1000 génomes), 14,7 millions de variations composent le jeu final négatif. Dès lors, le modèle de CADD basé sur une machine à support de vecteurs (*Support Vector Machine* ; SVM) a pu être entrainé sur un nombre important de données permettant d'évaluer n'importe quelle variation sur le génome selon un modèle évolutif original.

4.6 Métriques et évaluation

4.6.1 Métriques

Différentes métriques sont disponibles afin d'évaluer les performances des prédicteurs. Sans les énumérer et les décrire de manière exhaustive, on peut néanmoins citer la sensibilité et la spécificité comme briques élémentaires permettant de construire d'autres métriques telles que la précision, le rappel ou le F1-score (Hossin et Sulaiman 2015). Chaque métrique présente une propriété particulière permettant d'évaluer une caractéristique définie.

Si l'on conserve les exemples cités précédemment de jeux positifs associés à des variations délétères et de jeux négatifs associés à des variations bénignes, la sensibilité correspond dans ce cas, au taux de variations délétères correctement identifiées par le prédicteur. La spécificité présente la même propriété, mais concernant le jeu de variations bénin (Figure 30).

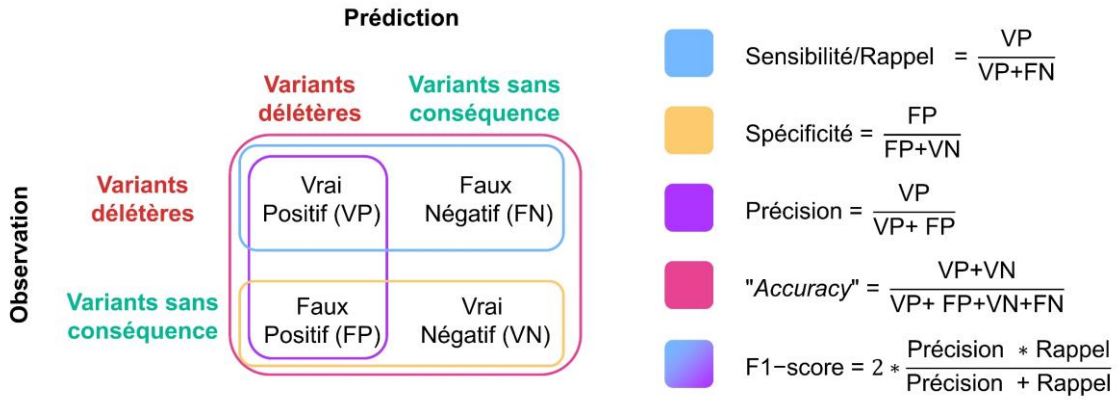


Figure 30 – Métriques d'évaluation des outils de prédiction

Source personnelle avec inspiration de newbiettn.github.io

4.6.2 Optimisation du seuil d'utilisation

Le seuil d'utilisation correspond à la valeur limite permettant de distinguer une variation délétère d'une variation bénigne. Un modèle de prédiction classique attribut un score de probabilité se distribuant de 0 à 1, 0 étant une probabilité nulle d'être une variation délétère et 1 étant la probabilité maximale d'en être une. Le seuil limite standard d'utilisation est de 0.5 (< 0.5 : bénin ; > 0.5 : délétère) comme c'est le cas pour REVEL (Figure 31). Dans un cadre d'analyse clinique, certains prédicteurs cherchent à maximiser leur sensibilité par différents moyens afin de minimiser les risques de non-identification des variations délétères (faux-négatifs). C'est le cas de M-CAP qui a optimisé son seuil d'utilisation. Afin d'obtenir des performances de l'ordre de 95% de sensibilité, M-CAP a abaissé son seuil d'utilisation, améliorant son pouvoir de détection des variations délétères, mais générant également un taux élevé de faux positifs (variations non-délétères prédites comme délétères) comme illustré sur la Figure 31. Une autre stratégie, employée par ClinPred, vise à calculer deux probabilités à l'aide de deux algorithmes distincts et à ne conserver que le score le plus élevé, augmentant ainsi la sensibilité de l'outil.

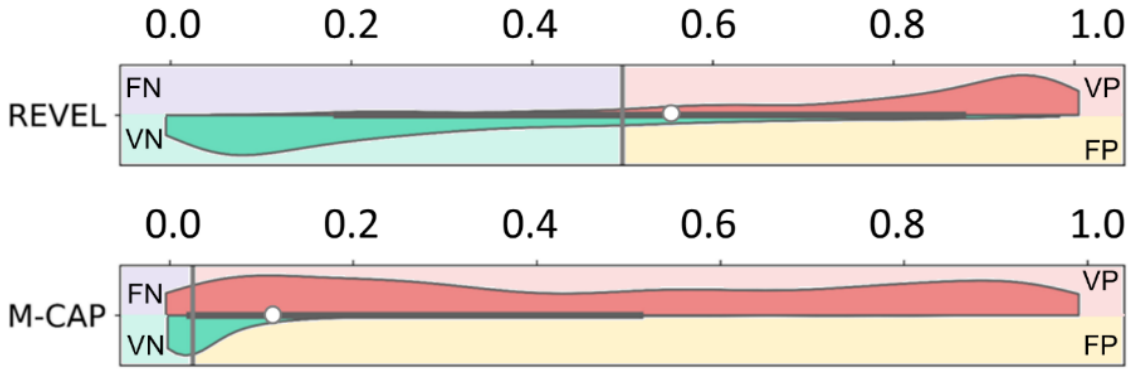


Figure 31 – Distribution des scores de REVEL et M-CAP

La partie haute des « violin plots » représente les scores attribués à l'ensemble d'un jeu d'évaluation de variations délétères (rouge) et la partie basse, les scores attribués à l'ensemble d'un jeu d'évaluation de variations non-délétères de population (vert). La barre grise représente le seuil limite d'utilisation recommandé par les auteurs. Le point blanc représente la médiane de la distribution de l'ensemble des scores attribués aux variations délétères et non-délétères.

4.6.3 *Benchmark* et jeux d'évaluation

Afin d'évaluer les performances d'un prédicteur, on utilise classiquement un sous-jeu d'évaluation extrait des données utilisées pour l'entrainement du modèle ou des jeux d'évaluation externes (*benchmark*) tels que les jeux construits dans le cadre de *Varibench* [(Nair et Vihinen 2013; Grimm et al. 2015) ; (Figure 32)]. Ce jeu vise à évaluer précisément le niveau de circularité des prédicteurs. La circularité, problème récurrent à tous types de prédicteurs, comporte deux niveaux. Le premier niveau de circularité (circularité de type 1) correspond à un biais induit par l'évaluation de variants déjà employés lors de l'entrainement

du prédicteur ou par l'un des prédicteurs intégrés en tant que descripteur dans les méta-prédicteurs. Le deuxième niveau de circularité (circularité de type 2) est lié à l'évaluation de variants présents sur des gènes identiques à ceux utilisés durant l'entrainement. On peut voir sur la Figure 32A que de nombreux outils sont concernés par ce problème, notamment FATHMM qui présente certainement un biais de circularité au vu de la chute importante de ses performances sur le jeu n°5.

Certains outils évaluent leur performance en analysant des exomes ou génomes complets. L'objectif est d'attribuer une probabilité d'être délétère à chaque variation de l'exome/génome, en respectant deux critères. Le premier est de trier par ordre croissant les variations selon leurs probabilités et d'identifier les rangs de variations délétères connues dans la liste ordonnée. Plus la position est proche de 1, plus la fiabilité de l'outil pour identifier correctement une variation causale est élevée. Le second critère est de générer la plus petite liste possible de variations avec un statut prédit comme délétère. Ceci répond au besoin de minimiser le nombre de faux positifs pouvant aboutir à des variations à statut inconnu (*VUS*, section 3.4). Cette méthode peut être réalisée sur des exomes cliniques pour lesquels la variation responsable de la pathologie des individus a été validée expérimentalement ou sur des exomes simulés dans lesquels on a injecté une variation délétère connue.

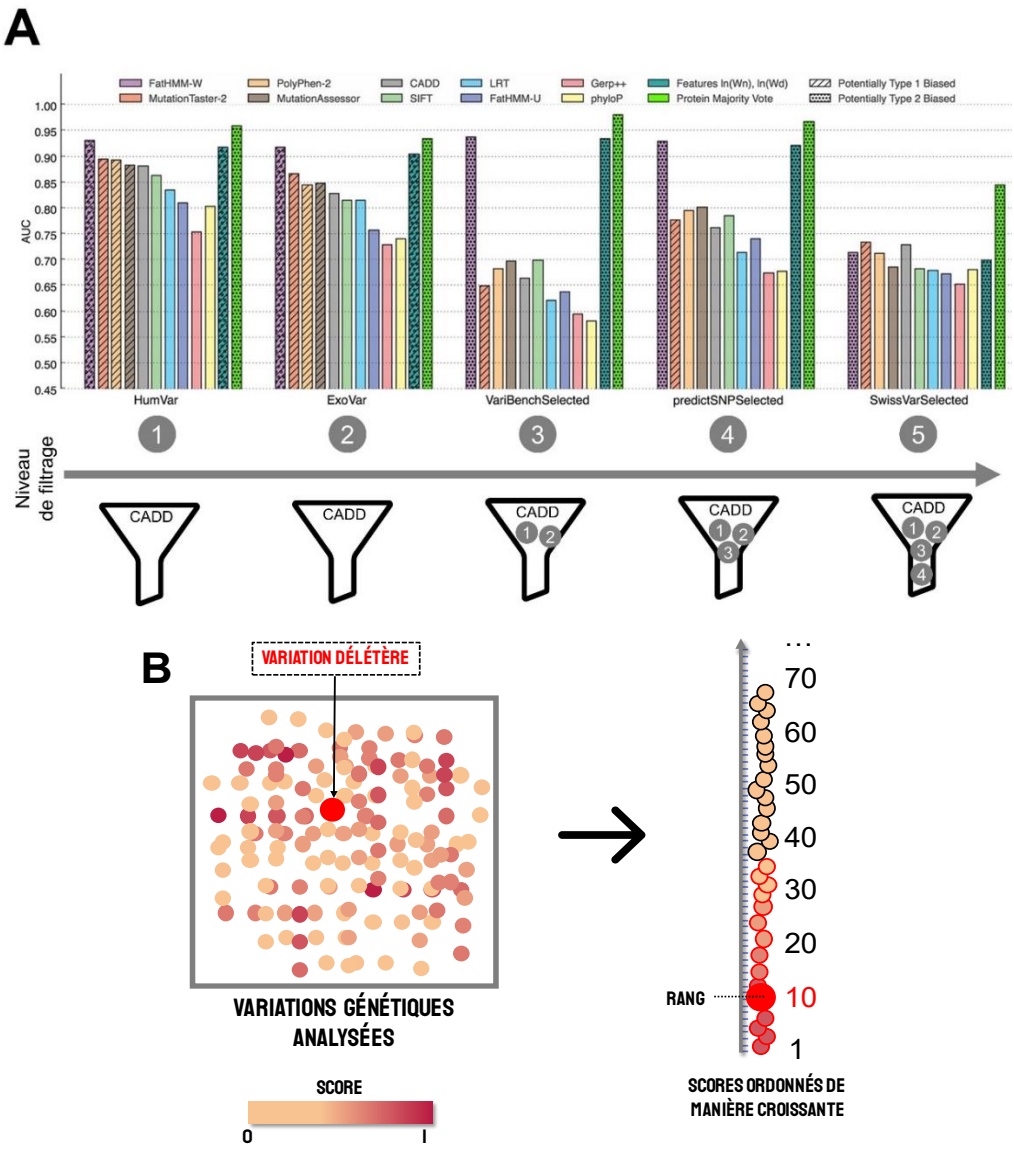


Figure 32 – Deux approches différentes d'évaluation des prédicteurs

A – Varibench est composé de 5 jeux indépendants correspondant à des niveaux de filtrage successifs. Les jeux 1 et 2 sont faciles à évaluer car seuls les variants utilisés lors de l'entrainement de CADD ont été filtrés. Les variants des jeux 3 à 5 ont été filtrés par le jeu d'entrainement de CADD et par l'ensemble des jeux précédemment utilisés comme illustré sur la figure. Ainsi, les jeux 1 et 2 présentent un risque plus élevé de circularité, tandis que le jeu 5 (plus haut niveau de filtrage) représente le jeu le plus « objectif ».
B – Illustration d'une analyse d'exome/ génome par un prédicteur. Sur la partie droite de la figure, les variations entourées par un cercle rouge sont prédites comme délétères par le prédicteur.
Sources : (A) (Grimm et al. 2015) & (B) source personnelle

Chapitre 5. Variations et transcriptome

Les variations génétiques sont classiquement analysées au regard de leur localisation dans les régions fonctionnelles d'un gène ou génome. Cependant, comprendre leurs relations aux maladies génétiques dans le cadre de l'expression des gènes reste un défi majeur.

Dans ce chapitre, je me focaliserai sur l'accès aux données de transcription à haut débit (transcriptome) et sur leur intégration récente dans le cadre de l'analyse des maladies génétiques. Cette intégration a entrainé de nouveaux développements de méthodes d'évaluation et de prédiction des impacts des variations génétiques.

5.1 Transcription et épissage alternatif

Les cellules d'un organisme eucaryotes multi-cellulaire se distinguent selon leur type et fonction particuliers (Kim-Hellmuth et al. 2020) qui découlent de divers mécanismes biologiques complexes à l'œuvre durant la différenciation et le maintien du stade différencié. Au cœur de ces processus, on trouve l'expression spécifique et régulée des gènes et de leurs **isoformes** (différentes formes de protéines issues d'un même gène). À cet égard, il n'est pas inutile de rappeler qu'avant le séquençage du génome humain et l'émergence de l'ère post-génomique, la communauté scientifique a longtemps cru que l'être humain possédait un nombre important de gènes [~120 000 en 2000 ; (Liang et al. 2000)]. Sans entrer dans les débats sur l'anthropocentrisme qui a sans doute guidé cette croyance, on peut concéder qu'elle essayait surtout d'expliquer l'incroyable complexité de la biologie humaine. À présent l'on sait qu'à l'instar de nombreux eucaryotes ou procaryotes, le génome humain ne renferme que 20 000 gènes (International Human Genome Sequencing Consortium 2004). Chez les eucaryotes, un des mécanismes qui peut réconcilier faible nombre de gènes et forte complexité des organismes est l'épissage alternatif.

L'épissage alternatif (Stamm et al. 2005; Modrek et Lee 2002) permet à un gène d'exprimer différents ARNm appelés isoformes de transcrits (*transcript isoforms*), chacune résultant d'une combinaison particulière d'exons donnant lieu, le cas échéant, à une protéine différente. (Figure 33). On sait à présent que l'expression de ces isoformes varie quantitativement et qualitativement dans les différents tissus et stades développementaux d'un organisme (Consortium 2020).

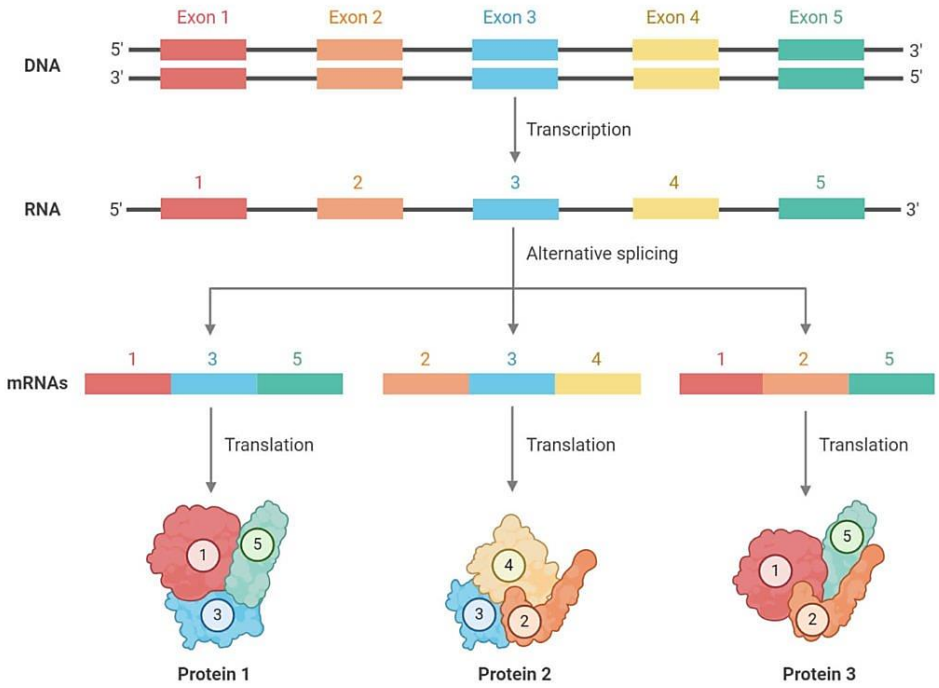


Figure 33 – Étapes de transcription et d'épissage alternatif

Source : (Sapkota 2020)

Bien qu'il a été estimé que la quasi-totalité (~95%) des gènes humains codants pour des protéines était soumis à l'épissage alternatif (Pan et al. 2008), un des résultats de mes travaux de thèse (9.1) montre que ce nombre risque également d'avoir été surévalué et qu'à ce jour, seuls 64% des gènes humains multi-exons codants pour des protéines présentent plusieurs isoformes validées selon des protocoles stricts.

Cet épissage alternatif peut se manifester par différents événements au niveau de l'architecture de l'ARN pré-messager ou hnRNA (*Heterogeneous nuclear RNA*) (Figure 34). Les événements les plus fréquents d'épissage alternatif sont le saut d'exon (*exon skipping* ; ES), la non-intégration d'un exon dans un transcrit, et les exons mutuellement exclusifs (*mutually exclusive exons* ; MXE), variante de l'ES ou deux exons ne sont jamais co-présents dans un même transcrit. On note aussi l'existence de rétention d'intron (*intron retention* ; IR) et de sites d'épissage alternatifs en 5' des sites donneurs ou en 3' des sites accepteurs (Figure 34).

Ces différents événements d'épissage alternatif montrent bien à quel point les notions d'exon et d'intron ne peuvent être considérées de façon univoque et figée. Certains exons, ou régions d'exons, peuvent être épissés et se comporter dès lors comme des introns, à l'inverse, certains introns, ou régions d'introns, peuvent être maintenus et se comporter comme des exons qui participeront à la constitution de l'ARN messager mature.

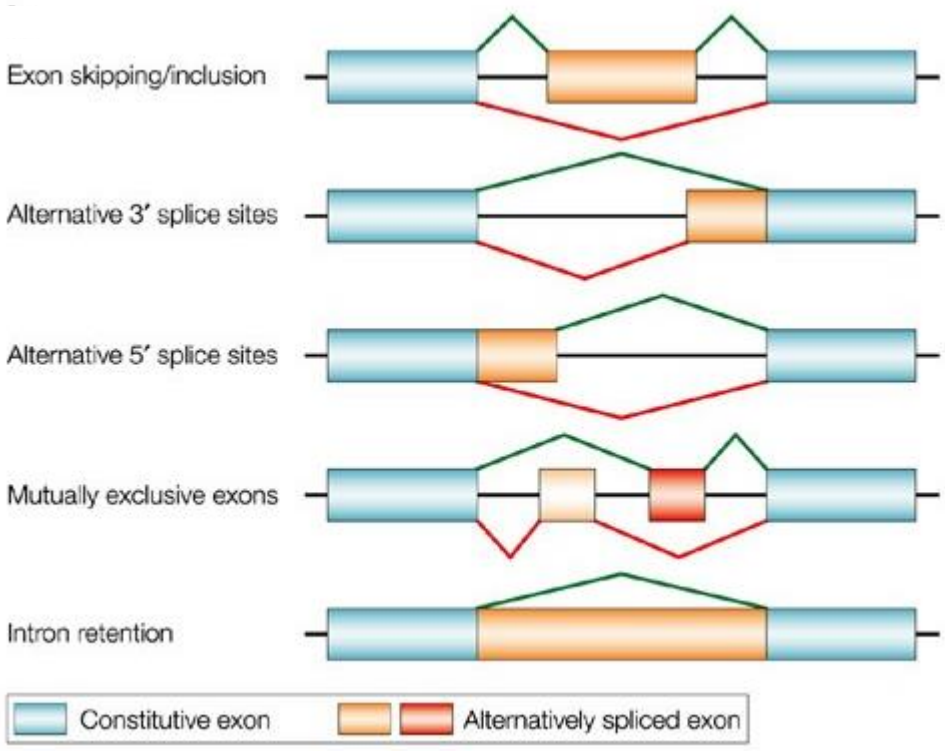


Figure 34 – Les différents types d'épissage alternatifs

Les exons bleus correspondent aux exons constitutifs tandis que les exons jaune et rouge correspondent aux exons alternatifs. Les traits noirs correspondent aux introns. Les traits verts et rouges correspondent à la formation de deux isoformes distincts. Source : (Cartegni, Chew, et Krainer 2002)

5.2 Expression et devenir des variations génétiques dans les transcrits

Deux cas de figure doivent être considérés pour pouvoir évaluer l'impact potentiel d'une variation au niveau des différents transcrits d'un gène. Le premier consiste à déterminer si la variation est localisée dans un exon (ou région d'exon) « **constitutif** », c'est-à-dire, présent dans l'ensemble des transcrits du gène ou dans un exon (ou région d'exon) « **alternatif** », présent seulement dans une partie des transcrits du gène. Le second niveau est d'évaluer les niveaux d'expression des différents transcrits dans les tissus du corps humain. C'est par une analyse fine des niveaux d'expression fournis par un séquençage de type RNA-Seq que l'on peut estimer : 1) la présence/absence d'un exon dans les transcrits et 2) sa fréquence ou niveau « d'utilisation » dans la population des transcrits d'un gène (Figure 35).

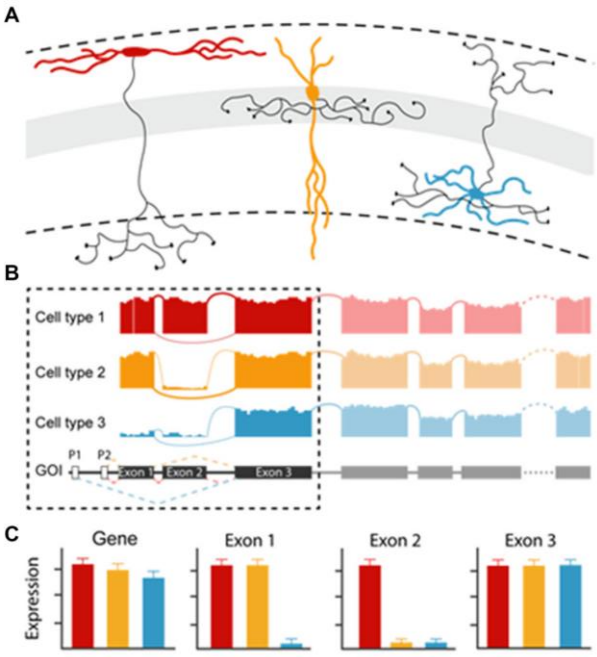


Figure 35 – Niveau d'expression du gène et des exons pour 3 types cellulaires neuronaux

(A) Exemples de différents types cellulaires d'interneurones GABAergic basés sur leur apparence morphologique. (B) Des modifications induites par l'épissage alternatif génèrent différentes isoformes dont les profils d'expression varient fortement. (C) Différences entre l'expression au niveau du gène et expression au niveau des exons selon les types cellulaires. Source : (Que, Winterer, et Földy 2019)

5.2.1 Prise en compte des transcrits dans le cadre des outils de prédiction de l'impact des nsSNV

Dans la logique de considération des SNV au niveau du transcrit, la version 4 de la base de données dbNSFP (section 4.4.2) a été spécifiquement construite afin de pouvoir lister l'ensemble des conséquences des nsSNV selon les différents transcrits codants d'un gène. Ainsi, 11 prédicteurs (ALoFT, DEOGEN2, FATHMM, LIST-S2, MPC, MutationAssessor, MVP, Polyphen2, PROVEAN, SIFT4G, VEST4) présents dans dnNSFP émettent une probabilité différente tenant compte de l'impact d'un nsSNV dans les différents transcrits d'un gène (localisation de la variation dans le transcrit/dans l'isoforme protéique, domaines, …). Pour chacun de ces 11 prédicteurs, la dispersion des scores [(score maximale – score minimale) / moyenne de l'ensemble des scores] a été étudiée pour un même nsSNV au sein des différents transcrits. Cette dispersion se révèle importante et rend compte de la nécessité de sélectionner le transcrit adéquate lors de la prédiction d'un nsSNV.

5.2.2 Intégration des données influençant l'expression des gènes dans les outils de prédiction des variations non-codantes

Dans le cadre des variations non-condantes, des données épigénomiques, obtenues à partir d'un nombre important de types cellulaires, ont été intégrées dans différents outils de prédiction.

Ces données proviennent de projets comme ENCODE et *Roadmap Epigenomics* (Kundaje et al. 2015) et ont été employées dans des prédicteurs tels que : ExPecto, FUN-LDA, GenoNet, GenoSkyline, deltaSVM (Tableau 5). Cette famille de données regroupe des informations obtenues à la suite d'expériences d'épigénomique renseignant sur les marquages d'histones, les sites de fixation des facteurs de transcription ou les profils d'accessibilité à la chromatine. Bien que n'étant pas des données d'expression à proprement parler, l'identification de ces paramètres a été réalisée pour un nombre important de types cellulaires (entre 100 et 220). Dès lors, ces informations permettent d'inférer l'impact transcriptionnel des variations dans les régions non-codantes, selon les différents types cellulaires étudiés.

Enfin, afin de vérifier les prédictions inférées pour les types cellulaires, la plupart des outils évaluent leur performance en comparaison aux eQTL (*expression Quantitative Trait Loci* ; section 2.2.4).

5.3 Ressources et prédicteurs basés sur l'expression des gènes

5.3.1 Ressources pour étudier l'expression des gènes

Dans le sillage de l'avènement du NGS, de nombreuses études ont exploré l'expression des gènes dans un ou plusieurs tissus particuliers. Schématiquement, on peut distinguer d'une part, les analyses réalisées à la manière d'une étude cas-contrôle, comparant un groupe d'individus malades à un groupe d'individus sains et d'autre part, des projets d'envergure afin de réaliser un atlas de l'expression des transcrits à travers différents tissus humains. Dans ce dernier cas, on peut citer, entre autres (Figure 36), le consortium **FANTOM5** (Andersson et al. 2014) se focalisant sur le lien entre éléments régulateurs et expression des gènes, le projet **HPA** pour *Human Protein Atlas* (Uhlén et al. 2015) évaluant les populations de transcrits codants (par RNA-Seq) et de protéines (par immunohistochimie) pour 44 tissus du corps humain ou encore, le consortium **GTEx** pour *Genotype-Tissue Expression* (GTEx Consortium 2015) dédiés à l'étude des transcrits au sein de 54 tissus du corps humain.

Dans le cadre du consortium GTEx, la version 8 du projet permet un accès à l'expression des transcrits de l'ensemble des gènes codants et non-codants humains ainsi qu'aux eQTL et sQTL (section 2.2.4) pour les 54 tissus humains cibles provenant d'environ 1 000 individus (Consortium 2020). Ces informations cruciales permettent une meilleure compréhension du lien entre variations génétiques, expression des gènes et impacts sur les traits phénotypiques humains.

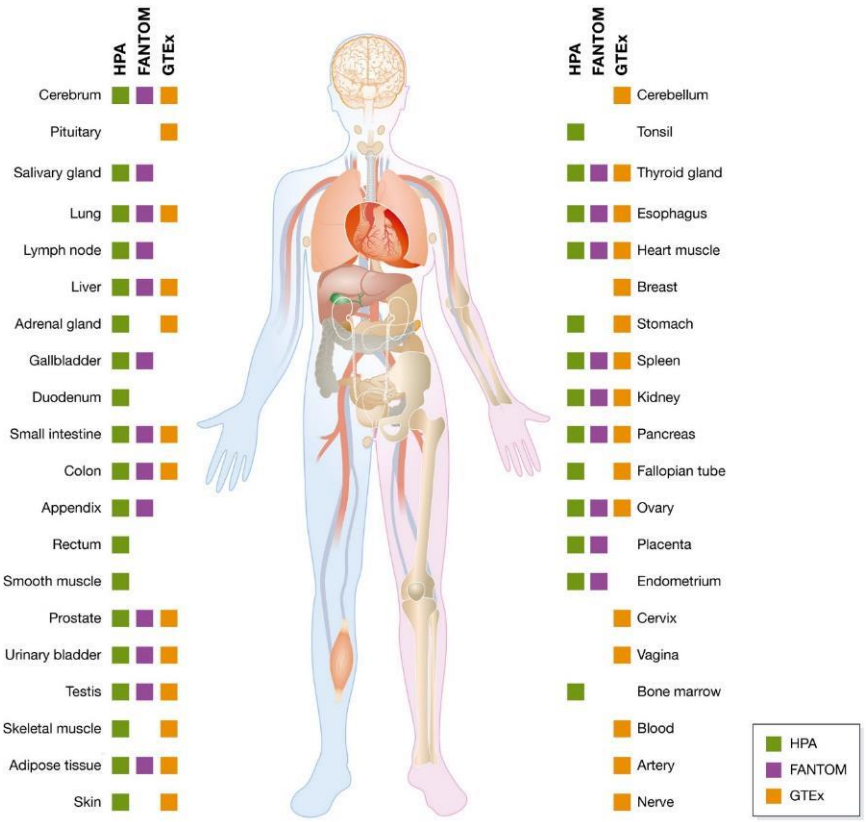


Figure 36 – Ressources transcriptomiques disponibles pour différents tissus du corps humain

Vue d'ensemble des tissus et organes analysés à l'aide de RNA-seq par le consortium Human Protein Atlas (HPA, vert), des tissus étudiés à l'aide de Cap Analysis Gene Expression (CAGE) au sein du consortium FANTOM (violet), et des tissus analysés par RNA-seq par le consortium Genome-based Tissue Expression (GTEx, orange). Au total, 22 tissus ou organes ont été étudiés à la fois avec les jeux de données HPA et FANTOM, tandis que 21 tissus se chevauchent entre les jeux de données HPA et GTEx. Source : (Uhlén et al. 2016)

5.3.2 Normalisation de l'expression au niveau de l'exon

À travers l'exploitation des données d'expression brutes issues du séquençage, des projets tels que GTEx, rendent disponibles des valeurs d'expression au niveau du gène et du transcrit. Celles-ci sont généralement normalisées (en TPM ; *transcrits per million* ; voir méthode de calcul section 6.4.3) afin de rendre comparables les niveaux d'expression entre plusieurs transcrits et plusieurs gènes.

Cependant, après épissage alternatif, un exon peut se retrouver exprimé ou non dans les différents transcrits d'un gène. En exploitant cette propriété, la méthodologie **pext** [*proportion expressed across transcripts ;* (Cummings et al. 2020) *;* section 0] permet d'évaluer le « niveau normalisé d'utilisation » d'un nucléotide dans un tissu donné. Contrairement au niveau d'expression qui n'est pas borné numériquement, la métrique pext se distribue entre 0 et 1, la valeur 1 correspondant à une présence dans l'ensemble des transcrits (exons constitutifs) s'exprimant dans un tissu donné (Figure 37). Cette normalisation facilite la manipulation des valeurs d'expression par nucléotide à travers les différents transcrits d'un gène, pour l'ensemble des 54 tissus analysés dans GTEx.

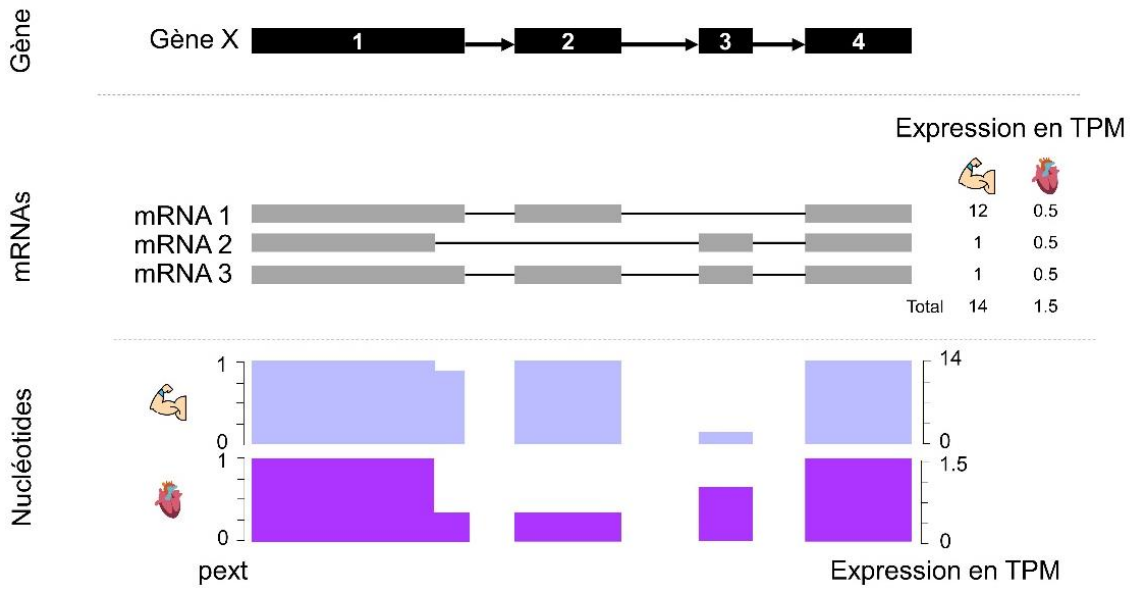


Figure 37 –Distinction des niveaux d'expression des transcrits et des exons d'un gène

Illustration des niveaux d'expression, des transcrits (en TMP) d'un gène (partie supérieure) et des nucléotides (en TPM et en score pext). La valeur pext correspond à une expression normalisée entre 0 et 1 au regard de l'apparition d'un nucléotide dans les transcrits, et du niveau d'expression de chacun de ces transcrits dans les tissus. TPM : transcript per million

5.3.3 Exploitation des données d'expression dans les prédicteurs d'impact des variations touchant les sites d'épissage

À ce jour, les outils exploitant les données d'expression des isoformes de transcrits inter-tissus s'appliquent tous à la prédiction de l'impact des SNV touchant les sites d'épissage. Le premier prédicteur ayant incorporé ces informations d'expression est l'outil SPANR/SPIDEX (Tableau 5Tableau 3). SPANR/SPIDEX intègre un nombre important de descripteurs (près de 1 400), incluant des données de RNA-Seq provenant de 16 tissus du projet *Human Body Map* (NCBI GSE30611), mais également des paramètres de contexte génomique et des éléments régulateurs, afin d'évaluer le pourcentage de transcrits dans lequel chaque exon est épissé pour chacun des différents tissus étudiés. À la suite de ce développement, d'autres outils tels que MMSplice et SpliceAI ont intégré les données d'expression provenant de GTEx. MMSplice a exploité les données de GTEx dans deux de ses modèles dédiés à la prédiction des sites d'épissage alternatifs 5' et 3' (section 5.1). SpliceAI, quant à lui, a utilisé les données provenant de GTEx afin d'améliorer son modèle distinguant exons épissés constitutivement et alternativement. CADD-Splice, variante du méta-prédicteur CADD (février 2021) dédiée à la prédiction de l'impact des variations sur les sites d'épissage, utilise un algorithme basé sur les réseaux de neurones profonds, qui intègre entre autres comme descripteurs les scores de MMSplice, SpliceAI et pext (section 5.3.2).

Enfin, MTSplice, autre outil récent (mars 2021) qui intègre MMSplice parmi ses descripteurs, est aussi le premier prédicteur à intégrer des données d'expression et à prédire l'impact de variations touchant les sites d'épissage de manière tissu-spécifique (dans 53 tissus issus de GTEx et 3 tissus issus de projets de RNA-Seq ciblés sur la rétine) à partir de réseaux de neurones profonds. MTSplice s'est basé sur le modèle MISO (Katz et al. 2010) afin de déterminer les niveaux d'abondance des différentes isoformes à partir des données brutes de GTEx. Le modèle MISO (Mixture-of-Isoforms) est un modèle statistique permettant d'évaluer les niveaux d'expression des exons épissés alternativement à partir de données de transcriptomique. Par l'utilisation de lectures pairées (*paired-end*), celui-ci améliore drastiquement la quantification des événements d'épissage alternatif, mais permet également la détection d'exons et d'isoformes dont l'expression est différentiellement régulée.

5.4 Utilisation de l'expression dans l'analyse des maladies génétiques rares

Bien que le RNA-Seq soit principalement utilisé pour la recherche des expressions différentielles dans le cadre d'études cas-témoins, il peut aussi augmenter le taux de diagnostic de manière significative, en permettant la validation de variations potentiellement reliées à des perturbations du mécanisme d'épissage. Différentes études (Gonorazky et al. 2016; Kremer et al. 2017; Kernohan et al. 2017; Murdock et al. 2021) ont montré l'intérêt croissant de l'usage du RNA-Seq afin de diagnostiquer des patients dont l'étiologie moléculaire n'avait pas été résolue. (Hartley et al. 2020).

Actuellement, le RNA-seq est envisagé comme piste d'amélioration du diagnostic de certaines maladies génétiques rares vérifiant différents critères, notamment des phénotypes bien définis (*e.g.* maladies musculaires ou troubles du dysfonctionnement mitochondrial) et un accès faiblement invasif aux tissus (sang, fibroblaste, peau…). Dès lors, une des limitations de l'emploi du RNA-seq réside dans la spécificité d'expression tissulaire. En effet, comme illustré sur la Figure 38 dans le cadre des maladies musculaires, de nombreux gènes impliqués dans ces maladies sont faiblement exprimés dans le sang ou les fibroblastes, deux tissus aisément accessibles au regard du muscle. Ceci illustre bien que l'analyse RNA-seq de tissus accessibles peut s'avérer insuffisante pour détecter les aberrations transcriptionnelles pertinentes de certains gènes.

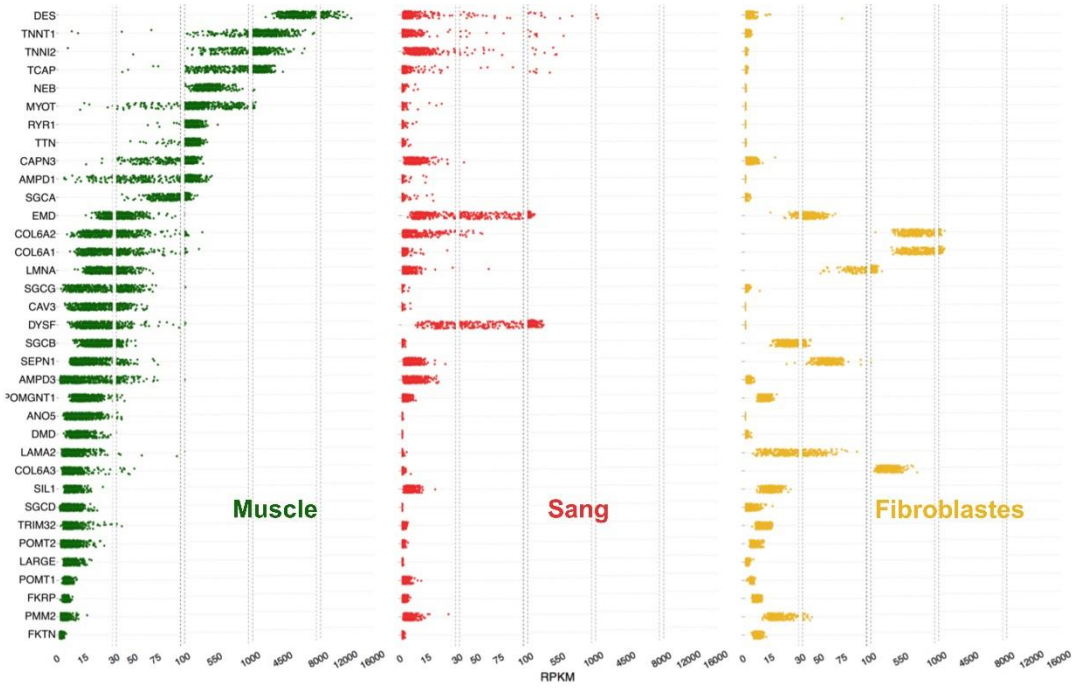


Figure 38 – Expression dans le muscle, le sang et les fibroblastes d'individus sains (GTEx)

L'expression de gènes connus de maladies neuromusculaires, dans 430 échantillons GTEx de muscle squelettique (vert), 393 de sang total (rouge) et 283 de fibroblastes (jaune) montre que ces gènes sont relativement peu exprimés dans les tissus sanguins et les fibroblastes plus facilement accessibles. RPKM : Reads per kilo base per million mapped reads. Source : (Cummings et al. 2017)

Un exemple marquant est l'utilisation de RNA-Seq sur 50 patients atteints de maladies musculaires rares sans diagnostic génétique après analyse par WES et WGS [(Cummings et al. 2017), Figure 39]. En analysant le transcriptome provenant du tissu musculaire de ces patients et en le comparant au transcriptome de patients seins (provenant de GTEx), de multiples variations ont été identifiés dans des régions exoniques ou introniques profondes (en dehors des sites d'épissage), conduisant à une altération importante de l'épissage et à la génération d'isoformes atypiques. Sur la Figure 39D, on peut voir que chez les patients atteints (en rouge), un pseudo-exon a été formé, entrainant l'introduction d'un codon stop prématuré. Ce pseudo-exon est lié d'une part, à la présence constitutive d'un dinucléotide AG (correspondant à un site donneur d'épissage potentiel) et d'autre part, à la présence, chez les patients, d'une variation dans l'intron aboutissant à un site accepteur d'épissage (GC > G**T**).

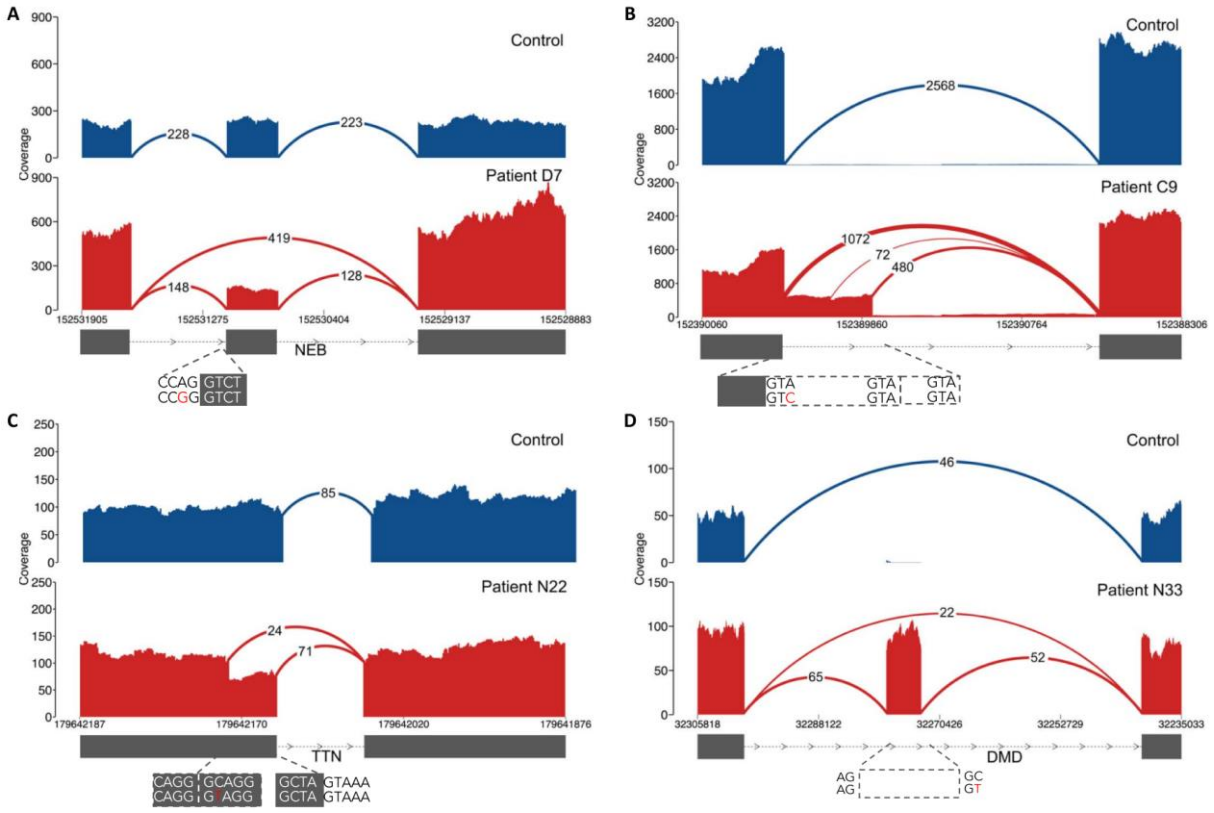


Figure 39 – Identification de variations entrainant des altérations d'épissage

(A) « exon skipping » causé par un variant de site d'épissage essentiel chez le patient D7, (B) extension d'exon (événement 3' alternatif) causée par un variant de site d'épissage donneur (cas particulier de site d'épissage donneur étendu) (C) création de site d'épissage donneur chez le patient N22 avec un contexte de séquence donneur, résultant en un motif d'épissage plus fort que le site d'épissage canonique existant, et (D) création d'un site d'épissage accepteur causé par un variant intronique entrainant une rétention partielle d'intron chez le patient N33. Les aberrations d'épissage (B-D) entrainent l'introduction d'un codon stop prématuré dans le transcrit. Les sashimi plots des figures A à D indiquent la couverture normalisée par nucléotide après assemblage des lectures de RNA-Seq. Les valeurs sur les arcs situés entre exons représentent le nombre de lectures ayant été cartographié sur ces jonctions exon-exon. La partie supérieure (bleue) de chaque figure représente les cas contrôles issus de GTEx et la partie inférieure (rouge) les cas atteints de maladies musculaires.

Source : (Cummings et al. 2017)

Ainsi, le RNA-Seq montre sa capacité à mieux caractériser les variations responsables de pathologies (souvent identifiées par WGS), à travers l'exploitation des données d'expression. L'ajout de ce niveau d'information biologique ouvre la voie à une compréhension plus fine des dysfonctionnements moléculaires ayant lieu chez des patients atteints de maladies génétiques rares, pour lesquels l'étude des variations génétiques au regard du génome seul ne peut suffire. Une des limitations actuelles reste néanmoins l'aspect invasif de l'accès aux tissus impactés.

MATÉRIELS ET MÉTHODES

Durant ma thèse, j'ai développé plusieurs outils bioinformatiques afin d'améliorer l'évaluation de l'impact des variations génétiques. Les ressources utilisées lors de ces développements (banques, bases de données, outils bioinformatiques, outils de programmation) sont présentées dans ce chapitre Matériels et Méthodes.

Chapitre 6. Ressources bioinformatiques

6.1 Banques de référence

6.1.1 RefSeq

RefSeq (*Reference Sequence*) (O'Leary et al. 2016) est une des nombreuses ressources du NCBI (*National Center for Biotechnology Information*). Elle offre l'accès à des séquences bien annotées et sans redondance. Chaque entrée est unique au regard de son contenu et labelisée afin de connaitre l'origine de son intégration dans la base (NP/NM/NC/NG : entrée d'origine manuelle, NT/NW/NZ : origine automatique associée à un séquençage en cours, XM/XR/XP/ZP : prédiction *via* l'annotation de génomes) comme illustré dans le Tableau 7.

Molécule	Identifiant	Commentaire
ADN	AC_	Molécule génomique complète, généralement assemblage alternatif
	NC_	**Molécule génomique complète, généralement assemblage de référence**
	NG_	Région génomique incomplète
	NT_	Contig ou "scaffold", basé sur des clones ou WGS
	NW_	Contig ou "scaffold", principalement WGS
	NZ_	Génomes complets et données WGS inachevés
ARN / ARNm	**NM_**	**Transcrits codant pour des protéines (généralement conservées)**
	XM_	Transcrits issus de modèles prédictifs (codant pour des protéines)
	NR_	**Transcrits non-codants**
	XR_c	Transcrits issus de modèles prédictifs (non-codants)
Protéine	AP_	Annoté sur l'assemblage alternatif AC_
	NP_	**Associé à une accession NM_ ou NC_.**
	YP_	Annoté sur des molécules génomiques sans transcrit instancié
	XP_	Modèle prédit, associé à une entrée XM_.
	WP_	Non redondant à travers plusieurs souches et espèces

Tableau 7 – Types d'accession disponibles dans RefSeq

Les identifiants en gras correspondent aux entrées curées manuellement.
Source : https://www.ncbi.nlm.nih.gov

RefSeq (https://www.ncbi.nlm.nih.gov/refseq/) est accessible *via* un site web ainsi qu'au travers d'un portail FTP (*File Transfert Protocol*) où sont disponibles les fichiers de séquences au

format FASTA et les fichiers d'annotations listant les gènes, transcrits et exons au format GFF (*Gene Feature Format)*.

Dans le cadre de ma thèse, je me suis intéressé aux fichiers d'annotation des gènes humains au format GFF [versions 109.20210514 (GRCh38) et 105.20190906 (GRCh37); Tableau 8] afin d'en extraire les gènes codant pour les protéines, leurs transcrits et les exons qui les composent. Ceci m'a permis d'établir le bilan statistique des exons alternatifs et d'identifier ceux qui composent les gènes MISOG (pour *Multiple transcript ISOform Genes*) (section 9.1.2).

Indice de position	Nom	Description
1	Séquence	Séquence (chromosome, scaffold, contig) de localisation de l'entrée.
2	Source	Source de provenance de d'entrée permettant par exemple d'identifier le niveau de validation (Exemple de sources dans le fichier RefSeq : *BestRefSeq, RefSeq, gnommon*).
3	Élément	Type d'entrée permettant de structurer le fichier à différents niveaux (Exemple : région (correspond à un chromosome ou un patch dans l'assemblage), gène, mRNA, exon (comprenant les UTR), CDS (partie codante des exons). Pour le format GFF3, le type d'élément doit être compatible aux normes publiées par la *Sequence Ontology*.
4	Début	Coordonnée génomique du début de l'élément.
5	Fin	Coordonnée génomique de fin de l'élément.
6	Score	Valeur numérique indiquant généralement la confiance de la source de l'élément annoté, ou son score. Une valeur de "." (un point) est utilisée pour définir une valeur nulle.
7	Brin	Caractère unique qui indique le brin codant de l'élément ; il peut prendre les valeurs "+" : 5 '→ 3', "-" : 3 '→ 5', ou "." : indéterminé.
8	Phase	Phase des éléments de séquence codante (CDS); il peut s'agir de 0, 1, 2 (pour les éléments CDS) ou "." (pour tout le reste).
9	Attributs	Toutes les autres informations relatives à cet élément. Exemple pour un gène : le nom, les identifiants associés dans des références croisées (Ensembl, UniProt, HGNC, OMIM), description, biotype du gène (codant, pseudogène, non-codant…).

Tableau 8 – Champs du format GFF

Source : Wikipédia consulté le 30/08/21

6.1.2 Ensembl

Ensembl (Howe et al. 2021) est une des nombreuses ressources de l'EBI (*European Bioinformatics Institute*), pendant européen du NCBI. Aujourd'hui, Ensembl permet l'accès à une quantité massive de données pour un nombre d'espèces important, dont plus de 50 000

génomes en 2020. Outre le navigateur (https://www.ensembl.org/) permettant d'explorer ces génomes annotés à l'aide de différentes sources d'informations (transcrits, conservation, variations génétiques, éléments de régulation), la ressource permet un accès aux données de génomique comparative (orthologues, paralogues), aux phénotypes associés aux gènes ou encore à leur séquence. De nombreuses références croisées pointant vers d'autres ressources sont également disponibles (HGNC, OMIM, CCDS, UniProt, RefSeq). Enfin, différentes informations sont associées à chacun des transcrits d'un gène permettant de connaitre son niveau d'authenticité (APPRIS, MANE select, TSL) à travers des méthodes informatiques ou expérimentales. Ce critère de certitude relatif au statut des transcrits permet de contrebalancer le nombre important de prédictions incorrectes [13% en 2013 contre 5% pour Refseq, (Nagy et Patthy 2013)].

Ensembl offre un accès aux données *via* le site web, une API (*Application Programming Interface*), un portail FTP et un outil BIOMART (Kinsella et al. 2011) (https://www.ensembl.org/biomart/martview). Durant ma thèse, je me suis servi de BIOMART (Ensembl Genes v103) pour plusieurs taches : extraire les références croisées des gènes humains vers d'autres ressources (*e.g.* HGNC ID, OMIM, RefSeq) ou encore, identifier l'ensemble des in-paralogues (gènes issus d'un événement de duplication au sein d'une même espèce) présents chez l'homme. BIOMART m'a permis de travailler avec des fichiers de sources hétérogènes de manière unifiée.

6.1.3 UCSC *Genome Browser*

L'UCSC (*University of California Santa Cruz*) *Genome Browser* (Navarro Gonzalez et al. 2021) est un outil web (https://genome-euro.ucsc.edu/cgi-bin/hgGateway) permettant une navigation dans le génome de différents organismes. Tout comme Ensembl, l'UCSC *Genome Browser* fait partie des ressources où le génome humain est consultable et exploitable. Il agrège de nombreuses ressources facilitant leur exploitation. Celui-ci permet notamment de comparer les gènes et transcrits présents sur les différents portails précédemment cités (RefSeq, Ensembl), d'afficher les phénotypes, les variants, les annotations associées à ceux-ci (section 3.3), l'expression des gènes, les éléments de régulation et de génomique comparative. Les données sont accessibles à travers le site web, un portail FTP, une API et l'outil UCSC *Table Browser*.

Ce navigateur m'a été utile pour explorer les gènes à plusieurs isoformes et afficher facilement les transcrits et les annotations associées.

6.2 Données associées au gène

6.2.1 Gene Ontology (GO)

Gene Ontology (GO ; http://geneontology.org/) (Mi et al. 2019) découle de l'effort d'un consortium international pour établir une ontologie (vocabulaire contrôlé) décrivant les produits des gènes sur trois plans : les processus biologiques, les composants cellulaires et les fonctions moléculaires. GO est inter-espèces et permet divers usages tels que : l'assignation de fonctions à des domaines protéiques ou la recherche de similarités fonctionnelles entre différentes listes de gènes (*e.g.* différentiellement exprimés ou co-exprimés dans une expérience de transcriptomique). La version actuelle (02/07/21) contient 43 917 termes ainsi que 7 908 721 annotations sur plus de 5 000 espèces. Les données sont accessibles *via* un site web permettant de réaliser des enrichissements en termes GO et *via* des fichiers d'ontologie au format OBO.

Dans le cadre de l'analyse exploratoire réalisée avec la métrique duxt, GO a été utilisée pour calculer les enrichissements sur les trois plans cités précédemment entre gènes à un transcrit codant unique (*Single transcript ISOform genes* ; **SISOG**) et gènes à multiples transcrits codants (*Multiple transcript ISOform genes* ; **MISOG**).

6.2.2 HUGO Gene Nomenclature Committee (HGNC)

La base de données HGNC (*HUGO Gene Nomenclature Committee*) recense les noms uniques des gènes humains, attribués par comité HUGO (Tweedie et al. 2021)]. Le site web (https://www.genenames.org/) référence plus de 33 000 gènes codants et non-codants, accessibles à travers différents services et fichiers téléchargeables.

La ressource HGNC a été utilisée notamment pendant le développement de duxt (Chapitre 9). La liste des 1 345 familles de gènes disponibles sur le site web a également été utilisée durant la comparaison entre MISOG et SISOG (9.1.3.2).

6.2.3 Contrainte & conservation

6.2.3.1 Constrained Coding Region (CCR)

La carte des régions codantes contraintes CCR (*constrained coding regions ;* *https://s3.us-east-2.amazonaws.com/ccrs/ccr.html*) du génome humain est une métrique exploitant la puissance de gnomAD pour identifier les régions appauvries en variants altérant les protéines (non-sens et faux-sens principalement) dans la population. La métaphore employée par les auteurs repose sur le concept du biais de survie, développé durant la Seconde Guerre mondiale par Abraham Wald et le *Statistical Research Group* (SRG). Pour consolider les avions sans poser

un blindage sur toute leur surface, le groupe a cartographié les impacts de balles sur les avions revenus des combats. Par contraste, les zones sans impact devaient correspondre aux avions abattus et nécessiter une protection supplémentaire (Figure 40A). Par analogie, le groupe de Quinlan (Havrilla et al. 2019) a considéré les régions géniques sans variations tronquantes (faux-sens, non-sens) comme intolérantes aux variations. Ces régions contraintes sont appauvries en variations faux-sens et non-sens dans la population saine, mais enrichies en variations délétères dans ClinVar. CCR est disponible sous la forme d'un fichier tabulé ainsi que d'un navigateur web.

CCR est un des différents descripteurs intégrés dans le modèle de MISTIC (Chapitre 8). La métrique a aussi été employée pour évaluer si une des deux classes d'exons constitutifs et alternatifs avait une pression supplémentaire de sélection (section 9.2.2.3) en utilisant les seuils (en percentile) recommandés par les auteurs : (0 - 20 ; 20 - 80 ; 80 - 90 ; 90 - 95 ; 95 - 99 ; 99 - 100).

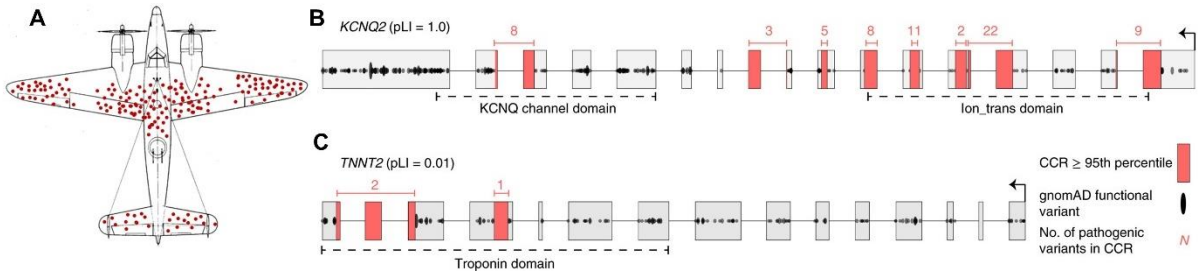


Figure 40 – Carte des régions contraintes (CCR) sur le génome humain

A – Illustration de l'idée pour renforcer le blindage des avions militaires durant la Seconde Guerre mondiale.
B – Cartographie CCR du gène KCNQ2 ayant un grand nombre de régions contraintes dont le percentile est supérieur à 95 (zones en rouge) et de nombreux variants responsables de maladies (chiffres en rouge).
C – Cartographie CCR du gène TNNT2 : faible nombre de régions contraintes et peu de variations délétères.
Source : (Havrilla et al. 2019)

*6.2.3.2 phylogenetic Codon Substitution Frequencies (*phyloCSF*)*

phyloCSF (*phylogenetic Codon Substitution Frequencies*) (Lin, Jungreis, et Kellis 2011) est un outil évaluant si une région génomique est conservée et donc, potentiellement codante en utilisant un alignement nucléotidique multi-espèces. Son originalité réside dans l'évaluation des signatures évolutives caractéristiques au travers, par exemple, des fréquences de substitution plus élevées de codons synonymes (sans changement d'acide aminé) ou d'une faible fréquence de variations faux-sens et non-sens. phyloCSF est utilisable au travers d'un exécutable (https://github.com/mlin/PhyloCSF/wiki).

Dans le cadre de l'étude des propriétés des exons constitutifs et alternatifs (section 9.2.2.3), j'ai utilisé phyloCSF afin d'évaluer si l'une des deux classes présentait un niveau de conservation supérieure. Les exons ayant un score phyloCSF supérieur à 1000 ont été

annotés comme étant hautement conservés tandis que ceux ayant un score inférieur à -100 ont été étiquetés comme peu conservés. Les exons présentant un score phyloCSF intermédiaire (- 100 < score < 1 000) n'ont pas été retenus durant les analyses.

6.3 Ressources biomédicales

6.3.1 Orphanet

Orphanet (Maiella et al. 2013) est une ressource d'origine française, devenue européenne, qui est le portail de référence concernant les maladies rares. Son objectif est de faciliter le diagnostic et le traitement des patients. La ressource répertorie plus de 7 800 liens entre maladies rares et gènes ainsi que 95 000 annotations phénotypiques. Des données d'épidémiologie, de test diagnostique, d'expérimentation clinique sont également disponibles. Orphanet (https://www.orpha.net/) a développé une nomenclature afin de mieux référencer les maladies, phénotypes et gènes associés et faciliter l'interopérabilité et l'intégration dans des systèmes d'information. Une ontologie appelée ORDO (*Orphanet Rare Diseases Ontology*) a également été mise en place pour développer un vocabulaire structuré sur les maladies rares. Orphanet met à disposition ses données au format XML *via* un site web et un portail (Orphadata).

Dans le cadre de la comparaison entre MISOG et SISOG (9.1.3.2), la ressource Orphanet a été utilisée afin d'extraire de la nomenclature, les différentes classes de maladies associées à chaque gène (*rare cardiac disease*, *rare neurologic disease* …).

6.3.2 *Human Phenotype Ontology* (HPO)

HPO (https://hpo.jax.org/) pour *Human Phenotype Ontology* (Köhler et al. 2021) est une ontologie dédiée aux variations phénotypiques du corps et à ses associations à des maladies. HPO contient actuellement 13 000 termes ordonnés dans un graphe dirigé acyclique où chaque terme parent représente un terme de spécificité moindre (ex : *hypertrophic cardiomyopathy* < *cardiomyopathy* < *abnormal myocardium morphology*). Le portail est accessible à travers différents concepts : gène, maladie ou phénotype. Chaque terme se caractérise par un identifiant, une définition, des synonymes et des références croisées vers d'autres banques (Orphanet, Pubmed, OMIM). Chaque entrée est reliée dans l'ontologie aux différents concepts s'y rattachant (ex : un phénotype est associé à plusieurs maladies et à plusieurs gènes). Les données sont accessibles *via* le site web et par téléchargement sous forme d'un fichier OBO (*Open Biomedical Ontologies*) ainsi que différents fichiers d'annotation au format TSV.

6.3.3 *Online Mendelian Inheritance in Man* (OMIM)

OMIM (https://omim.org/) (Amberger et al. 2019) est un catalogue listant les gènes impliqués dans les maladies génétiques humaines. Les informations référencées sont issues d'une revue de la littérature scientifique par des experts. Tout comme HPO, chaque entrée est liée à un concept unique (gène ou maladie). Dans le cadre d'un gène, on retrouve les différentes maladies liées à celui-ci ainsi qu'un tableau représentant un synopsis clinique attaché aux pathologies. Ce tableau est organisé par partie corporelle (H*ead & Neck, Cardiovascular, Respiratory, Muscle & soft tissues)*. Des informations sur la fonction du gène, les modes d'hérédité, les variants et leurs implications dans les phénotypes ainsi que les publications associées sont listées (Figure 41). Toutes ces données sont accessibles en utilisant une API (inscription soumise à réglementation).

Dans le cadre du développement de duxt (9.2), j'ai exploité OMIM afin d'en extraire l'ensemble des gènes présentant des particularités phénotypiques.

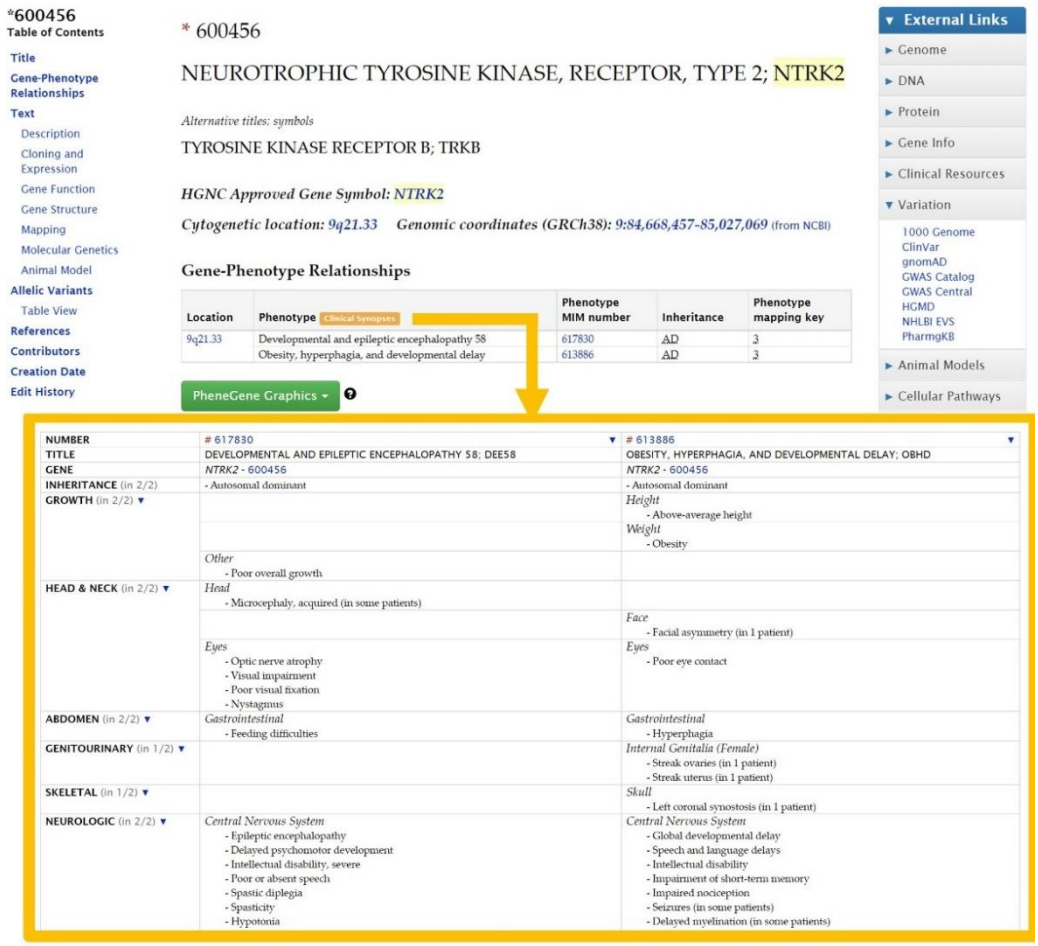


Figure 41 – Interface web de la ressource OMIM

Exemple pour le gène NTRK2. Pour chaque pathologie, un tableau clinique réalisé par un expert résume les phénotypes des patients étudiés pour différentes parties du corps humain (cœur, muscle, appareils génitaux …)
Source : capture d'écran du site omim.org, consulté le 30/09/21

6.3.4 ClinVar

ClinVar (Landrum et al. 2020) (https://www.ncbi.nlm.nih.gov/clinvar/), tout comme RefSeq est une ressource du NCBI. Créée en 2013, elle représente l'archive publique du NCBI des relations cliniques entre gènes, phénotypes et variations génétiques. ClinVar est la source de données publique la plus complète sur les variations associées à des maladies génétiques humaines. On dénombre (04/08/21) 970 183 entrées distinctes observées dans le cadre clinique, dont 136 615 avec un statut délétère ou suspecté comme tel (*clinical significance* : *Pathogenic/Likely_pathogenic*). Plusieurs niveaux d'information existent, notamment sur le statut d'interprétation clinique : variants à conséquence inconnue (VUS ; voir 3.4), bénins ou pathogènes contribuant à une maladie. On trouve aussi un statut de validation par variation (validé par un comité, par une seule personne/entité, avis conflictuels…), associé à un score de 0 à 4. La conséquence moléculaire est également disponible (faux-sens, non-sens, synonyme …). Enfin, on trouve des références croisées vers d'autres banques (OMIM, Orphanet, HPO) et vers les publications décrivant la variation. Les données sont accessibles *via* le site web et par portail FTP. Deux formats de fichiers sont disponibles, un format tabulé standard ainsi qu'un format VCF (*Variant Call Format*) au cœur de nombreuses ressources génomiques listant les variations génétiques (Tableau 9). Ce format comprend deux parties distinctes : un entête (dont chaque ligne est marquée par un double #) listant les champs référencés dans le fichier et un tableau à 9 colonnes (minimum) référençant les variations, leur position et divers champs informatifs.

Numéro colonne	***Nom***	***Description***
1	CHROM	Nom de la séquence/chromosome portant la variation détectée.
2	POS	Position de la variation dans la séquence CHROM
3	ID	Identifiant de la variation (*e.g.* identifiant dbSNP, ClinVar, HGMD…)
4	REF	Nucléotide dans la séquence de référence à la position POS
5	ALT	Nucléotide alternatif dans la séquence (liste si plusieurs allèles alternatifs)
6	QUAL	Score de qualité associé à la détection de la variation
7	FILTER	Label indiquant si la variation vérifie certains filtres lors du séquençage
8	INFO	Champ répertoriant les annotations associées à la variation (fréquence allélique, nombre d'homozygotes, signification clinique dans ClinVar/HGMD, score de prédiction, de conservation …). Les informations sont répertoriées par des clés : clé1=valeur1; clé2=valeur2.
9	FORMAT	Champ optionnel permettant de décrire les échantillons ou les individus (génotype, profondeur de séquençage, …).
+	SAMPLEs	Colonne supplémentaire pour chaque champ du champ FORMAT

Tableau 9 – Champs du format VCF (Variant Call Format)

Source : Wikipédia consulté le 30/08/21

ClinVar a été largement utilisée dans le cadre de ma thèse, notamment durant les phases d'entrainement et d'évaluation de MISTIC (Chapitre 8). La ressource a également été mise à contribution afin de répertorier les variations pathogènes présentes dans les exons alternatifs différentiellement exprimés identifiés par duxt (9.2).

6.3.5 *Human Gene Mutation Database* (HGMD)

La base de données HGMD (*Human Gene Mutation Database*) est la première base de données (1996) à répertorier les variations génétiques observées en clinique (Cooper et Krawczak 1996). Elle est disponible sous deux formes : une forme gratuite, téléchargeable depuis le site web (http://www.hgmd.cf.ac.uk/), mais avec un nombre limité d'entrées (210 341 dans la version 2021.2) et une forme commerciale sous licence (323 661 variations dans la version 2021.2). HGMD a été utilisé durant le développement de MISTIC afin de construire le jeu de variations positif constitué de variations délétères associées à des maladies génétiques.

6.3.6 *Database of Curated Mutations in cancer* (DoCM)

DoCM (*Database of Curated Mutations in cancer*) (Ainscough et al. 2016) est une base de données de variations génétiques délétères. Elle agrège, stocke et permet un suivi des variations impliquées dans les cancers à partir d'observations issues de la littérature. Elle se distingue de ClinVar par la nécessité d'une publication référençant une ou plusieurs variations avant inclusion dans la base, expliquant le nombre limité de variations répertoriées (inférieur à 1400).

DoCM (http://docm.info/) a été utilisée afin d'évaluer la capacité de MISTIC à prédire la pathogénicité de variations provenant d'une source externe à ClinVar et HGMD.

6.4 Bases de données de cohortes / génomique et transcriptomique

6.4.1 1000 Genomes (1000G)

Le projet 1000 génomes (https://www.internationalgenome.org/) (Auton et al. 2015) a démarré en janvier 2008 et constitue le premier projet NGS international visant à établir le catalogue le plus détaillé possible des variations génétiques humaines. La phase 1 a regroupé l'analyse de 1 092 individus et plus récemment, de 2 504 personnes de différentes régions du globe (phase 3). Globalement, 88 millions de SNV, 3,6 millions d'indels et 60 000 SV ont été identifiés au terme de cette troisième phase. L'une des limites de l'exploitation des données réside en l'utilisation d'une faible couverture de séquençage (moyenne de 8X) sur l'ensemble de la cohorte.

Lors de l'évaluation de MISTIC (Chapitre 8), nous avons exploité les fichiers bruts de variations provenant des 1 092 individus de la phase 1 de 1000 génomes afin de simuler des exomes synthétiques de malades (section 4.6.3). Pour ce faire, une variation faux-sens pathogène choisie aléatoirement a été insérée dans un exome « sain ». Ces exomes synthétiques ont servi à évaluer la capacité de MISTIC à identifier la variation « causale » (insérée) comme étant la plus délétère parmi tous les variants faux-sens d'un exome.

6.4.2 gnomAD

La ressource gnomAD (https://gnomad.broadinstitute.org/) constitue la plus grande ressource de génétique des populations humaines. Initié par Daniel G. MacArthur, gnomAD est la continuité du projet ESP (*Exome Sequencing Project*) regroupant 6 503 exomes d'individus « sains » distincts et du projet ExAC (*Exome Aggregation Consortium*), rassemblant 60 706 exomes pour un total de 10 millions de SNV découverts (Lek, Karczewski, et al. 2016). En basculant de l'exome au génome, le consortium a changé de nom (gnomAD ; *genome Aggregation Database*). Ainsi, la seconde version de la banque (Karczewski et al. 2020) regroupe 125 748 exomes (16 millions de SNV) et 15 708 génomes (229 millions de SNV). La version actuelle (v3.1 sortie fin 2020) est composée de 76 156 génomes pour un total de 760 millions variants identifiés à travers plus de 60 populations distinctes du globe (Figure 42).

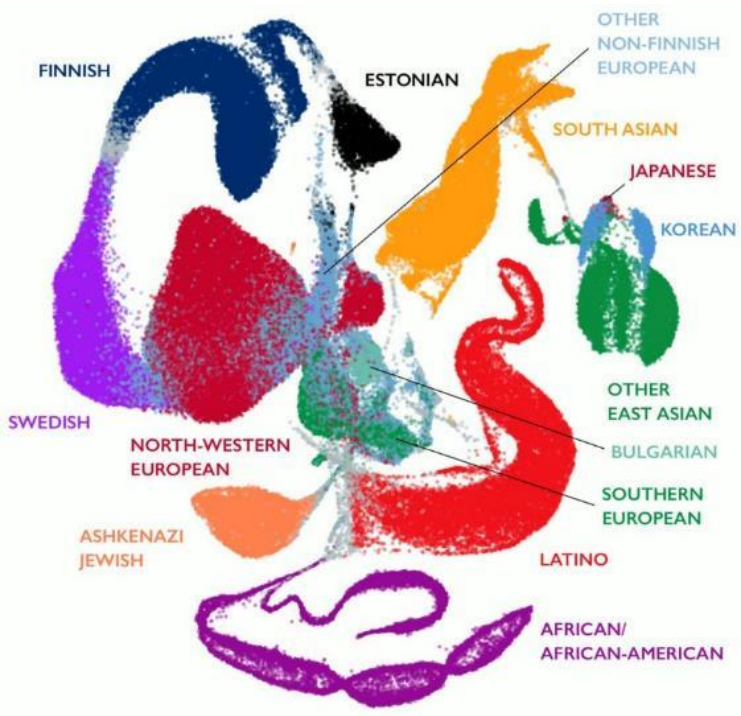


Figure 42 – Carte UMAP décrivant la diversité des individus présents dans gnomAD

L'UMAP (Uniform Manifold Approximation and Projection) est une approximation bidimensionnelle obtenue à partir de 7 composantes principales. L'échelle ne représente pas la distance génétique séparant les populations. Source : (Karczewski et al. 2020)

L'accès à la ressource web (Figure 43) se fait à 3 niveaux possibles : le gène, le transcrit ou le variant (avec une imbrication gène ◄ transcrits ◄ variations). Pour chaque gène, sa « carte d'identité » (abréviations, noms, locus) est disponible avec des références croisées vers d'autres bases ainsi que de multiples paramètres (contraintes, couverture par nucléotide). Chaque transcrit issu d'Ensembl est représenté ainsi que l'expression rationalisée (à travers les transcrits) pour chaque nucléotide (pext : section 6.4.3). Cette banque de données, tout comme son navigateur, a été un élément central lors des développements de MISTIC et de duxt.

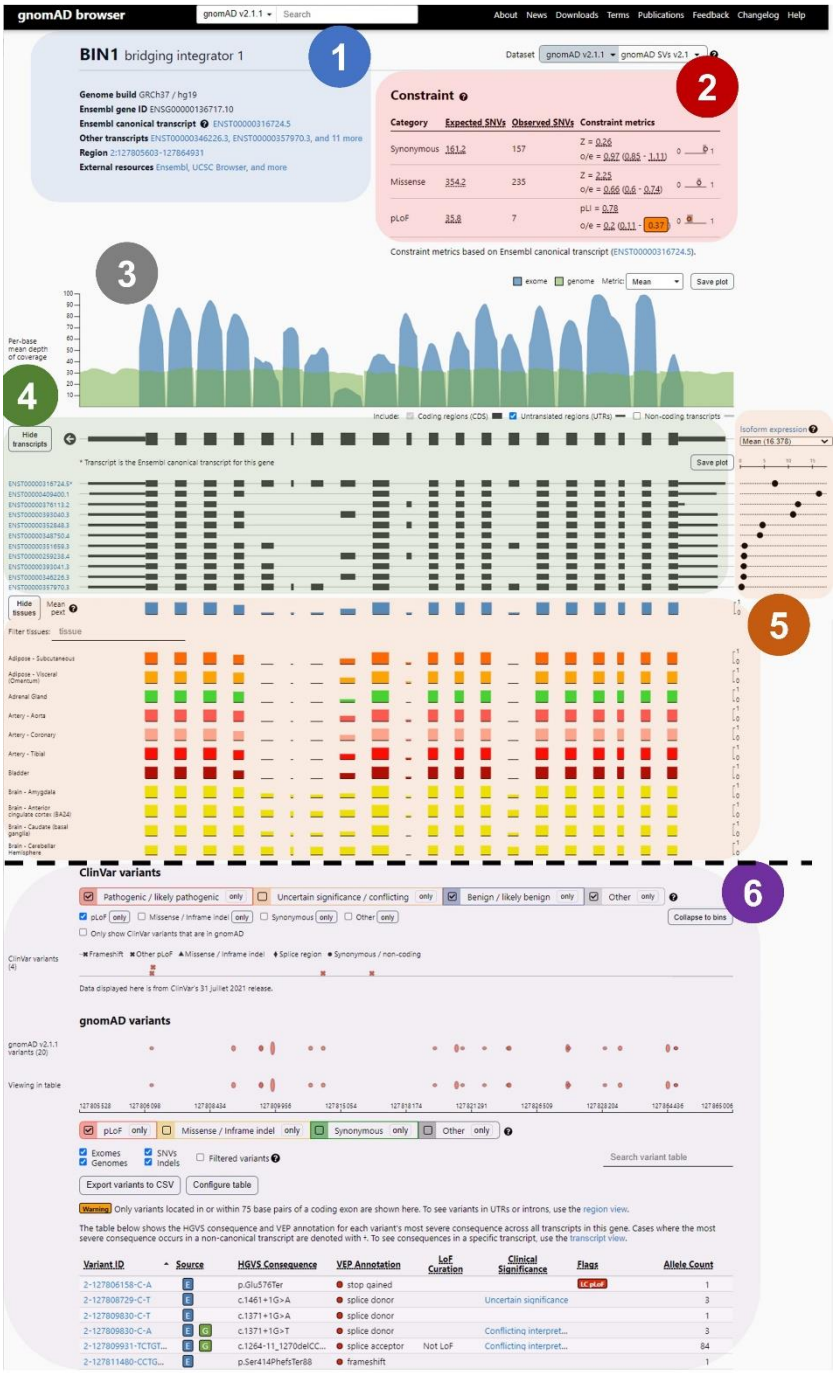


Figure 43 – Capture d'écran du gnomAD browser

1 – Informations au niveau du gène et des transcrits (identifiants, localisation)
2 – Mesures de contrainte calculées pour les variants synonymes, faux-sens et pLoF (putative loss of function ; corresponds aux non-sens et variations touchant les sites d'épissage)
3 – Couverture Exome (bleu) et Génome (vert)
4 – Carte de la structure du gène et de l'ensemble des transcrits disponibles sur Ensembl
5 – Expression (en TPM) des transcrits (en haut à droite) et expression normalisée des nucléotides/exons (pext) par tissu (53 tissus en tout)
6 – Variants recensés dans ClinVar et identifiés dans les individus 'sains' de gnomAD
Source : capture d'écran du site https://gnomad.broadinstitute.org avec annotation personnelle

6.4.3 *Genotype-Tissue Expression project* (GTEx)

À l'instar de gnomAD, le GTEx *portal* (https://www.gtexportal.org/) (Consortium 2020) représente un effort de coopération international en transcriptomique. Son but est de développer une ressource publique exhaustive dédiée à l'étude de l'expression et de la régulation des gènes dans les tissus humains. GTEx fournit des données inégalées d'expression des gènes humains ouvrant une nouvelle dimension des « Big Data » biomédicales. Une variation n'est plus considérée à travers une image figée du gène, mais *via* un profil dynamique dépendant des tissus dans lesquels elle s'exprime.

La version actuelle (v8) agrège les informations d'expression de 948 donneurs dans 54 tissus pour un total de 17 383 réplicats détaillés selon les tissus (Figure 44A). Une interface web permet d'accéder au niveau d'expression brute par exon, par jonction d'exon et par isoforme, ainsi qu'au profil d'expression normalisé du gène (en *Transcripts Per Million* ; TPM).

Le TPM est une méthode de normalisation utilisée en RNA-seq. Elle peut s'interpréter de la manière suivante : 10 TPM correspondent à 10 molécules du transcrit T sur 1 000 000 de molécules dans l'échantillon.

La normalisation est réalisée selon le protocole suivant (Équation 1) :

Méthode de calcul de l'expression normalisée en TPM

Pour chaque **gène G** présent dans l'échantillon :

Pour chaque **transcrit T** (du **gène G**) :

ratio A = Nb de lectures / longueur du transcrit (bp)

$$\text{Facteur F} = \frac{\sum_{1}^{T} \text{ratio A}}{1\,000\,000}$$

Pour chaque **transcrit T** (du **gène G**) :

$$\text{TPM}_{\text{Transcrit}} = \frac{\text{ratio A}}{\text{Facteur F}}$$

$$\text{TPM}_{\text{Gène}} = \sum_{1}^{T} \text{TPM}_{\text{Transcrit}}$$

Équation 1 – Méthode de calcul de l'expression normalisée en TPM (Transcripts Per Million)

Les génomes et transcriptomes de 838[2] donneurs ont été comparés afin, entre autres, d'identifier des traits quantitatifs associés à loci ou QTL (*Quantitative Trait Loci ;* Figure 44B). Ces QTL étudiés par les auteurs correspondent aux variations génétiques associées à une modification de l'expression de gènes chez un nombre statistiquement valide d'individus. Deux types de corrélation variation - altération de l'expression ont été ciblés : variations corrélant avec une altération de l'expression de certains transcrits (*expression QTL* ; eQTL) et variations corrélant avec une modification des populations de transcrits (*splicing QTL* ; sQTL) (section 2.2.4).

Outre le site web autorisant la consultation des données, GTEx offre un accès aux données brutes, prétraitées et normalisées *via* une page dédiée. Celles-ci sont essentiellement représentées par des matrices à deux dimensions de taille importantes [ex : 17 383 colonnes (une par réplica) pour près de 200 000 lignes (une par transcrit)] disponibles dans un format textuel compressé ou *Apache parquet*.

Dans le cadre du développement de duxt, j'ai exploité les données brutes provenant de GTEx au niveau des transcrits afin de filtrer ceux ne respectant pas les critères recommandés par le consortium (transcrit présent dans au moins 20% des échantillons et vérifiant à la fois une expression normalisée ≥ 0.1 TPM et une couverture ≥ 6 lectures). Ceci élimine les transcrits issus du bruit de fond inhérent aux expériences RNA-seq ou aux traitements logiciels. Les fichiers utilisés durant le développement de duxt sont les fichiers de comptage brut des lectures (données accessibles à cette adresse) et les valeurs d'expression normalisée en TPM (accessibles à cette adresse).

À l'heure où j'écris ces lignes, la version 9 de GTEx vient d'être publiée. Elle ne contient pas de nouveaux individus, mais une ré-exploitation d'échantillons disponibles par un séquençage utilisant la technologie du *single-nucleus* RNA-seq (séquençage du contenu ARN d'un noyau unique au lieu de l'ARN de l'ensemble de la cellule). Huit tissus ont été profilés afin d'identifier 43 types cellulaires y figurant (Figure 44C). Cette expérience préfigure le futur où le séquençage de cellule/noyau unique permettra de cibler dans un tissu, le/les types cellulaires dans lesquelles s'expriment une variation.

[2] seuls 838 génomes des 948 donneurs ont été séquencés

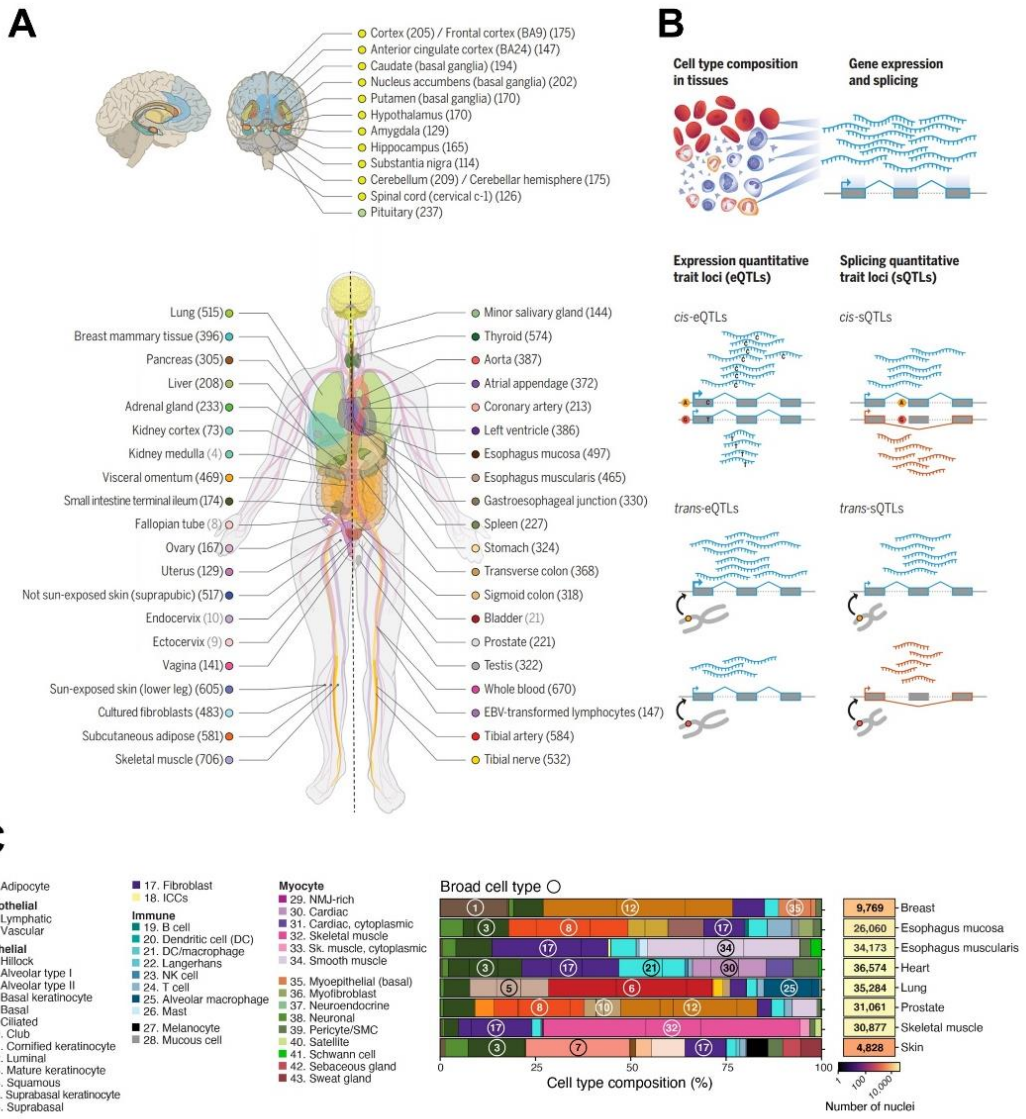


Figure 44 – Vue d'ensemble des composantes de GTEx

A – Présentation des 54 tissus prélevés et séquencés post-portem chez 948 donneurs
B – Objectifs du projet GTEx : évaluation de la composition en types cellulaires des tissus, détermination de l'expression des gènes et isoformes et identification des variations associées à un changement d'expression (eQTL) et/ou de population d'isoformes (sQTL) à courte (cis) ou longue (trans) distance
C – GTEx v9 : évaluation de 43 types cellulaires et de l'expression des gènes et isoformes pour 8 tissus (Breast, Esophagus mucosa, Esophagus muscularis, Heart, Lung, Prostate, Skeletal muscle et Skin) via snRNA-seq.
Source : (Consortium 2020)

6.4.4 *Proportion expressed across transcripts* (pext)

La métrique pext, développée conjointement par les équipes travaillant sur gnomAD et sur GTEx, a pour objectif de normaliser l'expression par nucléotide au travers des différents transcrits. L'idée est de réexploiter les valeurs d'expression des transcrits en TPM (méthode précédemment expliquée dans la section précédente) afin d'évaluer si un nucléotide présente un « taux d'utilisation » important dans un tissu donné en calculant un ratio allant de 0 à 1 (Figure 45). À travers cette métrique, on peut évaluer si un nucléotide est exprimé dans un tissu donné en regardant son « taux d'utilisation », indépendamment des valeurs d'expression en TPM qui ne sont pas bornées, ni normalisées. Seuls 53 des 54 tissus présents dans GTEx ont été estimés dans pext car le tissu « *kidney– medulla* » présente un faible nombre de réplicas (4 réplicas).

Dans le cadre du développement de duxt, j'ai développé une méthodologie orthogonale visant à établir si un exon est différentiellement utilisé dans les tissus, en me basant sur les valeurs précalculées par pext

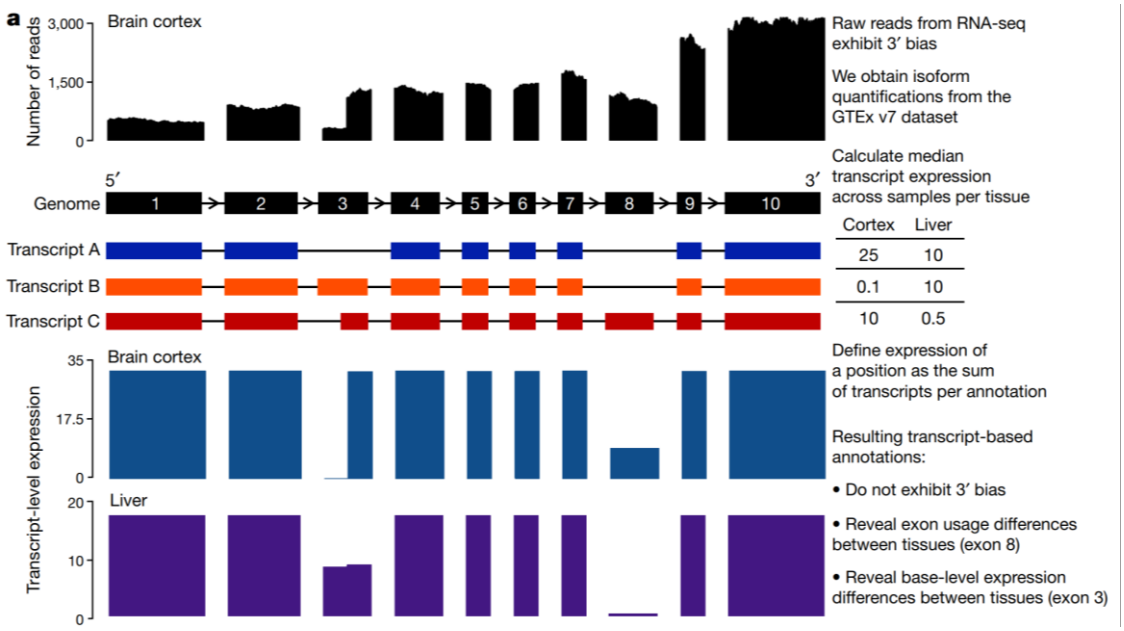


Figure 45 – Principe de l'expression normalisée au travers des transcrits par nucléotide (pext)

Exemple : pour un gène présentant 3 transcrits A, B et C avec, pour le foie, des valeurs d'expression respectives de 10, 10 et 0.5 TPM,, un nucléotide présent dans les transcrits B et C aura une valeur pext de : 10 + 0.5 / 10 + 10 + 0.5 soit pext = 0.512. Source : (Cummings et al. 2020)

6.5 Annotation des variations génétiques

6.5.1 *Variant Effect Predictor* (VEP)

Le *Variant Effect Predictor* (VEP) (McLaren et al. 2016) est une des ressources développées par Ensembl (6.1.2). VEP permet l'annotation des variations à travers l'apport de différentes informations telles que :

- Le(s) gène(s) et transcrit(s) portant la variation
- La position et la/les conséquence(s) fonctionnelle(s) sur les gènes/transcrits
- L'impact au niveau protéique (ID UniProt), acide aminé de référence et alternatif
- Les occurrences déjà observées dans des cohortes/bases de données (gnomAD, 1000 Genomes, ClinVar) et leurs propriétés (MAF, statut clinique)
- Des scores, métriques de conservation (phyloCSF, phastCons, phyloP), de contraintes (CCR), de prédiction de la pathogénicité (CADD, SIFT)

L'outil est utilisable *via* une interface web (pour une analyse manuelle d'un nombre restreint de variants, https://www.ensembl.org/info/docs/tools/vep/index.html), une API ainsi qu'un programme en ligne de commande. C'est ce dernier que j'ai utilisé dans le cadre du développement de MISTIC (version 96.3 basée sur l'assemblage GRCh37) pour uniformiser les jeux de variants avec différentes nomenclatures (*e.g.* niveau génomique/protéique) et dans différents formats de fichiers (*e.g.* VCF/TSV). Différents modules additionnels sont disponibles afin d'étendre les fonctionnalités de VEP à partir de ressources externes (*e.g.* CADD, dbNSFP).

6.5.2 Vcfanno

Tout comme VEP, Vcfanno (Pedersen, Layer, et Quinlan 2016) (https://github.com/brentp/vcfanno), est un programme qui permet l'annotation des variations génétiques. Celui-ci est utilisable sous la forme d'un exécutable. L'avantage de vcfanno est sa rapidité d'exécution due à la programmation et à l'exécution en langage Go (NB : langage *perl* pour VEP). De plus, vcfanno offre une flexibilité d'utilisation des annotations *via* de fichiers tabulés.

Vcfanno a été un outil complémentaire à VEP lors du développement de MISTIC afin d'apporter des annotations additionnelles non disponibles dans les différentes extensions de VEP. La version 0.3.2 a été utilisée lors du développement de MISTIC.

6.5.3 dbNSFP

dbNSFP (Liu et al. 2020) est une base de données d'annotation listant l'ensemble des SNV non-synonymes (*non-synonymous SNV ;* nsSNV). La version actuelle répertorie plus de 84 millions de variations auxquelles sont associés pour chacune : 37 scores de prédiction (SIFT, PolyPhen2, CADD, …), 9 scores de conservation (PhyloP, phastCons, …), les fréquences alléliques au sein des projets de séquençage (1000 génomes, gnomAD, …) ainsi que différentes descriptions fonctionnelles, identifiants de gènes... La ressource est utilisable uniquement *via* téléchargement de la base entière (https://sites.google.com/site/jpopgen/dbNSFP). La base est également exploitable *via* différents outils d'annotation des variations génétiques, dont VEP.

dbNSFP (v4.0b2) a été utilisée dans le développement de MISTIC à l'aide de VEP et de vcfanno afin d'obtenir l'ensemble des descripteurs nécessaires à la phase d'entrainement et d'évaluation du modèle.

Chapitre 7. Statistiques et programmation

7.1 Boite à outils en apprentissage automatique : scikit-learn

Scikit-learn (sklearn) (Pedregosa et al. 2011) est une librairie python permettant une utilisation simplifiée (à travers une API haut-niveau) de nombreux outils statistiques et informatiques dédiés à l'intelligence artificielle (IA) et plus particulièrement, à l'apprentissage automatique.

7.1.1 Méthodes d'apprentissage automatique implémentées

Sklearn (https://scikit-learn.org/stable/index.html) propose une large gamme de méthodes implémentées et prêtes à être utilisées, à partir d'apprentissage supervisé (exemples ci-dessous dans le Tableau 10) et d'apprentissage non-supervisé (intégrant notamment des méthodes de *clustering*). La librairie met à disposition des méthodes pour optimiser les hyperparamètres des méthodes statistiques, sélectionner les modèles les plus performants par des validations croisées (*cross validation* ; CV) ou encore, générer des métriques évaluant les performances (section 4.6.1).

Famille algorithmique (module dans scikit-learn)	Nom de la méthode	Classe associée dans scikit-learn
Ensemble `(sklearn.ensemble)`	*AdaBoost*	`AdaBoostClassifier`
	Gradient Boosting	`GradientBoostingClassifier`
	Random Forest	`RandomForestClassifier`
Modèle Linéaire `(sklearn.linear_regression)`	*Logistic Regression*	`LogisticRegression`
Classification naïve bayésienne `(sklearn.naive_bayes)`	*Gaussian Naive Bayes*	`GaussianNB`
	Bernoulli Naive Bayes	`BernouilliNB`
Plus proches voisins `(sklearn.neighbors)`	*K Nearest Neighbors*	`KNeighborsClassifier`
Réseaux neuronaux `(sklearn.neural_network)`	*Multi-Layer Perceptron*	`MLPClassifier`
Machine à Support de Vecteurs `(sklearn.svm)`	*Linear Support Vector Machine*	`LinearSVC`
	Support Vector Machine	`SVC`
Arbre de décisions `(sklearn.tree)`	*Extra Trees*	`ExtraTreesClassifier`
	Decision Trees	`DecisionTreesClassifier`

Tableau 10 – Exemples de méthodes de classification implémentées dans scikit-learn

7.1.2 *Recursive Features Elimination* (RFE)

« L'élimination récursive de descripteurs » (*Recursive Features Elimination* ; RFE) est une des méthodes de sélection de descripteurs implémentées dans scikit-learn. Le principe (Figure 46) est de supprimer de manière itérative le descripteur ayant eu le moins de poids statistique dans le modèle après évaluation de celui-ci, et de répéter l'opération jusqu'à ne garder qu'un seul descripteur dans le modèle final. À la fin du cycle, on regarde les performances obtenues pour identifier l'itération *i* pour laquelle les performances ont été les plus élevées et ainsi la combinaison optimale de descripteurs.

La RFE a été utilisée pour sélectionner la combinaison de descripteurs optimaux pour les deux algorithmes retenus dans MISTIC (*Random Forest* et *Logistic Regression*).

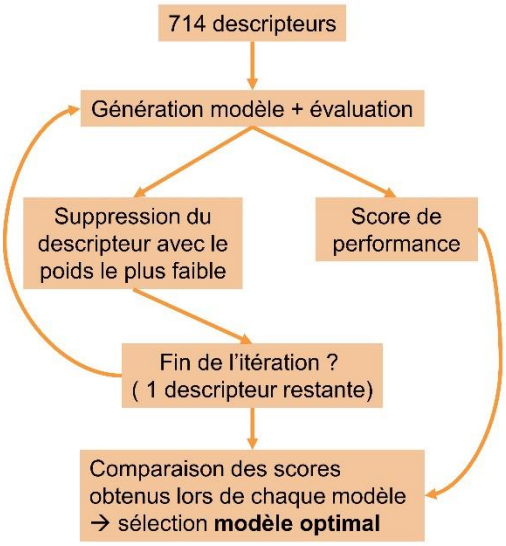


Figure 46 – Principe du RFE

A – principe de l'élimination récursive de descripteurs (Recursive Features Elimination ; RFE)
B – Évaluation et obtention des jeux de descripteurs présentant les meilleures performances pour la régression logistique et la forêt aléatoire.

7.2 Statistiques

7.2.1 Tests non-paramétriques

7.2.1.1 Test de Mann-Whitney U

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m} S(X_i, Y_j), \text{ avec } S(X,Y) = \begin{cases} 1, & if\ Y < X \\ \frac{1}{2}, & if\ Y = X \\ 0, & if\ Y > X \end{cases}$$

Équation 2 – Statistique du test de Mann-Whitney U

X et Y ainsi que leurs éléments respectifs sont indépendants

Le test de *Mann-Whitney U (MWU)* est un test statistique non-paramétrique permettant de comparer les « rangs » de valeurs présentes dans deux distributions. Le rang d'une observation correspond à l'ordre dans lequel apparait cette valeur dans une liste ordonnée. Dans le Tableau 11, la valeur 6 se situe au 4ème rang de la **liste A** tandis que la même valeur 6 se trouve au 1er rang de la **liste B**. Ainsi, le test de MWU permet de vérifier si les rangs des valeurs issues de deux listes distinctes sont statistiquement semblables ou non.

Rang	A	B
1	2	**6**
2	3	10
3	4	18
4	**6**	20
5	10	25
6	12	
7	14	

Tableau 11 – Rang des valeurs pour deux listes A et B de tailles différentes

Le test de Mann-Whitney U va comparer les rangs des différentes valeurs entre les listes A et B. Par exemple, la valeur 6 présente le rang 4 dans la liste A et le rang 1 dans la liste B.

7.2.1.2 Test binomial

Le test binomial permet de comparer une fréquence observée comparativement à une fréquence attendue. Dans le cadre d'un test binomial, l'hypothèse H_0 correspond au fait que la fréquence observée est proche de la fréquence attendue. Si l'on prend l'exemple d'un jeté de dés, la probabilité d'obtenir le chiffre 3 est de 1/6. Si après 100 tirages, le 3 est apparu 33 fois, la fréquence observée (33/100) est deux fois supérieure à la fréquence attendue (1/6 = 16.66/100).

$$p = \sum_{i=0}^{k} \Pr(X = i) = \sum_{i=0}^{k} \binom{n}{i} \pi_0^i (1 - \pi_0)^{n-i}$$

Équation 3 – Equation du test binomial unilatéral

Si l'on conserve l'exemple des dés : p : p-value du test binomial ; k : nombre de succès (33) π_0 : la fréquence attendue par l'utilisateur (1/6) ; n : nombre d'éléments (100 tirages)

7.2.2 Correction par la méthode de Benjamini-Hochberg

La méthode de Benjamini-Hochberg a été utilisée afin de corriger les résultats statistiques obtenus et de réduire le taux de faux positifs (*False Discovery Rate < 5%)*. Après correction, les *p-value* corrigées inférieures à 0.05 ont été conservées.

7.2.3 Rapports des côtes (Odds Ratio)

Les rapports de côtes (*Odds Ratio*) ont été utilisés pour étudier les enrichissements en variations délétères chez les TER (régions codantes d'exon) constitutives (Const) par rapport aux TER alternatives (Alt). Le rapport de côtes a été calculé sur la base du test exact de Fisher pour chaque condition en utilisant la formule :

$$Odds\ Ratio = \frac{a/b}{c/d}$$

Équation 4 – Formule du Odds Ratio

a : nombre de variations délétères dans les TER Const,
b : nombre de variations non-délétères (variations de population) dans les TER Const,
c : nombre de variations délétères dans les TER Alt,
d : nombre de variations non-délétères (variations de population) dans les TER Alt.

L'erreur standard (*standard error* ; SE) associée a été calculée selon la formule :

$$SE = \sqrt{\left((1/a) + (1/b) + (1/c) + (1/d)\right)}$$

Équation 5 – Calcul de l'erreur standard associé au Odds Ratio

Les limites inférieures et supérieures des intervalles de confiance (*confidence interval* ; CI) à 95 % ont été calculées en utilisant l'expression :

$$e^{\ln(Odds\ Ratio) \pm 1.96 \times SE}$$

Équation 6 – Calcul des limites inférieures/supérieures de l'intervalle de confiance du Odds Ratio

7.3 Environnement de programmation

7.3.1 code-server

Code-server (https://github.com/cdr/code-server) est un environnement de développement (*Integrated Development Environment* ; IDE) fonctionnant à distance basé sur *Visual Studio Code* (VScode), développé par Microsoft (https://code.visualstudio.com/) et reprenant la quasi-totalité de ses fonctionnalités. Comme tout logiciel nécessitant une connexion à distance, il requiert une exécution sur serveur et est accessible à travers un navigateur web, permettant son accès par n'importe quelle machine configurée. On profite ainsi d'un environnement de travail délocalisé, mettant à profit les ressources informatiques dont disposent les serveurs (nombreux cœurs CPU, mémoire vive de l'ordre du To) ainsi que de l'accès direct aux fichiers présents sur ceux-ci telles que les ressources mentionnées plus tôt (gnomAD v3 : 1,7 To de données). Un des grands avantages de code-server et VScode comparativement aux multiples IDEs existants est sa souplesse, sa réactivité ainsi que le nombre très important de personnalisations et d'extensions disponibles. La version 3.10.2 a été utilisée.

7.3.2 jupyter lab

Jupyter lab (https://jupyter.org/), à l'instar de code-server est un IDE fonctionnant *via* le principe serveur – client. Il présente des propriétés similaires à code-server, mais se distingue par une gestion facilitée des « notebooks », documents interactifs à la différence d'un script classique en programmation. Ceux-ci permettent d'afficher les résultats de fragments de code intermédiaire, autorisant ainsi la consultation de tableaux, figures ou résultats sans exécution de l'intégralité d'un script. La version 3.0.9 a été utilisée.

7.3.3 python

python (https://www.python.org/) est un langage de programmation interprété et multiparadigme. Sa philosophie le rend moins déclaratif que d'autres langages (C, Java), permettant des programmes de taille plus courte et plus lisible. Ceci est rendu possible grâce à une détection dynamique du type des variables déclarées (entier, réel, chaînes de caractères, liste) et une gestion automatique de la mémoire. En contrepartie, l'un des désavantages est sa faible vitesse d'exécution. Néanmoins, il reste avec le langage R, l'un des langages les plus utilisés par la communauté scientifique et la communauté des « *data sciences* ». Cette popularité est associée à un soutien technique qui offre des mises à jour régulières et de nouvelles applications rattachées. En effet, python s'emploie essentiellement au travers de librairies (Tableau 12), qui permettent de réaliser des manipulations complexes en peu de commandes (gestion de matrices par exemple). La majeure partie des outils d'intelligence artificielle sont notamment accessibles *via* une librairie python (scikit-learn, tensorflow, keras, pytorch). python ≥ 3.6 a été utilisée durant la thèse.

Nom	Version	Site	Utilisation	MISTIC	duxt
cyvcf2	0.11.7	https://github.com/brentp/cyvcf2	Manipulations de données (dédié variations)	✔	✘
pandas	1.1.4	https://pandas.pydata.org/	Manipulations de données	✔	✔
hail	0.2.64	https://hail.is/	Manipulations de données (dédié génomique)	✘	✔
scikit-lean	0.22.1	https://scikit-learn.org/	Apprentissage automatique	✔	✘
matplotlib	3.2.1	https://matplotlib.org/	Visualisation	✔	✔
seaborn	0.11.1	https://seaborn.pydata.org/	Visualisation	✔	✔
scipy	1.6.1	https://www.scipy.org/	Statistiques	✔	✔

Tableau 12 – Librairies majeures utilisées durant la thèse

7.3.4 conda

conda (https://www.anaconda.com/) est un système de gestion des paquets trans-plateformes (Windows, Linux, Mac OS) et de leur déploiement, faisant partie de la distribution Anaconda. Face au nombre important de librairies disponibles en Python, conda permet une installation et une gestion facilitées de ces dernières, empêchant notamment les conflits entre dépendances [exemple : j'installe les librairies A et B qui ont besoin de la librairie C (dont la version actuelle = v3) ; librairie A a besoin de librairie C en version 2 et librairie B a besoin de librairie C en version >=2 ; conda va installer librairie C en version 2 afin de correspondre aux besoins des librairies A et B]. Conda permet également la création d'environnements virtuels, intégrant pour chacun, le langage de programmation dans la version spécifiée (ex : Python 3.7) et les librairies (et versions adéquates) nécessaires à l'utilisation de l'environnement virtuel. La version 4.7.12 de Conda a été utilisée.

7.4 De la recherche manuelle à la requête distribuée

Durant ma thèse, j'ai fréquemment dû manipuler et analyser des fichiers au format VCF (format présenté en section 6.3.2). Ceci peut s'effectuer manuellement, *via* l'utilisation d'un programme à façon ou par les librairies python décrites ci-dessous.

7.4.1 cyvcf2

cyvcf2 (Pedersen et Quinlan 2017) (https://github.com/brentp/cyvcf2) est une des librairies permettant la manipulation (lecture, écriture) de fichiers VCF. Celle-ci est basée sur HTSlib (Bonfield et al. 2021), une librairie écrite en C permettant la gestion des fichiers issus du NGS avec rapidité et robustesse (à la base de SAMtools). À travers une itération sur le fichier VCF, cyvcf2 permet de gérer chaque variant comme un objet (*record*) et les différents champs du format VCF comme des attributs de cet objet (*record.CHROM, record.POS, record.INFO*). La version 0.11.7 de cyvcf2 a été utilisée.

7.4.2 pandas

pandas (https://pandas.pydata.org/) est une librairie qui s'est imposée comme le standard en matière de gestion de tableaux (*dataframe*) en python. Celle-ci permet une lecture et une écriture dans un grand nombre de formats de fichiers (CSV, Excel, Apache parquet, *feather*, hdf5) et une manipulation simplifiée des données présentes dans un tableau. À l'image d'une requête dans une base de données SQL, pandas permet de filtrer, de gérer les dupliquas, les valeurs manquantes, de réaliser des jointures entre tableaux, de les concaténer ou de modifier leur représentation. De plus, celui-ci est en étroite relation avec les différentes librairies de visualisation disponibles en python (matplotlib, seaborn notamment).

La majeure partie des données disponibles en biologie sont formatées dans une présentation similaire à celle d'un tableau (VCF, GFF, TSV). Dans le cadre de ma thèse, pandas a été l'outil central permettant de manipuler et analyser des données hétérogènes de diverses sources (variations, annotations, conservation, expression, données sur les maladies génétiques). La version 1.1.4 de pandas a été utilisée.

7.4.3 hail

hail (https://hail.is/) est une librairie récente dédiée à la génomique et développée par le *Broad Institute of MIT and Harvard* dans le but de faciliter l'accès au flux de données issues des projets de séquençage. Hail est en quelque sorte, l'alternative à pandas, dédiée à la génomique. Elle autorise des requêtes distribuées (*via* une optimisation de calculs sur serveur) et permet de questionner un fichier dont la taille peut aller jusqu'à l'ordre du pétabyte (1 000 000 de Go / 1 000 To). Elle permet la manipulation de tableaux (2D) ou de matrices (multi-dimensionnels). La librairie a été utilisée lors de l'analyses de différents projets à grande échelle (gnomAD, détection des QTL dans GTEx, mega-GWAS dans *UK BIOBANK*). La version 0.2.64 de Hail a été utilisée.

CONTRIBUTIONS

Chapitre 8. MISTIC : prédicteur robuste de l'impact des variations faux-sens

8.1 Contexte

Malgré l'apport déterminant du séquençage dans l'étude clinique des maladies génétiques rares, la recherche des variations délétères causales demeure complexe, compte tenu du nombre élevé de variations rares dans un génome (section 2.2.1). Dans ce contexte, de nombreux outils de prédiction de l'impact des variations ont été développés pour améliorer leur caractérisation (Chapitre 4 ; Tableau 5). Ces développements se sont concrétisés à la fois, par une utilisation croissante en clinique, mais également par la préconisation de bonnes pratiques par l'ACMG (section 3.3.2). La majeure partie de ces prédicteurs se focalisent sur l'étude des variations impactant les régions codantes des gènes, eu égard à la proportion élevée de variations délétères dans ces régions et dans les bases de données. Néanmoins, les outils actuels présentent des limites matérialisées par un nombre important de faux positifs engendrant des listes étendues de variations à statut inconnu (VUS ; section 3.4). Ces limites participent du taux de résolution limité oscillant entre 30 et 50% (section 3.1).

Pour pallier ces limites, nous avons développé MISTIC (*MISsense deleTeriousness predICtor*), un nouvel outil de prédiction de l'impact des variations faux-sens, classe la plus étudiée, mais aussi la plus difficile à interpréter compte tenu de leur nombre important chez un individu (section 2.1.1.1). MISTIC innove dans les trois grandes composantes de l'intelligence artificielle : les données d'entrainement du modèle, les descripteurs associés aux données et la méthode algorithmique utilisée. MISTIC devait répondre à trois objectifs: i) avoir une performance élevée aussi bien pour la détection/prédiction de l'impact des variations délétères (sensibilité) que des variations de population (spécificité); ii) attribuer un score à l'ensemble des 74 millions de faux-sens potentiels présents sur le génome humain, incluant environ 6 millions de faux-sens observés avec MAF et 68 millions sans MAF (non détectées par les consortiums de séquençage tel que gnomAD) ; iii) pouvoir s'inscrire dans une approche clinique et réaliser une analyse d'exome la plus précise possible, en identifiant la variation délétère, mais en générant peu de faux positifs.

8.2 Publication

MISTIC:

A PREDICTION TOOL TO REVEAL DISEASE-RELEVANT DELETERIOUS MISSENSE VARIANTS

PLOS ONE

RESEARCH ARTICLE

MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants

Kirsley Chennen[1,2☯]*, Thomas Weber[1☯], Xavière Lornage[2], Arnaud Kress[1], Johann Böhm[2], Julie Thompson[1], Jocelyn Laporte[2], Olivier Poch[1]*

1 Complex Systems and Translational Bioinformatics (CSTB), ICube laboratory – CNRS, Fédération de Médecine Translationnelle de Strasbourg (FMTS), University of Strasbourg, Strasbourg, France, **2** Institut de Génétique et de Biologie Moléculaire et Cellulaire, INSERM U1258, CNRS UMR7104, University of Strasbourg, Illkirch, France

☯ These authors contributed equally to this work.
* kchennen@unistra.fr (KC); poch@unistra.fr (OP)

Citation: Chennen K, Weber T, Lornage X, Kress A, Böhm J, Thompson J, et al. (2020) MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. PLoS ONE 15(7): e0236962. https://doi.org/10.1371/journal.pone.0236962

Editor: Miguel A Andrade-Navarro, Johannes Gutenberg Universitat Mainz, GERMANY

Received: March 27, 2020

Accepted: July 16, 2020

Published: July 31, 2020

Data Availability Statement: All data files and source code are available on the website http://lbgi.fr/mistic.

Funding: We thank the BISTRO and BICS bioinformatics platforms for their assistance. This work is supported by the Agence Nationale de la Recherche (BIPBIP: ANR-10-BINF-03-02; ReNaBi-IFB: ANR-11-INBS-0013, ELIXIR-EXCELERATE: GA-676559), Institute funds from the CNRS, the Université de Strasbourg and by the Fondation Maladies Rares within the frame of the

Abstract

The diffusion of next-generation sequencing technologies has revolutionized research and diagnosis in the field of rare Mendelian disorders, notably *via* whole-exome sequencing (WES). However, one of the main issues hampering achievement of a diagnosis *via* WES analyses is the extended list of variants of unknown significance (VUS), mostly composed of missense variants. Hence, improved solutions are needed to address the challenges of identifying potentially deleterious variants and ranking them in a prioritized short list. We present MISTIC (MISsense deleTeriousness predICtor), a new prediction tool based on an original combination of two complementary machine learning algorithms using a soft voting system that integrates 113 missense features, ranging from multi-ethnic minor allele frequencies and evolutionary conservation, to physiochemical and biochemical properties of amino acids. Our approach also uses training sets with a wide spectrum of variant profiles, including both high-confidence positive (deleterious) and negative (benign) variants. Compared to recent state-of-the-art prediction tools in various benchmark tests and independent evaluation scenarios, MISTIC exhibits the best and most consistent performance, notably with the highest AUC value (> 0.95). Importantly, MISTIC maintains its high performance in the specific case of discriminating deleterious variants from benign variants that are rare or population-specific. In a clinical context, MISTIC drastically reduces the list of VUS (<30%) and significantly improves the ranking of "causative" deleterious variants. Pre-computed MISTIC scores for all possible human missense variants are available at http://lbgi.fr/mistic.

Introduction

Next-Generation Sequencing technologies, such as Whole Exome Sequencing (WES) involving the targeted sequencing of exonic regions of all known protein-coding genes, have gradually replaced conventional approaches for the study of rare Mendelian disorders since 2010 [1]. Their usage is shifting from research investigations of disease-causing variants to routine clinical exome analysis for diagnosis of Mendelian disorders with known genetic aetiology

"Myocapture" sequencing project, the Fondation pour la Recherche Médicale, and the Association Française contre les Myopathies.

Competing interests: The authors have declared that no competing interests exist.

[2, 3]. However, with a diagnostic rate of ~40% for exome analyses, the identification of the deleterious variants, even in the coding regions, remains laborious [4–6]. The unsolved exomes usually result in extensive lists of variants, including numerous Variants of Unknown clinical Significance (VUS). The VUS are variants for which the pathogenicity (either benign or deleterious) could not be reliably determined given all available evidence (databases, collections of exomes etc), according to recommendation criteria from scientific communities, such as the Association for Molecular Pathology (AMP) [7] or the American College of Medical Genetics (ACMG) [8]. The VUS are mainly composed of missense variants, which make up to ~60% of 'Uncertain significance' variants in the ClinVar database [9].

The AMP and ACMG guidelines provide several criteria to classify deleterious/benign variants, in order to filter, prioritize or reduce the list of VUS into a shorter list of candidate variants that is amenable for expert review and additional experimental validation [10, 11]. For example, the minor allele frequency (MAF) (*e.g.* criteria PM2, BS1, BA1 of the ACMG) representing the observed frequency of a given variant in control healthy cohorts, has been demonstrated to be a very powerful filter. However, MAF values are often missing for deleterious or population-specific variants. To facilitate the evaluation of missense variants effects, several deleteriousness prediction tools have been developed that integrate a number of additional criteria [8, 12], such as the impact of the variant on the protein structure and/or function, the evolutionary conservation, or the physiochemical and biochemical properties of amino acids (*e.g.* SIFT [13], PolyPhen2 [14], VEST4 [15]). These tools have an accuracy ranging from 65 to 80% when benchmarked on known disease missense variants [16, 17]. Since individual tools tend to disagree on some missense variants, a novel type of ensemble prediction tools has recently emerged (*e.g.* Condel [18], CADD [19], MetaLR/MetaSVM [20], FATHMM-XF [21], Eigen [22], REVEL [23], M-CAP [24], ClinPred [25], and Primate AI [26]). The ensemble prediction tools combine the power of individual tools in order to achieve higher classification accuracies up to ~90% [17]. Nevertheless, the tools can still produce ambiguous predictions or even no prediction at all for some missense variants, contributing to the extended list of VUS (criteria PP3 of the ACMG) with a poor ranking of causative variants.

Here, we present MISTIC (MISsense deleTeriousness predICtor), a new supervised machine-learning model dedicated to the prediction of deleterious missense variants. MISTIC integrates a Soft Voting system [27] based on two optimized complementary machine-learning algorithms (Random Forest [28] and Logistic Regression [29]). The algorithms were trained to distinguish deleterious from benign missense variants based on a selection of 113 missense features, ranging from multi-ethnic MAF and evolutionary conservation constraints, to changes in amino acid physiochemical and biochemical properties. The performance of MISTIC is compared to other recent state-of-the-art prediction tools (Eigen, FATHMM-XF, REVEL, M-CAP, ClinPred and PrimateAI) in a series of benchmark tests designed to represent different variant analysis scenarios. We show that MISTIC has the best performance in predicting and ranking deleterious missense variants in coding regions. Moreover, in a clinical usage context, we demonstrate that MISTIC drastically reduces the list of VUS, and improves the ranking of the "causative" deleterious variants. To make MISTIC easily usable and accessible for future developments, we provide pre-computed scores for all possible human missense variants.

Materials and methods

Features

To describe missense variants, 714 features in 4 main categories were initially collected (S1 Table):

1. 8 multi-ethnic MAF [30]: all exomes (global MAF), African/African-American (AFR), Latino American (AMR), Ashkenazi Jewish (ASJ), East Asian (EAS), Finnish (FIN), Non-Finnish European (NFE), South Asian (SAS).
2. 8 conservation measures: PhastCons (primates, mammals, vertebrates) [31], PhyloP (primates, mammals, vertebrates) [32], SiPhy [33], GERP++ [32].
3. 690 functional measures: constrained coding regions (CCRs) [34], Missense badness, PolyPhen-2, and Constraint score (MPC) [35], physicochemical and biochemical properties from the AAindex databases (amino acid features) [36].
4. 7 pathogenicity predictors based on deleteriousness scores from different prediction tools: SIFT, PolyPhen2, VEST4, Condel, CADD, MetaSVM and MetaLR.

The features for the missense variants are based on the GRCh37 genome assembly and were extracted using Variant Ensembl Predictor (VEP) v96 [37] and VCFAnno v0.3.1 [38] from the CADD v1.4 and dbNSFP v4.0b2 [39] databases.

Training and test sets

MISTIC was trained and tested using variants from the VarData set, which is composed of (i) a positive set corresponding to rare deleterious missense variants, and (ii) a negative set corresponding to rare benign missense variants (S2 Table).

For the positive set, 38,565 deleterious missense variants with a "Pathogenic" clinical significance interpretation (CLNSIG) were selected from the ClinVar [9] VCF file (release of 30/09/2018). This list of variants was further filtered to select only 15,219 high confidence variants with a review status (CLNREVSTAT) "criteria provided" by the submitter, provided by "multiple submitters", a "reviewed by expert panel" or using "practice guideline", and "no conflicts" among the multiple interpretations submitted. Additionally, from the curated HGMD Pro [40] VCF file (version 2018.1), 76,523 missense variants with "Disease-Mutation" (DM) STATUS tag were selected as high-confidence deleterious missense variants. The resulting lists of variants from both ClinVar and HGMD Pro were then filtered to exclude:

1. any overlapping variants with the training set of the 7 prediction tools (PolyPhen-2 HUMVAR & PolyPhen-2 HUMDIV, SIFT, VEST4, Condel, CADD, MetaLR, MetaSVM) used as features in MISTIC to prevent type 1 circularity errors [17],
2. variants without a full annotation coverage of the features used in MISTIC (see Model definition section).

Finally, the VarData positive set contains 11,190 high confidence deleterious missense variants after merging the non-filtered variants from ClinVar and HGMD Pro databases.

For the negative set, rare benign missense variants were obtained from the gnomAD database, which combines variation data from over 125,000 exomes and over 15,000 genomes. Since no individuals in this database have any of the known severe childhood Mendelian disorders, it is assumed that highly penetrant disease-causing missense variants will be rare in this database (MAF $< 1\%$). The missense variants with a depth coverage $>30X$, were filtered to exclude (i) any overlapping variants in ClinVar and HGMD Pro databases, (ii) type I circularity error variants and (iii) variants without a full annotation coverage of MISTIC features, which resulted in 5,599,566 variants. The resulting list was divided into two sets: (i) Benign_VarData set comprising 11,190 randomly selected variants to match the size of the positive set for the training and testing of MISTIC, and (ii) Benign_EvalSet that contains the rest of the variants and serves as a negative set for the further evaluation of MISTIC (see below).

In order to train the supervised machine-learning models in MISTIC, 10,070 variants (~90% of VarData) were used from both positive and negative sets (denoted VarTrain). The remaining 992 variants (~10% left of VarData) in both positive and negative sets (denoted VarTest) were then used to test the performance of MISTIC.

Evaluation scenarios

To further evaluate the performance of MISTIC compared to other prediction tools, we collected six additional sets, including (i) two sets of deleterious variants, (ii) a set of rare benign variants and (iii) three population-specific variants (S3 Table).

1. Del_EvalSet contains two sets of deleterious variants:
   The ClinVarNew set was generated to assess the ability of the different tools to predict novel deleterious variants. We therefore identified recent deleterious missense variants present in the ClinVar database of April 2019 (release of 2019/04/03) and absent from the version of September 2018 (release of 2018/09/30 used to construct the VarTrain set). After applying the same filters for high confidence deleterious missense variants as described above for ClinVar in the VarData positive set, 437 "novel" high confidence deleterious missense variants were obtained. To avoid circularity errors, the variants overlapping with the training sets of the tools used in the benchmark study (PolyPhen-2, SIFT, VEST4, Condel, CADD, MetaLR, MetaSVM, VarTrain) were removed (referred to later as circularity error filter). After applying the circularity error filter, 388 variants were obtained. However, ClinPred and M-CAP did not provide any scores for 101 variants, so for fair comparison only the resulting 287 deleterious missense variants were used in the benchmark test.
   The DoCM set was generated by selecting deleterious missense variants from the Database of Curated Mutations (version 3.2), derived from the literature and composed of curated mutations observed in cancer [41]. The circularity error filter was applied to the initial 226 pathological missense variants and variants overlapping with the ClinVarNew set were removed, resulting in 126 deleterious missense variants.

2. Benign_EvalSet contains one set of benign variants with MAF data.
   The Benign_EvalSet was constructed to evaluate the ability of the tools to predict rare benign variants with different levels of MAF. As described above, the Benign_EvalSet comprises 4,974,224 missense variants, after applying the circularity error filter and removing the variants used in VarTrain and VarTest.

3. PopSpe_EvalSet contains thee sets of population-specific variants, for which no MAF information is available.
   The UK10K set was constructed by selecting population-specific variants present in 3,781 healthy individuals from two British cohorts of European ancestry present in the UK10K project [42], namely the Avon Longitudinal Study of Parents and Children (ALSPAC) [43] and TwinsUK [44]. Different filters were applied to the initial 295,218 missense variants: (i) a depth coverage >30X, (ii) the circularity error filter, (iii) the population specific filter, which removes variants with MAF data or present in the VarData set, evaluation sets (ClinVarNew, DoCM), or in the other population sets. Finally, 34,973 UK10K-population-specific variants were obtained.
   The SweGen set was constructed by selecting population-specific variants present in 1,000 healthy Swedish individuals from the SweGen project [45]. After applying the same filters as for UK10K, 25,635 SweGen-population-specific variants were obtained.
   The WesternAsia set was constructed by pooling variant sets from 269 healthy Kuwaiti natives (comprising 109 individuals of Saudi Arabian tribe ancestry, 126 individuals of

Persian ancestry, 34 individuals of Bedouin ancestry) [46] and 16 healthy Turkish individuals [47]. After applying the same filters as for UK10K, 14,594 WesternAsia-population-specific variants were obtained.

Different combinations of the six evaluation sets were then used to construct three prediction scenarios:

1. ClinVarNew and Benign_EvalSet, to distinguish novel deleterious variants from rare benign variants,
2. DoCM and Benign_EvalSet, to distinguish known deleterious variants from rare benign variants,
3. ClinVarNew/DoCM and PopSpe_EvalSet. to identify deleterious variants in population-specific datasets without MAF data.

Clinical context scenarios

We constructed different datasets representing both simulated and real disease exomes.

The 1KG set comprises simulated disease exomes, in which we introduce a randomly selected deleterious missense variant (from one of the deleterious sets described above) into the 1092 individual background exomes from the 1000 Genomes Project [48]. The simulated disease exomes were then annotated using VEP and VCFAnno. The variants were filtered according to community best practices, such as depth coverage >10X and MAF <1% in control healthy population databases. After applying the circularity filter, there was an average of 420 missense variants per simulated disease exome.

The MyoCapture set represents more than 1,200 clinical exomes from the French MyoCapture consortium on congenital myopathies [49]. The 15 selected resolved cases correspond to recently identified disease-causing deleterious variations, published after 2016 and not included in VarTrain. These cases were considered as solved if: (i) the disease-causing deleterious variant is in a known myopathy-causative gene and the gene associated phenotypes clinically match the patient's phenotypes; (ii) the disease-causing deleterious variant is in a novel disease gene with strong genetic validation (*e.g.* segregation analysis, multiple families with variants in the same gene, similar phenotype) and functional evidence (*e.g.* animal models reproducing the patient phenotypes) according to the ACMG's recommendations. The sequencing reads were mapped to the GRCh37/hg19 assembly of the human genome using BWA-MEM v0.7.10-r789 [50]. Variants were called using GATK v4.0.3.0 following the Haplotype Caller workflow from GATK best practices [51]. The procedures for the annotation and filtering steps, described above for the 1KG set were also applied here. After applying the circularity filter, there was an average of 1,566 missense variants per clinical exome.

Model definition

Using the python scikit-learn library v0.20.2, we trained Random Forest [28] and Logistic Regression [29] machine learning algorithms on the VarTrain missense variants, which includes 10,070 deleterious variants as the positive set and 10,070 benign variants as the negative set. The design of MISTIC was done in three main steps. First, a selection and implementation of the most informative variant features (detailed above) for each algorithm, the Recursive Feature Elimination method (RFE) was used [52]. RFE is a method that enables machine learning algorithms to perform feature selection by iteratively training a model, ranking features (by assigned weights or coefficients), and then removing the lowest ranking

features. Second, the predictions of the Random Forest and the Logistic Regression algorithms were then integrated in a Soft Voting system. In contrast to classical majority voting (Hard Voting), a Soft Voting system calculates the weighted average probabilities. Third, the optimized combination of parameters for the Random Forest and Logistic Regression algorithms and the hyper-parameters of their relative weights in the Soft Voting system was obtained after a grid search optimization of 20 iterations with 5 cross-validations each time.

The score generated by the Soft Voting system ranges from 0 to 1 and represents the probability of a given missense variant to be classified as deleterious. By default, missense variants with scores >0.5 are classified as deleterious and missense variants with scores <0.5 are classified as benign.

Benchmarking statistics

The performance of MISTIC was compared to six recent state-of-the-art tools for prediction of deleterious variants: Eigen, PrimateAI, FATHMM-XF, REVEL, M-CAP and ClinPred. However, since the deleteriousness scores from these tools were not always available for every missense variant (ranging from 3.6% of the missense variants for REVEL up to 9.4% for M-CAP), we excluded variants without scores. The thresholds recommended by the authors (S4 Table) were used to compare the prediction performance of the different tools on the evaluation sets. Furthermore, for clinically relevant applications, the prediction and ranking performances were compared on sets corresponding to simulated disease exomes (1KG) and real clinical exomes (MyoCapture).

To compare the performance of the prediction tools, we used several statistical metrics derived from a confusion matrix. To achieve this, we identified a correctly classified variant as a true positive (TP) if and only if the variant corresponded to the positive class (deleterious) and as a true negative (TN) if and only if the variant corresponded to the negative class (benign). Accordingly, a false positive (FP) is a negative variant (benign) that is classified as positive (deleterious) and a false negative (FN) is a positive variant (deleterious) classified as a negative one (benign). From these different classification statistics, we calculated 12 performance metrics (S5 Table) as described in the Human Mutation guidelines [53], notably:

1. Sensitivity—proportion of identified true deleterious variants compared to all the true deleterious variants.
2. Specificity—proportion of identified true benign variants compared to all the true benign variants.
3. Precision–proportion of identified true positive deleterious variants over all variants predicted as deleterious.
4. Area under the Receiver Operating Characteristics (ROC) curve (AUC). The AUC can take values between 0 and 1. A perfect tool has an AUC of 1 and the AUC of a random tool is 0.5.
5. F1 score—measure of prediction accuracy, with a balanced use of precision and sensitivity. The higher the F1 score, the higher the accuracy of the tool.
6. Matthews Correlation Coefficient (MCC)—considers true and false positives and negatives to represent the degree of correlation (range from -1 to 1) between the observed and predicted binary classifications. The MCC is generally regarded as a balanced method to evaluate tools. An MCC of -1 indicates a completely wrong binary tool, while an MCC of 1 indicates a completely correct binary tool.

7. Log Loss value—measures the divergence of a tool from the true variant labels (true deleterious or true benign), *i.e*. it measures the associated degree of uncertainty for a tool. The Log Loss value ranges from $+\infty$ to 0. In this case, a good tool will have a low Log Loss value, hence a low degree of uncertainty in its predictions.
8. Diagnostic Odd Ratio (DOR)–measures the effectiveness of a diagnostic binary classification test. It is defined as the ratio of the odds of the test being positive if the variant is deleterious relative to the odds of the test being positive if the variant is benign. The DOR value ranges from zero to $+\infty$ and hence higher DOR are indicative of better tool performance.

Results

Variant prediction model

In order to accurately classify deleterious and benign missense variants, we built the MISTIC model based on a Soft Voting system that combines predictions from Random Forest and Logistic Regression machine learning algorithms. We initially defined 714 features to fully characterize the missense variants (VarTrain dataset) used to train the model (S1 Table). However, a common problem of such high-dimensional data sets is the presence of correlated predictors, which impacts the ability of the algorithms to identify the strongest predictors. Hence, to reduce the dimensionality of our data, we identified the most important features for each of the Random Forest and Logistic Regression algorithms independently, using the RFE method.

The data in S1 Fig show that the performance of the Random Forest models increases as the number of features decreases, ranging from an AUC value of 0.852 for a model with 714 features to a peak AUC value of 0.895 for a model with 10 features. In contrast, the performance of the Logistic Regression models with less than 113 features are lower with a mean AUC value of 0.820, while models with more than 113 features have a stable performance with a mean AUC value of 0.826. Since the Soft Voting system requires that both algorithms have the same number of features, a cutoff was defined at 113 features for an optimised performance combining both algorithms. The 113 selected features cover 3 main categories: (i) multi-ethnic MAF values (6 features), (ii) functional and conservation measures (100 features), and (iii) scores from missense prediction tools (7 features) (Table 1, see detailed list in S6 Table).

Finally, the MISTIC model was trained on the VarTrain set, using the 113 selected features. Twenty iterations on a randomized grid search and a 5 cross-validation on VarTrain were used to obtain the hyper-parameters for the most optimized combination of the Random Forest and the Logistic Regression algorithms (S7 Table). Each algorithm calculated different weights for the individual features (See S2 Fig and S8 Table). For the Random Forest, the 5 most predominant features are the global MAF (19.73%), MetaSVM (9.44%), MetaLR (6.93%), VEST4 (5.84%) and Condel (5.39%). For the Logistic Regression, the 5 strongest features are VEST4 (16.16%), MetaLR (8.63%), MetaSVM (5.00%), PolyPhen (3.88%) and the AAindex matrix MIYS930101 [54] (3.22%) that evaluates contact frequencies in protein structures for all residues.

Comparison of MISTIC with individual component features and other prediction tools on VarTest set

The performance of the MISTIC Soft Voting system was compared with other prediction tools using the VarTest set. As might be expected, MISTIC globally outperforms each of its individual component features (MetaSVM, MetaLR, VEST4, Condel, CADD, PolyPhen2, SIFT). However, MISTIC also performs better than the state-of-the-art missense prediction tools

Table 1. Selected features included in the Soft Voting system of MISTIC.

Category	Name	Number of features	Description
Minor Allele Frequencies	minor allele frequencies	6	minor allele frequencies for 6 populations: all exomes (global MAF), African (AFR), American (AMR), East Asian (EAS), None Finnish European (NFE), South Asian (SAS)
Conservation measures	PhastCons	3	phastCons conservation score based on three categories of multiple alignments: (i) 100 vertebrate genomes, (ii) 30 mammalians, and (iii) 17 primates. The larger the score, the more conserved the site
	PhyloP	3	phyloP (phylogenetic p-values) conservation score based on three categories of multiple alignments: (i) 100 vertebrate genomes, (ii) 30 mammalians, and (iii) 17 primates. The larger the score. the more conserved the site
	SiPhy 29way logOdds	1	The estimated stationary distribution of A, C, G and T at the locus using SiPhy algorithm based on 29 mammalian genomes
	GERP++_RS	1	Identified constrained elements in multiple alignments
Functional measures	CCRS	1	The score reflects the intolerance of constrained coding regions of protein-coding genes for protein-altering variants
	MPC	1	A deleteriousness prediction score for missense variants based on regional missense constraints.
	AAindex	90	The AAindex substitution matrices for different physicochemical and biochemical properties of amino acids.
Pathogenicity predictors	SIFT	1	Prediction of the impact of an amino acid substitution on the protein function
	PolyPhen 2	1	Prediction of the impact of an amino acid substitution on the structure and function of a protein using straightforward physical and comparative considerations
	VEST4_score	1	Machine learning method predicting the functional significance of missense mutations based on the probability that they are pathogenic
	Condel	1	Weighted average of the normalized scores of five methods (SIFT, PolyPhen2, Logre, MAPP, MutationAssessor)
	CADD PHRED	1	Machine learning scoring model that integrates more than sixty annotation features into a single metric, to distinguish variants that survived naturel selection from simulated mutations
	MetaLR	1	Logistic Regression model combining multiple variant scoring metrics
	MetaSVM	1	Support Vector Machine model combining multiple variant scoring metrics

https://doi.org/10.1371/journal.pone.0236962.t001

(Eigen, PrimateAI, FATHMM-XF, REVEL, M-CAP and ClinPred) with the highest AUC value of 0.956 (S9 Table, Fig 1). M-CAP has the second-best overall performance, with an AUC value of 0.891. M-CAP has the highest sensitivity of 0.955, but this comes at the cost of a low specificity value of 0.547. In contrast, MISTIC has a balanced sensitivity of 0.863 and specificity of 0.901. Among the individual component features of MISTIC, the MetaLR score has the best performance with an AUC value of 0.859. We also calculated other metrics, such as the F1 score (measures accuracy based on the balance between precision and sensitivity), the Log Loss value (measures the degree of uncertainty associated with a prediction) and the Diagnostic Odds Ratio (measures the effectiveness of a deleterious prediction relative to the odds of a deleterious variant and the odds of a benign variant). Here, MISTIC has the highest F1-score of 0.881, the highest DOR value of 57.347, as well as the lowest Log Loss value of 4.082.

Evaluation of MISTIC in different variant analysis scenarios

The generalizability and relevance of MISTIC's prediction performance was further compared to the other prediction tools using datasets representing different scenarios. It is important to note that the variant sets used in these scenarios are independent from the variant sets used for the model training (VarTrain) and initial testing (VarTest) described in the previous section.

First, we tested the ability of the prediction tools to differentiate novel deleterious variants (ClinVarNew set) or known deleterious variants from diverse sources (DoCM set), from rare benign variants at 5 MAF levels (<0.01, <0.005, <0.001, <0.0001, singleton in Benign_Eval-Set). Since each MAF set does not have the same number of deleterious variants, the

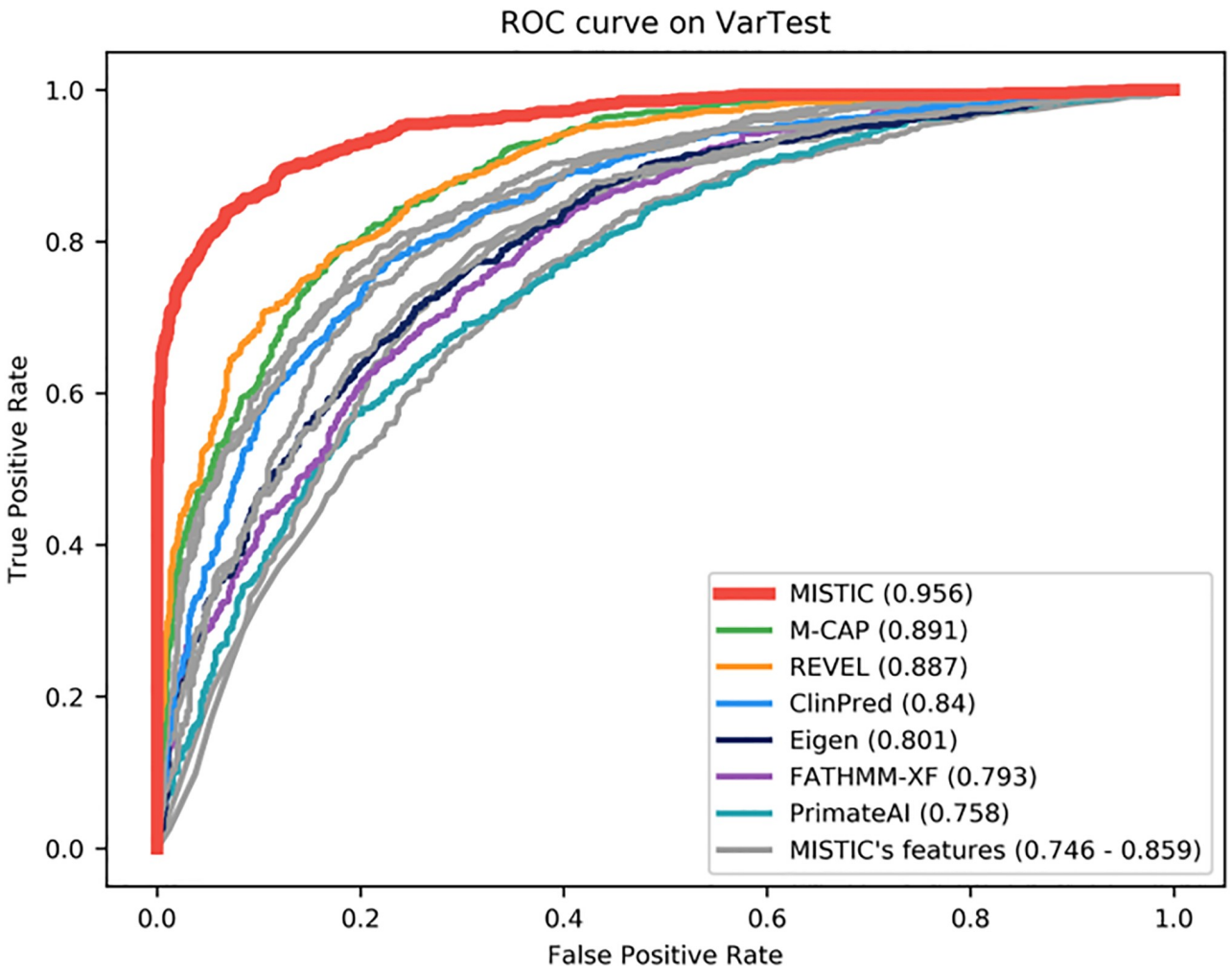


Fig 1. Performance of missense prediction tools on VarTest set. MISTIC was compared to individual component features (MetaSVM, MetaLR, VEST4, Condel, CADD, PolyPhen2, SIFT) used in its model (in grey) and the best-performing tools recently published (in color). The Area Under the receiver operating characteristics Curve (AUC) is shown in brackets.

https://doi.org/10.1371/journal.pone.0236962.g001

corresponding number of benign variants was randomly selected to obtain balanced pairs of deleterious-benign evaluation sets. This procedure was repeated 10 times and at each iteration, a different random set of benign variants was used.

Overall, MISTIC has the most consistent and best performance in discriminating deleterious variants from rare benign variants, with the highest mean AUC value on all the different scenarios (Fig 2). For the scenario involving novel deleterious variants (ClinVarNew set; Fig 2A, S10 Table) and rare benign variants, MISTIC has the highest mean AUC value of 0.963 ± 0.002, mean F1 score of 0.907 ± 0.002, mean DOR value of 92.548 ± 5.182, and the lowest mean Log Loss value of 3.332 ± 0.099. In terms of mean AUC and mean DOR values, M-CAP is the second best-performing tool with a mean AUC value of 0.930 ± 0.002 and a mean DOR value of 39.516 ± 1.650. However, in terms of mean F1 score, REVEL is the second-best performing tool (0.859 ± 0.004), as well as in terms of mean Log Loss value (5.048 ± 0.186).

For the scenario involving known deleterious variants from diverse sources (DoCM; Fig 2B, S11 Table) and rare benign variants, the same tendency was observed. Here, MISTIC has

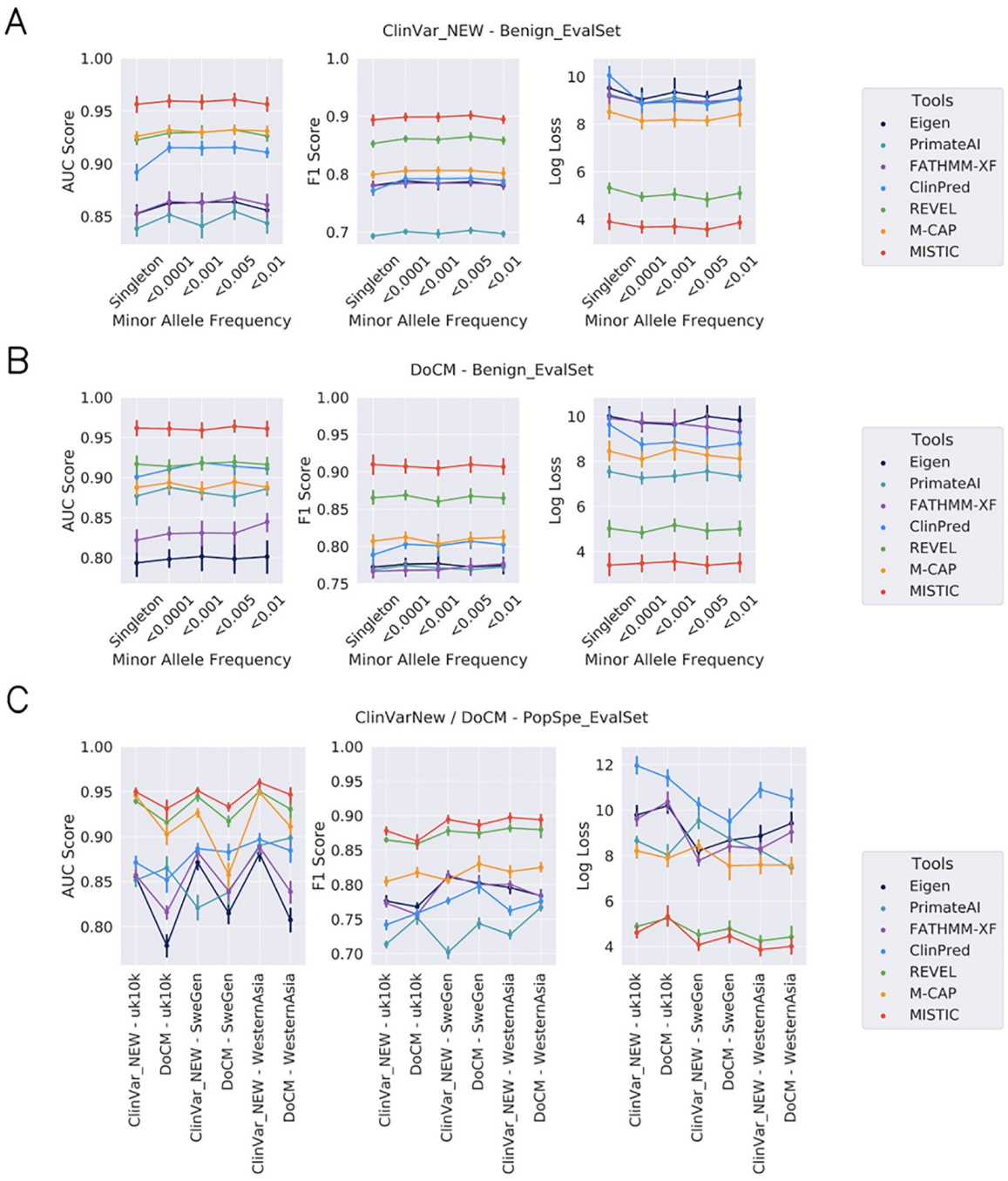


Fig 2. Evaluation of prediction tools on different variant analysis scenarios. The performance of MISTIC was compared to other missense prediction tools for the discrimination of deleterious variants from rare benign variants and population-specific missense variants. All prediction tools were evaluated using novel deleterious variants (Fig 2A - ClinVarNew and Benign_EvalSet set), known deleterious variants from diverse sources (Fig 2B - DoCM and Benign_EvalSet set), rare benign variants with MAF data (<0.01, <0.005, <0.001, <0.0001, singleton) or benign variants without MAF (ClinVarNew/DoCM and PopSpe_EvalSet: UK10K, SweGen, WesternAsia; Fig 2C).

https://doi.org/10.1371/journal.pone.0236962.g002

the best performance, with the highest mean AUC value of 0.968 ± 0.001, mean F1 score of 0.920 ± 0.003, mean DOR value of 125.642 ± 6.905, and the lowest mean Log Loss value of 2.981 ± 0.099.

Since the global MAF is an important feature in the MISTIC model (see S2 Fig), although MAF values are often missing for deleterious and population-specific benign variants, we evaluated the performance of MISTIC in discriminating deleterious variants from rare benign variants when no MAF data are available. To do this, benign population-specific variants were collected from three different populations, namely UK10K, SweGen and WesternAsia. For each deleterious set (ClinVarNew, DoCM), the corresponding number of benign variants was randomly selected from each population-specific set of variants. The deleterious and benign variants were scored by MISTIC and the six other missense prediction tools. This procedure was repeated 10 times, with a different random selection of the benign variants each time. For the six different combinations of two deleterious sets (ClinVarNew, DoCM) and three benign population-specific sets (UK10K, SweGen, WesternAsia), MISTIC has the best performance for three of them, with the highest mean AUC value of 0.945 ± 0.009 (Fig 2C, S12 Table). The second best-performing prediction tool is REVEL, with an overall mean AUC value of mean of 0.933 ± 0.012, F1 score 0.873 ± 0.007 and a mean Log Loss value of 4.688 ± 0.287. ClinPred has the highest DOR values for three of the combinations of variant sets and MISTIC has the highest DOR values for the other combinations. ClinPred has the highest sensitity (1) and DOR value (∞) in the combinations of variants based on the known deleterious variants (DoCM set). This is probably due to an overlap between the ClinPred training set and the DoCM set, leading to a problem of overfitting.

Performance on simulated disease exomes

In the context of a typical Mendelian disease exome analysis, even after most common benign variants have been removed with a standard allele frequency filter (MAF >1%), the challenge is to identify one or two rare causative deleterious variants among hundreds of predicted deleterious variants. Indeed, with current limited resources (time and cost), it not feasible to experimentally validate large numbers of candidate variants. To evaluate the ability of the prediction tools to prioritize the causative variants, we simulated Mendelian disease exomes by introducing one "causative" deleterious variant (from Del_EvalSet) in the background exomes of healthy individuals from the 1000 Genomes Project. The simulated disease exomes thus contained one "causative" variant and an average of ~420 missense variants per exome (see section Materials and methods).

First, we calculated the percentage of predicted deleterious variants obtained by the different tools, again using the authors' recommended threshold each time. The objective is to have the "causative" variants among the shortest list of predicted deleterious variants, that is trackable for a manual expert review. PrimateAI generated the shortest the list of variants by predicting only 5.393 ± 1.463% of the 1KG exomes variants as deleterious, while MISTIC's prediction was of 12.529 ± 3.195% (Fig 3A and S13 Table). Next, we evaluated the ability of the prediction tools to rank the "causative" variants among the top-scoring deleterious variants. We calculated the mean ranks of the "causative" variants introduced in the disease exomes after sorting the scores for each prediction tool (Fig 3B, and S13 Table). Overall, MISTIC has the best performance with a median rank value of 2, (mean rank: 14.092 ± 34.968) for the "causative" variants. The performance of MISTIC is significantly higher (Mann-Whitney $P < 1.21 \times 10^{-17}$) than the second-best tool, ClinPred, which has a median rank of 5 (mean rank: 11.155 ± 19.760).

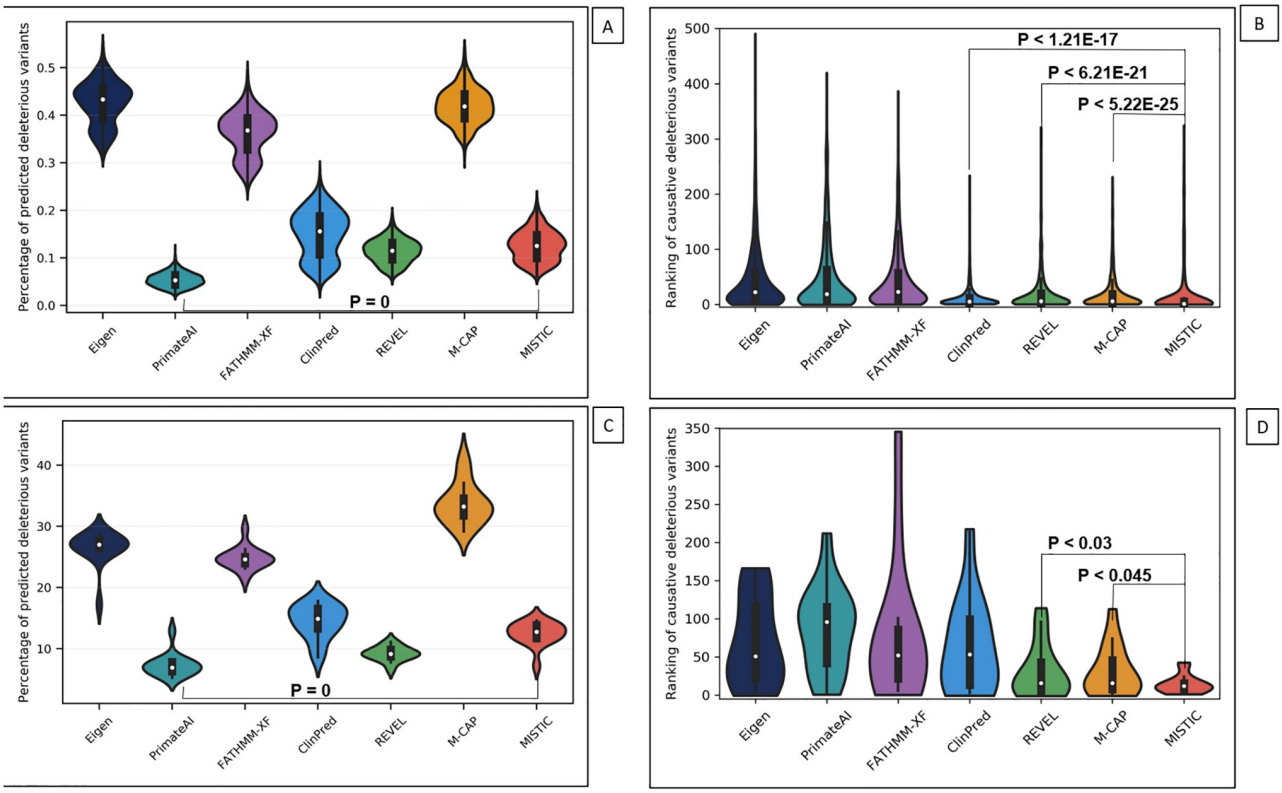


Fig 3. Evaluation of the different missense prediction tools using simulated and real disease exomes. A–Distribution of the percentage of predicted deleterious variants in the simulated disease exomes. B–Ranking of the "causative" deleterious variants introduced in simulated disease exomes. C–Distribution of the percentage of predicted deleterious variants on the exomes of the MyoCapture project. D–Ranking of the causative deleterious variants identified in real congenital myopathy exomes from the MyoCapture project.

https://doi.org/10.1371/journal.pone.0236962.g003

Performance on real clinical cases from a myopathy cohort

Finally, to represent a clinical practice scenario, we compared the performance of MISTIC to Eigen, PrimateAI, FATHMM-XF, ClinPred, M-CAP and REVEL, using 15 recently solved clinical exomes from the French Myocapture cohort on congenital myopathies. After applying the best-practice filtering procedures (See Material and Methods), the 1566 missense variants per exome were scored by the prediction tools (S14 Table).

As for the simulated disease exomes, PrimateAI achieves the largest reduction of the list of predicted deleterious variants (92.14% ± 2.19%), while MISTIC is the third best method (82.68% ± 2.36%) (Fig 3C, and S14 Table). However, in terms of ranking of the causative variants, MISTIC has the best performance with a median rank of 12 (mean: 14.67 ± 12.35) in the Myocapture exomes. M-CAP and REVEL were performed second-best with a median rank of 16 (M-CAP mean rank: 30.93 ± 32.59; REVEL mean rank: 31.07 ± 35.530). The ranking performance of MISTIC is significantly different from M-CAP and REVEL ($P < 0.045$ and $P < 0.030$ respectively).

Comparison of scores for deleterious and benign variants

To better understand the prediction behavior of MISTIC and the other tools, the score distribution of all variants in the pooled deleterious (Del_EvalSet) and benign sets (Benign_EvalSet, PopSpe_EvalSet) was visualized using violin plots (Fig 4, S3 Fig). Each tool provides a score and an associated class (deleterious or benign) based on the recommended threshold given by

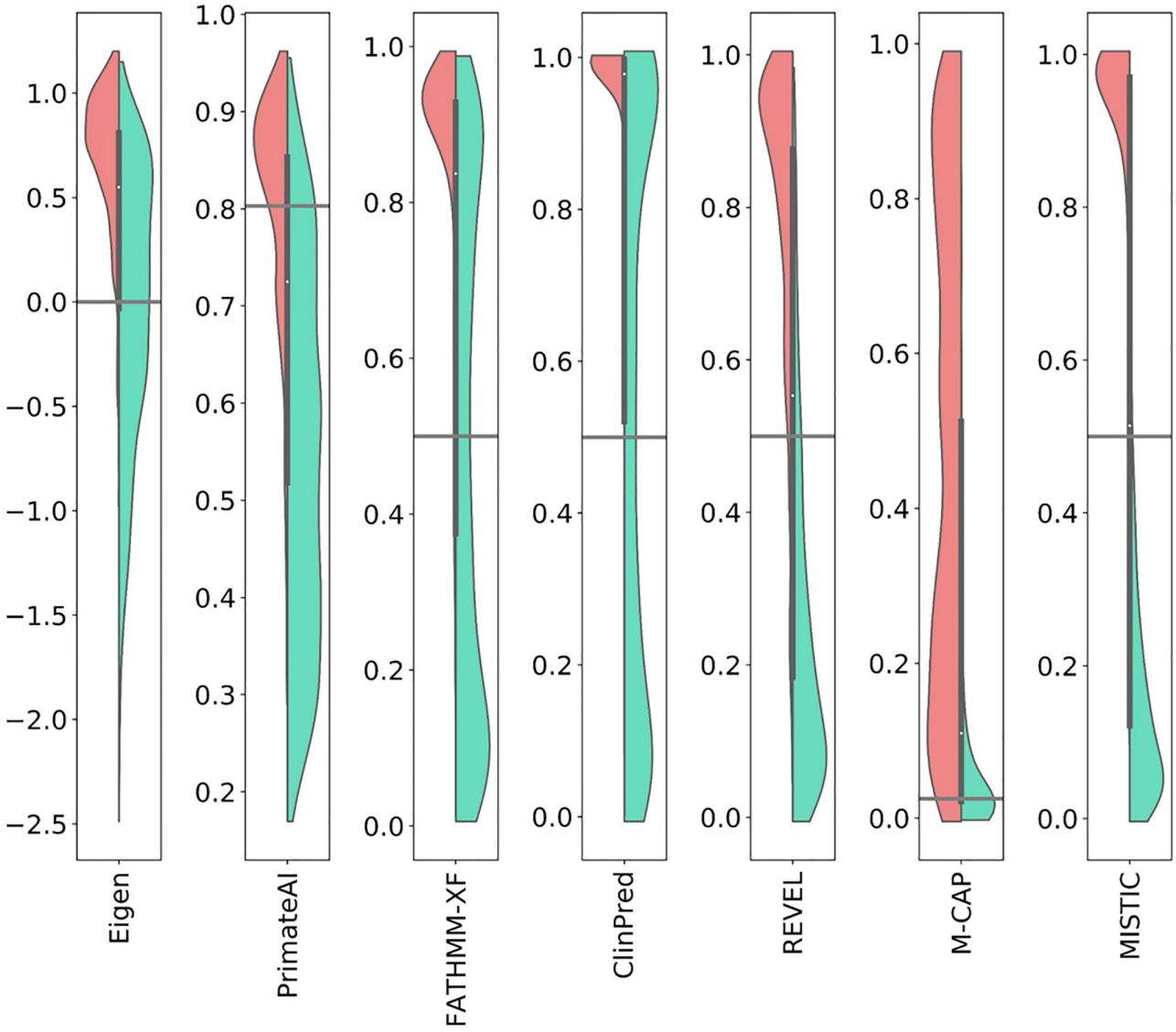


Fig 4. Distribution of scores for deleterious and benign variants. The variants of the deleterious (Del_EvalSet) and benign sets with MAF (Benign_EvalSet) were pooled and the distribution of the scores for deleterious and benign variants were represented using violin plots. Red area–distribution of scores for deleterious variants. Green area–distribution of scores for benign variants. Black line–recommended threshold.

https://doi.org/10.1371/journal.pone.0236962.g004

the authors (S4 Table). We therefore analyzed the score distributions for deleterious variants and benign variants with a MAF (Fig 4), and observed that tools (Eigen, PrimateAI, FATHMM-XF, ClinPred) that did not perform well (DOR value < 10) in our evaluation experiments generally have a poor performance in classifying benign variants. Around 50% of benign variants are misclassified as deleterious by these tools (44.8% for FATHMM-XF, 50.2% for ClinPred and 50.3% for Eigen). For M-CAP (DOR value < 15), we observed that its inherent hyper-sensitivity design (capacity to correctly categorize deleterious variants) comes at the cost of a poor specificity and consequently 46% of benign variants are misclassified as deleterious. It should be stressed that misclassified variants (benign as deleterious and *vice versa*) will contribute to the low resolution rate of exome analysis and to the generation of extended lists

of candidate variants, hence hindering the identification of the one or two "causative" deleterious variants.

Finally, we observed that MISTIC and REVEL both have a balanced sensitivity and specificity, *i.e*. a balanced ability to correctly classify both deleterious and benign variants. MISTIC misclassified 11% of the benign variants and 17% of the deleterious variants, while REVEL misclassified 16% of benign variants and 27% of deleterious variants. The same tendency was observed when comparing the distribution of variants without MAF (S3 Fig). This analysis provides further demonstration of the balanced ability of MISTIC to discriminate between deleterious and benign variants in comparison to other tools.

Effect of MISTIC design on its performance

To understand how the different factors incorporated in the original design of MISTIC (namely the Soft Voting system, the composition of the training set and the confidence in the status of the associated variants) contribute to its high performance and best ranking capacity, we generated different MISTIC models and compared their performance in the Del_EvalSet—Benign_EvalSet and the Del_EvalSet—PopSpe_Evalset scenarios.

Several commonly used individual classification models were trained using the same protocol as previsouly described for MISTIC (S4 Fig). On the Del_EvalSet—Benign_EvalSet, the Random Forest (RF) algorithm was the best model (mean AUC: 0.976 ± 0.004), followed by the Ada Boost (AB) model (mean AUC: 0.973 ± 0.003) and the Logistic Regression (LR) model (mean AUC: 0.954 ± 0.005). However, on the Del_EvalSet—PopSpe_Evalset, the LR model was the best model (mean AUC: 0.959 ± 0.010), followed by the RF model (mean AUC: 0.956 ± 0.014) and the AB model (mean AUC: 0.948 ± 0.017).

We further explored the concordance between these three models on the data from the evaluation scenarios (S5 Fig). Overall, on the Del_EvalSet—Benign_EvalSet data (S5A Fig), there is more than 80% of concordance among all the models (82.86% on benign variants; 89.45% on deleterious variants). The concordance is even higher among the two tree-based approaches (RF and AB models) with a concordance of 92.90% on benign variants and 92.21% on deleterious variants. The major difference between the LR model and the RF model is that it generated 11.62% more false positive deleterious predictions on benign variants with MAF. However, on the Del_EvalSet—PopSpe_EvalSet data (S5B Fig), while there is a 94.92% concordance among all models on deleterious variants, there is a concordance of only 10.68% for benign variants without MAF. The concordance between the tree-based approaches (RF and AB models) is of 99.44% on deleterious variants and 10.68% on benign variants. The AB model generated 29.50% more false positive deleterious predictions on benign variants without MAF. However, the concordance between the RF and the LR models is higher on benign variants without MAF, with a concordance of 40.18%. The LR additionally predicts 39.88% of true negative benign variants, which were mispredicted by the RF model. Hence, to avoid an unbalanced voting system towards false positive predictions on benign variants without MAF, we retained only the RF and the LR models in our Soft Voting system for MISTIC.

To investigate the contribution of the Soft Voting system (based on the weighted average of the RF and the LR models), we compared the full MISTIC model using Soft Voting to the RF and LR models in the different evaluation scenarious (S15 Table). The Soft Voting approach has the most balanced performance on both evaluation sets (Del_EvalSet-Benign_EvalSet: AUC of 0.969 and Del_EvalSet-PopSpe_EvalSet: AUC of 0.962). While the RF approach has the highest sensitivity, its specificity dropped from 0.902 on Benign_EvalSet to 0.466 on PopSpe_EvalSet. As for the LR approach, while it has also a balanced performance, its AUC and DOR values were systematically lower than the Soft Vorting system on both scenarios. On

average, over the two evaluation scenarios, the Soft Voting system has an improved performance for F1 score by 4.355 ± 10.932% and specificity by 12.423 ± 31.833%.

To investigate the potential contribution of combining deleterious and benign sets with a wide spectrum of variants in the training set, we compared the full MISTIC model to alternative models using a single source of deleterious variants (ClinVar only or HGMD only) or a source of benign variants with a reduced spectrum of variants in terms of ethnic groups or number of benign variants (UK10K or ClinVar). The full MISTIC model (using a training set of deleterious variants from both ClinVar and HGMD) exhibits a small improvement of AUC (1.223 ± 0.430%) compared to the models using only one source of deleterious variants (S16 Table). The source of benign variants had a greater impact on MISTIC performance (S17 Table), improving the AUC by 3.007 ± 3.026%, the Log Loss by 34.478 ± 22.739%, and the specificity by 34.314 ± 33.585% compared to the models using either UK10K or ClinVar benign variants only. This suggests that, although benign variants from curated databases such as ClinVar can be useful for improving the machine-learning definition of deleterious variants, these databases do not contain the full spectrum of benign variants that are present in population databases. This is also true for the model using benign variants from the UK10K set, which has a partial representation of the diverse ethnic groups.

Finally, to evaluate the impact of the high confidence training set, we compared the full MISTIC model to a model in which no high-confidence filtering criteria was applied (described in material and methods) for ClinVar variants (Pathogenic and Likely Pathogenic CLNSIG status) and HGMD variants (DM and DM? variants). The results in S18 Table show that the full MISTIC model using a high confidence training set increases the F1 score by 3.70 ± 0.97%, the Log Loss value by 31.95 ± 16.84%, and the specificity value by 8.83 ± 1.50%, compared to the model without high confidence filtering.

Discussion

With the widespread use of exome analyses for the study of rare Mendelian diseases, the major challenge hindering a complete transfer into routine clinical usage remains the interpretation of the list of VUS to identify the one or two "causative" deleterious variants. The list of VUS (mostly composed of missense variants) in unsolved exomes is generally too extensive to be screened manually or *via* experimental assays. Consequently, several tools have been developed to distinguish deleterious and benign variants and hence prioritize candidate variants for further validations assays. However, current solutions implement different strategies and can have large variations in performance.

MISTIC is a prediction tool combining a voting system with two complementary algorithms, which is dedicated specifically to the prediction of deleterious missense variants, in contrast to some generalist prediction tools aimed at predicting different types of variants (coding and noncoding) with diverse consequences (missense, nonsense, splice. . .). The performance of MISTIC and the other prediction tools were benchmarked on different evalution sets corresponding to diverse variant analysis scenarios ranging from evaluation of novel deleterious variants (ClinVarNew) and variants from different sources (DoCM), to rare benign variants with (Benign_EvalSet) or without (PopSpe_EvalSet) MAF information. Our results show that, in all the different evaluation scenarios, dedicated missense prediction tools (*e.g.* ClinPred, REVEL, M-CAP, PrimateAI and MISTIC) perform better than generalist ones (*e.g.* Eigen and FATHMM-XF). In this context, MISTIC exhibits the best performance compared to the other dedicated prediction tools. The results were obtained *via* objective analyses using independent evaluation sets (disjoint from the training set) to exclude any type I circularity error and selecting only variants with a score available for all the tools tested. Nevertheless, it is

important to note that the training sets of Eigen, ClinPred, M-CAP and REVEL were not readily accessible, and we could not exclude overlapping variants in the evaluation sets. In some cases, this might lead to an over-estimation of the performance for some tools. For instance, this was potentially the case for ClinPred on the DoCM set, where it had a sensitivity value of 1.

The improved performance of MISTIC can be attributed to the special care taken in its design. We evaluated and showed the impact of the different original design elements on MISTIC performance. First, this is, to our knowledge, the first usage of a combination of two different classes of machine-learning algorithms (Random Forest and Logistic Regression). In contrast, the other prediction tools use a single algorithm (Eigen, PrimateAI, FATHMM-XF, REVEL, M-CAP) or two similar ones (ClinPred uses two tree-based algorithms). Furthermore, MISTIC exploits the two machine learning algorithms in a Soft Voting system with optimized hyper parameters after a grid search with 20 iterations and 5 cross-validations. This synergic design results in a balanced sensitivity and specificity ratio (Fig 4, S3 Fig, and S15 Table) and thus a better classification of both deleterious and benign variants.

Second, MISTIC incorporates 113 features out of the initial 714 features, after a selection by Recursive Feature Elimination. This reduced set of features is used to characterize missense variants, ranging from the DNA level with the multi-ethnic MAF and evolutionary constraint features, to the amino-acid level with physiochemical and biochemical property changes. Since our training set is enriched in high confidence deleterious and benign variants, we expected that informative weighted features for distinguishing rare deleterious variants from rare benign variants could be identified. By studying the relative weights of the 113 features used by the two algorithms, we observed that the most predominant features for the Random Forest are the global MAF value and the MetaSVM score, while the MetaLR and VEST4 scores are the most predominant ones for the Logistic Regression. Overall, the integration of these features in two complementary machine algorithms may explain the overall best performance of MISTIC for the discrimination of deleterious variants from benign variants.

The third improvement in MISTIC's design is the constitution of its positive and negative training sets. The existing missense prediction tools used only one source of deleterious variants for the training of their model, either HGMD Pro (FATHMM-XF, REVEL and M-CAP) or ClinVar (ClinPred, Eigen). We showed that with a positive set composed of a wider spectrum of deleterious variants from multiple sources (ClinVar and HGMD Pro), MISTIC was able to improve its AUC value by 1%, its specificity by 3% and its Log Loss value by 6% (S16 Table). Moreover, to reduce the impact of misclassified deleterious variants, only the highest-confidence deleterious variants (with respect to each source) were used to train MISTIC, while tools like ClinPred also included variants with a 'likely pathogenic/deleterious' status in their training set. We showed that the use of a high-confidence positive set in MISTIC had the most impact on performance, increasing the specificity and Log Loss values by 34% (S18 Table). Concerning the negative training set, special attention was also taken to include a wide spectrum of rare benign variants from large control population databases. We also ensured that the negative set was distinct from the positive set, by filtering all the variants already present in the ClinVar and HGMD Pro databases, or other training sets (circularity error) in order to identify informative predictive features for rare benign variants. Our results show that this strategy improved MISTIC's AUC value by 3% and its specificity by 34% (S17 Table). The same tendency was observed for population specific variants, where other tools (REVEL, M-CAP) trained on negative sets from control population databases performed better than tools trained on a limited set of benign variants (*e.g.* ClinPred uses benign variants from ClinVar) (Fig 2C). Taken together, the constitution of a high confidence training set, with sources representing a

wider spectrum of variant profiles contributed to the performance of MISTIC in complex scenarios encountered in exome analyses.

The MAF feature, which is part of the ACMG recommendation, has previously been shown to be a powerful factor for filtering benign variants and it is already integrated in the other tools with various strategies. Hence, we construced evaluation scenarios using variants with/without MAF and in both cases we demonstrated that MISTIC had the best performance. MISTIC achieved an AUC improvement of 5% compared to the second-best performing tool on variants with MAF (VarTest) and an AUC improvement of 6% on variants without MAF.

Finally, in a context of routine clinical exome analysis, the major objective is to obtain a limited list of VUS (major challenge in 70% of unsolved exomes) with prioritized candidate variants that can be quickly screened experimentally with reasonable resources. The performance of some prediction tools on the simulated disease exomes (1KG) and real clinical exomes (MyoCapture) was contrasted with the previous evaluation results. Indeed, in the context of an exome analysis, PrimateAI obtained the best performance in terms of the smallest number of VUS (<20%), while M-CAP produced twice as many. However, in terms of ranking the causative deleterious variants, MISTIC achieved the best ranking performance on the simulated disease exomes ($P < 1.21E-17$) and the same tendency was observed on the real clinical exomes ($P < 0.045$). Taken together, these results illustrate that the balanced sensitivity and specificity of MISTIC in the different scenarios can also be applied in a context of personalized and precision medicine, in order to obtain a short list of prioritized candidate variants that is amenable to expert screening with reasonable resources.

In conclusion, MISTIC is a novel tool for prediction of deleterious of missense variants, based on a Soft Voting system of two complementary optimized supervised machine-learning algorithms. Among the 113 features integrated in MISTIC, multi-ethnic MAF are predominant for the classification of benign and deleterious variations. MISTIC consistently outperforms recent state-of-the-art prediction tools in the different scenarios tested. Finally, we provide a pre-computed score for all possible human missense variants (for canonical transcripts on the genome version GRCh37) in order to facilitate usage and integration in analysis pipelines. The source code of the method is available on the website http://lbgi.fr/mistic.

Future improvements will include additional informative features, such as multi-ethnic MAF from other population databases, genotype frequencies, and gene-based calibration of the different scores. Moreover, our approach could be applied for the design of dedicated prediction tools for other categories of variants, such as splice variants or non-coding variants, to prepare the transition from exome to complete genome analyses.

Supporting information

S1 Fig. Selection of MISTIC features. The pruning of the initial 714 missense features was performed using the Recursive Feature Elimination method for the Random Forest and Logistic Regression models and the VarTrain set. The red dotted line indicates the cutoff for the selected features in the final Soft Voting system.
(TIF)

S2 Fig. Relative normalized weights of the individual missense features integrated in MISTIC. The histograms show the relative weights of the individual missense features for the Random Forest and Logistic Regression models integrated in MISTIC. Key: MAF—minor allele frequency, AFR—African American population, AMR—Latino American population,

ASJ–Ashkenazi Jewish population, EAS—East Asian population, FIN—Finnish population, NFE—Non-Finnish European population, SAS–South Asian population, CCRS—Constrained-Coding RegionS.
(TIF)

S3 Fig. Distribution of scores for deleterious and benign variants without MAF. The variants of the deleterious (Del_EvalSet) and benign sets (PopSpe_EvalSet) without MAF were pooled and the distribution of the scores for deleterious and benign variants were represented using violin plots. Red area–distribution of scores for deleterious variants. Green area–distribution of scores for benign variants. Black line–recommended threshold.
(TIF)

S4 Fig. Performance of classifier models on data from the evaluation scenarios. Different individual classifier models were evaluated for their ability to discriminate deleterious variants from rare benign variants and population-specific missense variants. All classifier models were evaluated using: A—Del_EvalSet-Benign_EvalSet corresponding to novel deleterious variants, known deleterious variants from diverse sources and rare benign variants with MAF data (<0.01, <0.005, <0.001, <0.0001, singleton). B—Del_EvalSet-PopSpe_Evalset corresponding to novel deleterious variants, known deleterious variants from diverse sources and benign variants without MAF data.
(TIF)

S5 Fig. Concordance among classification models on data from the evaluation scenarios. Binary predictions made by the 3 classification models for each benign or deleterious variant in the Del_EvalSet, Benign_EvalSet and PopSpe_Evalset scenarios are shown in the upper and lower panels. Each variant is represented by a *row* and a *red or green* tile depicts a deleterious or benign prediction, respectively, by the corresponding classification model. A–Prediction of the classification models for Del_EvalSet–Benign_EvalSet variants (1990 deleterious variants; 1990 benign variants). B–Prediction of the classification models for Del_EvalSet–PopSpe_EvalSet variants (983 deleterious variants; 1062 benign variants).
(TIF)

S1 Table. List of missense features.
(XLSX)

S2 Table. Generation of VarData set for the training and testing of MISTIC. Different filters were applied in order to generate balanced positive set (high-confidence deleterious variants) and negative set (benign variants). a–Selection of variants with a "Pathogenic" information in the clinical significance (CLNSIG) INFO tag in ClinVar VCF file. b–Selection of at least two-stars high-confidence variants with either 'criteria_provided', '_multiple_submitters', 'reviewed_by_expert_panel', 'practice_guideline' or 'no_conflicts' information in the clinical review status (CLNREVSTAT) INFO tag in ClinVar VCF file. c–Selection of high-confidence missense variants with a Disease-Mutation (DM) STATUS INFO tag in HGMD Pro VCF file. d–Selection of missense variants with a depth coverage $>$ 30X and absent from ClinVar and HGMD Pro databases. e–Filtering of variants that overlap any of the training set variants of SIFT, PolyPhen-2, Condel, VEST4, CADD, MetaLR/MetaSVM.
(XLSX)

S3 Table. Generation of evaluation sets for the benchmark of prediction tools. N/A: Not Applicable. *: Mean value per exome.
(XLSX)

S4 Table. List of recommended thresholds used for the prediction tools. *—There was no recommended threshold for Eigen. Hence, by default the threshold of Eigen was set to 0. N/A: Not Applicable.
(XLSX)

S5 Table. List of metrics used to compare the performance of the prediction tools.
(XLSX)

S6 Table. Selection of features for MISTIC models. The features for the Random Forest and the Logistic Regression models, integrated in the MISTIC Soft Voting system, were selected after applying a Recursive Feature Elimination method.
(XLSX)

S7 Table. List of grid-searched optimized hyper-parameters for the Soft Voting systems of MISTIC. These values were obtained after 20 iterations and 5 cross-validations.
(XLSX)

S8 Table. Features integrated in the Soft Voting system of MISTIC. The cells highlighted in green correspond to the most important features for each algorithm.
(XLSX)

S9 Table. Benchmark metrics of missense prediction tools on VarTest set. Best scores are in bold.
(XLSX)

S10 Table. Benchmark of prediction tools on ClinVarNew deleterious missense variants and BenignEvalSet rare benign missense variants. Best scores are in bold.
(XLSX)

S11 Table. Benchmark of prediction tools on DoCM deleterious missense variants and BenignEvalSet rare benign missense variants. Best scores are in bold.
(XLSX)

S12 Table. Benchmark of prediction tools on deleterious sets of missense variants and population-specific benign missense variants. Best scores are in bold.
(XLSX)

S13 Table. Evaluation of missense prediction tools on simulated disease exomes. The p-value results from the statistical test comparing the results of MISTIC to other prediction tools. The best performances are in bold. N/A: Not Applicable.
(XLSX)

S14 Table. Evaluation of prediction tools on real clinical exomes from the MyoCapture project. The best performances are in bold.
(XLSX)

S15 Table. Evaluation of the contribution of the Soft Voting system in the performance of MISTIC. The best performances are in bold.
(XLSX)

S16 Table. Evaluation of the contribution of the source of deleterious variants in the performance of MISTIC. The best performances are in bold.
(XLSX)

S17 Table. Evaluation of the contribution of the source of benign variants in the performance of MISTIC. The best performances are in bold.
(XLSX)

S18 Table. Evaluation of the impact of the high confidence training set of deleterious variants on the performance of MISTIC. The best performances are in bold.
(XLSX)

Acknowledgments

We thank the PolyPhen-2, CADD, VEST4, Condel and MetaLR/MetaSVM authors for making their training and testing data readily available.

Author Contributions

Conceptualization: Kirsley Chennen, Thomas Weber, Olivier Poch.

Data curation: Kirsley Chennen, Thomas Weber.

Formal analysis: Kirsley Chennen.

Funding acquisition: Julie Thompson, Jocelyn Laporte, Olivier Poch.

Resources: Xavière Lornage, Arnaud Kress, Johann Böhm, Jocelyn Laporte.

Software: Thomas Weber, Arnaud Kress.

Supervision: Julie Thompson, Jocelyn Laporte, Olivier Poch.

Validation: Xavière Lornage, Johann Böhm, Jocelyn Laporte.

Visualization: Kirsley Chennen, Thomas Weber.

Writing – original draft: Kirsley Chennen.

Writing – review & editing: Julie Thompson, Jocelyn Laporte, Olivier Poch.

References

1. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. Am J Hum Genet. 2015; 97:199–215. https://doi.org/10.1016/j.ajhg.2015.06.009 PMID: 26166479
2. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet. 2013; 14:295–300. https://doi.org/10.1038/nrg3463 PMID: 23478348
3. Rehm HL. Evolving health care through personal genomics. Nature Reviews Genetics. 2017; 18:259–67. https://doi.org/10.1038/nrg.2016.162 PMID: 28138143
4. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward PA, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med. 2013; 369:1502–11. https://doi.org/10.1056/NEJMoa1306555 PMID: 24088041
5. Gahl WA, Markello TC, Toro C, Fajardo KF, Sincan M, Gill F, et al. The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. Genet Med. 2012; 14:51–9. https://doi.org/10.1038/gim.0b013e318232a005 PMID: 22237431
6. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nature Reviews Genetics. 2018; 19:253–68. https://doi.org/10.1038/nrg.2017.116 PMID: 29398702
7. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, Abecasis GR, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014; 508:469–76. https://doi.org/10.1038/nature13127 PMID: 24759409
8. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, 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. Genetics in Medicine. 2015; 17:405–23. https://doi.org/10.1038/gim.2015.30 PMID: 25741868

9. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014; 42 Database issue:D980–985. https://doi.org/10.1093/nar/gkt1113 PMID: 24234437
10. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–6. https://doi.org/10.1038/nature08250 PMID: 19684571
11. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nature Genetics. 2010; 42:30–5. https://doi.org/10.1038/ng.499 PMID: 19915526
12. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics. 2011; 12:628–40. https://doi.org/10.1038/nrg3046 PMID: 21850043
13. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4:1073–81. https://doi.org/10.1038/nprot.2009.86 PMID: 19561590
14. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–9. https://doi.org/10.1038/nmeth0410-248 PMID: 20354512
15. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013; 14 Suppl 3:S3.
16. Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat. 2011; 32:358–68. https://doi.org/10.1002/humu.21445 PMID: 21412949
17. Grimm DG, Azencott C-A, Aicheler F, Gieraths U, MacArthur DG, Samocha KE, et al. The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. Human Mutation. 2015; 36:513–23. https://doi.org/10.1002/humu.22768 PMID: 25684150
18. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet. 2011; 88:440–9. https://doi.org/10.1016/j.ajhg.2011.03.004 PMID: 21457909
19. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46:310–5. https://doi.org/10.1038/ng.2892 PMID: 24487276
20. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015; 24:2125–37. https://doi.org/10.1093/hmg/ddu733 PMID: 25552646
21. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018; 34:511–3. https://doi.org/10.1093/bioinformatics/btx536 PMID: 28968714
22. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics. 2016; 48:214–20. https://doi.org/10.1038/ng.3477 PMID: 26727659
23. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016; 99:877–85. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373
24. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nature Genetics. 2016; 48:1581–6. https://doi.org/10.1038/ng.3703 PMID: 27776117
25. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. The American Journal of Human Genetics. 2018; 103:474–83. https://doi.org/10.1016/j.ajhg.2018.08.005 PMID: 30220433
26. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nature Genetics. 2018; 50:1161. https://doi.org/10.1038/s41588-018-0167-z PMID: 30038395
27. Goldman SA, Warmuth MK. Learning binary relations using weighted majority voting. Mach Learn. 1995; 20:245–71.
28. Breiman L. Random Forests. Machine Learning. 2001; 45:5–32.

29. Collins M, Schapire RE, Singer Y. Logistic Regression, AdaBoost and Bregman Distances. Machine Learning. 2002; 48:253–85.

30. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536:285–91. https://doi.org/10.1038/nature19057 PMID: 27535533

31. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15:1034–50. https://doi.org/10.1101/gr.3715005 PMID: 16024819

32. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; 20:110–21. https://doi.org/10.1101/gr.097857.109 PMID: 19858363

33. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics. 2009; 25:i54–62. https://doi.org/10.1093/bioinformatics/btp190 PMID: 19478016

34. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nature Genetics. 2019; 51:88. https://doi.org/10.1038/s41588-018-0294-6 PMID: 30531870

35. Samocha KE, Kosmicki JA, Karczewski KJ, O'Donnell-Luria AH, Pierce-Hoffman E, MacArthur DG, et al. Regional missense constraint improves variant deleteriousness prediction. preprint. Genomics; 2017. https://doi.org/10.1101/148353

36. Kawashima S, Kanehisa M. AAindex: amino acid index database. Nucleic Acids Res. 2000; 28:374. https://doi.org/10.1093/nar/28.1.374 PMID: 10592278

37. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17:122. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795

38. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016; 17:118. https://doi.org/10.1186/s13059-016-0973-5 PMID: 27250555

39. 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. 2016; 37:235–41. https://doi.org/10.1002/humu.22932 PMID: 26555599

40. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet. 2014; 133:1–9. https://doi.org/10.1007/s00439-013-1358-4 PMID: 24077912

41. Ainscough BJ, Griffith M, Coffman AC, Wagner AH, Kunisaki J, Choudhary MN, et al. DoCM: a database of curated mutations in cancer. Nat Methods. 2016; 13:806–7. https://doi.org/10.1038/nmeth.4000 PMID: 27684579

42. UK10K Consortium, Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. The UK10K project identifies rare variants in health and disease. Nature. 2015; 526:82–90. https://doi.org/10.1038/nature14962 PMID: 26367797

43. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2013; 42:111–27. https://doi.org/10.1093/ije/dys064 PMID: 22507743

44. Moayyeri A, Hammond CJ, Hart DJ, Spector TD. The UK Adult Twin Registry (TwinsUK Resource). Twin Res Hum Genet. 2013; 16:144–9. https://doi.org/10.1017/thg.2012.89 PMID: 23088889

45. Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, et al. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. European Journal of Human Genetics. 2017; 25:1253–60. https://doi.org/10.1038/ejhg.2017.130 PMID: 28832569

46. John SE, Antony D, Eaaswarkhanth M, Hebbar P, Channanath AM, Thomas D, et al. Assessment of coding region variants in Kuwaiti population: implications for medical genetics and population genomics. Sci Rep. 2018; 8:1–30.

47. Alkan C, Kavak P, Somel M, Gokcumen O, Ugurlu S, Saygi C, et al. Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. BMC Genomics. 2014; 15:963. https://doi.org/10.1186/1471-2164-15-963 PMID: 25376095

48. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. https://doi.org/10.1038/nature11632 PMID: 23128226

49. Böhm J, Schneider R, Malfatti E, Schartner V, Lornage X, Nelson I, et al. Integrated analysis of the large-scale sequencing project "Myocapture" to identify novel genes for myopathies. Neuromuscular Disorders. 2017; 27:S195.

50. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio]. 2013. http://arxiv.org/abs/1303.3997. Accessed 19 May 2019.

51. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11.10.1–33.

52. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning. New York, NY: Springer New York; 2013. https://doi.org/10.1007/978-1-4614-7138-7

53. Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics. 2012; 13 Suppl 4:S2.

54. Miyazawa S, Jernigan RL. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng. 1993; 6:267–78. https://doi.org/10.1093/protein/6.3.267 PMID: 8506261

8.3 Contraintes et précisions supplémentaires sur les résultats obtenus

Dans cette section j'ai souhaité détailler certains résultats présentés dans les *Supplementary Materials* de la publication de MISTIC, et présenter certaines difficultés rencontrées durant son développement.

8.3.1 Données d'entrainement et d'évaluation

8.3.1.1 Construction des jeux d'évaluation

Un effort tout particulier a porté sur la construction de plusieurs jeux d'entrainement et d'évaluation. Ainsi, notre jeu d'entrainement VarTrain est composé de faux-sens issus de gnomAD (jeu négatif) et de deux sources complémentaires de variants délétères : ClinVar et HGMD (jeu positif). De même, nous avons cherché à multiplier les sources et jeux d'évaluation pour bien cerner les forces et limites de MISTIC. Cette volonté a abouti à la création d'un jeu de variations délétères (Del_EvalSet) et de deux jeux distincts de variations de population : le premier avec MAF (Benign_EvalSet) et le second sans MAF (PopSpe_EvalSet). Del_EvalSet est composé de ClinVarNew, jeu de variations récentes, mais surtout jamais utilisées par aucun prédicteur, ni méta-prédicteur et de DoCM, variations observées en cancer. Benign_EvalSet vise à évaluer la performance des algorithmes sur des variations supposées bénignes et à MAF variables (Figure 47A) tandis que PopSpe_EvalSet permet d'évaluer la performance sur des variations observées dans des projets de séquençage annexes (UK10K, SweGen, WesternAsia), mais sans MAF disponible (Figure 47B).

8.3.1.2 Contraintes rencontrées

La construction de ces jeux a été une étape aussi importante que délicate sur les plans méthodologiques et informatiques. En effet, afin de limiter au maximum la circularité de type 1 (section 4.6.3), nous devions impérativement accéder aux jeux d'entrainement de l'ensemble des prédicteurs et outils. Ceci a été réalisé en cherchant directement les jeux dans les publications et archives, lorsqu'ils étaient disponibles, ou en contactant les auteurs lorsque cela n'était pas le cas. Nous avons ensuite dû filtrer l'intégralité des jeux de données déjà utilisés, mais également dû procéder à une vérification de l'absence de recouvrement totale entre notre jeu d'entrainement et les différents jeux d'évaluation constitués.

Une autre difficulté, qui constitue une limitation inhérente aux méta-prédicteurs, est le nombre limité de données disponibles après avoir filtré l'ensemble des données « récursives ». C'est pour cette raison que parmi les quelques 91 000 variations faux-sens

délétères répertoriées, seules 11 190 ont pu être exploitées dans MISTIC à la fois pour l'entrainement et en tant que jeu de test résiduel.

Enfin, la dernière contrainte est d'ordre purement informatique. Durant le développement de MISTIC, je n'avais pas connaissance de la librairie *hail* (7.4.3), outil de traitement de génomique apte à manipuler des millions de variations de façon optimale. Cette non-connaissance a entrainé de nombreuses heures d'accès et de traitement des données pour parcourir des ressources telles que gnomAD, CADD ou dbNSFP qui comportent toutes plusieurs dizaines de millions de lignes.

8.3.2 Algorithmes employés

8.3.2.1 Sélection des deux algorithmes utilisés dans MISTIC

L'utilisation de boites à outils en intelligence artificielle (section 4.1 et 7.1) permet un accès aisé à un grand nombre d'algorithmes de classification. Comme présenté dans la publication, les deux algorithmes retenus dans notre système de vote ont été la forêt aléatoire (*Random Forest* ; RF) et la régression logistique (*Logistic Regression* ; LR). Ce choix s'est révélé délicat, étant donné notre objectif de performance sur des variants faux-sens en présence et en absence d'un critère aussi sélectif et déterminant que la MAF.

Dès lors, notre approche a été de retenir les algorithmes ayant les meilleures performances, dont les valeurs d'aire sous la courbe ROC dans chacun des deux scénarios, en l'occurrence : la forêt d'arbre (*random forest* ; RF), le AdaBoost (AB) et la régression logistique (*logistic regression* ; LR) (Figure 47). Puis, pour identifier le nombre d'outils à incorporer dans un système de vote, nous avons évalué la concordance et la complémentarité des trois algorithmes (RF, AB, LR). Finalement, pour maintenir un maximum de stabilité à notre outil, nous avons retenu les algorithmes RF et LR et exclu le AB, ce dernier apportant un grand nombre de faux-positifs en absence de MAF. Cette combinaison permet de tirer profit des capacités du RF pour évaluer les variants en présence de MAF avec une sensibilité élevée (aptitude d'identifier les variations délétères) et la capacité du LR à identifier une variation de population en présence et en absence de MAF (Figure 47D).

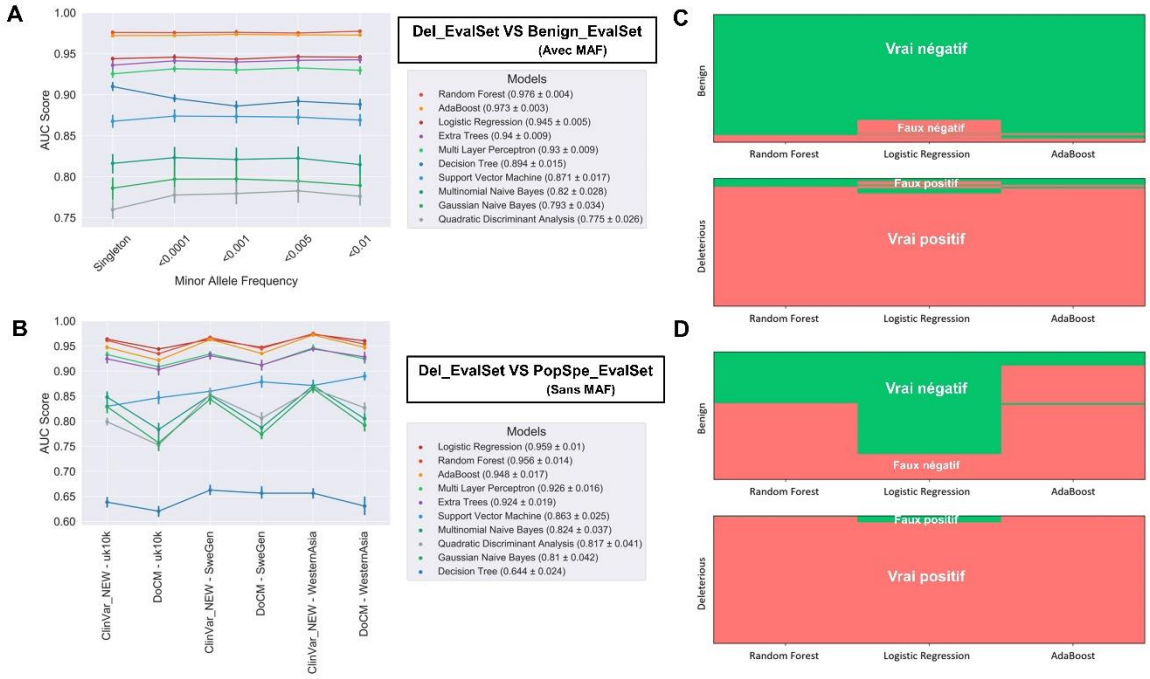


Figure 47 – Performance des différents algorithmes évalués sur différents scénarios

A et B – AUC (Area Under Cuve) de la courbe ROC (Receiver Operating Curve) évaluant la combinaison sensibilité/spécificité de 10 algorithmes à différents niveaux de seuil de classification, en présence de MAF sur le jeu Del_EvalSet versus Benign_EvalSet (A) et en absence de MAF sur le jeu Del_EvalSet versus PopSpe_EvalSet (B), (C) et (D) – concordance entre les 3 algorithmes les plus performants sur l'évaluation réalisée en A et B sur les jeux Del_EvalSet-Benign_EvalSet (C) et Del_EvalSet-PopSpe_EvalSet (D)

8.3.2.2 Contraintes rencontrées

La principale difficulté rencontrée lors de la sélection des algorithmes est reliée au temps de calcul. En effet, chaque algorithme annoté de classification présente des temps d'exécution très différents, de l'ordre de quelques secondes pour la LR contre plus d'une heure pour un SVM sur notre jeu de données. Bien que ces temps paraissent raisonnables, il est important de prendre en compte ces paramètres lors de l'optimisation des hyperparamètres sur une grille (nécessitant la génération de plusieurs dizaines de modèles) ou lors de la sélection de paramètres par RFE.

8.3.2.3 Perspectives d'amélioration dans la sélection des algorithmes

De plus en plus d'outils utilisent des méthodes d'ensemble dans des buts de prédiction de problèmes biologiques transverses (Mirza et al. 2019) mais, peu ont combiné des algorithmes différents pour améliorer leurs performances (Tableau 5). Par exemple, un des outils récents les plus performants, ClinPred, a exploité de manière découplée deux méthodes d'ensemble (section 4.1 ; RF et XGBoost) afin d'améliorer sa sensibilité (section 4.6.2). MISTIC, grâce à son système de vote (en soi également une méthode d'ensemble) combine de manière originale deux approches algorithmiques très distinctes et

complémentaires (LR et RF) et atteste par ses performances que cette démarche peut résoudre des situations complexes.

Malgré la combinatoire que cela implique, il serait utile de mettre en place les moyens de réaliser des études comparatives approfondies évaluant l'ensemble des combinaisons de descripteurs et d'algorithmes en présence/absence d'un critère sélectif important afin d'identifier les combinaisons les plus performantes face à des situations tranchées. Cela pourrait être réalisé par un système d'identification automatique qui réponde à un besoin transverse de performance. Ainsi, dans le cadre particulier de la prédiction de l'impact des variations, un protocole est présenté en Figure 48 où l'ensemble des combinaisons algorithmes/descripteurs serait testé deux fois, une première fois sur des variants avec le critère sélectif (en l'occurrence, la MAF) et une seconde fois, sur des variants sans MAF. Différentes méthodes de sélection des descripteurs détaillées dans la section 8.3.3.2 pourraient être utilisées. Enfin, on pourrait comparer les performances des deux combinaisons d'associations après couplage dans un système de vote final (scénario 1 dans la figure) ou de manière indépendante (scénario 2). Ce système d'identification se place dans la continuité des problèmes liés notamment, à la gestion des données manquantes (Mirza et al. 2019). En effet, l'apprentissage d'un algorithme ne peut se réaliser sur des valeurs manquantes, c'est pourquoi nous avons dû attribuer par défaut une valeur MAF nulle (0) aux variations non documentées dans gnomAD (variation sans MAF). En distinguant clairement les deux situations avec et sans MAF dans la constitution des modèles, ce système éviterait cet ajout de valeurs nulles, qui peut pour certaines méthodes faire office « d'étiquette ».

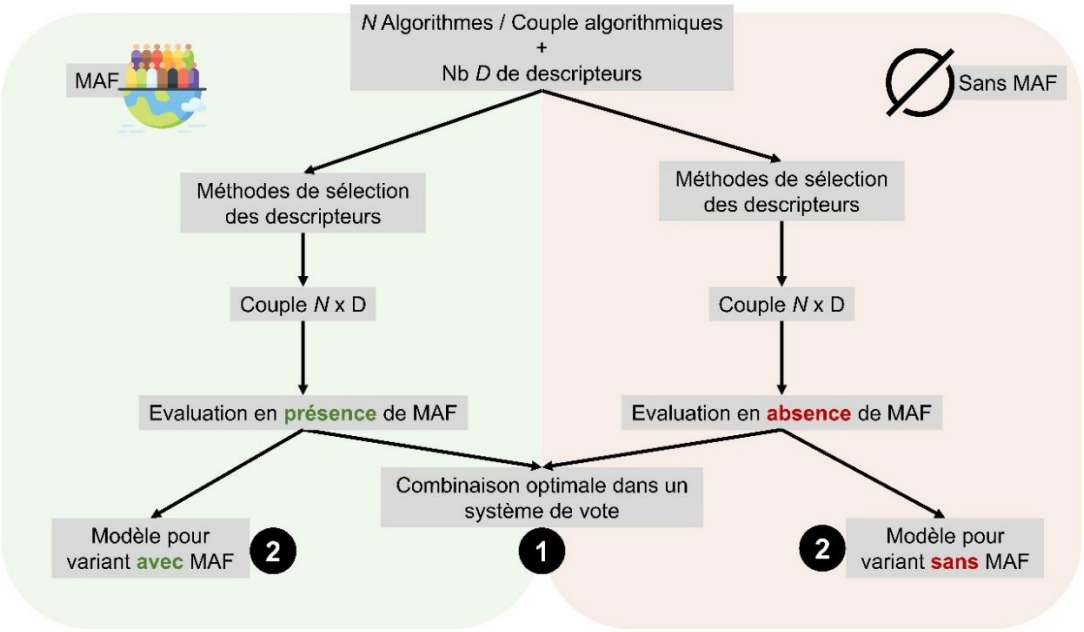


Figure 48 –Protocole proposé permettant une sélection automatisée des couples algorithmiques

L'objectif de ce protocole est d'évaluer l'ensemble des possibilités algorithmiques face à un critère sélectif : dans notre cas la MAF. Bien que cela nécessite un temps de calcul important, on pourrait ainsi évaluer la quasi-totalité des couples algorithmes/descripteurs permettant de répondre de la manière optimale à ces deux situations tranchées et potentiellement s'affranchir de l'ajout de valeurs nulles en absence de MAF. Le scénario 1 correspond à un système de vote final, qui pourrait être pondéré, et intégrerait les deux modèles optimaux générés de manière indépendante. Le scénario 2, quant à lui, correspond à l'utilisation de deux modèles (pouvant eux-mêmes intégrés des systèmes de vote intermédiaires) qui répondrait de manière optimale au critère présence/absence de MAF.

8.3.3 Descripteurs

8.3.3.1 Sélection de descripteurs via RFE

Il est connu que l'intégration d'un grand nombre de descripteurs très corrélés et redondants peut entrainer des chutes de performance et une baisse de la stabilité lors de l'entrainement de certains types de modèles en intelligence artificielle, ce qui est parfois appelé le « fléau de la dimensionalité » (Mirza et al. 2019; D. Chen et al. 2013). Afin de limiter ce problème et d'améliorer les performances du modèle, nous avons eu recours à la technique « d'élimination récursive de descripteurs » (*Recursive Features Elimination* ; RFE ; section 7.1.2) appliquée aux 714 descripteurs préalablement retenus. L'utilisation du RFE a été possible car les algorithmes sélectionnés (LR et RF) renseignent sur le poids attribué à chacun des descripteurs utilisés lors de l'entrainement. Ces deux algorithmes étant combinés dans un système de vote souple, nous avons pu utiliser l'union des deux ensembles de descripteurs retenus (109 pour la LR et 16 pour la RF) afin d'obtenir une liste finale de 113 descripteurs non redondants intégrés dans le modèle de MISTIC (Figure 49). La RF s'est montrée la plus impactée par la redondance avec une amélioration des performances de l'ordre de 4% des F1 Scores entre le jeu comprenant la totalité des descripteurs (714) et le jeu optimal de 16 descripteurs.

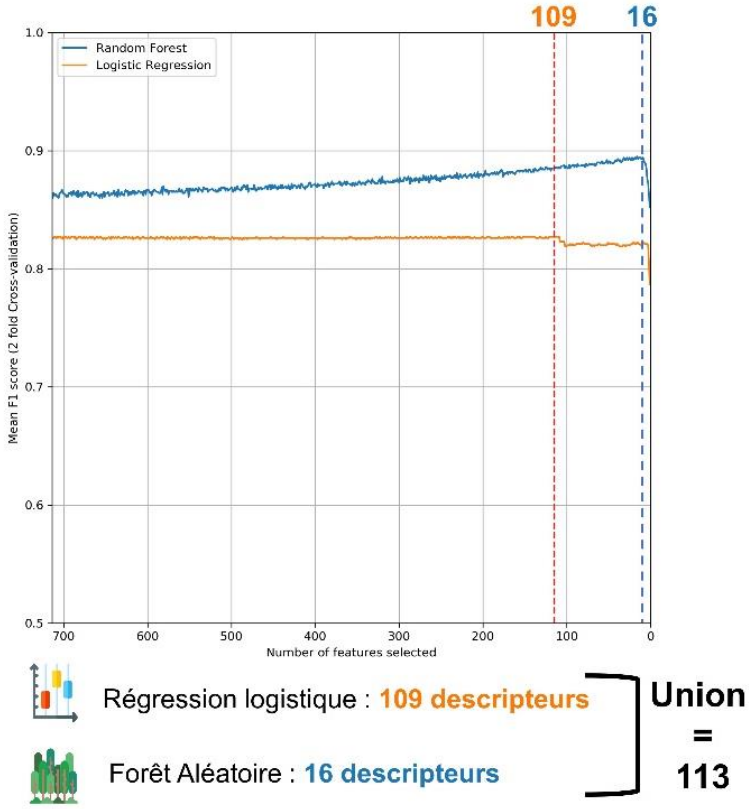


Figure 49 – Sélection du nombre optimal de paramètres par la méthode RFE

La forêt aléatoire (RF) et la régression logistique (LR) ont été utilisées dans le cadre du RFE. Le système de vote finale implémenté dans scikit-learn nécessitant un nombre de paramètres identique, nous avons réalisé l'union des deux combinaisons de descripteurs optimales respectives du RF et du LR. Ceci a donné lieu à un ensemble final de 113 descripteurs intégré dans le modèle de MISTIC.

8.3.3.2 Autres pistes pour la recherche de descripteurs

Dans le cadre du développement d'une nouvelle version de MISTIC, il serait intéressant de tester d'autres méthodes de sélection des descripteurs. Ces méthodes telles que : « la sélection séquentielle de descripteurs » (*Sequential Feature Selector* ; SFS) qui est l'inverse du RFE, avec un ajout itératif du descripteur le plus informatif ou bien encore, « la sélection exhaustive de descripteur » (*Exhaustive Feature Selector* ; EFS ; Figure 50) où chaque combinaison de descripteurs est évaluée. Cette dernière approche nécessite de restreindre au préalable le nombre de descripteurs face à l'explosion de la combinatoire. Il serait également intéressant d'évaluer des méthodes de réduction de dimensionnalité, telles que : l'analyse en composantes principales (*Principal Component Analysis* ; PCA), l'analyse discriminante linéaire (LDA) ou les auto-encodeurs (Hira et Gillies 2015; Z. Chen et al. 2018; Ippolito 2019).

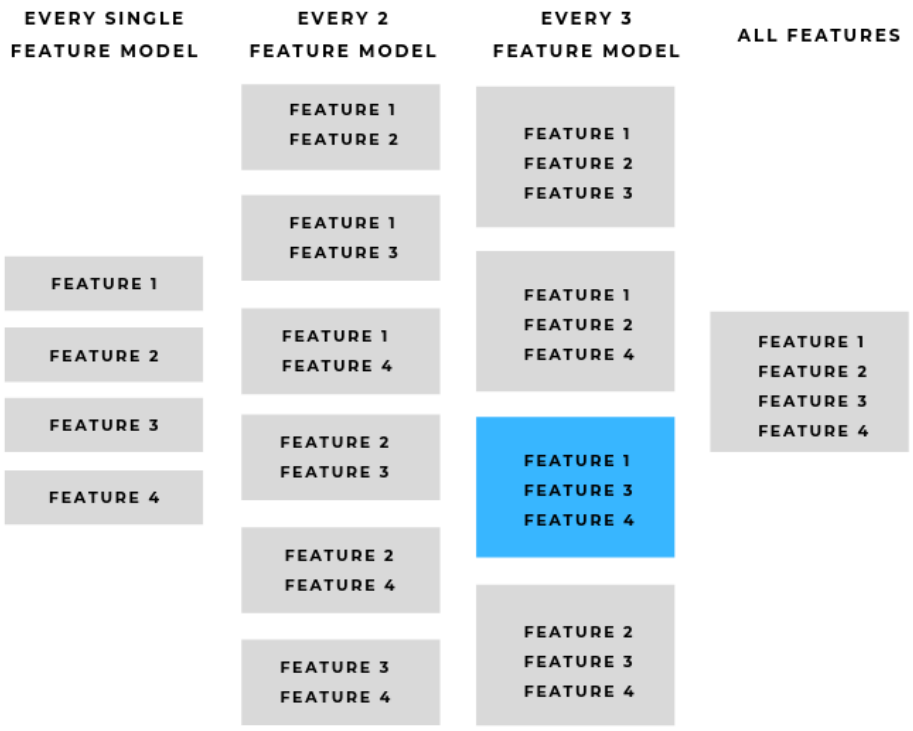


Figure 50 – Principe de la sélection exhaustive de descripteurs

Toutes les combinaisons sont réalisées à partir de 4 descripteurs. La solution optimale est colorée en bleu.

8.3.3.3 Importance des descripteurs pour les deux algorithmes combinés

La capacité de MISTIC à distinguer les variations délétères des variations bénignes dans de nombreux scénarios s'explique à la fois par la complémentarité algorithmique (décrite dans la section 8.3.2.1 avec la concordance illustrée Figure 47C-D), mais également par la complémentarité des descripteurs sélectionnés pour les algorithmes RF et LR. En effet, seuls quatre paramètres ont un poids relatif important pour les deux algorithmes retenus (VEST4, MetaLR, MetaSVM, MPC ; Figure 51). Ces descripteurs étant eux-mêmes des prédicteurs, il est logique que l'apprentissage identifie leur importance prédominante pour discriminer les deux classes de variations. Cependant, certaines familles de descripteurs sont spécifiques à un des deux algorithmes employés. Concernant la RF, on observe que la MAF globale, les MAF de certains groupes ethniques ainsi que les mesures de conservation et de contrainte (GERP++, SiPhy, CCR, PhyloP) présentent un poids élevé. À l'inverse, pour la LR, les propriétés biochimiques et physico-chimiques issues de la base de données AAIndex présentent un poids spécifique à cet algorithme. Une étude approfondie des causes d'une distribution de poids aussi tranchée nous renseignerait sans doute sur la structure des algorithmes et des descripteurs afin d'en déduire des règles utilisables dans d'autres projets.

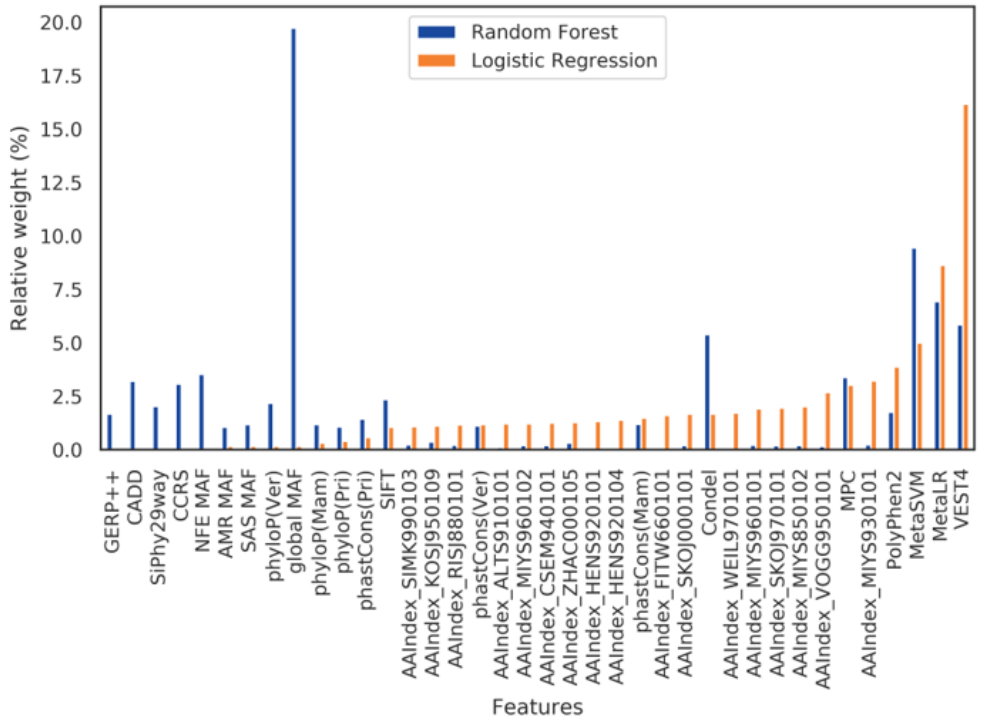


Figure 51 – Poids relatif normalisé des différents descripteurs intégrés dans MISTIC

MAF—minor allele frequency, AMR—Latino American population, NFE—Non-Finnish European population, SAS–South Asian population, CCRS—Constrained-Coding RegionS.

8.3.3.4 Contraintes rencontrées

Concernant les descripteurs, les difficultés rencontrées ont essentiellement porté sur l'annotation des variations à partir de différentes sources de données. Avec plus de 74 millions de variations associées et plus de 600 de scores et informations connexes, dbNSFP a été la ressource principale à manipuler, tandis que certaines métriques, e.g. CCR, devaient être gérées indépendamment. Il a été important de toujours veiller aux assemblages employés par les différents descripteurs pour ne pas réaliser de mélange entre la version GRCh37, majoritairement utilisée dans le développement de MISTIC et la version GRCh38.

8.4 Nouveaux résultats

Une partie des résultats obtenus en juillet 2020 dans le cadre de la publication de MISTIC était basée sur ClinVarNew, un jeu de données validées composé de variants récemment répertoriés dans ClinVar (entre septembre 2018 et avril 2019). Ainsi, nous avons pu comparer la performance de MISTIC et des autres outils sur des variants jamais employés par aucun prédicteur, ni méta-prédicteur.

Avec un objectif identique, nous avons réitéré ce protocole en travaillant sur les variations délétères répertoriées dans ClinVar entre avril 2019 et août 2021. Ainsi, nous avons identifié 8 846 faux-sens délétères pour lesquels les prédicteurs les plus performants (M-CAP, ClinPred, REVEL) et deux nouveaux prédicteurs publiés depuis (LIST-S2 et MetaRNN, Tableau 5) fournissent un score de prédiction. Concernant les variations supposées non-délétères, nous avons constitué deux jeux afin de reproduire le protocole de la publication : un premier jeu constitué de variations de population avec MAF, et un second jeu constitué de variations sans MAF. L'ensemble des variations pour lesquelles un recouvrement entre jeu positif et négatif a été observé ont été exclues de l'évaluation.

Après évaluation sur les variations délétères et les variations de population avec MAF (Figure 52A et C), on constate que MISTIC présente encore aujourd'hui la meilleure performance, tant au niveau de l'aire sous la courbe ROC (aire = 0.923), que sous la courbe de *Precision-Recall* (aire = 0.939). REVEL et ClinPred sont toujours les outils les plus performants derrière MISTIC dans les deux métriques. En ce qui concerne l'évaluation sur les variations délétères et les variations sans MAF (Figure 52B et D), MISTIC se classe deuxième (ROC AUC : 0.849, PR AUC : 0.854) juste derrière REVEL (ROC AUC : 0.857, PR AUC : 0.866). Ainsi, MISTIC reste l'outil le plus performant en termes de classification, un an après sa publication et malgré l'arrivée de nouveaux prédicteurs de faux-sens.

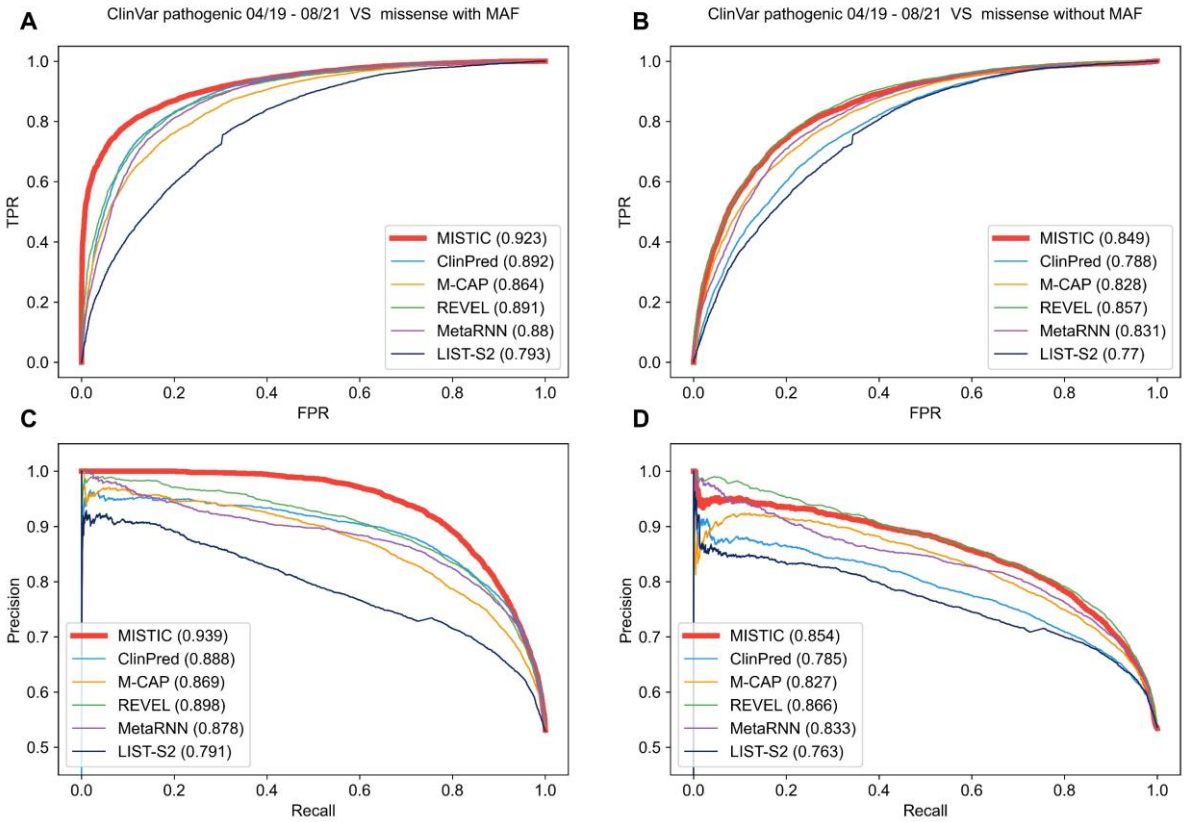


Figure 52 – Évaluation de MISTIC sur des faux-sens récents de ClinVar

A, C – Évaluation sur des faux-sens récents (04/19 – 08/21) contre des variants de population avec MAF.
B, D – Évaluation sur des faux-sens récents contre des variants sans MAF.
A, B – courbe ROC (Receiver Operating Characteristic) évaluant la performance des modèles de classification binaires à différents seuils de classification en fonction du taux de vrais positifs (True Positive Rate ; TPR) et du taux de faux positifs (False Positive Rate ; FPR).
C, D – courbe PR (Precision Recall) évaluant la performance des modèles de classification binaires à différents seuils de classification en fonction de la précision et du rappel/sensibilité.
Métriques présentées en Figure 30.

Enfin, dans le cadre de la 6ème édition du *Critical Assessment for Genome Interpretation* (CAGI) visant à évaluer la pertinence des scores des prédicteurs, nous avons eu la chance de participer à la compétition « *Annotate all missense* » et d'attribuer une probabilité à l'ensemble des 74 278 013 variants faux-sens « potentiels » du génome listés dans dbNSFP. On remarque sur la Figure 53 que MISTIC est un des rares outils à présenter une couverture complète de l'ensemble des faux-sens théoriques. Certains outils comme ClinPred ou MutationAssessor n'attribuent pas de score à 22% des faux-sens, soit plus de 16 millions de variations.

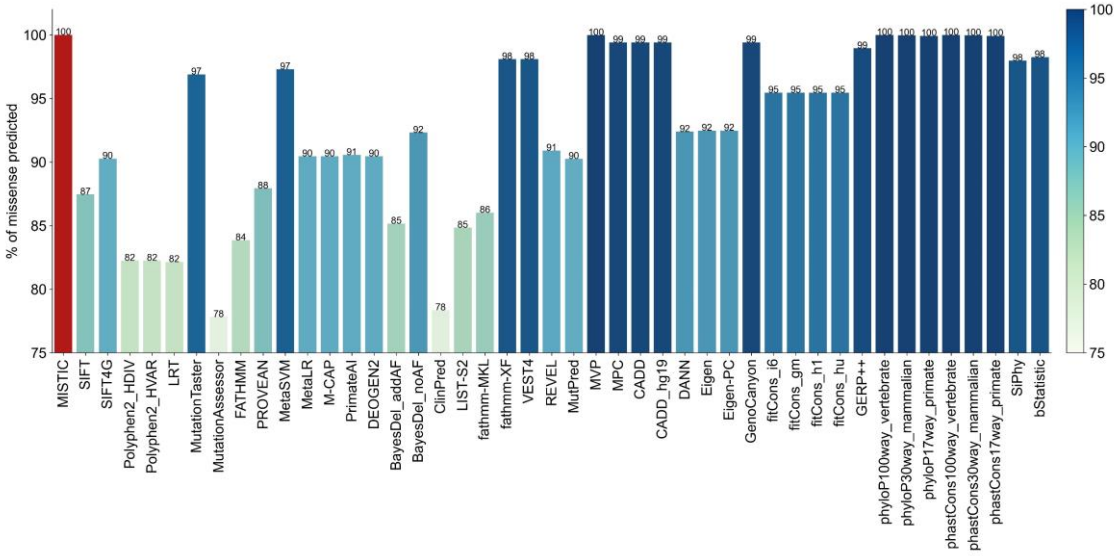


Figure 53 – Pourcentage de faux-sens évalués par prédicteur

MISTIC est identifié en rouge. La couleur des autres prédicteurs dépend de leur niveau de couverture selon les 74 278 013 faux-sens du génome humain disponible dans dbNSFP.

8.5 Conclusion et perspectives

8.5.1 La sensibilité à tout prix ?

Grâce à son architecture originale, MISTIC possède une performance élevée et balancée dans différents scénarios (variants avec/sans MAF, analyse d'exome…). Il répond ainsi à deux problématiques récurrentes, clairement identifiées par l'ACMG, qui sont liées aux faibles spécificités des prédicteurs : un nombre élevé de faux-positifs et des listes étendues de variations à signification inconnue. À titre d'exemple, on peut rappeler qu'un outil comme M-CAP (section 4.6.2) possède une sensibilité très élevée, optimisation souhaitée par les auteurs pour détecter un maximum de variations délétères, mais au prix d'une faible spécificité et d'un nombre élevé de faux positifs. À mes yeux, ce type d'approche s'éloigne des logiques de diagnostic clinique (section 3.5), où l'objectif majeur est d'identifier, rapidement si possible, la variation responsable de la pathologie, en conservant un maximum de fiabilité et sans générer une liste importante de faux positifs.

8.5.2 Les méta-prédicteurs, solution universelle ?

Un des points qui me paraît essentiel à approfondir est le développement des méta-prédicteurs construits à partir de méthodes de classification par ensemble. Dans le cadre de MISTIC, nous avons réalisé un travail poussé de traitement des données d'entrainement afin de minimiser au maximum la circularité de type 1 (variants déjà utilisés dans l'entrainement de prédicteurs incorporés ; section 4.6.3). Ceci a certainement contribué à la limitation du surapprentissage et aux performances robustes du modèle. Cependant, le nombre de méta-prédicteurs continue d'augmenter et l'imbrication récursive et circulaire (section 4.3.3) de ces méta-prédicteurs dans le modèle de nouveaux outils tend certainement vers une limite. On sait, en analysant les poids statistiques associés aux descripteurs selon les deux algorithmes de MISTIC (Figure 51), que les méta-prédicteurs tiennent une part importante dans le score final fourni. Ainsi, il serait intéressant d'évaluer si la juxtaposition de différents types d'information reste judicieuse à l'avenir ou si une stratification des informations s'impose. Au regard des poids relatifs de certains descripteurs (MAF, contraintes/conservation, AAindex…), on peut se demander si ces informations ne devraient pas être présentés à une intelligence artificielle différemment qu'un score issu d'un prédicteur ou méta-prédicteur. Ceci pourrait être réalisé *via* l'utilisation d'un système d'apprentissage par renforcement à base de règles (par exemple : système de classeurs développé au sein de l'équipe) ou par une méthode d'apprentissage automatique permettant de stratifier et ordonner les classes et niveaux d'information (Urbanowicz et Moore 2009).

8.5.3 La puissance de la MAF comme descripteur

Il a été évoqué en introduction (sections 3.3.3 et 2.2.2) que la MAF globale dans la population mondiale ne représente qu'une image approximative des fréquences des variations génétiques. Tout d'abord, il est connu que la composition de gnomAD (section 2.2.1) est largement biaisée par la prépondérance des populations européennes non-finlandaises (50% des individus étudiés) vis-à-vis d'autres ethnies. Concernant les variations responsables de maladies génétiques rares, les disparités de fréquences inter-ethniques nécessitent de considérer, non seulement, la MAF globale, mais également les MAF des groupes ethniques (section 3.3.3). Ces valeurs ont déjà été intégrées dans MISTIC et représentent un poids conséquent, mais il pourrait sans doute constituer un élément déterminant dans le système à base de règles mentionné au-dessus, à même de mieux stratifier les informations.

Sans conteste, l'importance de la MAF va s'amplifier au fil des années et du nombre d'individus séquencés. Comme mentionné en section 2.2.1, plus de la moitié des variations séquencées dans ExAC (plus de 60 000 exomes issus d'individus sains) ont été identifiées comme privées et spécifiques à un individu. Ces proportions ont vocation à évoluer fortement à l'avenir améliorant au passage, l'identification des régions fortement contraintes (privées de variations) et la précision des MAF *via* une distribution plus fine des fréquences. J'ai la conviction que ces améliorations découlant directement des technologies et projets de séquençage vont profondément rebattre les cartes à l'avenir. Cette simplification du paysage des descripteurs face à la puissance d'une MAF revisitée facilitera la recherche de nouvelles sources d'informations réellement orthogonales à intégrer dans les modèles. Ces nouvelles sources pourront provenir de données génomiques comme nous l'avons fait dans MISTIC avec l'ajout des CCR ou de données générées par les approches de génomiques fonctionnelles ou « omiques » que nous-mêmes et d'autres auteurs sommes actuellement en train d'explorer et évaluer (section 5.3.3, Chapitre 9). À n'en pas douter, ces innovations couplées à une meilleure exploitation des outils d'intelligence artificielle permettront très prochainement une meilleure identification des variations causales des maladies génétiques rares et une compréhension accrue de leur impact sur les réseaux biologiques.

8.5.4 Vers un changement de paradigme ?

Il est de plus en plus évoqué la possibilité de réaliser des outils ou modèles, pour chacun des gènes codant pour les protéines (Tian et al. 2019). Cette hypothèse de travail peut s'envisager de différentes manières. On peut imaginer, pour les gènes présentant suffisamment de variations, entrainer un modèle spécifique à chaque gène, ou bien,

entrainer un modèle pour l'ensemble des gènes dont les probabilités seraient recalibrées à façon de manière individuelle selon différents critères (haploinsuffisance du gène, contrainte, section 4.4.1) (Accetturo, Bartolomeo, et Stella 2020).

8.5.5 MISTIC et la gestion des variations au travers des « omiques »

À l'image des protocoles employés par la quasi-totalité des prédicteurs, les prédictions de MISTIC ont été calculées pour fournir une prédiction unique par position génomique, et par voie de conséquence, pour chaque position des exons. Cependant, il est également envisageable d'émettre une prédiction pour les variations faux-sens au regard de leurs « usages » dynamiques dans les cellules. Comme on l'a déjà évoqué, cette évolution s'inscrit dans une exploitation accrue des données « omiques ».

La première étape du passage du génome aux « omiques » se matérialise par l'intégration des données de transcriptomique, permettant d'évaluer si un nucléotide est exprimé et peut potentiellement impacter un tissu (Chapitre 5). Dans ce contexte, on peut noter les développements d'outils récents (CADD-Splice ou MTSplice) intégrant des données d'expression pour prédire l'impact des variations touchant les sites d'épissage. Au niveau de la protéine, on peut également penser à des intégrations futures reprenant les travaux réalisés par des outils tels que MAPP afin de prédire l'impact structural sur les protéines en exploitant des modèles de prédiction obtenus à partir d'*AlphaFold* (Jumper et al. 2021). Enfin, à terme, on peut envisager d'évaluer l'impact d'une variation dans les interactions dynamiques à l'œuvre dans les complexes ou réseaux (N. Zhao et al. 2014).

Chapitre 9. Variations et épissage alternatif

Les variations génétiques sont classiquement étudiées selon leur conséquence au niveau des gènes ou des éléments fonctionnels du génome. Ainsi, on peut caractériser une variation faux-sens au regard du changement d'acide aminé et des modifications induites sur la structure 3D de la protéine ou du complexe, ou évaluer une variation non-codante touchant un *enhancer* ou un site de fixation d'un facteur de transcription selon les perturbations induites sur l'expression de gènes cibles (Chapitre 5). Avec la disponibilité croissante de données omiques, les variations sont de plus en plus considérées sous l'angle de leur impact dans le contexte fonctionnel cellulaire ou tissulaire.

Concernant les variations dans les régions non-codantes, l'exploitation des éléments fonctionnels génomiques et épigénomiques identifiés dans ENCODE (Moore et al. 2020) pour plus de 200 types cellulaires a permis de prédire l'impact potentiel des variations dans ces types cellulaires (Zhou et al. 2018). Concernant les variations touchant les sites d'épissage, grâce au projet de RNA-Seq massif **GTEx** (section 6.4.3), l'exploitation des données d'expression des différentes isoformes de transcrits d'un gène a permis d'évaluer l'impact d'une variation sur les populations tissulaires de transcrits (Cheng et al. 2021). Cependant, à notre connaissance, aucun outil n'évalue l'impact d'une variation non-synonyme (nsSNV) selon l'expression des exons au sein des isoformes de transcrits.

Suite aux multiples projets de séquençage, on sait que de nombreux gènes humains codants pour des protéines sont soumis à l'épissage alternatif dans divers tissus (Pan et al. 2008). Ces travaux ont permis de distinguer, pour chaque gène codant, les exons constitutifs (présents dans tous les transcrits) des exons alternatifs (expression transitoire selon les tissus, stades ou environnements) (section 5.1). Dès lors, on peut aisément comprendre qu'au regard d'une variation nsSNV présente dans un exon, le statut constitutif ou alternatif peut profondément modifier son impact potentiel et les éventuels phénotypes associés. En effet, la variation située dans un exon constitutif suivra le profil d'expression de tous les transcrits et produira des effets dans toutes les situations. À l'inverse, une variation située dans un exon alternatif verra ses effets et phénotypes potentiels dépendre de la présence/absence de l'exon alternatif. C'est cette question, apparemment simple, de la prédiction de l'impact d'une variation en fonction de son expression dans des transcrits d'isoformes que j'ai voulu aborder en développant duxt.

Cependant, il est très vite apparu que poser cette question impliquait de vérifier que l'architecture et les caractéristiques des gènes à multiples transcrits d'isoformes (*Multiple transcript-ISOform Genes* ou MISOG) étaient comparables à celle des gènes sans épissage alternatif (*Single transcript-ISOform Genes* ou SISOG).

9.1 Étude comparative de l'architecture des gènes codant pour des protéines au regard de l'épissage alternatif

9.1.1 Contexte

Plusieurs publications ont estimé que la quasi-totalité (95%) des gènes humains composés de plusieurs exons présentaient des isoformes de transcrits (Wang et al. 2008; Pan et al. 2008). Cependant, divers travaux ont questionné cette proportion face aux événements d'épissage alternatif dus au « bruit transcriptionnel » (Sorek, Shamir, et Ast 2004; Melamud et Moult 2009; Pickrell et al. 2010), aux problèmes techniques liés à la formation de cDNA par transcription inverse/amplification ou au séquençage RNASeq par lectures courtes qui peut biaiser l'identification précise des jonctions exon-exon (Jiang et Chen 2021).

Notre but étant d'identifier les variations présentes au sein d'exons alternatifs avérés, nous avons retenus les transcrits ayant subi une curation dans ressource RefSeq (section 6.1.1). Sur cette base, nous avons effectué une étude statistique comparant gènes à transcrit unique (SISOG) et gènes à multiples transcrits d'isoforme (MISOG). Cette analyse exploratoire a conduit à un résultat surprenant concernant la proportion des MISOG humains (64%), proportion différente des 95% rapportés dans la littérature. Mais surtout, nous avons observé que les architectures des MISOG et SISOG différaient significativement, au regard des nombres et tailles d'exons ainsi que de la structure des régions 5'.

Ces résultats sont présentés dans le manuscrit ci-dessous, en cours d'évaluation par la revue *Genome Biology*, à l'heure où j'écris ces lignes.

9.1.2 Manuscrit

DO 5' REGIONS OF HUMAN PROTEIN-CODING GENES CONTAIN THE BLUEPRINTS FOR ALTERNATIVE SPLICING?

Do 5’ regions of human protein-coding genes contain the blueprints for alternative splicing?

Thomas Weber,[1*] Luc Moulinier,[1,2] Nicolas Scalzitti,[1] Julie D. Thompson,[1] Kirsley Chennen,[1*†] Olivier Poch[1*†]

1 - Complex Systems and Translational Bioinformatics (CSTB), ICube laboratory – CNRS UMR7357, Centre de Recherche en Biomédecine de Strasbourg (CRBS), 1 rue Eugène Boeckel, 67000, France

2 - BiGEst-ICube platform, ICube laboratory, UMR7357, 1 rue Eugène Boeckel, 67000, Strasbourg, France

* - Correspondence: thomas.weber@unistra.fr (T.W.), kchennen@unistra.fr (K.C.), poch@unistra.fr (O.P.)

† - co-last author

Abstract

To investigate alternative splicing capacity, we statistically compared the properties of human protein-coding genes with multiple transcript isoforms (MISOG) and single transcript isoforms (SISOG). Apart from global exon content, differential features are concentrated in the 5’ gene regions, with MISOG presenting complex 5’ untranslated region architecture and a distinctive flanking environment around first 5’ intron. Importantly, we found that 5’ exons are more prone to alternative splicing in MISOG. These results unravel previous observations indicating the importance of 5’ gene regions in some transcriptional processes and call for their reassessment in light of the MISOG/SISOG profiles.

Background

Alternative splicing (AS) (1,2) in eukaryotes is a powerful process to increase a gene's functional diversity by combining different structures and functions into distinct protein isoforms. Thanks to next generation sequencing advances (3), many human isoforms are now available, resulting in an extensive and comprehensive catalogue (4) suitable for the study of isoforms and the understanding of their importance in fundamental processes such as diseases and environmental responses (5). However, to our knowledge, there is no systematic study evaluating the complete set of Human Protein Coding Genes (HPCG) in terms of their ability to produce multiple transcript isoforms. Here, we present a robust multi-level statistical comparison of gene properties, such as the number and length of exons, untranslated regions (UTR), translated exonic regions (TER) and introns, to evaluate whether MISOG and SISOG have different profiles.

Results and discussion

To focus on transcript isoforms of the 19,285 HPCG in the RefSeq database (4), we excluded 1,983 single-TER genes, including 1,610 SISOG and 373 MISOG with AS only in UTRs. The remaining 17,302 genes are composed of 10,995 MISOG and 6,307 SISOG, implying that MISOG are 1.7 times more frequent than SISOG. It is to note that, while 95% of genes have been estimated to undergo alternative splicing (6), only 64% of HPCG present multiple transcript isoforms curated by experts. The final gene set corresponds to 59,290 distinct transcripts (52,983 MISOG / 6,307 SISOG) comprising 230,664 exons (165,585 / 65,079), 213,004 TERs (150,276 / 62,728) and 223,564 introns (164,826 / 58,738) (Tables S1-3).

In terms of median gene lengths, MISOG (36,988 bp) are significantly longer than SISOG (20,617 bp) (Fig. 1, Table S4) and the total genomic length of MISOG (933 Mbp) is 2.9 times greater than SISOG (322 Mbp). This indicates that the larger genomic length of MISOG is not only related to their greater number (1.7 times) but also to distinct MISOG/SISOG gene architecture.

We then performed a detailed statistical analysis of the exon, TER, UTR and intron components. The median number of exons (Fig. 1, Table S5) is significantly greater for MISOG (12 exons) than

SISOG (7 exons). For SISOG, the median number of TERs is also 7, while for MISOG, the median of 11 TERs suggest additional 5' or 3' UTR exons. Considering median lengths (Table S6), MISOG exons (134 bp) and TERs (119 bp) are only slightly shorter than SISOG exons (136 bp) and TERs (126 bp). Thus, the difference observed between mature mRNA lengths of MISOG (3,341 bp) and SISOG (2,688 bp) (Table S1) can be attributed to the larger number of MISOG exons. On the other hand, MISOG introns (1,963 bp) were significantly ~1.5 times longer than SISOG (1,336 bp) suggesting that the intronic regions, which are known to be longer than exonic regions in eukaryotes (7), are key MISOG/SISOG discriminative elements.

As previous studies on human and other eukaryotic genes have shown that the first intron is significantly longer than following ones (8–10), we performed an ordinal MISOG *versus* SISOG analysis for the 5' and 3' gene components (Fig. S1A, Table S7 and S8).

Both MISOG and SISOG exhibit a first 5' intron (Fig. S1A; MISOG: 4,901 bp; SISOG: 2,953 bp) significantly longer than following introns (MISOG: 1,847-3,032 bp; SISOG: 1,341-1,823 bp). However, although introns lengths are divergent, MISOG displayed a much longer first 5' intron than SISOG, especially concerning the upper quartile (MISOG: 19,992 bp; SISOG: 10,314 bp). We observed that the first MISOG and SISOG exons (Fig. S1B, Table S7) are longer (MISOG: 164 bp; SISOG: 195 bp) than following exons (MISOG: 123-125 bp; SISOG: 130-141 bp) and that the first MISOG and SISOG TERs (Fig. S1C, Table S3) are shorter (MISOG: 94 bp; SISOG: 117 bp) than the following TERs (MISOG: 119-124 bp; SISOG: 124-133 bp), suggesting that 5' UTRs might account for the size discrepancies. Concerning the 5' UTRs (Fig. S1D, Table S7), a decreasing length between first and second UTR exons is observed for both MISOG and SISOG. However, the first two 5' UTR exons were found to have significantly longer lengths for MISOG (first and second exons, respectively 129 bp and 70 bp) compared to SISOG (first and second exons, respectively 97 bp and 39 bp). In addition, we noted a huge decrease in SISOG UTR' numbers for positions 3-5 (241 / 67 / 26) compared to positions 1-2 (6,197 / 1,546).

Concerning the last five 3' components, no major differences between MISOG and SISOG were observed for introns regardless of the position (Fig. S1E, Table S8). Both MISOG and SISOG exhibit longer last exons (Fig. S1F; MISOG: 1,537 bp; SISOG: 1,221 bp) and last TERs (Fig. S1G; MISOG:

135 bp; SISOG: 158 bp) than the preceding ones (exons: MISOG: 122-124 bp; SISOG: 129-132 bp; TER: MISOG: 120-121 bp; SISOG: 125-130 bp). Finally, we noted that both MISOG and SISOG exhibit a single long 3' UTR exon (Fig. S1H). However, 3' UTR median exon length was 40% longer for MISOG (1,285 bp) compared to SISOG (918 bp).

To better understand the disparities between 5' and 3' non-coding regions, we analyzed the UTR in more detail and found that 75% of SISOG and only 41% of MISOG exhibit a single 5' UTR exon attached to the CDS (Fig. 2A). Thus, most MISOG (59%) present multiple 5' UTR exons, with the first one being the longest. At the opposite 3' end, almost all MISOG and SISOG present a single long 3' UTR exon attached to the CDS (Fig. 2B), confirming previous observations (10,11). These results are confirmed by considering the total UTR exon length for all genes (Fig. S2), where an 84% increase is observed for MISOG (MISOG: 204 bp; SISOG: 111 bp) at the 5' UTR. Additionally, the length separating the TSS from the start codon was found to be 11 times longer for MISOG (1,577 bp) than for SISOG (143 bp) (Fig. 2C, Table S9).

Finally, by analyzing the exon and TER frequencies across transcript isoforms in MISOG, we found that the closer these elements are to the 5' end, the more likely they are alternatively spliced (Fig. S3 & Table S10). Indeed, the first alternative exon and the first alternative TER are present in only 50% of the isoforms. Moreover, second and third TER are also more alternatively spliced (resp 67% and 86%). Concerning the 3' end, the last TER is found in 67% of isoforms and the last exon in 75%, corresponding to known 3' UTR modulation (10).

Conclusions

In this study, we have shown that 64% of HPCG in the RefSeq database are MISOG, representing 74% of the total genome length covered by coding genes (~1,25 Gbp) and despite a large distribution, the detailed statistical analysis of the number and length of MISOG *versus* SISOG components (exons, TER, UTR and introns) revealed discriminative profiles (Fig. 1). Unsurprisingly, SISOG (n=7) have fewer exons than MISOG (n=12), but all SISOG exons are totally or partially coding while MISOG present at least one 5' UTR exon without any coding part. Thus, the first 5' SISOG and MISOG introns are clearly distinct, since in SISOG it is embedded in the coding region while in

MISOG it is flanked by UTR. Such a distinct flanking environment may echo our observation that 5' MISOG exons are statistically more prone to alternative splicing than the following exons. Finally, we show that the length of the first MISOG intron is almost 2 times longer than the SISOG one. The complex 5' UTR architecture combined with a longer 5' intron induces a fundamental difference of the TSS to start codon distance, which is generally 11 times longer in MISOG than in SISOG.

Various statistical studies have described the specificity of the 5' eukaryotic gene region, notably concerning the properties of the first introns in term of length (8,9), conservation (11) or UTR inclusion (12,13) as well as the presence of *cis*-regulatory elements (14–16) and varying promoter number (17,18). For instance, it has already been noted that intron length impacts transcriptional time delay (19–21) or that the presence of 5′-UTR introns influences the choice of the alternative mRNA export pathway and the final mRNA targeting to endoplasmic reticulum and mitochondria (12). Thus, reanalyzing these distinctive properties in the light of the MISOG/SISOG profiles may allow a better discrimination of the processes and evolution linked to the splicing from those associated to alternative splicing.

Finally, this survey took advantage of the extended and high-quality of human data, and it remains to be seen whether applying the same protocol to other eukaryotic organisms with comprehensive isoform information may corroborate our results. Nevertheless, the finding of discriminatory features is promising as it opens the way to new *in silico* approaches to better predict the transcript isoform gene status and improve genome annotation.

All HPCG data are free to access and structured according to MISOG/SISOG status and all programing notebooks are provided and reusable (see Availability of data and materials).

Methods

All definitions and methods described in the following sections are illustrated in Fig. S4.

Data loading and processing

The RefSeq database (5) was selected as source of the genomic data for its associated quality (good annotation, non-redundancy, and low number of errors). The Human GFF file based on GRCh38p.13

assembly (version 109.20210514) was downloaded and filtered to retain only protein-coding genes and their features.

Using a python (3.7) script (*script: prepare_refseq.py*) based on the pandas library, four intermediate files were created to separately capture the four types of biological concepts (Fig. S4A) associated with protein-coding genes from the GFF files: genes (*feature=gene*), mRNAs (*feature=mRNA*), exons (*feature=exon*), and Translated Exonic Regions (TER) (*feature=CDS*).

Three main filtering steps were applied to curate the data. First, only entries having an identifier prefix of *NC_* were retained. Second, only protein-coding genes were selected (*gene_biotype=protein_coding*). Third, for mRNA, exons and TER features, only curated entries with accession prefixes or parent entities starting with *NM_* or *NP_* were retained.

Of the 42,793 gene entries (file G), 19,456 genes had a prefix NC_ and the attribute *gene_biotype=protein_coding*. After filtering 171 genes with no validated mRNAs (prefix NM_), the total dataset included 19,285 genes. We also excluded 1,983 single-TER genes, comprising 1,610 SISOG (of which 995 were single-exon genes) and 373 MISOG where the AS occurs uniquely in the UTRs.

Gene, mRNA, exon and TER lengths were calculated using the retrieved genomic coordinates. Gene start and end coordinates correspond to Transcription Start (TSS) and Transcription Termination (TTS) Sites respectively.

Definition of MISOG and SISOG

We identified MISOG and SISOG based on the number of mRNAs per gene (respecting previously defined filtering steps). MISOG have at least two non-identical mRNAs while SISOG have only a single mRNA as presented in Fig S4B.

Characterization of gene elements: exons, TER, introns and UTRs

Exons and TER lengths were computed using their respective start and end genomic coordinates. Intron boundaries were computed from exon positions as presented in the Fig. S4C. Based on previous observation (7), introns shorter than 26 bp were not analyzed.

Identification of 5’ / 3’ UTR exons was determined using the transcription strand (+ = forward; - = reverse). UTRs were retrieved for each transcript by comparing exons and TER using a nested loop join algorithm conditioned by the presence of an overlap between an exon and a TER element. If so, UTR boundaries (excluding TER part of the exon) were computed as presented in the Fig. S4D. UTR boundaries were used as a new column in the file listing exons and TER for each mRNA.

For MISOG, gene elements may appear multiple times in RefSeq GFF file due to the data format structuration. To produce robust statistics, we identified and did not consider redundant elements (see Fig. S4E).

Gene element ordinal position

For each gene element, the ordinal position was defined as presented in Fig. S4F with respect to the transcription strand (+ = forward; - = reverse) and its position in the mRNA.

Exons and TER usage frequencies across isoforms

For MISOG, we identified and counted the constitutive (present in all mRNAs) exons and TERs and the alternative ones (present in a subset of mRNAs). Frequencies were computed as presented in Fig. S4G and defined as the number of transcripts where the exon/TER is present divided by the total number of transcripts in the gene.

Statistical analyses

As parametric statistics (mean and standard deviation) are highly sensitive for extreme values and due to the widespread distribution of gene element lengths across all HPCG, all comparisons were performed using non-parametric statistics (median and quartiles). Statistically significant differences were evaluated using standard Mann-Whitney U tests.

Abbreviations

CDS: Coding DNA Sequence; TSS: Transcription Start Site; TTS: Transcription Termination Site; UTR: UnTranslated Region; HPCG: Human Protein Coding Genes; MISOG: Multiple transcript ISOform Genes; SISOG: Single transcript ISOform Genes; NMD: Non-sense Mediated Decay

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

We thank the BiGEst-ICube bioinformatics platforms for their assistance. This work is supported by the Agence Nationale de la Recherche (ELIXIR-EXCELERATE: GA-676559), Institute funds from the CNRS, the Institut Français de Bioinformatique and the Université de Strasbourg.

Availability of data and materials

The code used in this study can be found at https://github.com/weber8thomas/MISOG_SISOG under MIT License. RefSeq raw GFF file was retrieved from FTP site (GCF_000001405.39). The data sets are accessible in compressed CSV, Apache parquet and XLSX formats at https://zenodo.org/record/5546587. All developed programs and notebook were specifically designed to allow facilitated updates on new versions of RefSeq GFF files.

Authors' contributions

T.W., L.M, K.C., and O.P. designed the study. K.C. and O.P. supervised the work. T.W. and N.S. produced the visualizations. T.W., K.C. wrote the manuscript. J.T. and O.P. contributed to revision of the manuscript.

Acknowledgements

No statement

References

1. Zhao S. Alternative splicing, RNA-seq and drug discovery. Drug Discov Today. 2019 Jun;24(6):1258–67.
2. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, et al. Function of alternative splicing. Gene. 2013 Feb 1;514(1):1–30.
3. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019 Nov;20(11):631–56.
4. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016 Jan 4;44(D1):D733-745.
5. Kim HK, Pham MHC, Ko KS, Rhee BD, Han J. Alternative splicing isoforms in health and disease. Pflüg Arch - Eur J Physiol. 2018 Jul;470(7):995–1016.
6. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008 Dec;40(12):1413–5.
7. Piovesan A, Antonaros F, Vitale L, Strippoli P, Pelleri MC, Caracausi M. Human protein-coding genes and gene feature statistics in 2019. BMC Res Notes. 2019 Jun 4;12(1):315.
8. Bradnam KR, Korf I. Longer First Introns Are a General Property of Eukaryotic Gene Structure. PLOS ONE. 2008 Aug 29;3(8):e3093.
9. McCoy MJ, Fire AZ. Intron and gene size expansion during nervous system evolution. BMC Genomics. 2020 May 14;21(1):360.
10. Zhu L, Zhang Y, Zhang W, Yang S, Chen J-Q, Tian D. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics. 2009;12.
11. Park S, Hannenhalli S, Choi S. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics. 2014;15(1):526.
12. Bicknell AA, Cenik C, Chua HN, Roth FP, Moore MJ. Introns in UTRs: Why we should stop ignoring them. BioEssays. 2012;34(12):1025–34.
13. Hong X, Scofield DG, Lynch M. Intron Size, Abundance, and Distribution within Untranslated Regions of Genes. Mol Biol Evol. 2006 Aug 30;23(12):2392–404.
14. Gaunitz F, Deichsel D, Heise K, Werth M, Anderegg U, Gebhardt R. An intronic silencer element is responsible for specific zonal expression of glutamine synthetase in the rat liver. Hepatol Baltim Md. 2005 Jun;41(6):1225–32.
15. Beaulieu E, Green L, Elsby L, Alourfi Z, Morand EF, Ray DW, et al. Identification of a novel cell type-specific intronic enhancer of macrophage migration inhibitory factor (MIF) and its regulation by mithramycin. Clin Exp Immunol. 2011 Feb;163(2):178–88.
16. Jo B-S, Choi SS. Introns: The Functional Benefits of Introns in Genomes. Genomics Inform. 2015 Dec;13(4):112–8.
17. Xin D, Hu L, Kong X. Alternative Promoters Influence Alternative Splicing at the Genomic Level. PLOS ONE. 2008 Jun 18;3(6):e2377.
18. Kolathur KK. Role of promoters in regulating alternative splicing. Gene. 2021 May 25;782:145523.
19. Seoighe C, Korir PK. Evidence for intron length conservation in a set of mammalian genes associated with embryonic development. BMC Bioinformatics. 2011 Oct 5;12(9):S16.
20. Swinburne IA, Silver PA. Intron delays and transcriptional timing during development. Dev Cell. 2008 Mar;14(3):324–30.
21. Carmel L, Chorev M. The Function of Introns. Front Genet. 2012;3:55.

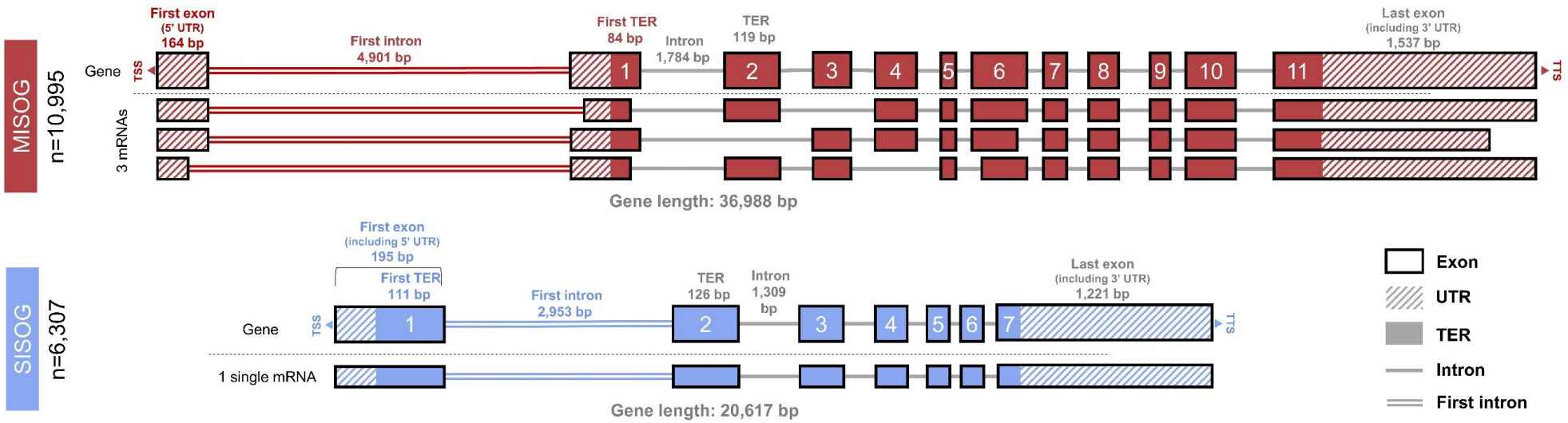


Figure 1 – MISOG and SISOG archetypal gene architecture

Major values discussed in the manuscript are summarized graphically (all values are available in the supplementary Tables). The median lengths of each gene component (exon, TER, intron, UTR) are described according to the values observed for the MISOG (red; median number of 3 mRNAs as transcript isoforms) and SISOG (blue). The major MISOG and SISOG discriminant elements (first 5' introns, 5' exons and 5' TER) are indicated by colored labels. Non-discriminant elements are indicated by grey labels.

Gene lengths was calculated from TSS (Transcription Start Site) to TTS (Transcription Termination Site). TER stands for Translated Exonic Regions, the coding part for amino acids of exons.

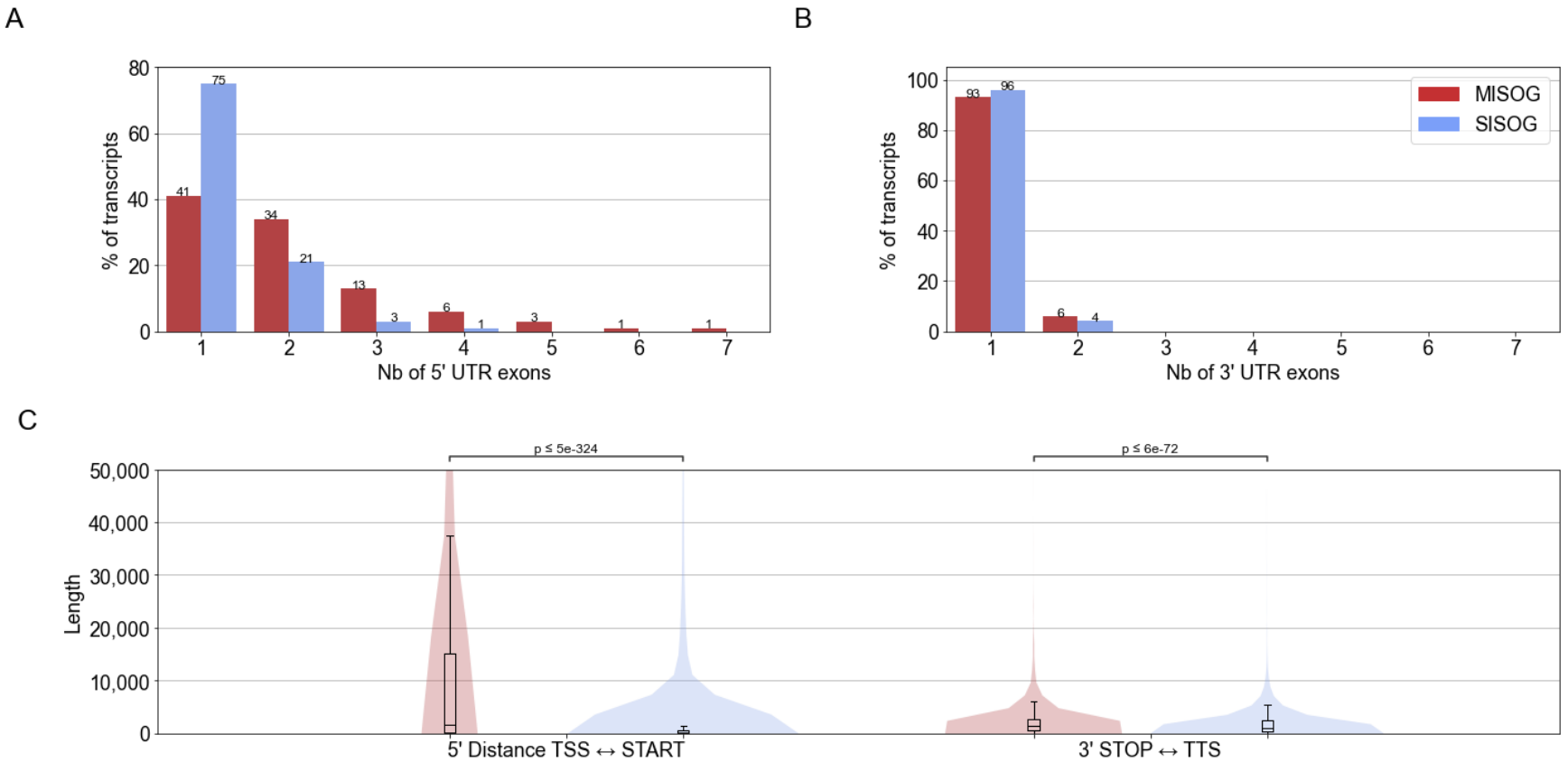


Figure 2 – Comparative analysis of MISOG and SISOG 5' and 3' non-coding regions.

(A-B) barplots indicate the percentage of MISOG/SISOG transcripts exhibiting from one up to seven 5' or 3' UTR exons. (C) Comparison of the MISOG *versus* SISOG median distances separating TSS from START codon (left) and STOP codon from TTS (right) (all values are available in Table S9).

(A-B) represent the percentage of transcript according to the number of exons in 5' (A) or 3' UTR(B).

Supplementary Material

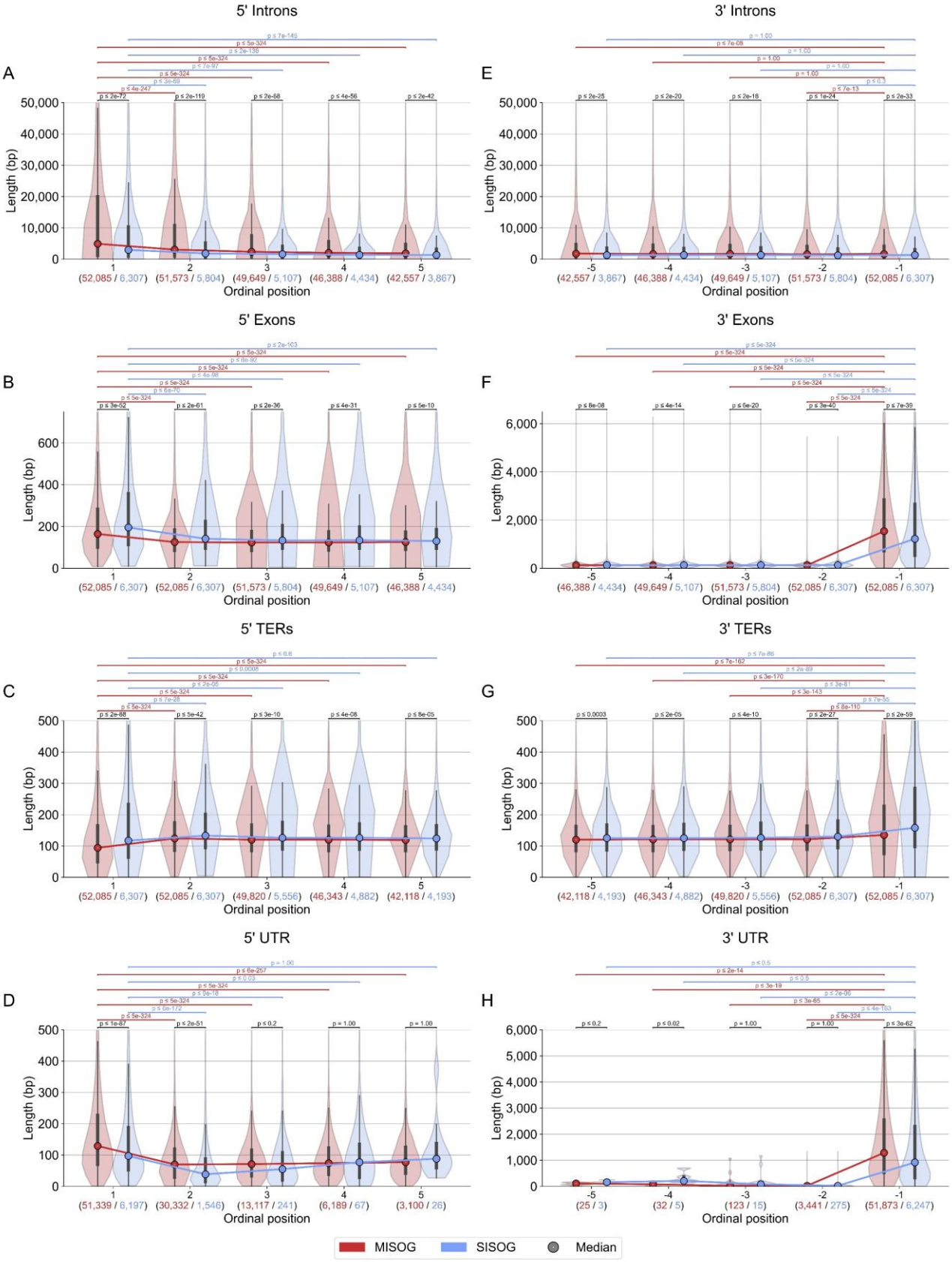


Figure S1 – Length distribution of gene elements (Intron, Exon, TER and UTR exons) according to ordinal positions.

All plots represent the length distribution of gene elements according to ordinal position (as illustrated in Fig. S4F). Plots A-D correspond to the first five 5' elements in the genes (1 up to 5), plots E-H correspond to the last five 3' elements (-5 to -1). The p-values obtained through Mann-Whitney U tests are displayed above the violin plots. Black p-values correspond to comparisons between MISOG and SISOG while colored p-values represent comparisons between first elements and following ones (MISOG: red, SISOG: blue). Values under each plot represent the numbers of observed elements at the ordinal position for MISOG (red) and SISOG (blue).

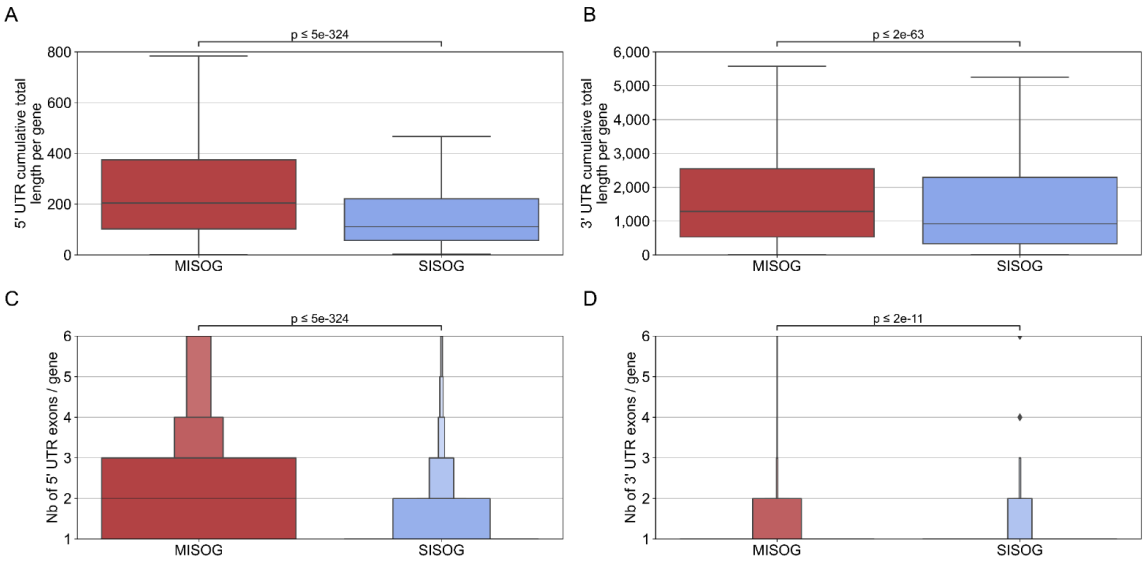


Figure S2 – Cumulative length and number of 5’ and 3’ UTR exons by gene

Cumulative total lengths of 5’ (A) and 3’ (B) UTR regions per gene. Number of 5’ and 3’ UTR exons per gene are displayed in (C, D). The p-values obtained through Mann-Whitney U tests are displayed above the boxplots (A, B) and boxenplots (C, D).

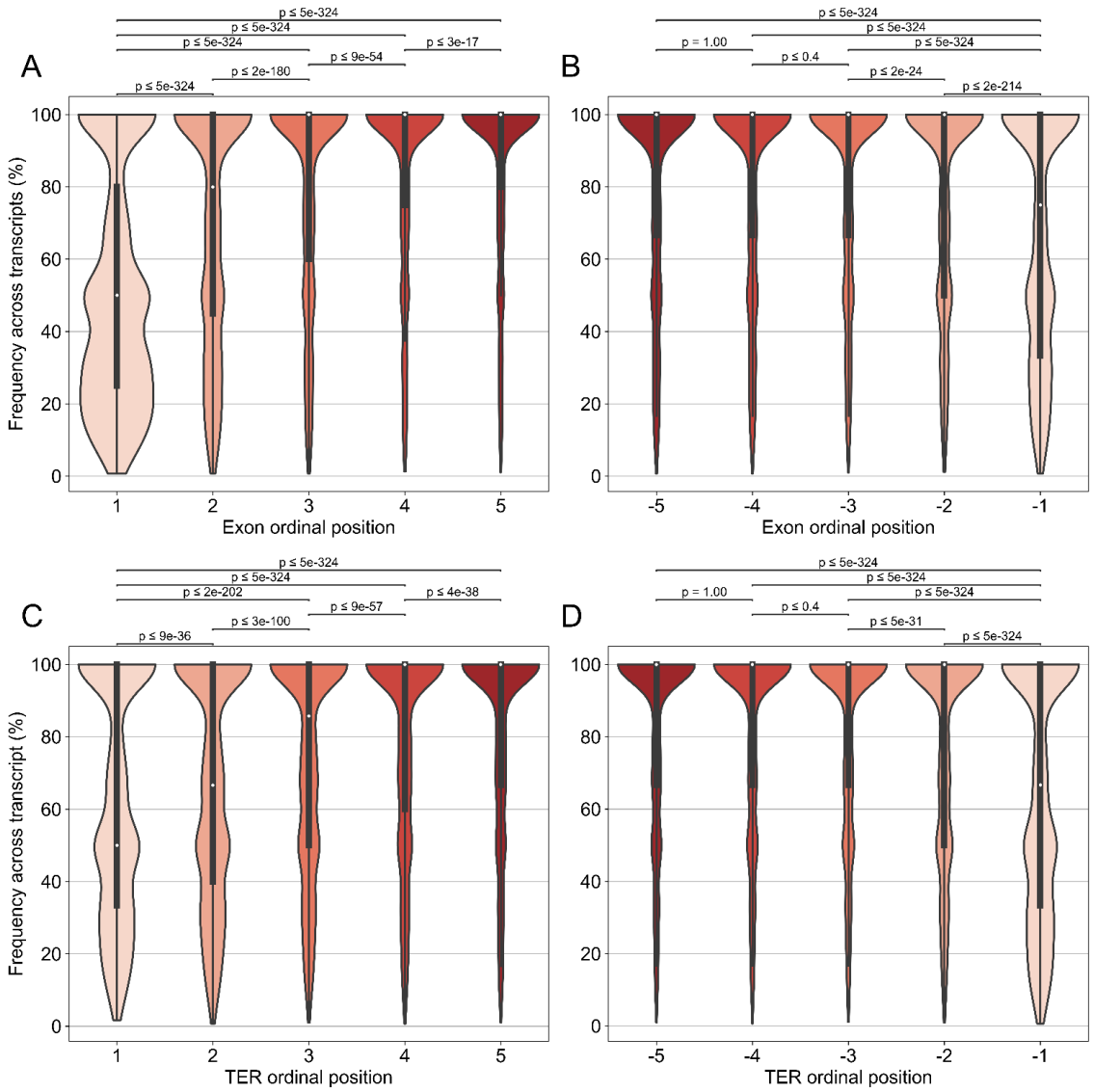


Figure S3 – Percentage of alternative exons and TERs according to their ordinal positions.

All plots show calculated frequencies of alternative exons (A, B) and alternative TERs (C, D) as illustrated in Fig. S4G. Plots (A, C) correspond to the first five 5' elements in the genes (1 up to 5), while plots (B,D) correspond to the last five 3' elements (-5 to -1). The p-values obtained through Mann-Whitney U tests are displayed above the violin plots.

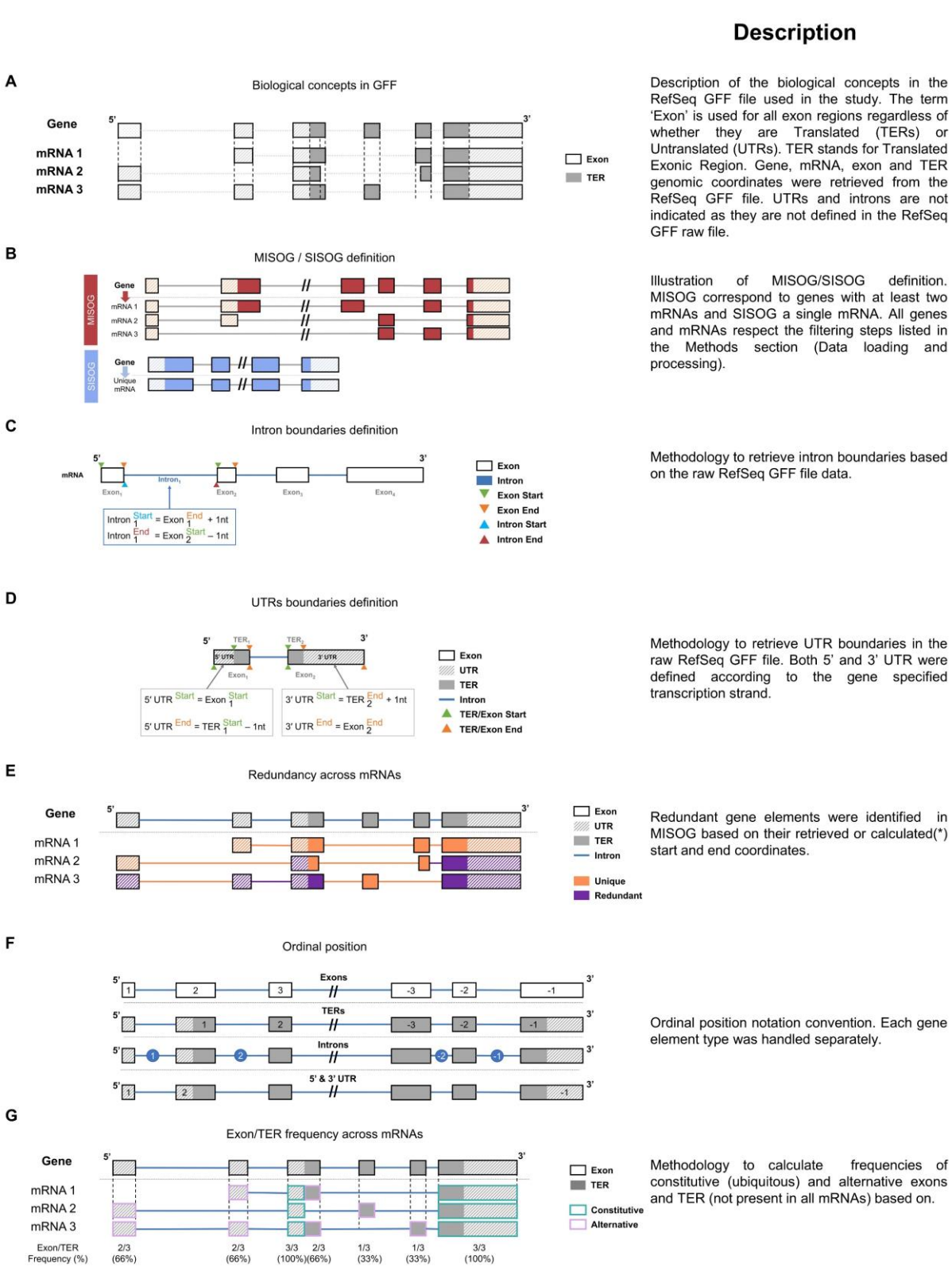


Figure S4 – Material and Methods schemas

9.1.3 Résultats complémentaires

9.1.3.1 Comparaison MISOG/SISOG dans le cadre des maladies génétiques

Une première analyse a été réalisée afin d'évaluer si une des deux catégories de gènes était surreprésentée dans le cadre des maladies génétiques. Pour cela, nous avons identifié les gènes ayant au moins une variation délétère validée dans ClinVar (section 6.3.4), soit un total de 4 554 gènes. Parmi ceux-ci, deux tiers sont associés à des MISOG (3 015) contre un tiers pour les SISOG (1 539) (Figure 54). Malgré la proportion de MISOG plus importante, cet écart reste significatif (p-value test binomial = 1,4e-37, test binomial présenté en section 7.2.1.2), mais ne semble pas constituer un élément discriminant majeur.

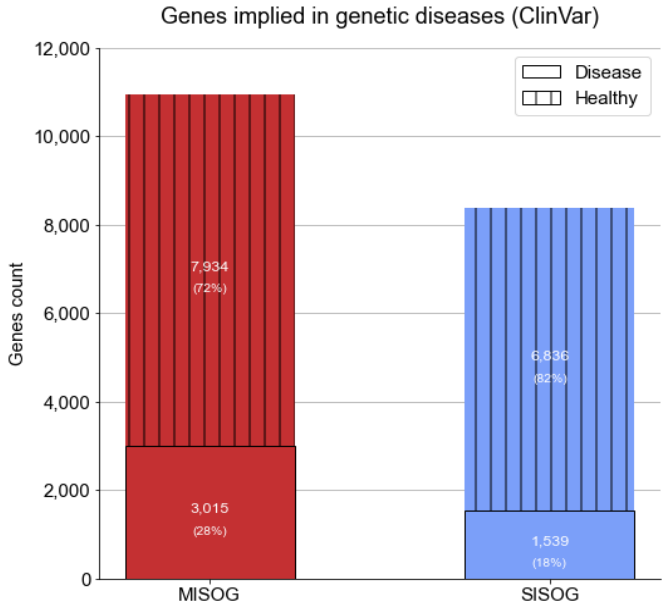


Figure 54 –Proportion de gènes MISOG/SISOG impliqués dans les maladies génétiques

Les gènes impliqués dans les maladies génétiques ont été identifiés à partir de ClinVar. Bien que les MISOG (10 949 gènes) sont 1,3 fois plus nombreux que les SISOG (8 375 gènes), le nombre de gènes impliqués dans les maladies génétiques est lui près de 2 fois supérieur (MISOG : 3 015 ; SISOG : 1 539).

Nous avons ensuite évalué si les MISOG ou SISOG était surreprésentés dans une ou plusieurs classes de maladies. Pour cela nous avons récupéré, *via* l'API d'OMIM, les gènes et phénotypes associés et extraits de la nomenclature Orphanet (section 6.3.1) l'ensemble des phénotypes (avec leur identifiant OMIM ; section 6.3.3). En combinant ces deux sources, nous avons établi des relations Gène → Maladie génétique → Classe de maladies. Pour chaque classe de maladies, nous avons évalué si les SISOG ou MISOG était enrichis par un test binomial (section 7.2.1.2) corrigé par la méthode de Benjamini-Hochberg (section 7.2.2) en prenant comme référence le nombre total de relations (SISOG=3 980 ; MISOG=8 618).

Les résultats présentés dans le Tableau 13 révèlent que certaines classes de maladies semblent enrichies dans l'une des deux catégories. Cependant, seules deux classes (en rouge dans le tableau) sont enrichies de manière significative (*p-value* < 0,05), les maladies du système immunitaire (1,58 fois plus de relations MISOG) et les maladies neurologiques (1,25 fois plus de relations MISOG), en accord avec des observations pointant le tissu cérébral comme enrichi en événements d'épissage alternatif (Weyn-Vanhentenryck et al. 2018).

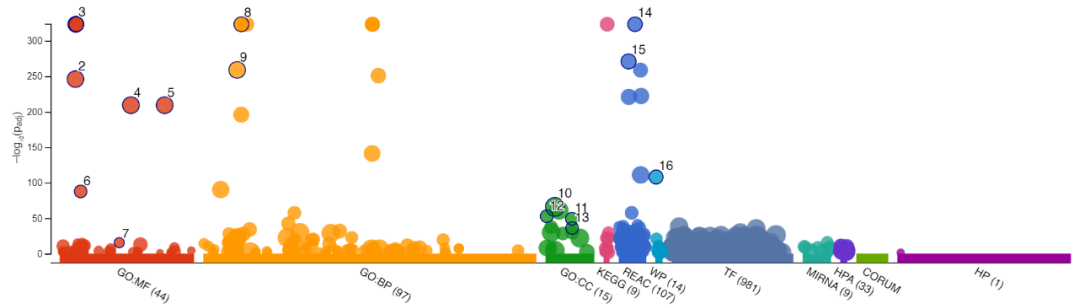
	SISOG			MISOG		
Orphanet Disorder classification	Nombre	Enrich.	p-value ajustée	Nombre	Enrich.	p-value ajustée
rare immunological diseases	63	0,63	1,00	216	1,58	**0,01**
rare surgical maxillo-facial diseases	32	0,71	1,00	97	1,40	0,47
rare systemic or rheumatologic diseases	13	0,78	1,00	36	1,28	0,97
rare neurological diseases	529	0,80	1,00	1434	1,25	**0,00**
rare neoplastic diseases	116	0,80	1,00	313	1,25	0,26
rare sucking/swallowing disorders	87	0,84	1,00	223	1,18	0,56
rare otorhinolaryngological diseases	117	0,91	1,00	278	1,10	0,97
rare cardiac diseases	70	0,92	1,00	165	1,09	0,97
rare surgical thoracic diseases	14	0,92	1,00	33	1,09	0,97
rare respiratory diseases	39	0,93	1,00	91	1,08	0,97
rare hepatic diseases	50	0,94	1,00	115	1,06	0,97
rare renal diseases	128	0,95	1,00	292	1,05	0,97
rare transplant-related disorders	299	0,96	1,00	676	1,04	0,97
rare gynecological & obstetric diseases	52	0,97	1,00	116	1,03	0,97
rare abdominal surgical diseases	42	1,02	0,95	89	0,98	1,00
rare inborn errors of metabolism	291	1,06	0,69	594	0,94	1,00
rare systemic & rheumatological diseases	43	1,07	0,85	87	0,93	1,00
rare gastroenterological diseases	55	1,07	0,85	111	0,93	1,00
rare bone diseases	214	1,09	0,60	426	0,92	1,00
rare infertility disorders	74	1,10	0,83	146	0,91	1,00
rare developmental anomalies	609	1,10	0,21	1194	0,91	1,00
rare odontological diseases	41	1,12	0,83	79	0,89	1,00
rare ophthalmic disorders	377	1,13	0,21	725	0,89	1,00
rare endocrine diseases	158	1,16	0,32	296	0,87	1,00
rare skin diseases	235	1,24	0,16	410	0,81	1,00
rare hematological diseases	114	1,26	0,21	196	0,79	1,00
rare circulatory system diseases	43	1,37	0,32	68	0,73	1,00
rare cardiac malformations	36	1,39	0,32	56	0,72	1,00
rare urogenital diseases	39	1,51	0,21	56	0,66	1,00
Total	3980	/	/	8618	/	/

Tableau 13 – Enrichissement en MISOG et SISOG dans les classes de maladies génétiques

Les classes significativement enrichies en MISOG (p-value corrigée < 0,05) sont colorées en rouge.

9.1.3.2 Analyse des gènes présentant une région exonique codante unique

Comme indiqué dans la publication, 1 983 MISOG ont été exclus, car les exons alternatifs sont hors des régions traduites en protéine (*Translated Exonic Region* ; TER). Ces 1 983 gènes (373 MISOG et 1 610 SISOG) comportent 995 gènes à un seul exon (TER + UTR inclut). L'analyse d'enrichissement des termes GO (*Gene Ontology* ; section 6.2.1) a révélé une forte relation de ces gènes avec des fonctions liées aux récepteurs couplés aux protéines G (*G Protein-Coupled Receptor* ; GPCR), notamment ceux liés à l'olfaction, l'odorat et le goût (Figure 55), confortant des observations déjà réalisées dans la littérature (Jorquera et al. 2018). Parmi ces GPCR, sur les 399 récepteurs olfactifs connus dans le génome humain, 374 (94%) sont associés à un gène ayant un seul transcrit validé actuellement. À noter qu'on identifie également certains termes GO statistiquement enrichis tels que : nucléosome, complexe d'empaquetage de l'ADN ou filament de kératine.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0004984	olfactory receptor activity	4.941×10^{-324}
2	GO:MF	GO:0004888	transmembrane signaling receptor activity	9.025×10^{-247}
3	GO:MF	GO:0004930	G protein-coupled receptor activity	4.941×10^{-324}
4	GO:MF	GO:0038023	signaling receptor activity	4.988×10^{-210}
5	GO:MF	GO:0060089	molecular transducer activity	4.988×10^{-210}
6	GO:MF	GO:0005549	odorant binding	2.006×10^{-88}
7	GO:MF	GO:0033038	bitter taste receptor activity	3.447×10^{-16}
8	GO:BP	GO:0007608	sensory perception of smell	4.941×10^{-324}
9	GO:BP	GO:0007186	G protein-coupled receptor signaling pathway	1.098×10^{-259}
10	GO:CC	GO:0016021	integral component of membrane	1.326×10^{-66}
11	GO:CC	GO:0044815	DNA packaging complex	8.887×10^{-50}
12	GO:CC	GO:0000786	nucleosome	1.836×10^{-53}
13	GO:CC	GO:0045095	keratin filament	7.422×10^{-37}
14	REAC	REAC:R-HSA-38...	Olfactory Signaling Pathway	4.941×10^{-324}
15	REAC	REAC:R-HSA-41...	G alpha (s) signalling events	1.056×10^{-271}
16	WP	WP:WP455	GPCRs, Class A Rhodopsin-like	8.759×10^{-109}

Figure 55 – Enrichissement fonctionnel sur les 1 983 gènes humains à région TER unique
Les termes fortement enrichis et pointés sur la figure sont détaillés dans le tableau. Outil utilisé : g-profiler

Ce nombre élevé de gènes dans une même famille fonctionnelle nous a amené à penser que l'absence d'épissage alternatif pouvait être liée à un autre mécanisme aboutissant à une amplification fonctionnelle, la paralogie (lien évolutif entre deux gènes issus d'un événement de duplication). En effet, un des rôles premiers de l'épissage alternatif est également d'augmenter le répertoire fonctionnel d'un gène.

Pour tester cette hypothèse, nous avons cherché les familles de gènes enrichies en MISOG et SISOG en exploitant la liste de 1 345 familles de gènes regroupant l'ensemble des gènes humains (source HGNC, section 6.2.2). Après avoir réalisé un test binomial corrigé par la méthode de Benjamini-Hochberg, nous avons identifié certaines familles de gènes enrichies en MISOG ou SISOG à TER unique, puis évalué le taux de gènes paralogues présent dans chaque famille d'intérêt. La majorité des groupes fonctionnels précédemment cités est enrichie en SISOG (Figure 56). On retrouve les récepteurs olfactifs, du goût (*Taste receptors* 2) ou encore, les histones. Cependant, il est surtout intéressant de noter que toutes ces familles ont un nombre élevé de paralogues, confortant notre hypothèse reliant la paralogie et absence d'épissage alternatif.

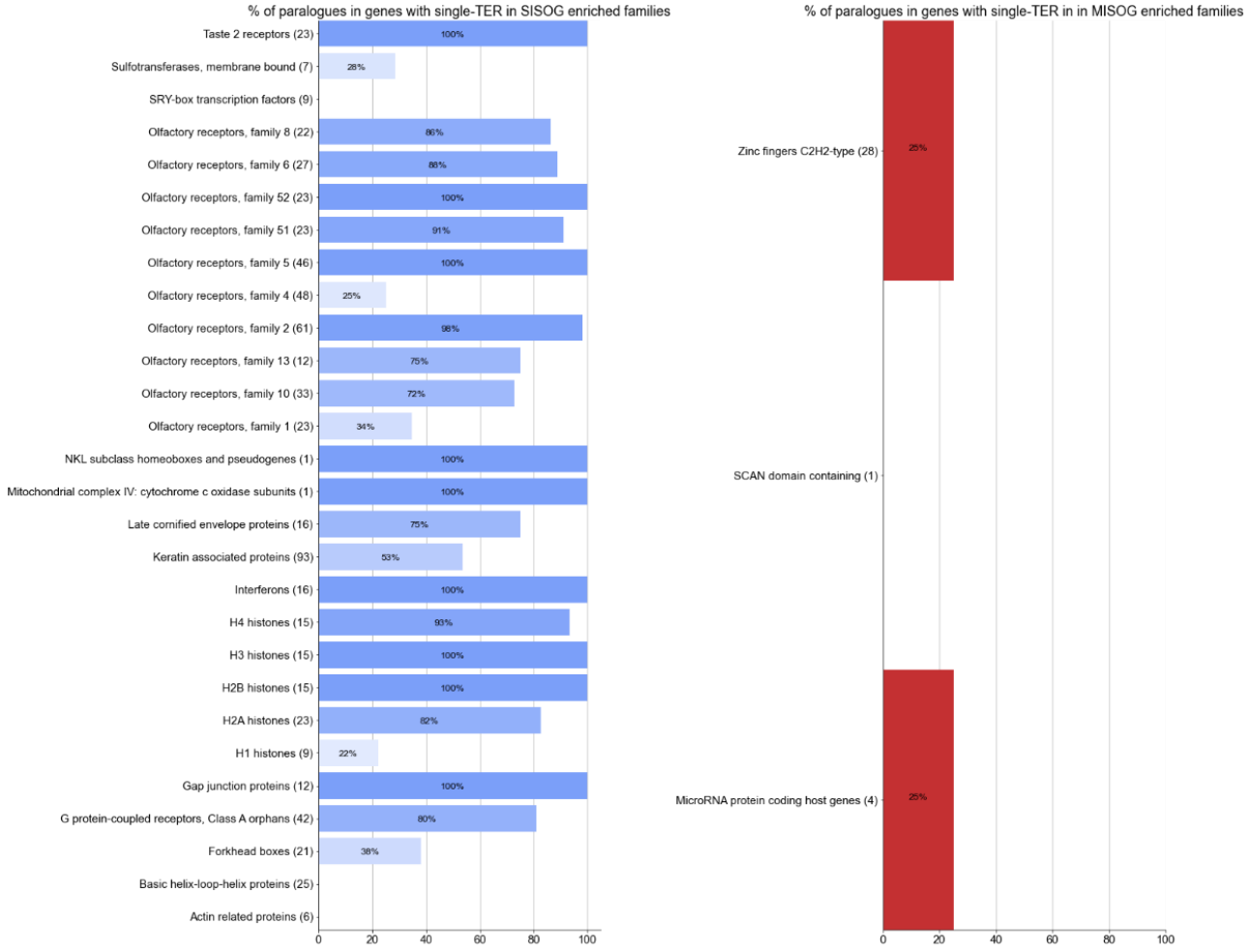


Figure 56 – Pourcentage de paralogues chez les familles de gènes enrichies en SISOG/MISOG

L'ensemble des familles enrichies en SISOG et MISOG à TER unique ont été évaluées selon leur proportion en gènes paralogues. On constate qu'aucun résultat intéressant ne ressort chez les MISOG, résultat partiellement expliqué du faible nombre de gènes à TER unique (373). Chez les SISOG à TER unique, on constate que 28 familles de gènes sont enrichies et que la majorité d'entre elles présente un niveau élevé de paralogues. Les fonctions des familles collent également avec l'analyse d'enrichissement fonctionnel réalisé précédemment.

9.1.4 Applications futures et perspectives

Les résultats prometteurs de cette analyse ouvrent différentes perspectives concernant les gènes humains codant pour les protéines et leurs isoformes. Il sera tout d'abord intéressant d'utiliser les données d'expression comme moyen de validation de notre approche selon un protocole à définir. La réanalyse d'observations publiées concernant les mécanismes d'export nucléaire ou la multiplicité des promoteurs permettrait de mieux comprendre certains de ces mécanismes au regard du statut MISOG/SISOG. Les éléments mis en évidence dans notre analyse ouvrent également la voie à une meilleure prédiction des gènes et de leurs isoformes, certainement *via* une intelligence artificielle exploitant les propriétés architecturales distinctes des MISOG/SISOG.

Compte tenu des résultats présentés dans le manuscrit sur l'importance supposée du premier intron et des régions UTR adjacentes, il apparaît comme important de mieux réévaluer certaines variations présentes dans ces régions non-codantes pouvant potentiellement perturber l'épissage alternatif et être la cause de maladies génétiques encore mal caractérisées. Enfin, toujours dans le cadre des maladies génétiques, l'analyse de la sévérité d'une variation entrainant une maladie pourra se faire, d'une part pour les MISOG, à travers les phénomènes compensatoires entre isoformes au sein d'un tissu, et d'autre part pour les SISOG, à travers la complémentarité des différents paralogues identifiés.

9.2 duxt : une nouvelle métrique pour déchiffrer l'utilisation différentielle d'un exon alternatif dans les tissus

9.2.1 Contexte

Dans le cadre des gènes codant pour les protéines, les isoformes de transcrits diffèrent au niveau de leur architecture et composition exonique, après épissage alternatif. On distingue ainsi, les exons constitutifs, présents dans l'ensemble des transcrits d'un gène, des exons alternatifs, présents dans un sous-ensemble de transcrits (Figure 57). Comme évoqué dans le Chapitre 5, chaque transcrit peut avoir un profil d'expression particulier selon les tissus et types cellulaires qui va aboutir à une **« utilisation » différentielle** des exons. Dans ce contexte, nous avons cherché à développer une nouvelle métrique : duxt (*differential usage across tissues*) qui permette d'identifier et évaluer les différentiels d'utilisation d'exons alternatifs à travers différents tissus humains. Cette métrique s'inscrit dans la logique d'une caractérisation plus fine des variations génétiques délétères intégrant l'expression tissulaire alternative afin de mieux prédire, voire expliquer, des phénotypes complexes.

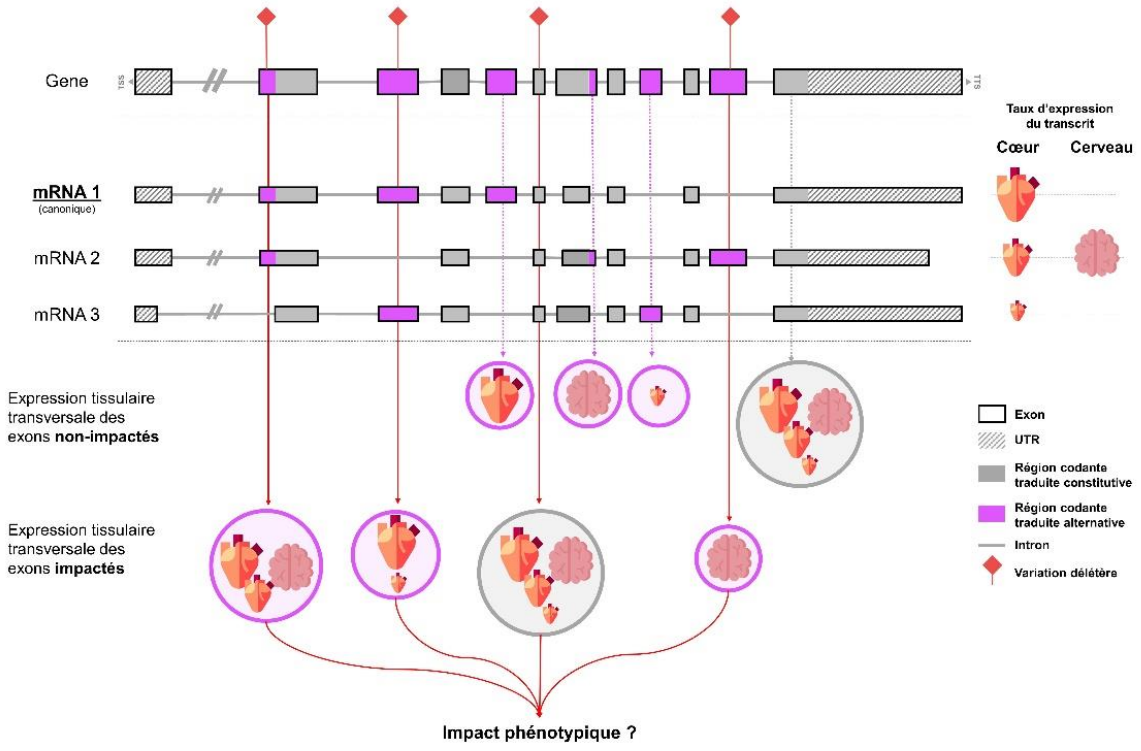


Figure 57 – Contexte de la métrique duxt

Illustration de l'utilisation différentielle des exons alternatifs selon les tissus du cœur et du cerveau. Certains exons se retrouvent spécifiquement dans un des tissus, ou dans plusieurs tissus, mais en proportion variable.

9.2.2 Caractérisation des MISOG et exons exprimés

9.2.2.1 Identification des MISOG et SISOG

Parmi les 31 283 gènes référencés dans RefSeq (section 6.1.1), 21 657 codent pour des protéines dont 19 306 associées à la source *BestRefSeq* et à des identifiants d'accession basés sur des méthodes de validation manuelle (NC, NM et NP ; section 6.1.1). Sur cet ensemble de 19 306 gènes codant humains, nous avons identifié 8 369 gènes à isoforme unique (*Single transcript ISOform genes* ; SISOG) et 10 937 gènes présentant plusieurs isoformes de transcrits (*Multiple transcript ISOform genes* ; MISOG). Nous avons ensuite appliqué les filtres recommandés par le Consortium GTEx (section 6.4.3), pour retenir les transcrits exprimés à un seuil significatif. Ceci a permis de retenir 5 657 SISOG (2 712 SISOG sous les seuils) et 9 696 MISOG (1 241 sous les seuils). Certains MISOG ont été reclassés en SISOG (1 516 gènes), car seul un de leurs transcrits était exprimé à un seuil significatif aboutissant à un jeu composé de 7 173 SISOG et 8 180 MISOG, soit un total de 15 353 Gènes Humains Codant pour des Protéines (HPCG) exprimés (Figure 58).

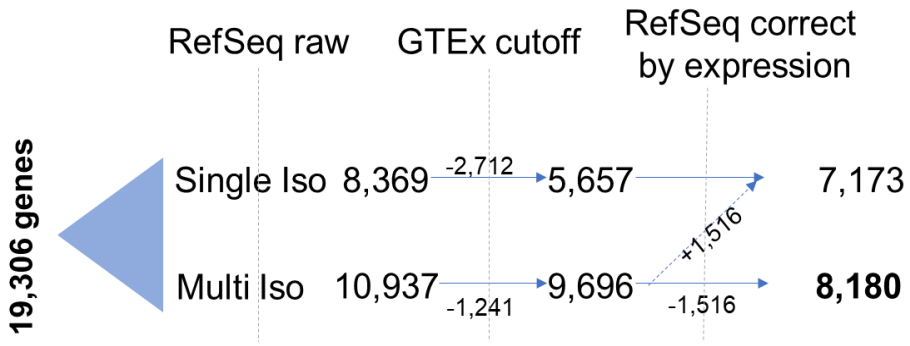


Figure 58 – HPCG dans la base RefSeq avant et après filtrage par le niveau d'expression

9.2.2.2 Comparaison du protocole appliqué aux données RefSeq et Ensembl

Pour vérifier que nos résultats n'étaient pas biaisés par la sélection de la base de données RefSeq, le même traitement a été appliqué aux données d'Ensembl (version 103). Ceci a abouti à des résultats similaires, avec 15 848 HCPG exprimés provenant d'Ensembl (Tableau 14), dont 15 249 gènes communs aux deux ressources. En comparant ces valeurs après exploitation des données d'expression, à celles obtenues *in silico* et présentées dans le manuscrit 9.1.2, nous obtenons un pourcentage de MISOG exprimés (53%) inférieur aux 65% *in silico* et toujours très inférieur aux 95 % de gènes avec épissage alternatif rapportés dans la littérature. Enfin, il convient de noter que, malgré la masse de données d'expression répertoriées dans GTEx (54 tissus disponibles chez près d'un millier de personnes), 3 953 gènes validés manuellement n'ont aucun transcrit avec un seuil d'utilisation exploitable dans les tissus séquencés. Parmi ces gènes, plus de 68 % sont des gènes à transcrit unique

(SISOG). Cependant, ce résultat est sans doute faussé par le faible nombre de personnes et de tissus analysés, et il conviendra de le réévaluer périodiquement au gré des résultats de séquençage massif de transcriptomes provenant de nouveaux types cellulaires (He et al. 2020; Kim-Hellmuth et al. 2020; Eraslan et al. 2021), de différentes étapes du développement (Mazin et al. 2021; GTEx 2021) ou de groupes ethniques plus étendus (J.-W. Li et al. 2014).

	Ensembl		**RefSeq**	
	Genes	**Transcripts**	**Genes**	**Transcripts**
HPCG	22,492	95,340	19,306	53,866
HPCG *Present in GTEx*	18,924	72,911	18,240	48,304
HPCG *Present in GTEx & passing filters*	15,848	42,094	15,353	37,323

Tableau 14 – HPCG dans Ensembl & RefSeq après étapes de filtrage liées à l'expression

9.2.2.3 Propriétés des exons constitutifs et alternatifs au sein des MISOG exprimés

Dans toutes les sections suivantes, afin d'étudier l'éventuel impact différentiel des variations dans les isoformes d'un MISOG exprimé, nous ne nous intéresserons qu'aux exons ou régions d'exons traduits en protéines en excluant les régions UTR.

Tout d'abord, nous avons comparé les exons constitutifs (Const) et alternatifs (Alt) selon différentes propriétés telles que : le nombre et la longueur par gène, la conservation-contrainte (phyloCSF, CCR, section 6.2.3), la distribution des variations délétères et non-délétères selon leur conséquence moléculaire (synonyme, faux-sens, …) ou leur MAF.

9.2.2.3.1 Comparaison des longueurs et du nombre d'exons par gène

Parmi les 8 180 MISOG exprimés, nous avons identifié 65 322 exons Const et 36 584 exons Alt (Figure 59a) impliquant que les exons Alt représentent plus d'un tiers des exons totaux. Un MISOG présente de manière médiane 7 exons Const de longueur 118 pb et 3 exons Alt de longueur 96 pb (Figure 59b et c) et la différence des tailles entre exons Const et exons Alt est significative comme en atteste la valeur de *p-value* proche de 0 obtenue par le test Mann-Whitney U (P_{MWU}).

Nous avons ensuite analysé le nombre et la longueur des exons Alt en fonction de leurs fréquences dans les isoformes. La catégorie la plus représentée correspond aux exons Alt avec une fréquence de 40% à 60% dans les transcrits avec 13 712 exons, soit 37% des 36 584 exons Alt retenus. Cette sur-représentation de la catégorie 40-60% peut en partie s'expliquer par le nombre élevé de gènes à 2 et 3 transcrits (plus de 60 % des MISOG, Figure 60) qui aboutit à des fréquences d'exon oscillant entre 33% et 66% (Figure 59d). On

constate également un faible nombre d'exons très rares (0-20%) ou très fréquents (80-100%) totalisant 15% des exons Alt (3261 et 2248 respectivement ; Figure 59d). À noter que la longueur est corrélée à la fréquence des exons Alt, ce qui semble indiquer que plus les exons Alt sont fréquents et s'approchent du statut d'exon constitutif, plus leurs longueurs se rapprochent de celles observées pour ces derniers (Figure 59f).

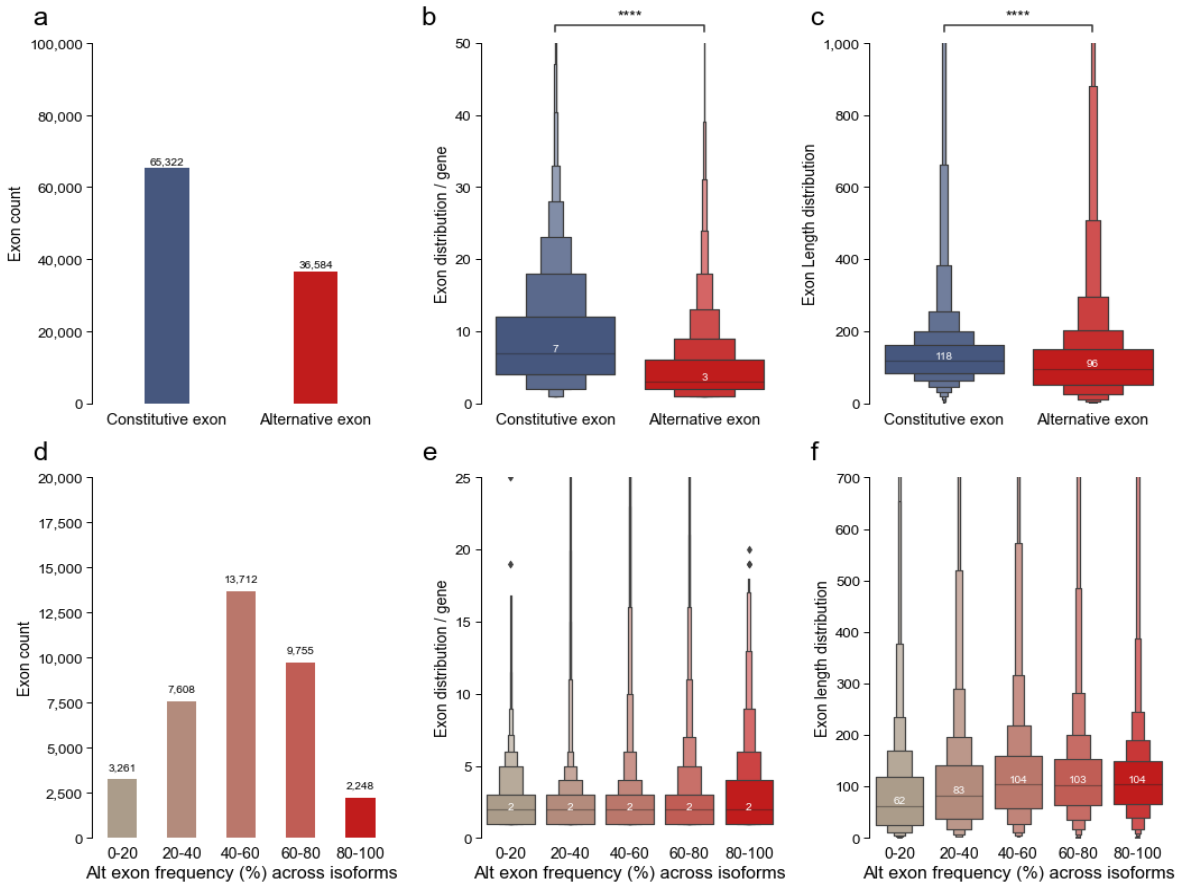


Figure 59 – Nombre d'exons, nombre par gène et longueur pour les exons constitutifs et alternatifs

a), b) et c), comparent les exons constitutifs et alternatifs
d), e), f), comparent les exons alternatifs selon leur fréquence in silico dans les isoformes de transcrits.

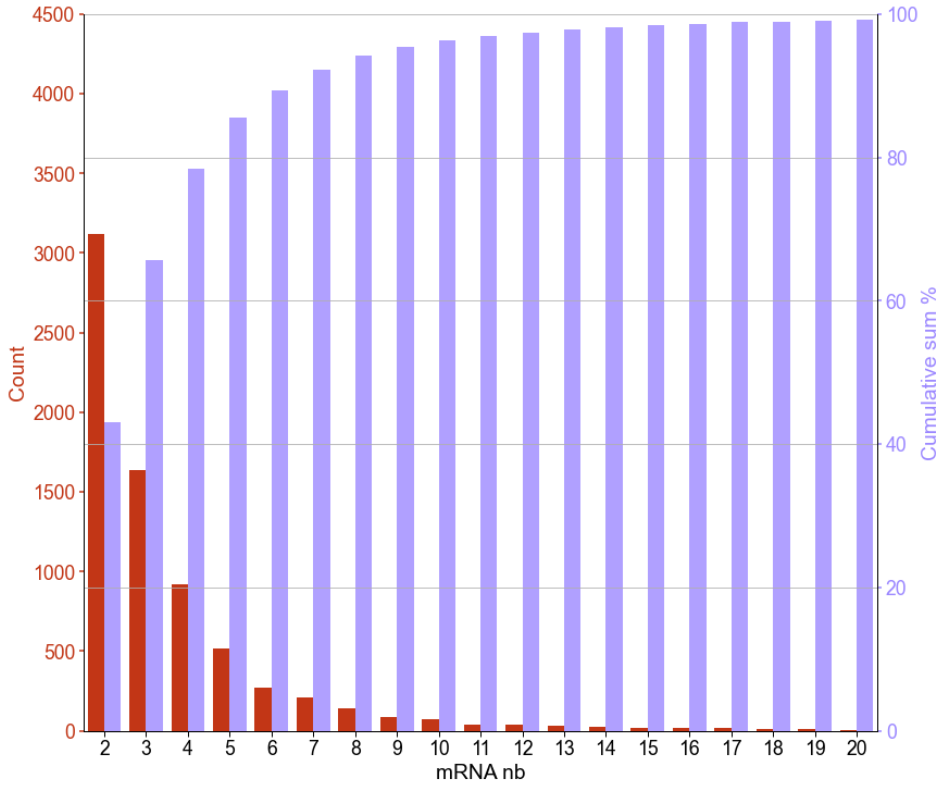


Figure 60 – Distribution du nombre d'isoformes de transcrit chez les gènes MISOG

La partie rouge de l'histogramme représente la quantité de gènes présentant un nombre donné de mRNA. La partie violette représente la somme cumulative en pourcentage. On constate que près de 80% des gènes MISOG ont entre 2 et 4 mRNAs.

9.2.2.3.2 Évaluation de l'enrichissement selon des propriétés estimées

La répartition, au sein des exons Const et Alt, de différentes propriétés provenant de méthodes d'estimation a été analysée : conservation/contrainte (phyloCSF, CCR), impact délétère, conséquences moléculaires et MAF des variations.

Concernant les conservations (phyloCSF et CCR), les seuils décrits en section 6.2.3 ont été utilisés. Concernant les variations, différentes étapes d'extraction et de filtrage ont été appliquées. Pour les SNV non-délétères situées dans les exons des gènes MISOG, ces variations ont été extraites de la base de données gnomAD (v2.1.1) (section 6.4.2). Après avoir supprimé les SNV dont le nombre d'allèles corrigés était nul (AC=0), 2 810 853 SNV non délétères ont été identifiées dans 7918 gènes MISOG. Pour les SNV délétères, sur les 789 579 SNV extraites de la base de données ClinVar (section 6.3.4 ; v20210123), seules 24 402 SNV délétères ont été retenues après élimination selon quatre critères :

1) Présence dans 4 921 MISOG répertoriés dans ClinVar = 256 927 variations codantes ;

2) Absence d'information dans des champs nécessaires : identifiant de l'allèle (ALLELEID), signification clinique (CLNSIG), statut de révision clinique (CLNREVSTAT) ou information sur le gène (GENEINFO) = 255 975 SNV dans 4 911 MISOG.
3) Entrées avec statut clinique (CLNSIG) *"pathogenic/likely_pathogenic"* et sans mention *"conflicting_interpretations"* = 29 506 SNV dans 2 042 MISOG.
4) Absence des SNV dans gnomAD = 24 402 SNV dans 1 844 MISOG.

Comme illustré sur la Figure 61, les exons Const sont plus conservés (*Odds Ratio=2,85*) et contraints (*Constraint*=0-20 : *Odds Ratio*=0,92 ; *Constraint*=95-100 : *Odds Ratio*=1,45) que les exons Alt et présentent un faible appauvrissement en variations fréquentes à travers la population (MAF=1e-06 – 1e-05 : *Odds Ratio*=1,05 ; MAF=0.1 – 1 : *Odds Ratio*=0,87). De même, les variations très impactantes (stop perdu, start perdu) sont quasiment absentes des exons Const, mais sur-représentées dans les exons Alt (*start lost* : *Odds Ratio*=0,40 ; stop lost : *Odds Ratio*=0,54). Ceci semble en accord avec les fortes conservation et contrainte des exons Const, tant il paraît cohérent qu'ils soient plus « protégés » en raison de leur caractère ubiquitaire. Enfin, on note la quasi-absence d'enrichissement statistique des variations délétères et non-délétères (*pathogenic* : Odds Ratio=1,09).

Globalement, nous pouvons conclure qu'à l'exception notoire de la conservation/contrainte, les exons Const et Alt partagent des propriétés assez similaires, ce qui était attendu. Cette faible conservation (calculée sur de nombreuses espèces) des exons Alt confirme des résultats de la littérature démontrant que 76 % des exons Alt humains sont apparus chez les primates, au cours des 90 derniers millions d'années (Rodriguez et al. 2020).

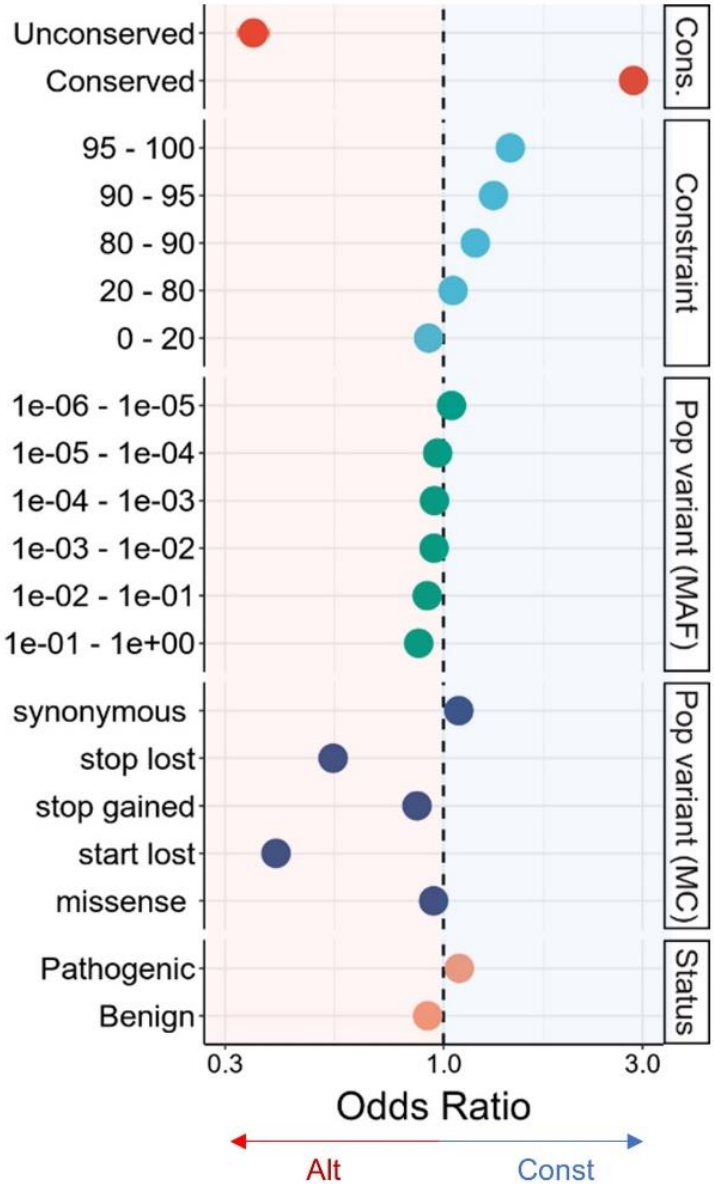
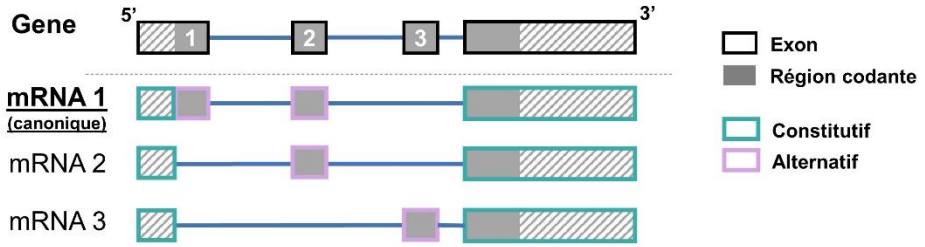


Figure 61 – Analyse de plusieurs paramètres pour comparer exons constitutifs et alternatifs

L'échelle utilisée est logarithmique. La partie rouge à gauche représente l'enrichissement du critère sélectionné pour les exons Alt et la partie bleue à droite, l'enrichissement du critère sélectionné pour les exons Const.

Enfin, si on considère la notion de mRNA canonique, correspondant au mRNA de référence dans la grande majorité des banques et qui constitue souvent le mRNA auquel les biologistes vont se référer, nous avons observé que 33% des 36 584 exons Alt retenus sont absents de ces mRNAs et exclusivement exprimés dans des transcrits non-canoniques (Figure 62). Cela souligne l'importance de mieux considérer ces régions alternatives souvent ignorées.



Exon	Association transcrit non-canonique	Fraction / 3 exons
Exon 2 et 3	Au moins 1 transcrit non-canonique	66%
Exon 3	Uniquement transcrit non-canonique	33%

Figure 62 – Illustration de la fréquence des exons alternatifs dans les transcrits non-canoniques

Le mRNA canonique, ici le mRNA 1 est le mRNA de référence dans les bases de données. Nous avons observé qu'un exon sur trois était uniquement présent dans des mRNAs non-canoniques.

9.2.3 Identification des exons alternatifs présentant des différentiels d'utilisation tissulaire

Par définition, l'exon Alt d'un MISOG est présent dans un sous-ensemble des transcrits et est lié à un profil d'expression particulier. La métrique pext (*proportion expressed across transcripts* ; section 6.4.3) normalise l'expression de chaque nucléotide au regard de son apparition dans les différents transcrits d'un gène, fournissant ainsi une valeur d'expression de 0 à 1 pour chaque nucléotide dans les 53 tissus cibles.

Avec duxt (*differential usage across tissues*), nous proposons une métrique complémentaire permettant d'identifier des différentiels d'utilisation d'exon trans-tissus chez les isoformes des transcrits de MISOG. En s'appuyant sur les 53 tissus exploités par pext, notre protocole permet d'identifier et évaluer l'exon alternatif responsable d'un différentiel d'expression ainsi que le tissu où l'effet est majeur.

À ce stade, il est capital de rappeler que, les deux métriques s'appuient sur des notions d'expression normalisée relative à l'expression globale d'un gène. Ces métriques n'abordent pas l'aspect quantitatif de l'expression (nombre de transcrits par tissu), mais plutôt la sur-/sous-représentation d'un nucléotide/exon au sein d'une population de transcrits quelle que soit le niveau d'expression. Ces normalisations étaient indispensables pour réaliser des analyses des niveaux d'expression des exons trans-tissus. Cependant, cela laisse ouvert le problème de l'évaluation des quantités effectives des divers transcrits d'un gène indissociable de la fonction du gène au sein de la cellule/tissu. Pour illustrer ce propos, on peut se demander si dix transcrits en plus ou en moins d'un régulateur faiblement exprimé (protéine kinase, facteur de transcription…) auront un impact plus/moins important sur le devenir de la cellule que mille transcrits différentiellement exprimés impliqués dans le ribosome.

9.2.3.1 Protocole développé

Pour chaque exon, la valeur pext d'expression a été convertie en un Z-score (**Erreur ! Source du renvoi introuvable.**) où x_{tissue} est la valeur pext du nucléotide pour un tissu donné, μ : la moyenne des valeurs pext dans l'ensemble des tissus pour ce nucléotide et σ : l'écart-type des valeurs pext pour l'ensemble des tissus pour ce nucléotide.

Le Z-score présentant une sensibilité élevée, nous avons défini une pondération (**Erreur ! Source du renvoi introuvable.**) correspondant à l'écart entre chaque valeur et la moyenne de l'ensemble des valeurs restantes. Le Z-score couplé à cette pondération (**Erreur ! Source du renvoi introuvable.** $*$ **Erreur ! Source du renvoi introuvable.**) permet de distinguer les situations (Figure 63 *Exon* 1) où les valeurs d'expression diffèrent légèrement d'une distribution uniforme avec un faible écart-type (moins de 5% de déviation par rapport à la moyenne), des situations (Figure 63 *Exon* 2) présentant une diminution/augmentation significative de l'expression pour un tissu spécifique.

c

Nous avons ensuite appliqué une fonction sigmoïde (Équation 7), qui distend la distribution et les valeurs extrêmes facilitant l'analyse des différentiels. Puis, les valeurs ont été mises à l'échelle entre]-1,1[autour d'une ligne de base fixée à 0 (Équation 8), permettant la distinction entre sous-utilisation (valeurs négatives) et sur-utilisation (valeurs positives) (Figure 63).

$$(3)\ sigmoid = \frac{1}{1 + e^{-(Z-score * weight)}}$$

Équation 7 – Fonction sigmoïde

$$(4)\ duxt = (2 * Z - score\ * weight) - 1$$

Équation 8 – Équation finale de duxt

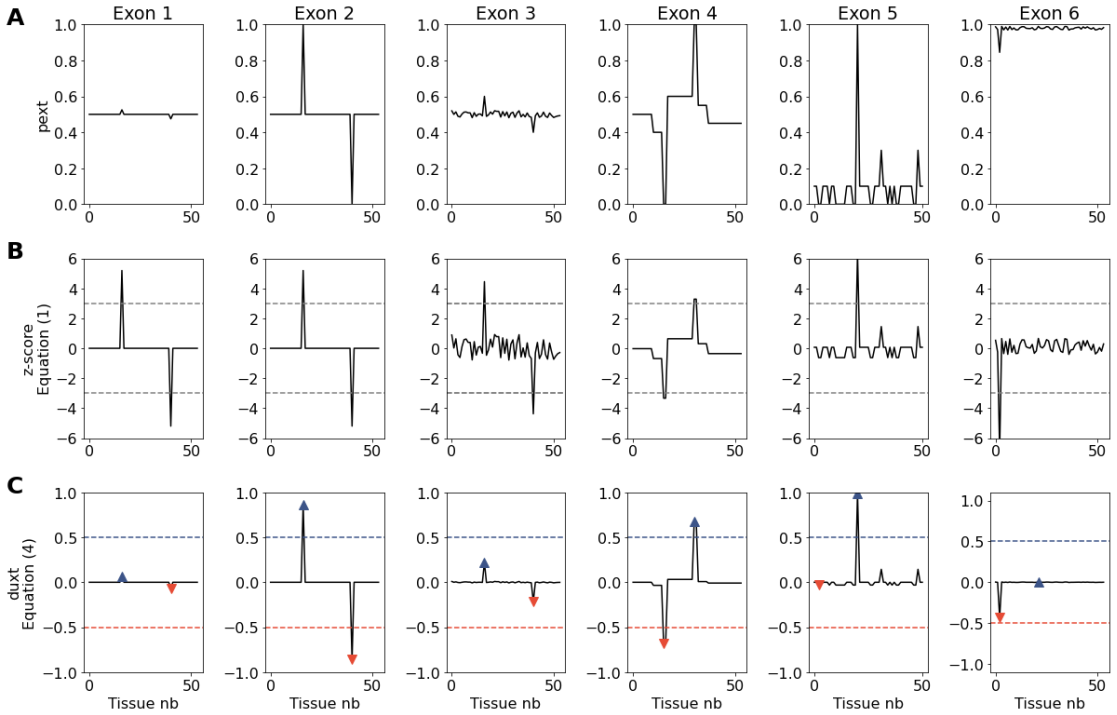


Figure 63 – Comparaison des valeurs de pext, de z-score et de duxt pour différents exemples

Les lignes pointillées (B) correspondent à une valeur limite de z-score = 3 (valeur supérieure/inférieure à 3 fois l'écart-type). Les exons 1 et 2 illustrent la sensibilité importante du z-score. Dans les deux cas, le z-score possède une valeur maximale tandis que la pondération appliquée dans duxt permet de distinguer la situation de l'exon 1 de celle de l'exon 2. duxt (C) réalise une « mise à l'échelle » des valeurs pext en tirant parti de la propriété du z-score pour créer une ligne de base à 0. Ceci permet une comparaison plus aisée entre exon à faible utilisation dans de nombreux tissus (Exon 5) et exons utilisés dans la quasi-totalité des tissus (Exon 6). Dans la figure C, la ligne pointillée bleue représente le seuil limite supérieur pour détecter une sur-utilisation, et la ligne pointillée rouge, la limite inférieure pour détecter une sous-utilisation d'exon. Les triangles rouge et bleu correspondent respectivement aux valeurs duxt-up et duxt-down, seules les valeurs supérieures/inférieures (étoiles) aux seuils sont retenues.

Finalement, nous avons introduit deux valeurs supplémentaires correspondant respectivement à la valeur duxt maximale (**duxt-up**) et minimale (**duxt-down**) d'un exon. Plus la valeur absolue de duxt-up ou |duxt-down| est élevée, plus la spécificité de sur/sous-utilisation liée à un tissu est importante.

Pour la suite de l'étude, en examinant la distribution des exons en fonction des valeurs duxt-up et |duxt-down| (Figure 64), nous avons choisi de définir comme exons différentiellement sur- et/ou sous-utilisés, les exons correspondants à duxt-up ou |duxt-down| > 0,5, ce qui correspond à 11% (percentile = 0,89) et 13% (percentile = 0,87) des exons différentiellement utilisés, respectivement pour les catégories duxt-up et duxt-down. Les exons répondant à ces critères stricts ont été appelés duxtExons et les gènes au sein desquels ils se trouvaient, duxtGenes.

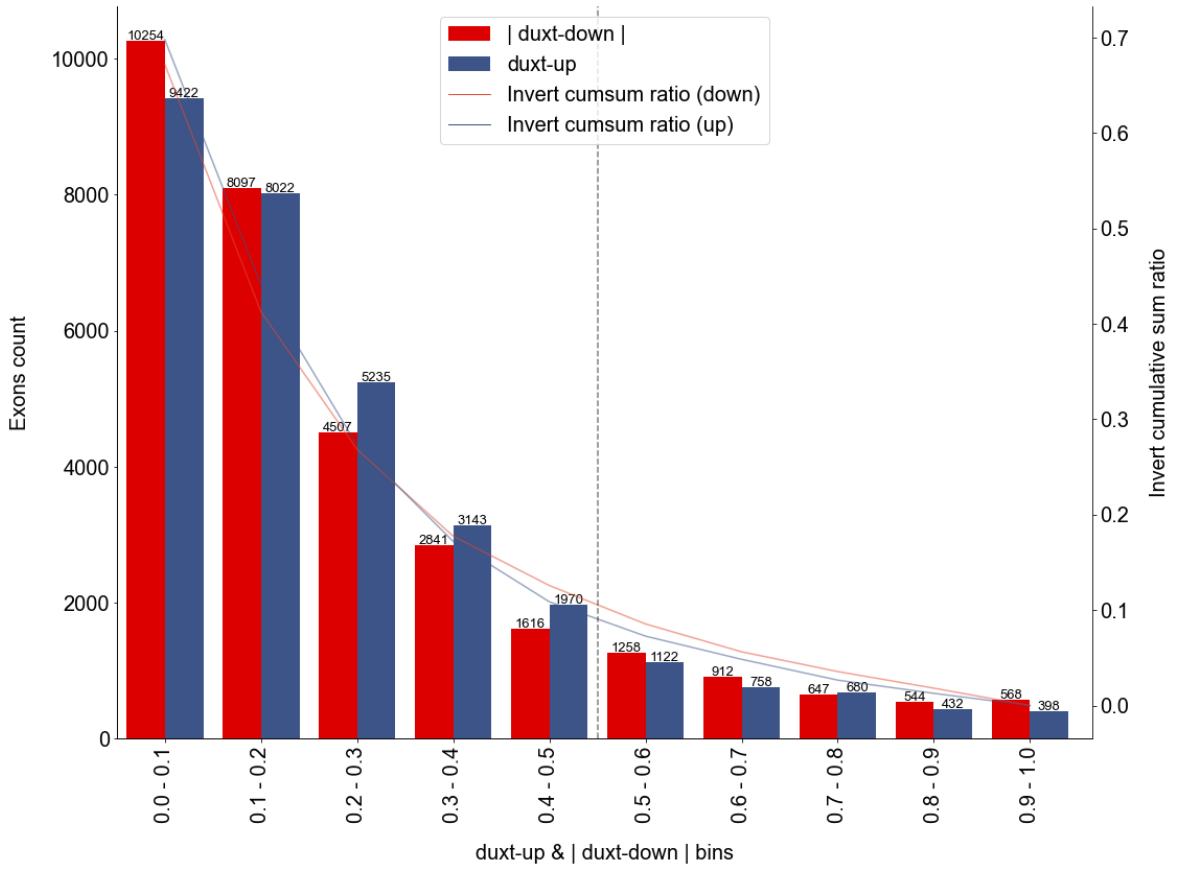


Figure 64 – Distribution du nombre d'exons selon les scores |duxt-down| et duxt-up

La ligne pointillée représente le seuil fixée de 0,5 pour identifier les exons à forte sur- ou sous-utilisation différentielle.

9.2.3.2 Bilan d'étape

Parmi les 8 180 MISOG identifiés, un score duxt a pu être attribué à 6 904 MISOG impliquant 31 606 exons des 36 584 exons Alt totaux Tableau 15). Cette première caractérisation des exons alternatifs permet d'envisager adjoindre une information nouvelle aux variations situées dans ces régions qui regroupent 7 919 variants délétères, 36 902 variants à statut inconnue (VUS ; section 3.4) et 810 118 variations de population selon la base gnomAD. Ces résultats sont encourageants, car ils témoignent que, même en appliquant des filtres très stricts, l'épissage alternatif influence un nombre non négligeable de gènes, exons ou variations.

Pour la suite de notre étude, nous avons introduit les notions de duxt-up et duxt-down pour mieux identifier les exons dont l'utilisation serait particulièrement atypique au sein de gènes ou tissus. Dans ce cadre, nous avons défini un seuil strict de 0,5 qui a permis de distinguer 7 233 exons fortement sur- ou sous-utilisés (duxtExon) dans 2 272 gènes [duxtGenes ; Figure 64]. Ces duxtExons se répartissent en 3 390 duxtExons sur-utilisés (duxtExon-up) dans 1 349 gènes et 3 929 duxtExons sous-utilisés (duxtExon-down) dans 1 381 gènes.

Filter	**Genes Number**	**Alt Exons Number**
RAW	8,180	36,584
Merge pext	6,979	31,944
duxt score available	6,904	31,606
duxt score > 0.5	2,272	7,233

Tableau 15 – Étapes du filtrage appliqué durant le développement de duxt

Un récapitulatif du principe de fonctionnement de duxt est présenté en Figure 65.

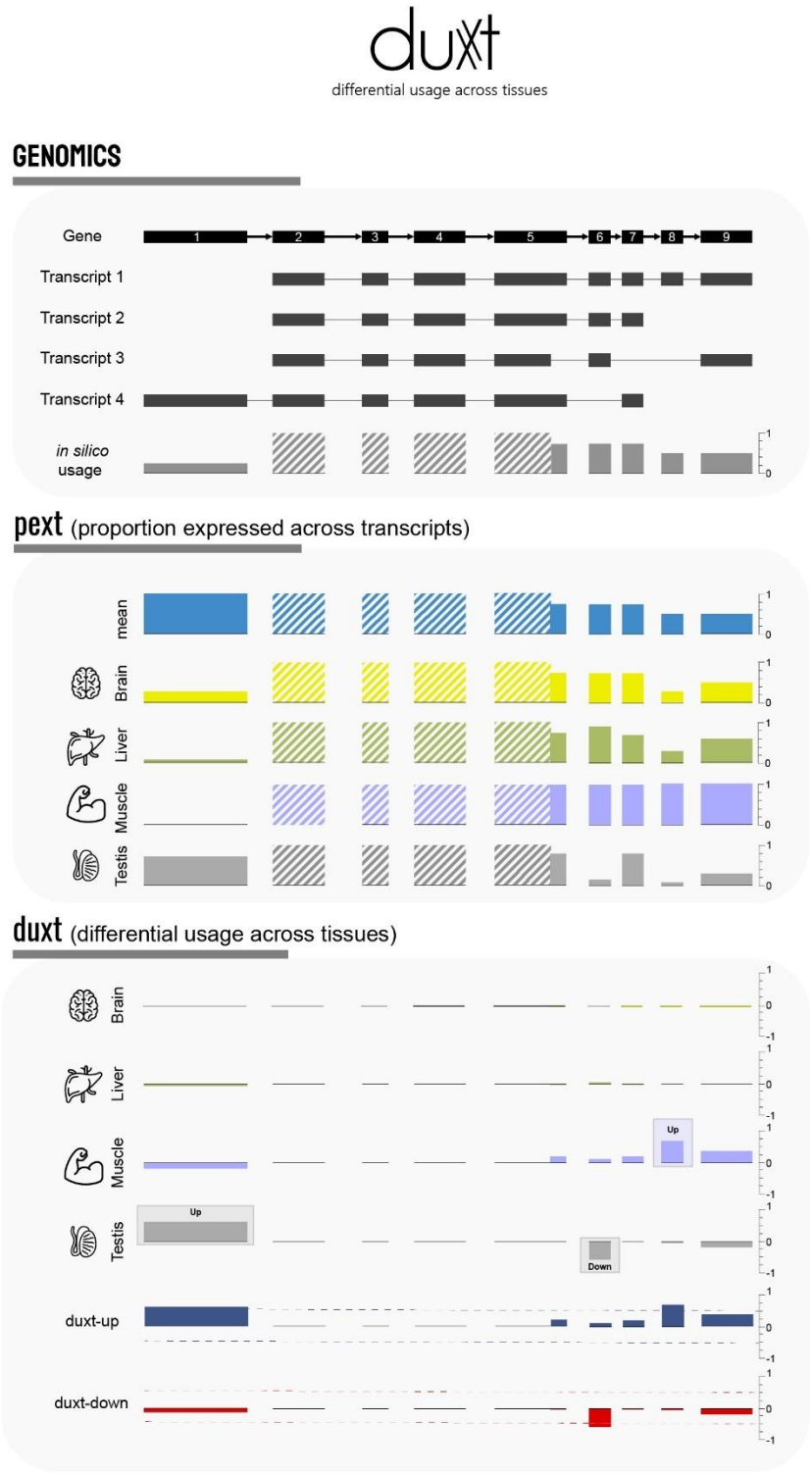


Figure 65 – Figure récapitulative du fonctionnement de duxt

Pour chaque gène MISOG, nous avons d'abord identifié l'utilisation in silico des exons le composant. L'ensemble des exons présentant un ratio d'utilisation inférieur à 1 ont été définis comme exons alternatifs. Nous avons ensuite utilisé les valeurs de pext pour pouvoir identifier des différentiels d'utilisation à partir du protocole présenté précédemment dans le cadre de duxt. Dans cet exemple, le muscle et le tissu testiculaire présentent des différentiels d'utilisation pour trois exons alternatifs (Exon 1 : sur-utilisation dans le tissu testiculaire, Exon 6 : sous-utilisation dans le tissu testiculaire, Exon 7 : sur-utilisation dans le tissu musculaire).

9.2.4 Études exploratoires : applications de la métrique duxt

9.2.4.1 Tissus enrichis en sur- ou sous-utilisation différentielle

En analysant la distribution des duxtExon et duxtGenes par tissu (Figure 66), nous avons identifié plusieurs tissus enrichis en utilisation différentielle d'exons : les testicules (1 702 exons, 667 gènes), le sang (1 123 exons, 437 gènes), les lymphocytes transformés par EBV (1 034 exons, 389 gènes), les fibroblastes en culture (856 exon, 271 gènes) et le muscle squelettique (721 exons, 277 gènes).

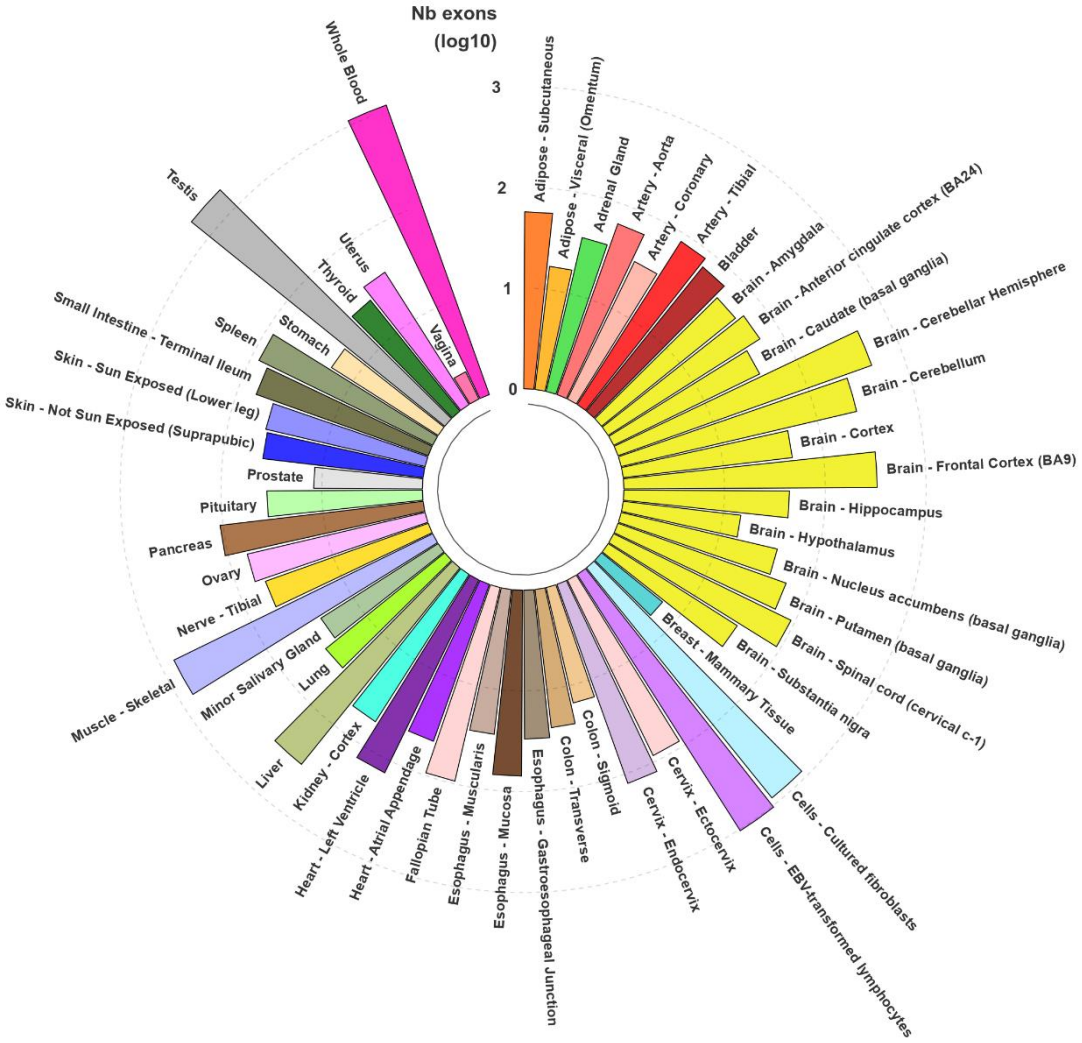


Figure 66 – Nombre d'exons différentiellement utilisés par tissu

Figure récapitulative du nombre d'exons différentiellement utilisé par tissus. L'échelle employée est logarithmique. On constate que certains tissus présentent un nombre très élevé d'exons différentiellement utilisé (testis, whole blood, muscle skeletal) alors qu'à l'inverse certains sont fortement appauvris (vagina, breast – mammary tissue).

Pour mettre en évidence les tissus enrichis en sur- ou sous-utilisation différentielle, nous avons calculé le pourcentage d'utilisation différentielle selon l'Équation 9. Après avoir éliminé les tissus présentant moins de 30 exons utilisés de manière différentielle, nous avons examiné ceux présentant un pourcentage d'utilisation différentielle > ± 33% (c'est-à-dire les tissus dont ⅔ des exons sont soit sur-exprimés, soit sous-exprimés).

$$Differential\ Usage\ Percentage\ (DUP) = 100\ *\ (\frac{Nb_{duxtExons-up} - Nb_{duxtExons-down}}{Nb_{duxtExons-up} + Nb_{duxtExons-down}})$$

Équation 9 – Formule du pourcentage de différentiel d'utilisation par tissus

En ce qui concerne les tissus enrichis en sous-utilisation différentielle (duxtExon-down), l'hypophyse (*pituitary*) (DUP= -88,6%, 33 duxtExons-down) est le tissu dont le DUP est le plus élevé. Le tissu testiculaire (DUP = -44,8%, 1 232 duxtExons-down) et le foie (DUP= -35%, 216 duxtExons-down) sont les tissus les plus fortement représentés en termes d'exons et de gènes (Figure 67).

Concernant les enrichissements en sur-utilisation différentielle, le tissu ayant le DUP le plus élevé est l'artère coronaire (DUP = +57,6%, 26 duxtExons-down). Les tissus observés comme étant les plus représentés en termes d'exons et de gènes sont les fibroblastes en culture (DUP = +43%,612 duxtExons-up) ainsi que le muscle squelettique (DUP = +36,7%, 493 duxtExons-up). Enfin, il est à noter que 4 tissus cérébraux (*Hippocampus, Cortex, Anterior cingulate cortex (BA24), Nucleus accumbens (basal ganglia)*) sont enrichis en sur-expression différentielle.

Ces résultats concernant les tissus enrichis en sous-utilisation (testicules, foie) et sur-utilisation d'exons (fibroblastes, muscle squelettique, certaines parties du cerveau) sont en accord avec des observations antérieures pointant ces mêmes tissus comme enrichis en événements d'épissage alternatif (Xu, Modrek, et Lee 2002; G. Yeo et al. 2004). Néanmoins, la métrique duxt offre un nouveau champ de réflexion en suggérant que l'épissage alternatif, au sein du foie ou des testicules, aura tendance à enrichir la population de transcrits en « transcrits sans un, ou des, exons ubiquitaires » tandis que ce mécanisme tendra à faire apparaitre des transcrits ayant des exons spécifiques statistiquement absents dans les autres tissus.

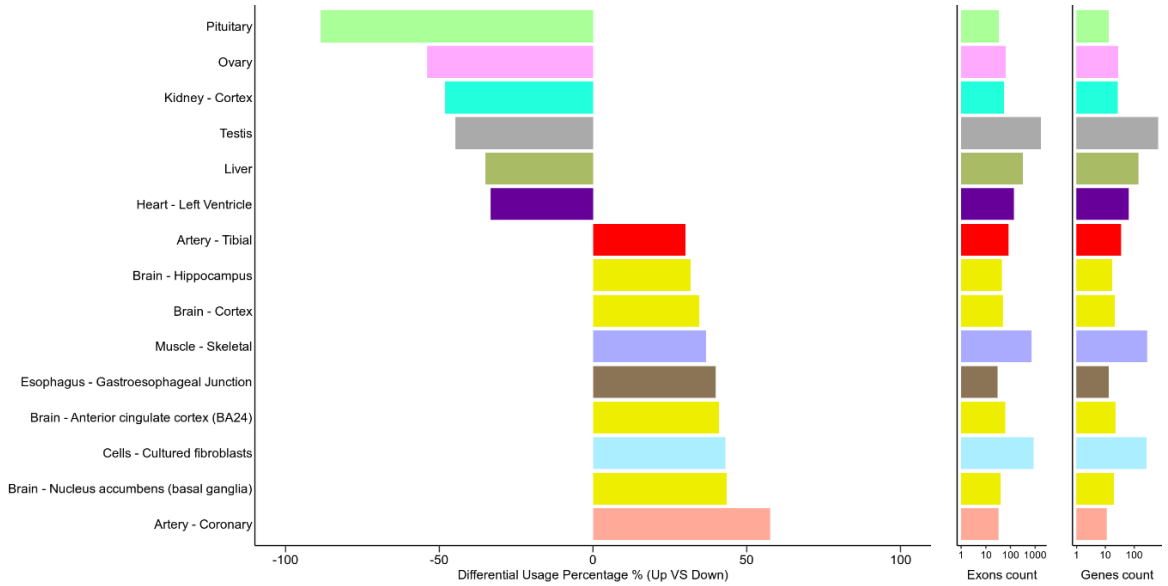


Figure 67 – Tissus présentant un enrichissement en différentiel de sur- ou sous-utilisation

La figure centrale représente le Differential Usage Percentage. Les figures de droites représentent respectivement le nombre d'exons et de gènes différentiellement utilisés par tissu.

9.2.4.2 Utilisation différentielle et maladies génétiques rares

Le spectre phénotypique associé à différents individus présentant la même maladie génétique est souvent très hétérogène. Bien que les conséquences moléculaires au niveau du gène et de l'isoforme protéique canonique soient assez bien étudiées, l'étude de l'impact d'une variation sur les différentes isoformes au regard de leur expression tissulaire est peu abordée. Dans ce contexte, nous avons cherché à observer si les utilisations différentielles d'exons détectées par duxt pouvaient être corrélées à des phénotypes atypiques rapportés dans la littérature.

9.2.4.2.1 Protocole exploratoire de recherche des utilisations différentielles en relation avec des phénotypes atypiques

Pour rechercher les différences phénotypiques, nous avons utilisé l'API OMIM (section 6.3.3) afin de récupérer les informations se rapportant aux entrées des gènes et aux phénotypes associés, aboutissant à un ensemble de 4 071 gènes reliés à 5 260 entrées phénotypiques distinctes. Après identification des groupes et sous-groupes anatomiques référencés dans la section *ClinicalSynopsis* d'OMIM, les données ont été reformatées en une matrice présence-absence de phénotypes en fonction des groupes anatomiques pour chaque gène associé. Une cartographie préliminaire entre les 53 tissus analysés par pext/duxt et les groupes anatomiques référencés dans le synopsis clinique d'OMIM est présentée dans le Tableau 16. La plupart des tissus présentent une correspondance directe avec des groupes anatomiques (*e.g.* foie, pancréas, cœur…). Cependant, en absence de correspondance explicite, les entrées OMIM qui nous ont semblées les plus adaptées aux tissus cibles ont été sélectionnées.

pext/duxt tissues	OMIM anatomical groups
Adipose - Subcutaneous	growthWeight
Adipose - Visceral (Omentum)	growthWeight
Adrenal Gland	endocrineFeatures
Artery - Aorta	cardiovascularVascular
Artery - Coronary	cardiovascularVascular
Artery - Tibial	neurologicPeripheralNervousSystem
Bladder	genitourinaryBladder
Brain - Amygdala	headAndNeck*, neurologic*
Brain - Anterior cingulate cortex (BA24)	headAndNeck*, neurologic*
Brain - Caudate (basal ganglia)	headAndNeck*, neurologic*
Brain - Cerebellar Hemisphere	headAndNeck*, neurologic*
Brain - Cerebellum	headAndNeck*, neurologic*
Brain - Cortex	headAndNeck*, neurologic*
Brain - Frontal Cortex (BA9)	headAndNeck*, neurologic*
Brain - Hippocampus	headAndNeck*, neurologic*
Brain - Hypothalamus	headAndNeck*, neurologic*
Brain - Nucleus accumbens (basal ganglia)	headAndNeck*, neurologic*
Brain - Putamen (basal ganglia)	headAndNeck*, neurologic*
Brain - Spinal cord (cervical c-1)	headAndNeck*, neurologic*
Brain - Substantia nigra	headAndNeck*, neurologic*
Breast - Mammary Tissue	chestBreasts
Cells - Cultured fibroblasts	/
Cells - EBV-transformed lymphocytes	immunology
Cervix - Ectocervix	genitourinaryExternalGenitaliaFemale
Cervix - Endocervix	genitourinaryInternalGenitaliaFemale
Colon - Sigmoid	abdomenGastrointestinal
Esophagus - Gastroesophageal Junction	abdomenGastrointestinal
Esophagus - Mucosa	/
Esophagus - Muscularis	/
Fallopian Tube	genitourinaryInternalGenitaliaFemale
Heart - Atrial Appendage	cardiovascularHeart
Heart - Left Ventricle	cardiovascularHeart
Kidney - Cortex	genitourinaryKidneys
Liver	abdomenLiver
Lung	respiratoryLung
Minor Salivary Gland	headAndNeckMouth, headAndNeckTeeth
Muscle - Skeletal	muscleSoftTissue
Nerve - Tibial	neurologicPeripheralNervousSystem
Ovary	genitourinaryInternalGenitaliaFemale
Pancreas	abdomenPancreas
Pituitary	endocrineFeatures
Prostate	genitourinaryInternalGenitaliaMale
Skin - Sun Exposed (Lower leg)	skinNailsHairSkin
Skin - Not Sun Exposed (Suprapubic)	skinNailsHairSkin
Small Intestine - Terminal Ileum	abdomenGastrointestinal
Spleen	abdomenSpleen
Stomach	abdomenGastrointestinal
Testis	genitourinaryExternalGenitaliaMale
Thyroid	endocrineFeatures
Uterus	genitourinaryInternalGenitaliaFemale
Vagina	genitourinaryExternalGenitaliaFemale
Whole Blood	hematology

Tableau 16 – Associations Tissus d'expression (pext/duxt) – groupes anatomiques (OMIM)

** : groupes englobant les sous-groupes anatomiques : neurologic (BehavioralPsychiatricManifestations, CentralNervousSystem) ; headAndNeck (Ears, Eyes, Face, Head, Mouth, Neck, Nose, Teeth)*

Dans cette étude exploratoire, nous nous sommes focalisés sur l'étude des variations délétères présentes dans des exons différentiellement sur-utilisés (duxtExon-up), c'est-à-dire des exons fortement utilisés dans un/quelques tissus. Ce choix a été guidé par l'hypothèse qu'il serait plus facile d'identifier l'apparition d'un phénotype « atypique » liée à la sur-utilisation d'un exon porteur de variation délétère dans un/quelques tissus plutôt que la disparition d'un phénotype « fréquent » lié à une faible utilisation d'un exon portant une variation délétère (duxtExon-down) dans un/quelques tissus.

Les gènes présentant des phénotypes atypiques ont été sélectionnés par un protocole automatisé de filtrage selon les critères suivants :

1) gènes associés à plusieurs pathologies (un tableau clinique par pathologie) dans OMIM et présentant un organe/tissu spécifiquement atteint dans une de ces pathologies : 375 gènes
2) gènes pour lesquels l'organe/tissu spécifique d'une pathologie (OMIM) correspond au tissu où un duxtExon est observé : 219 gènes
3) gènes présentant au moins une variation délétère référencée dans ClinVar à l'intérieur du duxtExon cible : 9 gènes

9.2.4.2.2 Identification de cas répertoriés et analyse approfondie

Après application des filtres précédemment mentionnés, nous avons identifié 17 variations délétères dans 14 duxtExon-up dans 9 gènes. L'ensemble de ces variations étaient référencées à la fois dans ClinVar et OMIM, avec une description contenant un résumé des publications où ces variations étaient mentionnées.

Suite à la lecture des publications, nous présentons un cas retenu comme intéressant dans le cadre de cette recherche exploratoire (Tableau 17). Ce cas correspond au gène NTRK2 (Récepteur tyrosine kinase neurotrophique) ou BDNF (Facteur neurotrophique dérivé du cerveau), dont plusieurs exons de la partie C-terminale de la protéine (3' du gène) sont différentiellement utilisés dans l'hémisphère cérébelleux (cervelet). Par comparaison des tableaux cliniques entre les deux pathologies associées au gène dans OMIM, notre protocole a identifié la microcéphalie comme spécifique du phénotype DEE58 (*Developmental And Epileptic Encephalopathy* 58). À l'intérieur des duxtExons, une variation faux-sens délétère (p.Thr720Ile) a été référencée dans une publication (Hamdan et al. 2017). La publication associée à cette variation décrit l'analyse de cinq patients atteints de DEE58, quatre portant une variation faux-sens (p.Tyr434Cys) dans un exon constitutif et la cinquième (p.Thr720Ile) dans un des exons alternatifs identifiés par duxt, toutes étant des variations *de novo*.

Après analyse des caractéristiques cliniques des patients, nous nous sommes aperçus que notre protocole automatique présentait une limite, car la microcéphalie acquise, identifiée dans la comparaison des tableaux cliniques OMIM est associée chez 2 des 4 patients au faux-sens Tyr434Cys présent dans les exons constitutifs.

Cependant, plusieurs différences phénotypiques entre les deux types de patients ont pu être relevées. Premièrement, le handicap mental (*Intellectual Disability* ; ID), et le retard global de développement (*Global Developmental Delay* ; GDD) sont sévères chez les 4 patients porteurs de Tyr434Cys, tandis que la patiente Thr720Ile présente une ID et un GDD modérés. Deuxièmement, les 4 patients Tyr434Cys présentent une hypoplasie du nerf optique non retrouvée chez la patiente Thr720Ile. Enfin, la patiente Thr720Ile est la seule à présenter une hyperphagie et une obésité précoce (3 ans), phénotypes retrouvés pour une variation adjacente [p.Tyr722Cys ; (G. S. H. Yeo et al. 2004)] et qui semblent liés aux mécanismes de signalisation entre NTRK2 et la neurotrophine-4.

Gene	*Gene name*	*NTRK2*
	OMIM gene ID	*600456*
duxt	*up/down*	*duxt-up*
	duxt score	*0,66*
	Tissue	*Brain - Cerebellar Hemisphere*
Phenotype & Disease	*PMID*	*29100083*
	Patient(s)	*HSJ0335*
	OMIM Phenotype (ID)	*Developmental and epileptic encephalopathy 58 (617830)*
Variant	*Genomic variant coordinates (GRCh37)*	*9-87570419-C-T*
	HGVS Nucleotide / Protein	*NM_006180.4:c.2159C>T NP_006171.2:p.Thr720Ile*
	OMIM variant ID	*600456.0004*
	ClinVar Variation ID	*487685*
	Variation type	*missense*
	gnomAD global MAF	*No value*
	MISTIC	*0,832*
	Inheritance	*Dominant*

Tableau 17 –Tableau récapitulatif de la variation délétère et des phénotypes associés

9.2.4.3 Les variations génétiques affectant l'épissage alternatif sont liées à une utilisation différentielle.

L'épissage alternatif à l'origine des différentes populations de transcrits au sein des tissus peut être affectée par des variations génétiques bénignes ou pathogènes. Ces variations affectant l'épissage sont appelées sQTL (*splicing Quantitative Trait Loci* ; section 2.2.4). Ainsi, lors de la détection des sQTL réalisée dans GTEx à partir des 54 tissus de 1000 individus sains, chaque sQTL a été caractérisé selon sa position génomique et le tissu où la population des transcrits était altérée. Dans ce cadre, nous avons cherché à évaluer si ces sQTL affectaient préférentiellement des utilisations différentielles d'exons identifiées par duxt.

Les sQTL ont été extraits du catalogue établi par l'outil sqtlseeker2 (Garrido-Martín et al. 2021) sur la base des données GTEx. Sur les 344 211 sQTL identifiés dans 9 051 gènes de GTEx V7, 41 670 sQTL sont présents dans 1 036 MISOG avec au moins un exon alternatif fortement différentiellement utilisé (duxtExon). Parmi ces 1 036 gènes, deux types de sQTLs ont été distingués : (1) les sQTLs affectant l'expression dans le même tissu que celui où une utilisation différentielle d'exon a été identifiée par duxt = **duxt-sQTLs** et (2), les sQTLs affectant l'expression dans d'autres tissus que ceux identifiés par duxt = **others-sQTLs**.

Les paramètres de Garrido-Martin et collaborateurs (Garrido-Martín et al. 2021) pour évaluer l'importance des sQTL sur la transcription, sont : le score *md* qui correspond à « la différence maximale absolue d'expression relative ajustée des transcrits entre groupes de génotypes » et trois catégories de scores discrets de md : md de faible importance, md modérée et md de forte importance (Figure 68). Ce paramètre *md* reflète l'ampleur de la transition entre deux populations de transcrits associées à des individus présentant des génotypes différents.

Pour comparer les populations de duxt-sQTL et others-sQTL, l'impact des sQTL, estimé par le score md, a été pris en compte en comparant la médiane statistique des md et l'enrichissement dans les trois catégories discrètes : md de faible importance, md modérée et md de forte importance (Figure 68). De plus, les situations de sous-utilisation (duxtExon-down) et de sur-utilisation des exons (duxtExon-up) ont été distinguées.

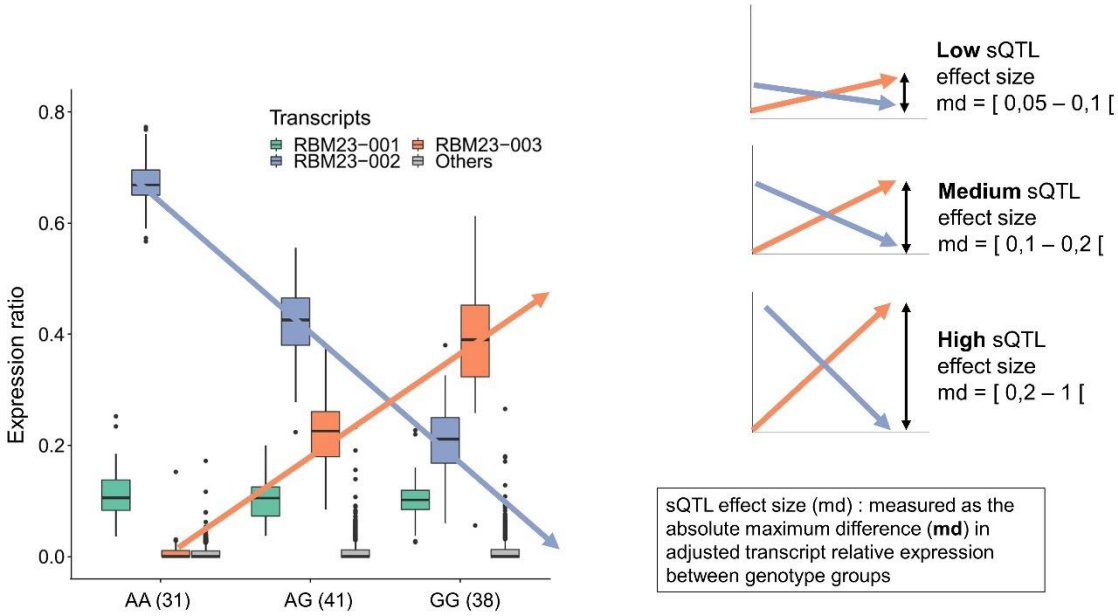


Figure 68 – Notion d'importance de l'effet des sQTL

Source : (Garrido-Martín et al. 2021)

L'analyse concernant les exons différentiellement sur-utilisés (duxtExons-up) (Figure 69a) a identifié une légère augmentation de la médiane du md entre les others-sQTLs (0,087) et les duxt-sQTLs (0,103) ainsi qu'un enrichissement statistique en duxt-sQTLs à md modéré (*Odds Ratio* = 2,99) et un appauvrissement en duxt-sQTLs à md élevé (*Odds Ratio* = 0,05). Concernant les exons différentiellement sous-utilisés (duxtExons-down) (Figure 69b), on observe une augmentation de la médiane md entre les duxt-sQTLs (*Odds Ratio* = 0,149) et les others-sQTLs (*Odds Ratio* = 0,109) ainsi qu'un enrichissement important des dsQTLs à md élevé (*Odds Ratio* = 9,99) et une diminution des duxt-sQTLs à md modéré (*Odds Ratio* = 0,60).

Ainsi, de manière générale, on note une importance d'effet plus élevé pour les sQTL dont le tissu est identique au tissu présentant un exon différentiellement sous-utilisé (Figure 70). Cependant, on note que les duxt-sQTL associés à des exons différentiellement sous-utilisés (duxtExons-down) sont associés à un changement de population de transcrits beaucoup plus marqué (lié au paramètre md).

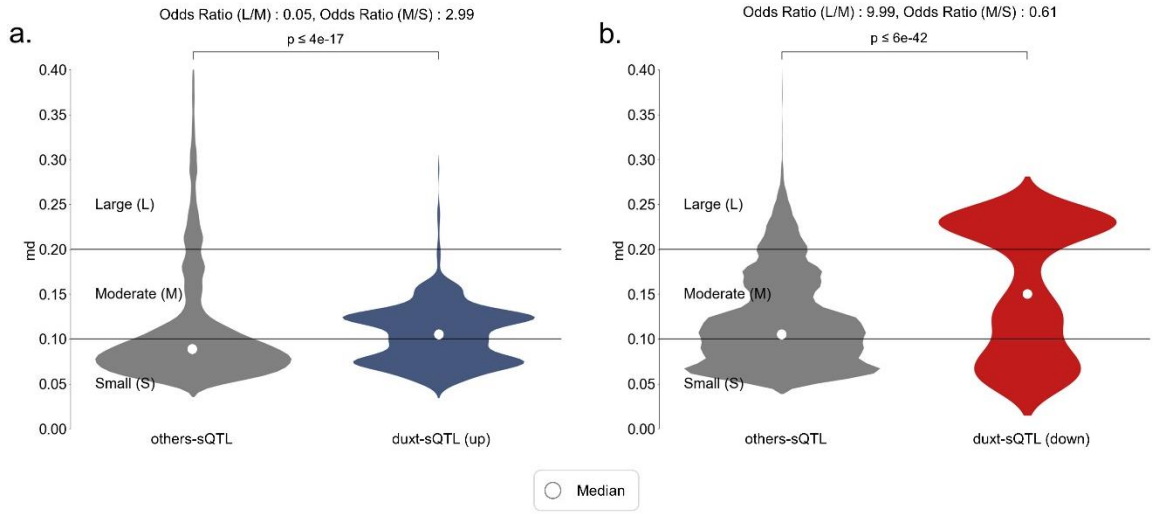


Figure 69 – Comparaison de l'importance de l'effet des sQTL

La figure (a) compare la distribution de l'importance de l'effet des sQTL entre sQTL associés à des exons différentiellement sur-utilisés dans un tissu (duxt-sQTL(up)) et les sQTL détectés dans d'autres tissus. La figure (b) illustre la même comparaison, mais concernant les exons différentiellement sous-utilisés pour un tissu (duxt-sQTL(down)). Les seuils utilisés par l'auteur ont été réutilisés pour définir des catégories discrètes (Small : 0,05 – 0,1 ; Moderate : 0,1 – 0,2 ; Large : > 0,2). Les Odds Ratio ont été utilisés afin de comparer les enrichissements en catégories discrètes.

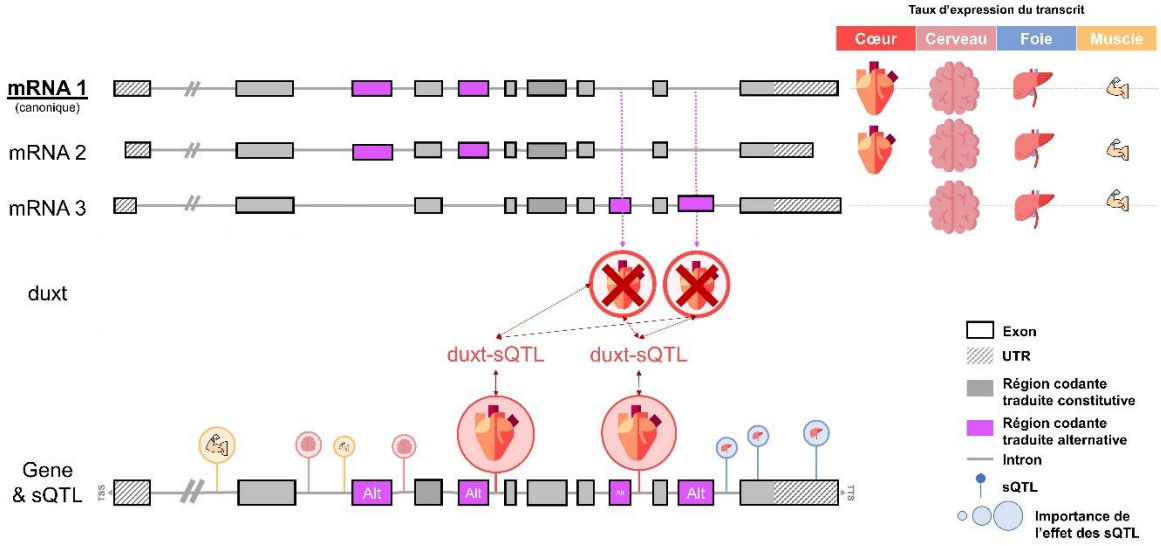


Figure 70 – Illustration de l'association entre sQTL et duxtExons

Pour un gène présentant 3 mRNAs et 4 exons alternatifs, une absence du mRNA 3 dans le cœur conduit à une sous-utilisation différentielle des exons alternatifs spécifique de ce mRNA et donc à leur absence du cœur. Les sQTL détectés dans le cœur présentent une importance d'effet plus élevée que dans les sQTL détectés dans les autres tissus du même gène.

9.2.5 Conclusion et perspectives

Les travaux présentés dans ce Chapitre visaient à mieux évaluer l'impact potentiel de variations génétiques situées dans des exons alternatifs. Ceci s'est révélé beaucoup plus laborieux qu'attendu. En effet, il m'a fallu, dans un premier temps, m'éloigner des ressources de références qui amalgament souvent données expérimentales et prédictions, pour identifier les exons constitutifs et alternatifs sur la base de données transcriptomiques provenant d'une série de tissus humains. Puis, j'ai dû vérifier que les propriétés de ces deux populations (taille, nombre, conservation-contrainte, distribution des variations…) étaient comparables. Cette comparaison a fait ressortir deux éléments : les exons alternatifs sont moins contraints et moins conservés que les exons constitutifs, mais ne sont pas sous-représentés en quantité de variations délétères qu'ils abritent. Dès lors, la prise en compte des variations présentes dans des exons alternatifs pourrait constituer un réel enjeu pour un meilleur diagnostic moléculaire et une évaluation plus précise des phénotypes associés.

Pour aborder cette problématique, nous avons été amenés à intégrer les données d'expression en développant duxt, une nouvelle métrique permettant de caractériser les différentiels d'utilisation transcriptionnel d'exons alternatifs dans les 53 tissus disponibles dans la base de données GTEx. Grâce à la métrique duxt, nous avons pu identifier des tissus enrichis en exons à utilisation différentielle, c'est-à-dire des exons fortement présents ou absents dans un ou quelques tissus comparativement aux autres tissus. Les tissus identifiés (testicules, cerveau, foie, muscle…) correspondent aux tissus à fort taux d'épissage alternatif déjà identifiés dans la littérature. Néanmoins, duxt nous a permis de distinguer les tissus où des événements « d'extinction » d'exons quasi-constitutifs avaient fréquemment lieu des tissus « d'apparition » fréquente d'exons par ailleurs absents. Cette distinction pourrait s'avérer majeure pour l'évaluation d'un VUS ou d'une variation délétère, voire bénigne, si l'on considère les complexes ou réseaux protéiques au sein desquels une isoforme mutée serait susceptible d'interagir. En effet, dans les tissus à fort taux « d'apparition » d'exons, tel que le muscle, il existe statistiquement de nombreux gènes ayant des isoformes spécifiques sur-représentées au sein des populations d'isoformes (Rodriguez et al. 2020). Il serait donc intéressant de vérifier si ces différentes isoformes ne participent pas aux mêmes complexes ou réseaux susceptibles d'être particulièrement impactés par la présence d'une variation. À l'inverse, il semble plus difficile d'évaluer les conséquences d'une variation présente sur un exon quasi-constitutif absent spécifiquement de certains tissus, si ce n'est, peut-être, sous l'angle de la disparition des phénotypes associés aux tissus riches en « extinction » d'exons.

Par la suite, nous avons cherché à identifier des cas de phénotype atypique pouvant être associés à des variations délétères présentes dans des exons différentiellement utilisés. La

première version de notre protocole a permis d'identifier 9 gènes, dont un (NTRK2) présentant des phénotypes divergents entre quatre patients touchés par une variation délétère dans un exon constitutif et une patiente impactée par une variation délétère dans un exon alternatif ayant un différentiel d'utilisation important dans l'hémisphère cérébelleux.

Ces résultats préliminaires sont encourageants, mais indiquent clairement que le protocole établi est à améliorer. Ainsi, le seuil strict de 0,5 pour définir un duxtExon représente sans doute un facteur limitant pour la recherche des associations duxtExons – phénotypes. De même, en absence de référence, nous avons choisi de tester uniquement des exons alternatifs de type duxtExon-up en imaginant que la sur-représentation, au sein des populations des transcrits, de transcrits avec un exon atypique porteur d'une variation délétère serait plus impactante au niveau phénotypique et donc, plus aisée à identifier. Une analyse systématique des associations duxtExons – phénotypes pour les deux types de duxtExons pourrait questionner le bien-fondé de notre choix et mieux définir le seuil à utiliser dans de futures études. De même, la méthode d'identification d'un phénotype atypique chez un patient est à revoir. Dans cette première version, nous avons comparé les tableaux cliniques des différentes pathologies référencés dans OMIM pour un même gène (Tableau 16). Cette comparaison nécessiterait d'être améliorée en fixant des règles plus détaillées afin de s'affranchir d'associations tissus – groupes anatomiques non pertinentes. La structuration actuelle des bases de données et leur consultation fait que la majorité des informations phénotypiques disponibles sont associés au gène et rarement à l'impact précis d'une variation. De plus, le nombre de cas documentés disponibles limite ce type d'analyse, qui deviendra d'autant plus pertinente avec l'expansion continue des bases de données biomédicales. Certaines des limites évoquées pourraient être levées par l'intégration dans notre protocole, d'outils de *text mining*, tel que PubTator (Wei et al. 2019), cependant, pour le moment, la lecture détaillée des publications demeure une étape limitante et incontournable.

Enfin, nous avons observé une corrélation entre les variations affectant la machinerie d'épissage alternatif (sQTL) et des exons différentiellement utilisés. Ces sQTL sont identifiés en associant statistiquement la présence d'une variation génétique (chez des individus sains) à un changement drastique de population d'isoformes dans un tissu spécifique. La plus forte corrélation duxt-sQTL a été observée pour les sQTL détectés dans des tissus où des duxtExons étaient spécifiquement absents. On peut donc émettre l'hypothèse que cette « extinction » d'exon dans un tissu donné pourrait être perturbée par la variation/sQTL, entrainant l'expression d'un isoforme ne devant pas s'exprimer de manière classique dans le tissu cible. Ce résultat intermédiaire, néanmoins encourageant, pourrait laisser supposer un lien aux mécanismes fortement contraints que suppose l'épissage alternatif dans un tissu.

DISCUSSION

Chapitre 10. Discussion ouverte et perspectives

La séquence complète du génome humain, esquissée en 2001 et complétée au fil de ces deux décennies, a ouvert la voie à d'innombrables avancées en biologie (Gates et al. 2021). Ces avancées se rapportent aussi bien à la découverte de l'ensemble des gènes non codant et codant pour des protéines que l'obtention, pour de nombreux tissus humains, d'informations sur l'expression des gènes ou sur les interactions entre leurs produits. Une autre avancée majeure a concerné l'identification d'un nombre sans précédent de variations génétiques, fruits du hasard, de l'évolution et du brassage génétique nécessaire à la stabilité et à la pérennité de notre espèce (Auton et al. 2015). Ces variations génétiques, source de l'hétérogénéité interindividuelle humaine, sont aujourd'hui au cœur des plus grandes questions de la recherche biomédicale, allant de la compréhension de leurs rôles dans les maladies génétiques ou communes jusqu'à l'importance de certains traits ou susceptibilités face à des environnements complexes et en plein dérèglement [section 2.2 ; (Frazer et al. 2009)].

10.1 Prédiction de l'impact des variations et intégration des « omiques »

Dès à présent, grâce aux biotechnologies à haut débit, l'analyse d'un simple changement de nucléotide à une position du génome mobilise une avalanche de données observées, estimées ou prédites. Ces données, longtemps limitées aux seuls champs génique et génomique explorés sur la base de milliers d'individus (fréquences alléliques, conservations, conséquences moléculaires…), commencent à intégrer de nouveaux niveaux de complexité liés notamment, à l'expression des gènes et aux populations de transcrits disparates qui en découlent, selon les individus, les cellules ou les stades de développement. Dès lors, on peut aisément penser qu'à l'avenir, chaque variation génétique s'appréciera en fonction de nouveaux niveaux d'organisation du vivant.

Ainsi, par-delà la prise en compte d'un nombre croissant d'individus et de situations, la variation génétique pourra s'examiner sous l'angle des répercussions éventuelles sur la fixation de la machinerie transcriptionnelle, sur la formation de structures atypiques d'ARN (Gaither et al. 2021) ou sur les vitesses de traduction ribosomale, *pausing* et décrochage (*ribosomal stall*) inclus (Collart et Weiss 2020). Les dynamiques de création et de maintien des complexes et réseaux biologiques (Swaney et al. 2021) seront sans doute également mis à profit afin de comprendre intimement les multiples mécanismes impactés par la

présence d'une variation délétère au sein d'une protéine ou d'un ARN. Bien entendu, ces analyses complexes amélioreront aussi la caractérisation d'autres types de variations depuis celles dont le statut clinique est inconnu (VUS) jusqu'aux variants de population ou structuraux. Ces avancées préfigurent l'avenir d'une médecine génomique, toujours plus personnalisée, à même de considérer chaque variation dans le paysage des millions de variations individuelles et de leur relation au sein des divers niveaux d'organisation biologique.

10.2 Prédiction et explicabilité

Il semble évident qu'un horizon aussi ambitieux ne peut s'envisager sans une évolution à la hauteur de l'informatique et de la bioinformatique. Sans risque de se tromper, on peut prédire que l'intelligence artificielle (IA) sera au cœur de cette évolution, notamment grâce à sa capacité à optimiser l'emploi des ressources informatiques pour faire émerger un signal fiable à partir de masses de données exponentielles. Cependant, le domaine de la santé, comme celui des sciences sociales (Zeng, Ustun, et Rudin 2017), ne saurait se contenter du seul critère de performance d'une IA et, d'ores et déjà, le besoin s'impose d'explications compréhensibles par l'humain sur le « raisonnement statistique » qui a amené un modèle à proposer une solution, fût-elle optimale.

Cette notion d'explicabilité, qui est un domaine de recherche de plus en plus actif en IA, sera au cœur des futurs prédicteurs de l'impact des variations génétiques. En effet, la plupart des modèles d'IA développés et plus particulièrement, en apprentissage profond, fonctionnent aujourd'hui sur le modèle d'une « boite noire », fournissant peu d'informations sur les modalités de genèse du modèle ou d'élaboration d'une réponse (Zednik 2021). Néanmoins, divers développements, notamment au sein du laboratoire, sont en cours afin de construire des systèmes robustes et explicables, permettant d'apporter à l'expert et à l'utilisateur humain, un raisonnement détaillé sur les éléments qui ont orienté le modèle pour distinguer les situations et proposer une solution (Orhand et al. 2019; Lauritsen et al. 2020).

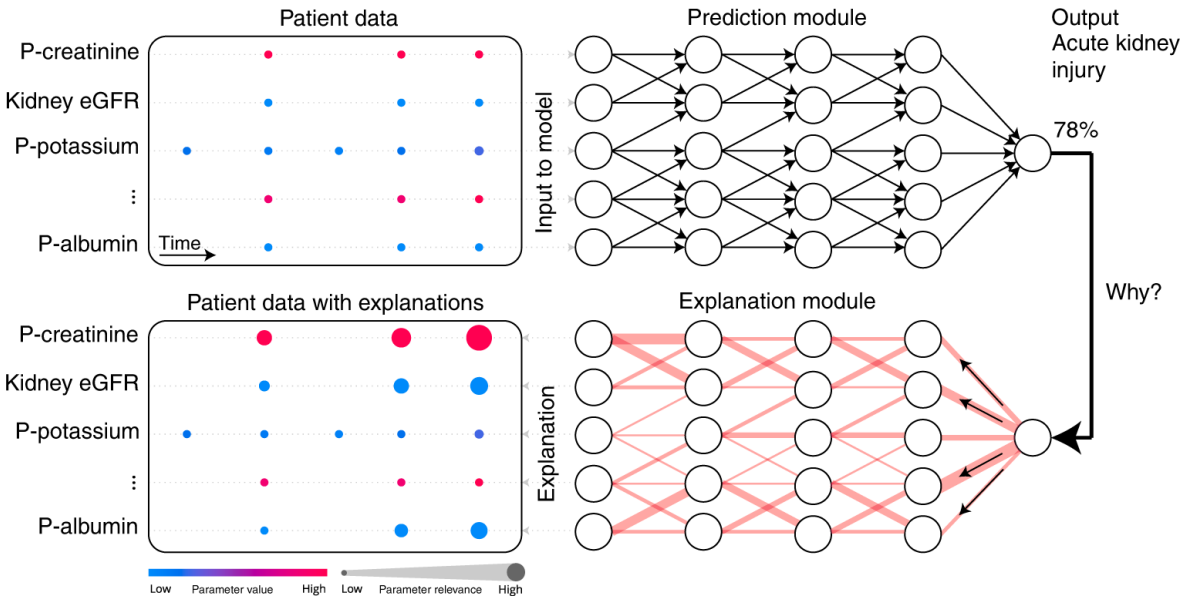


Figure 71 – Exemple d'explicabilité

Cette figure illustre l'explicabilité dans un modèle de « deep learning » dont l'objectif est de prédire les maladies graves aiguës à partir des données de santé électronique (Electronic Health Records ; EHR). Les données EHR du patient ont été utilisées comme entrée dans le module de prédiction d'un réseau de neurone convolutif temporel (Temporal Convolutional Network ; TCN). Sur la base de ces données, le modèle établit une prédiction, par exemple ici un risque de 78 % de présenter une insuffisance rénale aiguë. Le module d'explication de type DTD (Deep Taylor Decomposition) explique ensuite les prédictions du réseau à partir des valeurs des variables d'entrée. P, plasma; eGFR, estimated Glomerular Filtration Rate. Source : (Lauritsen et al. 2020)

10.3 Vers une logique de segmentation des problèmes

Les futurs prédicteurs de l'impact des variations nécessiteront également une architecture particulière afin de gérer de manière adéquate l'intrication grandissante entre des propriétés descriptives des variations (fréquence allélique, conservation, propriétés physico-chimiques…), des prédicteurs intégrant ces mêmes informations et des méta-prédicteurs intégrant ces mêmes prédicteurs.

Dans ce cadre, une autre problématique émerge reliée aux prédicteurs et à leur futur champ d'application. Étant donné les écarts de performance séparant d'une part, les approches de , les logiques de prédiction de l'impact des variations tendent à se « génaliser » par le développement de modèles spécifiques à chaque gène ou groupes de gènes ou *via* une re-calibration à façon de modèles généraux (Lali et al. 2021; Price et al. 2010; van der Velde et al. 2017).

10.4 Protection des données et éthiques

Enfin, dans le cadre d'exploitations futures des variations génétiques, il me semble crucial d'aborder les notions de protection des données et d'éthique. En effet, la société numérique dans laquelle nous vivons a connu, et connaît toujours, des innovations technologiques que les institutions gouvernementales ont du mal à anticiper et à réglementer. Dans le cadre du Règlement Général de la Protection des Données (RGPD), chaque organisation se doit de protéger les données de ses utilisateurs/clients dans un objectif de confidentialité et de protection de la vie privée. Ces données comprennent aussi bien des informations généralistes (identité, localisation, identifiants de connexion…) que des données ayant trait au domaine de la santé (Amselem et al. 2021). Compte tenu de l'évolution fulgurante du séquençage génomique individualisé et des traitements afférents, on peut se demander comment concilier, demain, le besoin impérieux de confidentialité face à des usages malintentionnés et les bénéfices multiples d'accès au génome d'un patient comme complément des bilans médicaux. Là encore, on peut penser que l'IA participera sans doute à la résolution de cet épineux problème en segmentant, masquant ou cryptant judicieusement des informations identifiantes tout en fournissant au demandeur autorisé les informations pertinentes et circonscrites en adéquation précise à la question posée.

Références

Abadi, Martın, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, et al. 2015. « TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems », 19.

Accetturo, Matteo, Nicola Bartolomeo, et Alessandro Stella. 2020. « In-silico Analysis of NF1 Missense Variants in ClinVar: Translating Variant Predictions into Variant Interpretation and Classification ». *International Journal of Molecular Sciences* 21 (3): 721. https://doi.org/10.3390/ijms21030721.

Ahmed, Shabbir, Zhan Zhou, Jie Zhou, et Shu-Qing Chen. 2016. « Pharmacogenomics of Drug Metabolizing Enzymes and Transporters: Relevance to Precision Medicine ». *Genomics, Proteomics & Bioinformatics*, SI: Big Data and Precision Medicine, 14 (5): 298-313. https://doi.org/10.1016/j.gpb.2016.03.008.

Ainscough, Benjamin J., Malachi Griffith, Adam C. Coffman, Alex H. Wagner, Jason Kunisaki, Mayank NK Choudhary, Joshua F. McMichael, et al. 2016. « DoCM: A Database of Curated Mutations in Cancer ». *Nature Methods* 13 (10): 806-7. https://doi.org/10.1038/nmeth.4000.

Albert, Thomas J., Michael N. Molla, Donna M. Muzny, Lynne Nazareth, David Wheeler, Xingzhi Song, Todd A. Richmond, et al. 2007. « Direct Selection of Human Genomic Loci by Microarray Hybridization ». *Nature Methods* 4 (11): 903-5. https://doi.org/10.1038/nmeth1111.

Alirezaie, Najmeh, Kristin D. Kernohan, Taila Hartley, Jacek Majewski, et Toby Dylan Hocking. 2018. « ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants ». *American Journal of Human Genetics* 103 (4): 474-83. https://doi.org/10.1016/j.ajhg.2018.08.005.

Amberger, Joanna S., Carol A. Bocchini, Alan F. Scott, et Ada Hamosh. 2019. « OMIM.Org: Leveraging Knowledge across Phenotype-Gene Relationships ». *Nucleic Acids Research* 47 (D1): D1038-43. https://doi.org/10.1093/nar/gky1151.

Ameur, Adam, Johan Dahlberg, Pall Olason, Francesco Vezzi, Robert Karlsson, Marcel Martin, Johan Viklund, et al. 2017. « SweGen: A Whole-Genome Data Resource of Genetic Variability in a Cross-Section of the Swedish Population ». *European Journal of Human Genetics* 25 (11): 1253-60. https://doi.org/10.1038/ejhg.2017.130.

Amselem, Serge, Sonia Gueguen, Jérôme Weinbach, Annick Clement, Paul Landais, et for the RaDiCo Program. 2021. « RaDiCo, the French national research program on rare disease cohorts ». *Orphanet Journal of Rare Diseases* 16 (1): 454. https://doi.org/10.1186/s13023-021-02089-5.

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, et al. 2014. « An Atlas of Active Enhancers across Human Cell Types and Tissues ». *Nature* 507 (7493): 455-61. https://doi.org/10.1038/nature12787.

Auton, Adam, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, Gonçalo R. Abecasis, David R. Bentley, Aravinda Chakravarti, et al. 2015. « A Global Reference for Human Genetic Variation ». *Nature* 526 (7571): 68-74. https://doi.org/10.1038/nature15393.

Baker, Monya. 2012. « Structural Variation: The Genome's Hidden Architecture ». *Nature Methods* 9 (2): 133-37. https://doi.org/10.1038/nmeth.1858.

Benson, Dennis A, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, James Ostell, Kim D Pruitt, et Eric W Sayers. 2018. « GenBank ». *Nucleic Acids Research* 46 (D1): D41-47. https://doi.org/10.1093/nar/gkx1094.

Beyter, Doruk, Helga Ingimundardottir, Asmundur Oddsson, Hannes P. Eggertsson, Eythor Bjornsson, Hakon Jonsson, Bjarni A. Atlason, et al. 2021. « Long-Read Sequencing of 3,622 Icelanders Provides Insight into the Role of Structural Variants in Human Diseases and Other Traits ». *Nature Genetics* 53 (6): 779-86. https://doi.org/10.1038/s41588-021-00865-4.

Boeck, Kris De. 2020. « Cystic Fibrosis in the Year 2020: A Disease with a New Face ». *Acta Paediatrica* 109 (5): 893-99. https://doi.org/10.1111/apa.15155.

Bonfield, James K., John Marshall, Petr Danecek, Heng Li, Valeriu Ohan, Andrew Whitwham, Thomas Keane, et Robert M. Davies. 2021. « HTSlib: C Library for Reading/Writing High-Throughput Sequencing Data ». *GigaScience* 10 (2): giab007. https://doi.org/10.1093/gigascience/giab007.

Brandt, Margot, et Tuuli Lappalainen. 2017. « SnapShot: Discovering Genetic Regulatory Variants by QTL Analysis ». *Cell* 171 (4): 980-980.e1. https://doi.org/10.1016/j.cell.2017.10.031.

Cadet, Jean, et J. Richard Wagner. 2013. « DNA Base Damage by Reactive Oxygen Species, Oxidizing Agents, and UV Radiation ». *Cold Spring Harbor Perspectives in Biology* 5 (2): a012559. https://doi.org/10.1101/cshperspect.a012559.

Carr, Steven. 2014. « Transitions vs transversions ». 2014. https://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html.

Cartegni, Luca, Shern L. Chew, et Adrian R. Krainer. 2002. « Listening to Silence and Understanding Nonsense: Exonic Mutations That Affect Splicing ». *Nature Reviews. Genetics* 3 (4): 285-98. https://doi.org/10.1038/nrg775.

Caspar, S. M., N. Dubacher, A. M. Kopps, J. Meienberg, C. Henggeler, et G. Matyas. 2018. « Clinical Sequencing: From Raw Data to Diagnosis with Lifetime Value ». *Clinical Genetics* 93 (3): 508-19. https://doi.org/10.1111/cge.13190.

Chaisson, Mark J. P., Ashley D. Sanders, Xuefang Zhao, Ankit Malhotra, David Porubsky, Tobias Rausch, Eugene J. Gardner, et al. 2019. « Multi-Platform Discovery of Haplotype-Resolved Structural Variation in Human Genomes ». *Nature Communications* 10 (1): 1784. https://doi.org/10.1038/s41467-018-08148-z.

Chatterjee, Nimrat, et Graham C. Walker. 2017. « Mechanisms of DNA Damage, Repair, and Mutagenesis: DNA Damage and Repair ». *Environmental and Molecular Mutagenesis* 58 (5): 235-63. https://doi.org/10.1002/em.22087.

Chen, Dong, Xudong Cao, Fang Wen, et Jian Sun. 2013. « Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification ». In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 3025-32. Portland, OR, USA: IEEE. https://doi.org/10.1109/CVPR.2013.389.

Chen, Peng, Jiajun Gu, Eric Brandin, Young-Rok Kim, Qiao Wang, et Daniel Branton. 2004. « PROBING SINGLE DNA MOLECULE TRANSPORT USING FABRICATED NANOPORES ». *Nano letters* 4 (11): 2293-98. https://doi.org/10.1021/nl048654j.

Chen, Zhen, Pei Zhao, Fuyi Li, André Leier, Tatiana T Marquez-Lago, Yanan Wang, Geoffrey I Webb, et al. 2018. « IFeature: A Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences ». Édité par Alfonso Valencia. *Bioinformatics* 34 (14): 2499-2502. https://doi.org/10.1093/bioinformatics/bty140.

Cheng, Jun, Muhammed Hasan Çelik, Anshul Kundaje, et Julien Gagneur. 2021. « MTSplice predicts effects of genetic variants on tissue-specific splicing ». *Genome Biology* 22 (1): 94. https://doi.org/10.1186/s13059-021-02273-7.

Choudhury, Ananyo, Shaun Aron, Laura R. Botigué, Dhriti Sengupta, Gerrit Botha, Taoufik Bensellak, Gordon Wells, et al. 2020. « High-Depth African Genomes Inform Human Migration and Health ». *Nature* 586 (7831): 741-48. https://doi.org/10.1038/s41586-020-2859-7.

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

Claussnitzer, Melina, Judy H. Cho, Rory Collins, Nancy J. Cox, Emmanouil T. Dermitzakis, Matthew E. Hurles, Sekar Kathiresan, et al. 2020. « A Brief History of Human Disease Genetics ». *Nature* 577 (7789): 179-89. https://doi.org/10.1038/s41586-019-1879-7.

Collart, Martine A, et Benjamin Weiss. 2020. « Ribosome pausing, a dangerous necessity for co-translational events ». *Nucleic Acids Research* 48 (3): 1043-55. https://doi.org/10.1093/nar/gkz763.

Consortium, The GTEx. 2020. « The GTEx Consortium Atlas of Genetic Regulatory Effects across Human Tissues ». *Science* 369 (6509): 1318-30. https://doi.org/10.1126/science.aaz1776.

Cooper, D. N., et M. Krawczak. 1996. « Human Gene Mutation Database ». *Human Genetics* 98 (5): 629. https://doi.org/10.1007/s004390050272.

Cummings, Beryl B., Konrad J. Karczewski, Jack A. Kosmicki, Eleanor G. Seaby, Nicholas A. Watts, Moriel Singer-Berk, Jonathan M. Mudge, et al. 2020. « Transcript Expression-Aware Annotation Improves Rare Variant Interpretation ». *Nature* 581 (7809): 452-58. https://doi.org/10.1038/s41586-020-2329-2.

Cummings, Beryl B., Jamie L. Marshall, Taru Tukiainen, Monkol Lek, Sandra Donkervoort, A. Reghan Foley, Veronique Bolduc, et al. 2017. « Improving genetic diagnosis in Mendelian disease with transcriptome sequencing ». *Science Translational Medicine* 9 (386): eaal5209. https://doi.org/10.1126/scitranslmed.aal5209.

Dahm, Ralf. 2005. « Friedrich Miescher and the Discovery of DNA ». *Developmental Biology* 278 (2): 274-88. https://doi.org/10.1016/j.ydbio.2004.11.028.

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

Dehghan, Abbas. 2018. « Genome-Wide Association Studies ». *Methods in Molecular Biology (Clifton, N.J.)* 1793: 37-49. https://doi.org/10.1007/978-1-4939-7868-7_4.

Desouky, Omar, Nan Ding, et Guangming Zhou. 2015. « Targeted and non-targeted effects of ionizing radiation ». *Journal of Radiation Research and Applied Sciences* 8 (2): 247-54. https://doi.org/10.1016/j.jrras.2015.03.003.

Dunnen, Johan T. den, et Stylianos E. Antonarakis. 2000. « Mutation Nomenclature Extensions and Suggestions to Describe Complex Mutations: A Discussion ». *Human Mutation* 15 (1): 7-12. https://doi.org/10.1002/(SICI)1098-1004(200001)15:1<7::AID-HUMU4>3.0.CO;2-N.

Durland, Justin, et Hamid Ahmadian-Moghadam. 2021. « Genetics, Mutagenesis ». In *StatPearls*. Treasure Island (FL): StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK560519/.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. « Real-Time DNA Sequencing from Single Polymerase Molecules ». *Science* 323 (5910): 133-38. https://doi.org/10.1126/science.1162986.

Eilbeck, Karen. 2017. « Settling the Score: Variant Prioritization and Mendelian Disease ». *Nature Reviews Genetics* 18 (10): 599-612. https://doi.org/10.1038/nrg.2017.52.

Eilbeck, Karen, Suzanna E. Lewis, Christopher J. Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, et Michael Ashburner. 2005. « The Sequence Ontology: a tool for the unification of genome annotations ». *Genome Biology* 6 (5): R44. https://doi.org/10.1186/gb-2005-6-5-r44.

Ellard, Sian. 2020. « ACGS Best Practice Guidelines for Variant Classification in Rare Disease 2020 ». Preprint. Genomics. https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf.

Eraslan, Gokcen, Eugene Drokhlyansky, Shankara Anand, Ayshwarya Subramanian, Evgenij Fiskin, Michal Slyper, Jiali Wang, et al. 2021. « Single-Nucleus Cross-Tissue Molecular Reference Maps to Decipher Disease Gene Function ». https://doi.org/10.1101/2021.07.19.452954.

Falconer. 1996. « Falconer: Introduction to quantitative genetics - Google Scholar ». 1996. https://scholar.google.com/scholar_lookup?&title=Introduction%20to%20Quantitative%20Genetics&publication_year=1996&author=Falconer%2CDS&author=Mackay%2CTFC.

Fauman, Eric. 2020. « Eric Fauman sur Twitter ». Twitter. 2020. https://twitter.com/Eric_Fauman/status/1261786563102625794.

Frazer, Kelly A., Sarah S. Murray, Nicholas J. Schork, et Eric J. Topol. 2009. « Human Genetic Variation and Its Contribution to Complex Traits ». *Nature Reviews Genetics* 10 (4): 241-51. https://doi.org/10.1038/nrg2554.

Fu, Wenqing, Timothy D. O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M. Leal, Stacey Gabriel, et al. 2013. « Analysis of 6,515 Exomes Reveals the Recent Origin of Most Human Protein-Coding Variants ». *Nature* 493 (7431): 216-20. https://doi.org/10.1038/nature11690.

Gaither, Jeffrey B S, Grant E Lammi, James L Li, David M Gordon, Harkness C Kuck, Benjamin J Kelly, James R Fitch, et Peter White. 2021. « Synonymous variants that disrupt messenger RNA structure are significantly constrained in the human population ». *GigaScience* 10 (4). https://doi.org/10.1093/gigascience/giab023.

Garrido-Martín, Diego, Beatrice Borsari, Miquel Calvo, Ferran Reverter, et Roderic Guigó. 2021. « Identification and Analysis of Splicing Quantitative Trait Loci across Multiple Tissues in the Human Genome ». *Nature Communications* 12 (1): 727. https://doi.org/10.1038/s41467-020-20578-2.

Garrison, Erik, et Gabor Marth. 2012. « Haplotype-based variant detection from short-read sequencing ». *arXiv:1207.3907 [q-bio]*, juillet. http://arxiv.org/abs/1207.3907.

Gates, Alexander J., Deisy Morselli Gysi, Manolis Kellis, et Albert-László Barabási. 2021. « A Wealth of Discovery Built on the Human Genome Project — by the Numbers ». *Nature* 590 (7845): 212-15. https://doi.org/10.1038/d41586-021-00314-6.

« gnomAD v2.1 | gnomAD news ». 2018. 2018. https://gnomad.broadinstitute.org/news/2018-10-gnomad-v2-1/.

Gonorazky, Hernan, Minggao Liang, Beryl Cummings, Monkol Lek, Johann Micallef, Cynthia Hawkins, Raveen Basran, et al. 2016. « RNAseq Analysis for the Diagnosis of Muscular Dystrophy ». *Annals of Clinical and Translational Neurology* 3 (1): 55-60. https://doi.org/10.1002/acn3.267.

Greener, Joe G., Shaun M. Kandathil, Lewis Moffat, et David T. Jones. 2021. « A Guide to Machine Learning for Biologists ». *Nature Reviews Molecular Cell Biology*, septembre, 1-16. https://doi.org/10.1038/s41580-021-00407-0.

Grimm, Dominik G., Chloé-Agathe Azencott, Fabian Aicheler, Udo Gieraths, Daniel G. MacArthur, Kaitlin E. Samocha, David N. Cooper, et al. 2015. « The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity ». *Human Mutation* 36 (5): 513-23. https://doi.org/10.1002/humu.22768.

GTEx. 2021. « NIH Will Expand Existing Gene Expression Resources to Include Developmental Tissues ». Genome.Gov. 2021. https://www.genome.gov/news/news-release/NIH-will-expand-existing-gene-expression-resources-to-include-developmental-tissues.

GTEx Consortium. 2015. « Human Genomics. The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans ». *Science (New York, N.Y.)* 348 (6235): 648-60. https://doi.org/10.1126/science.1262110.

Gusella, James F., Nancy S. Wexler, P. Michael Conneally, Susan L. Naylor, Mary Anne Anderson, Rudolph E. Tanzi, Paul C. Watkins, et al. 1983. « A Polymorphic DNA Marker Genetically Linked to Huntington's Disease ». *Nature* 306 (5940): 234-38. https://doi.org/10.1038/306234a0.

Hamdan, Fadi F., Candace T. Myers, Patrick Cossette, Philippe Lemay, Dan Spiegelman, Alexandre Dionne Laporte, Christina Nassif, et al. 2017. « High Rate of Recurrent De Novo Mutations in Developmental and Epileptic Encephalopathies ». *American Journal of Human Genetics* 101 (5): 664-85. https://doi.org/10.1016/j.ajhg.2017.09.008.

Hartley, Taila, Gabrielle Lemire, Kristin D. Kernohan, Heather E. Howley, David R. Adams, et Kym M. Boycott. 2020. « New Diagnostic Approaches for Undiagnosed Rare Genetic Diseases ». *Annual Review of Genomics and Human Genetics* 21 (1): 351-72. https://doi.org/10.1146/annurev-genom-083118-015345.

Havrilla, James M., Brent S. Pedersen, Ryan M. Layer, et Aaron R. Quinlan. 2019. « A Map of Constrained Coding Regions in the Human Genome ». *Nature Genetics* 51 (1): 88-95. https://doi.org/10.1038/s41588-018-0294-6.

He, Shuai, Lin-He Wang, Yang Liu, Yi-Qi Li, Hai-Tian Chen, Jing-Hong Xu, Wan Peng, et al. 2020. « Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs ». *Genome Biology* 21 (1): 294. https://doi.org/10.1186/s13059-020-02210-0.

Heijl, Stephan, Bas Vroling, Tom van den Bergh, et Henk-Jan Joosten. 2020. « Mind the Gap: Preventing Circularity in Missense Variant Prediction ». Preprint. Bioinformatics. https://doi.org/10.1101/2020.05.06.080424.

Hershey, A. D., et M. Chase. 1952. « Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage ». *The Journal of General Physiology* 36 (1): 39-56. https://doi.org/10.1085/jgp.36.1.39.

Hiemenz, Matthew C., Stephan Kadauke, David B. Lieberman, David B. Roth, Jianhua Zhao, Christopher D. Watt, Robert D. Daber, et Jennifer J. D. Morrissette. 2016. « Building a Robust Tumor Profiling Program: Synergy between Next-Generation Sequencing and Targeted Single-Gene Testing ». *PLOS ONE* 11 (4): e0152851. https://doi.org/10.1371/journal.pone.0152851.

Hira, Zena M., et Duncan F. Gillies. 2015. « A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data ». *Advances in Bioinformatics* 2015 (juin): 1-13. https://doi.org/10.1155/2015/198363.

Hirschhorn, Joel N., et Mark J. Daly. 2005. « Genome-Wide Association Studies for Common Diseases and Complex Traits ». *Nature Reviews Genetics* 6 (2): 95-108. https://doi.org/10.1038/nrg1521.

Ho, Steve S., Alexander E. Urban, et Ryan E. Mills. 2020. « Structural Variation in the Sequencing Era ». *Nature Reviews Genetics* 21 (3): 171-89. https://doi.org/10.1038/s41576-019-0180-9.

Holley, Robert W., Jean Apgar, George A. Everett, James T. Madison, Mark Marquisee, Susan H. Merrill, John Robert Penswick, et Ada Zamir. 1965. « Structure of a Ribonucleic Acid ». *Science* 147 (3664): 1462-65. https://doi.org/10.1126/science.147.3664.1462.

Hossin, M, et M.N Sulaiman. 2015. « A Review on Evaluation Metrics for Data Classification Evaluations ». *International Journal of Data Mining & Knowledge Management Process* 5 (2): 01-11. https://doi.org/10.5121/ijdkp.2015.5201.

Howe, Kevin L, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, et al. 2021. « Ensembl 2021 ». *Nucleic Acids Research* 49 (D1): D884-91. https://doi.org/10.1093/nar/gkaa942.

Huang, Ni, Insuk Lee, Edward M Marcotte, et Matthew E Hurles. 2010. « Characterising and Predicting Haploinsufficiency in the Human Genome ». *PLoS Genetics* 6 (10): e1001154. https://doi.org/10.1371/journal.pgen.1001154.

International Human Genome Sequencing Consortium. 2004. « Finishing the Euchromatic Sequence of the Human Genome ». *Nature* 431 (7011): 931-45. https://doi.org/10.1038/nature03001.

Ioannidis, Nilah M., Joseph H. Rothstein, Vikas Pejaver, Sumit Middha, Shannon K. McDonnell, Saurabh Baheti, Anthony Musolf, et al. 2016. « REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants ». *American Journal of Human Genetics* 99 (4): 877-85. https://doi.org/10.1016/j.ajhg.2016.08.016.

Ippolito, Pier Paolo. 2019. « Feature Extraction Techniques ». Medium. 11 octobre 2019. https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be.

Ittisoponpisan, Sirawit, Suhail A. Islam, Tarun Khanna, Eman Alhuzimi, Alessia David, et Michael J. E. Sternberg. 2019. « Can Predicted Protein 3D Structures Provide Reliable Insights into Whether Missense Variants Are Disease Associated? » *Journal of Molecular Biology* 431 (11): 2197-2212. https://doi.org/10.1016/j.jmb.2019.04.009.

Jain, Miten, Sergey Koren, Karen H. Miga, Josh Quick, Arthur C. Rand, Thomas A. Sasani, John R. Tyson, et al. 2018. « Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads ». *Nature Biotechnology* 36 (4): 338-45. https://doi.org/10.1038/nbt.4060.

Jian, Xueqiu, Eric Boerwinkle, et Xiaoming Liu. 2014. « In silico prediction of splice-altering single nucleotide variants in the human genome ». *Nucleic Acids Research* 42 (22): 13534-44. https://doi.org/10.1093/nar/gku1206.

Jiang, Wei, et Liang Chen. 2021. « Alternative Splicing: Human Disease and Quantitative Analysis from High-Throughput Sequencing ». *Computational and Structural Biotechnology Journal* 19 (janvier): 183-95. https://doi.org/10.1016/j.csbj.2020.12.009.

Jorquera, Roddy, Carolina González, Philip Clausen, Bent Petersen, et David S Holmes. 2018. « Improved ontology for eukaryotic single-exon coding sequences in biological databases ». *Database: The Journal of Biological Databases and Curation* 2018 (septembre): bay089. https://doi.org/10.1093/database/bay089.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. « Highly Accurate Protein Structure Prediction with AlphaFold ». *Nature* 596 (7873): 583-89. https://doi.org/10.1038/s41586-021-03819-2.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. « The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans ». *Nature* 581 (7809): 434-43. https://doi.org/10.1038/s41586-020-2308-7.

Katz, Yarden, Eric T. Wang, Edoardo M. Airoldi, et Christopher B. Burge. 2010. « Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation ». *Nature Methods* 7 (12): 1009-15. https://doi.org/10.1038/nmeth.1528.

Kawashima, S., et M. Kanehisa. 2000. « AAindex: Amino Acid Index Database ». *Nucleic Acids Research* 28 (1): 374. https://doi.org/10.1093/nar/28.1.374.

Kernohan, Kristin D., Laure Frésard, Zachary Zappala, Taila Hartley, Kevin S. Smith, Justin Wagner, Hongbin Xu, et al. 2017. « Whole-transcriptome sequencing in blood provides a diagnosis of spinal muscular atrophy with progressive myoclonic epilepsy ». *Human Mutation* 38 (6): 611-14. https://doi.org/10.1002/humu.23211.

Kiefer, Jürgen. 2007. « Effects of Ultraviolet Radiation on DNA ». In *Chromosomal Alterations: Methods, Results and Importance in Human Health*, édité par Günter Obe et Vijayalaxmi, 39-53. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-71414-9_3.

Kim-Hellmuth, Sarah, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, Valentin Wucher, Stephane E. Castel, et al. 2020. « Cell type–specific genetic regulation of gene expression across human tissues ». *Science* 369 (6509): eaaz8528. https://doi.org/10.1126/science.aaz8528.

Kinsella, Rhoda J., Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, et al. 2011. « Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space ». *Database: The Journal of Biological Databases and Curation* 2011: bar030. https://doi.org/10.1093/database/bar030.

Koboldt, Daniel C. 2020. « Best practices for variant calling in clinical sequencing ». *Genome Medicine* 12 (1): 91. https://doi.org/10.1186/s13073-020-00791-w.

Köhler, Sebastian, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, David Lewis-Smith, Nicole A. Vasilevsky, Daniel Danis, et al. 2021. « The Human Phenotype Ontology in 2021 ». *Nucleic Acids Research* 49 (D1): D1207-17. https://doi.org/10.1093/nar/gkaa1043.

Kremer, Laura S., Daniel M. Bader, Christian Mertes, Robert Kopajtich, Garwin Pichler, Arcangela Iuso, Tobias B. Haack, et al. 2017. « Genetic Diagnosis of Mendelian Disorders via RNA Sequencing ». *Nature Communications* 8 (1): 15824. https://doi.org/10.1038/ncomms15824.

Krokan, Hans E., et Magnar Bjørås. 2013. « Base Excision Repair ». *Cold Spring Harbor Perspectives in Biology* 5 (4): a012583. https://doi.org/10.1101/cshperspect.a012583.

Kundaje, Anshul, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, et al. 2015. « Integrative Analysis of 111 Reference Human Epigenomes ». *Nature* 518 (7539): 317-30. https://doi.org/10.1038/nature14248.

Lali, Ricky, Michael Chong, Arghavan Omidi, Pedrum Mohammadi-Shemirani, Ann Le, Edward Cui, et Guillaume Paré. 2021. « Calibrated Rare Variant Genetic Risk Scores for Complex Disease Prediction Using Large Exome Sequence Repositories ». *Nature Communications* 12 (1): 5852. https://doi.org/10.1038/s41467-021-26114-0.

Lamnidis, Thiseas C., Kerttu Majander, Choongwon Jeong, Elina Salmela, Anna Wessman, Vyacheslav Moiseyev, Valery Khartanovich, et al. 2018. « Ancient Fennoscandian Genomes Reveal Origin and Spread of Siberian Ancestry in Europe ». *Nature Communications* 9 (1): 5018. https://doi.org/10.1038/s41467-018-07483-5.

Lander, Eric S., Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, et al. 2001. « Initial Sequencing and Analysis of the Human Genome ». *Nature* 409 (6822): 860-921. https://doi.org/10.1038/35057062.

Landrum, Melissa J., Shanmuga Chitipiralla, Garth R. Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, et al. 2020. « ClinVar: Improvements to Accessing Data ». *Nucleic Acids Research* 48 (D1): D835-44. https://doi.org/10.1093/nar/gkz972.

Lauritsen, Simon Meyer, Mads Kristensen, Mathias Vassard Olsen, Morten Skaarup Larsen, Katrine Meyer Lauritsen, Marianne Johansson Jørgensen, Jeppe Lange, et Bo Thiesson. 2020. « Explainable Artificial Intelligence Model to Predict Acute Critical Illness from Electronic Health Records ». *Nature Communications* 11 (1): 3852. https://doi.org/10.1038/s41467-020-17431-x.

Lek, Monkol, Timothy Fennell, Anne H. O'Donnell-Luria, James S. Ware, Andrew J. Hill, Beryl B. Cummings, Taru Tukiainen, et al. 2016. « Analysis of Protein-Coding Genetic Variation in 60,706 Humans ». *Nature* 536 (7616): 285-91. https://doi.org/10.1038/nature19057.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O'Donnell-Luria, et al. 2016. « Analysis of Protein-Coding Genetic Variation in 60,706 Humans ». *Nature* 536 (7616): 285-91. https://doi.org/10.1038/nature19057.

Li, Guo-Min. 2008. « Mechanisms and Functions of DNA Mismatch Repair ». *Cell Research* 18 (1): 85-98. https://doi.org/10.1038/cr.2007.115.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et 1000 Genome Project Data Processing Subgroup. 2009. « The Sequence Alignment/Map format and SAMtools ». *Bioinformatics* 25 (16): 2078-79. https://doi.org/10.1093/bioinformatics/btp352.

Li, Jing-Woei, Keng-Po Lai, Arthur K. K. Ching, et Ting-Fung Chan. 2014. « Transcriptome Sequencing of Chinese and Caucasian Population Identifies Ethnic-Associated Differential Transcript Abundance of Heterogeneous Nuclear Ribonucleoprotein K (HnRNPK) ». *Genomics* 103 (1): 56-64. https://doi.org/10.1016/j.ygeno.2013.12.005.

Liang, F., I. Holt, G. Pertea, S. Karamycheva, S. L. Salzberg, et J. Quackenbush. 2000. « Gene Index Analysis of the Human Genome Estimates Approximately 120,000 Genes ». *Nature Genetics* 25 (2): 239-40. https://doi.org/10.1038/76126.

Libbrecht, Maxwell W., et William Stafford Noble. 2015. « Machine Learning Applications in Genetics and Genomics ». *Nature Reviews Genetics* 16 (6): 321-32. https://doi.org/10.1038/nrg3920.

Lin, Michael F., Irwin Jungreis, et Manolis Kellis. 2011. « PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions ». *Bioinformatics* 27 (13): i275-82. https://doi.org/10.1093/bioinformatics/btr209.

Lindahl, T., et D.E. Barnes. 2000. « Repair of Endogenous DNA Damage ». *Cold Spring Harbor Symposia on Quantitative Biology* 65 (0): 127-34. https://doi.org/10.1101/sqb.2000.65.127.

Lindahl, Tomas. 1993. « Instability and Decay of the Primary Structure of DNA ». *Nature* 362 (6422): 709-15. https://doi.org/10.1038/362709a0.

Lindblad-Toh, Kerstin, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, et al. 2011. « A High-Resolution Map of Human Evolutionary Constraint Using 29 Mammals ». *Nature* 478 (7370): 476-82. https://doi.org/10.1038/nature10530.

Liu, Xiaoming, Chang Li, Chengcheng Mou, Yibo Dong, et Yicheng Tu. 2020. « dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs ». *Genome Medicine* 12 (1): 103. https://doi.org/10.1186/s13073-020-00803-9.

Loeb, Lawrence A., et Raymond J. Monnat. 2008. « DNA Polymerases and Human Disease ». *Nature Reviews Genetics* 9 (8): 594-604. https://doi.org/10.1038/nrg2345.

Low, Wai Yee, Rick Tearle, Derek M. Bickhart, Benjamin D. Rosen, Sarah B. Kingan, Thomas Swale, Françoise Thibaud-Nissen, et al. 2019. « Chromosome-Level Assembly of the Water Buffalo Genome Surpasses Human and Goat Genomes in Sequence Contiguity ». *Nature Communications* 10 (1): 260. https://doi.org/10.1038/s41467-018-08260-0.

Lupski, James R., Roberto Montes de Oca-Luna, Susan Slaugenhaupt, Liu Pentao, Vito Guzzetta, Barbara J. Trask, Odila Saucedo-Cardenas, et al. 1991. « DNA Duplication Associated with Charcot-Marie-Tooth Disease Type 1A ». *Cell* 66 (2): 219-32. https://doi.org/10.1016/0092-8674(91)90613-4.

MacArthur, D. G., T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, et al. 2014. « Guidelines for Investigating Causality of Sequence Variants in Human Disease ». *Nature* 508 (7497): 469-76. https://doi.org/10.1038/nature13127.

Mackay, Trudy F. C. 2001. « Quantitative Trait Loci in Drosophila ». *Nature Reviews Genetics* 2 (1): 11-20. https://doi.org/10.1038/35047544.

Maiella, S., A. Rath, C. Angin, F. Mousson, et O. Kremp. 2013. « [Orphanet and its consortium: where to find expert-validated information on rare diseases] ». *Revue Neurologique* 169 Suppl 1 (février): S3-8. https://doi.org/10.1016/S0035-3787(13)70052-3.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. « Finding the Missing Heritability of Complex Diseases ». *Nature* 461 (7265): 747-53. https://doi.org/10.1038/nature08494.

Marshall, Jessica. 2014. « The Genetic Code ». *Proceedings of the National Academy of Sciences* 111 (16): 5760-5760. https://doi.org/10.1073/pnas.1405652111.

Matsuo, Hirotaka, Ken Yamamoto, Hirofumi Nakaoka, Akiyoshi Nakayama, Masayuki Sakiyama, Toshinori Chiba, Atsushi Takahashi, et al. 2016. « Genome-Wide Association Study of Clinically Defined Gout Identifies Multiple Risk Loci and Its Association with Clinical Subtypes ». *Annals of the Rheumatic Diseases* 75 (4): 652-59. https://doi.org/10.1136/annrheumdis-2014-206191.

Maxam, A. M., et W. Gilbert. 1977. « A New Method for Sequencing DNA ». *Proceedings of the National Academy of Sciences* 74 (2): 560-64. https://doi.org/10.1073/pnas.74.2.560.

Mazin, Pavel V., Philipp Khaitovich, Margarida Cardoso-Moreira, et Henrik Kaessmann. 2021. « Alternative Splicing during Mammalian Organ Development ». *Nature Genetics* 53 (6): 925-34. https://doi.org/10.1038/s41588-021-00851-w.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, et Fiona Cunningham. 2016. « The Ensembl Variant Effect Predictor ». *Genome Biology* 17 (1): 122. https://doi.org/10.1186/s13059-016-0974-4.

Melamud, Eugene, et John Moult. 2009. « Stochastic noise in splicing machinery ». *Nucleic Acids Research* 37 (14): 4873-86. https://doi.org/10.1093/nar/gkp471.

Mendel, Gregor. 1865. « EXPERIMENTS IN PLANT HYBRIDIZATION (1865) », 41.

Merker, Jason D, Aaron M Wenger, Tam Sneddon, Megan Grove, Zachary Zappala, Laure Fresard, Daryl Waggott, et al. 2018. « Long-Read Genome Sequencing Identifies Causal Structural Variation in a Mendelian Disease ». *Genetics in Medicine* 20 (1): 159-63. https://doi.org/10.1038/gim.2017.86.

Meyer, Axel, Siegfried Schloissnig, Paolo Franchini, Kang Du, Joost M. Woltering, Iker Irisarri, Wai Yee Wong, et al. 2021. « Giant Lungfish Genome Elucidates the Conquest of Land by Vertebrates ». *Nature* 590 (7845): 284-89. https://doi.org/10.1038/s41586-021-03198-8.

Meyer, Matthias, et Martin Kircher. 2010. « Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing ». *Cold Spring Harbor Protocols* 2010 (6): pdb.prot5448. https://doi.org/10.1101/pdb.prot5448.

Mi, Huaiyu, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, et Paul D Thomas. 2019. « PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools ». *Nucleic Acids Research* 47 (D1): D419-26. https://doi.org/10.1093/nar/gky1038.

Mirza, Bilal, Wei Wang, Jie Wang, Howard Choi, Neo Christopher Chung, et Peipei Ping. 2019. « Machine Learning and Integrative Analysis of Biomedical Big Data ». *Genes* 10 (2): 87. https://doi.org/10.3390/genes10020087.

Miescher. 1871. *Medicinisch-chemische Untersuchungen*.

Modrek, Barmak, et Christopher Lee. 2002. « A Genomic View of Alternative Splicing ». *Nature Genetics* 30 (1): 13-19. https://doi.org/10.1038/ng0102-13.

Moore, Jill E., Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, et al. 2020. « Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes ». *Nature* 583 (7818): 699-710. https://doi.org/10.1038/s41586-020-2493-4.

Moritz, Dirk. 2020. « What's the Definition of a "Rare Disease?" It Depends on Who You Ask... » *Blue Matter Consulting* (blog). mars 2020. https://bluematterconsulting.com/rare-disease-definition/.

Murdock, David R., Hongzheng Dai, Lindsay C. Burrage, Jill A. Rosenfeld, Shamika Ketkar, Michaela F. Müller, Vicente A. Yépez, et al. 2021. « Transcriptome-Directed Analysis for Mendelian Disease

Diagnosis Overcomes Limitations of Conventional Genomic Testing ». *The Journal of Clinical Investigation* 131 (1). https://doi.org/10.1172/JCI141500.

Nagy, Alinda, et László Patthy. 2013. « MisPred: A Resource for Identification of Erroneous Protein Sequences in Public Databases ». *Database* 2013 (janvier). https://doi.org/10.1093/database/bat053.

Nair, Preethy Sasidharan, et Mauno Vihinen. 2013. « VariBench: A Benchmark Database for Variations ». *Human Mutation* 34 (1): 42-49. https://doi.org/10.1002/humu.22204.

Navarro Gonzalez, Jairo, Ann S Zweig, Matthew L Speir, Daniel Schmelter, Kate R Rosenbloom, Brian J Raney, Conner C Powell, et al. 2021. « The UCSC Genome Browser database: 2021 update ». *Nucleic Acids Research* 49 (D1): D1046-57. https://doi.org/10.1093/nar/gkaa1070.

Nirenberg, Marshall W., et J. Heinrich Matthaei. 1961. « The Dependence of Cell-Free Protein Synthesis in E. Coli upon Naturally Occurring or Synthetic Polyribonucleotides ». *Proceedings of the National Academy of Sciences* 47 (10): 1588-1602. https://doi.org/10.1073/pnas.47.10.1588.

Nurk, Sergey, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, et al. 2021. « The Complete Sequence of a Human Genome ». Preprint. Genomics. https://doi.org/10.1101/2021.05.26.445798.

O'Leary, Nuala A., Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, et al. 2016. « Reference Sequence (RefSeq) Database at NCBI: Current Status, Taxonomic Expansion, and Functional Annotation ». *Nucleic Acids Research* 44 (D1): D733-45. https://doi.org/10.1093/nar/gkv1189.

Orhand, Romain, Anne Jeannin-Girardon, Pierre Parrend, et Pierre Collet. 2019. « DeepExpert : vers une Intelligence Artificielle autonome et explicable ». In *Rencontres des Jeunes Chercheurs en Intelligence Artificielle 2019*, édité par Maxime Lefrançois, MINES Saint-Étienne, et France, 63-65. Actes des Rencontres des Jeunes Chercheurs en Intelligence Artificielle 2019. Toulouse, France. https://hal.archives-ouvertes.fr/hal-02161105.

Owen, Mallory J., Anna-Kaisa Niemi, David P. Dimmock, Mark Speziale, Mark Nespeca, Kevin K. Chau, Luca Van Der Kraan, et al. 2021. « Rapid Sequencing-Based Diagnosis of Thiamine Metabolism Dysfunction Syndrome ». *New England Journal of Medicine* 384 (22): 2159-61. https://doi.org/10.1056/NEJMc2100365.

Pachter, Lior. 2013. « *Seq ». *Bits of DNA* (blog). 25 août 2013. https://liorpachter.wordpress.com/seq/.

Pan, Qun, Ofer Shai, Leo J. Lee, Brendan J. Frey, et Benjamin J. Blencowe. 2008. « Deep Surveying of Alternative Splicing Complexity in the Human Transcriptome by High-Throughput Sequencing ». *Nature Genetics* 40 (12): 1413-15. https://doi.org/10.1038/ng.259.

Pedersen, Brent S., Ryan M. Layer, et Aaron R. Quinlan. 2016. « Vcfanno: fast, flexible annotation of genetic variants ». *Genome Biology* 17 (1): 118. https://doi.org/10.1186/s13059-016-0973-5.

Pedersen, Brent S, et Aaron R Quinlan. 2017. « cyvcf2: fast, flexible variant analysis with Python ». *Bioinformatics* 33 (12): 1867-69. https://doi.org/10.1093/bioinformatics/btx057.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. « Scikit-Learn: Machine Learning in Python ». *MACHINE LEARNING IN PYTHON*, 6.

Pegg, A. E. 1990. « DNA Repair and Carcinogenesis by Alkylating Agents ». In *Chemical Carcinogenesis and Mutagenesis II*, édité par Colin S. Cooper et Philip L. Grover, 103-31. Handbook of Experimental Pharmacology. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-74778-6_5.

Petrovski, Slavé, Quanli Wang, Erin L. Heinzen, Andrew S. Allen, et David B. Goldstein. 2013. « Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes ». Édité par Scott M. Williams. *PLoS Genetics* 9 (8): e1003709. https://doi.org/10.1371/journal.pgen.1003709.

Pickrell, Joseph K., Athma A. Pai, Yoav Gilad, et Jonathan K. Pritchard. 2010. « Noisy Splicing Drives MRNA Isoform Diversity in Human Cells ». *PLOS Genetics* 6 (12): e1001236. https://doi.org/10.1371/journal.pgen.1001236.

Pinto, Navin, et M. Eileen Dolan. 2011. « Clinically Relevant Genetic Variations in Drug Metabolizing Enzymes ». *Current drug metabolism* 12 (5): 487-97.

Poczai, Péter, Neil Bell, et Jaakko Hyvönen. 2014. « Imre Festetics and the Sheep Breeders' Society of Moravia: Mendel's Forgotten "Research Network" ». *PLOS Biology* 12 (1): e1001772. https://doi.org/10.1371/journal.pbio.1001772.

Pommier, Yves, Juana M. Barcelo, V. Ashutosh Rao, Olivier Sordet, Andrew G. Jobson, Laurent Thibaut, Ze-Hong Miao, et al. 2006. « Repair of Topoisomerase I-Mediated DNA Damage ». *Progress in Nucleic Acid Research and Molecular Biology* 81: 179-229. https://doi.org/10.1016/S0079-6603(06)81005-6.

Poplin, Ryan, Pi-Chuan Chang, David Alexander, Scott Schwartz, Thomas Colthurst, Alexander Ku, Dan Newburger, et al. 2018. « A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks ». *Nature Biotechnology* 36 (10): 983-87. https://doi.org/10.1038/nbt.4235.

Poplin, Ryan, Valentin Ruano-Rubio, Mark A. DePristo, Tim J. Fennell, Mauricio O. Carneiro, Geraldine A. Van der Auwera, David E. Kling, et al. 2018. « Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples ». https://doi.org/10.1101/201178.

Price, Alkes L., Gregory V. Kryukov, Paul I. W. de Bakker, Shaun M. Purcell, Jeff Staples, Lee-Jen Wei, et Shamil R. Sunyaev. 2010. « Pooled Association Tests for Rare Variants in Exon-Resequencing Studies ». *The American Journal of Human Genetics* 86 (6): 832-38. https://doi.org/10.1016/j.ajhg.2010.04.005.

Que, Lin, Jochen Winterer, et Csaba Földy. 2019. « Deep Survey of GABAergic Interneurons: Emerging Insights From Gene-Isoform Transcriptomics ». *Frontiers in Molecular Neuroscience* 12: 115. https://doi.org/10.3389/fnmol.2019.00115.

Rahit, K. M. Tahsin Hassan, et Maja Tarailo-Graovac. 2020. « Genetic Modifiers and Rare Mendelian Disease ». *Genes* 11 (3): 239. https://doi.org/10.3390/genes11030239.

Reardon, Sara. 2021. « A Complete Human Genome Sequence Is Close: How Scientists Filled in the Gaps ». *Nature* 594 (7862): 158-59. https://doi.org/10.1038/d41586-021-01506-w.

Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, et al. 2015. « 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 ». *Genetics in Medicine* 17 (5): 405-23. https://doi.org/10.1038/gim.2015.30.

Rimmer, Andy, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R. F. Twigg, Andrew O. M. Wilkie, Gil McVean, et Gerton Lunter. 2014. « Integrating Mapping-, Assembly- and Haplotype-Based Approaches for Calling Variants in Clinical Sequencing Applications ». *Nature Genetics* 46 (8): 912-18. https://doi.org/10.1038/ng.3036.

Robertson. 1967. « Robertson: Heritage from Mendel - Google Scholar ». 1967. https://scholar.google.com/scholar_lookup?&title=Heritage%20From%20Mendel&pages=265-280&publication_year=1967&author=Robertson%2CA.

Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, et Jill P. Mesirov. 2011. « Integrative Genomics Viewer ». *Nature Biotechnology* 29 (1): 24-26. https://doi.org/10.1038/nbt.1754.

Rodenburg, Richard J. 2018. « The functional genomics laboratory: functional validation of genetic variants ». *Journal of Inherited Metabolic Disease* 41 (3): 297-307. https://doi.org/10.1007/s10545-018-0146-7.

Rodriguez, Jose Manuel, Fernando Pozo, Tomas di Domenico, Jesus Vazquez, et Michael L. Tress. 2020. « An Analysis of Tissue-Specific Alternative Splicing at the Protein Level ». *PLOS Computational Biology* 16 (10): e1008287. https://doi.org/10.1371/journal.pcbi.1008287.

Sanger, F., S. Nicklen, et A. R. Coulson. 1977. « DNA Sequencing with Chain-Terminating Inhibitors ». *Proceedings of the National Academy of Sciences* 74 (12): 5463-67. https://doi.org/10.1073/pnas.74.12.5463.

Sapkota, Anupama. 2020. « RNA Splicing- Definition, Process, Mechanism, Types, Errors, Uses ». Microbe Notes. 8 octobre 2020. https://microbenotes.com/rna-splicing/.

Sastre, Leandro. 2014. « Exome Sequencing: What Clinicians Need to Know ». *Advances in Genomics and Genetics*, mars, 15. https://doi.org/10.2147/AGG.S39108.

Schärer, Orlando D. 2013. « Nucleotide Excision Repair in Eukaryotes ». *Cold Spring Harbor Perspectives in Biology* 5 (10): a012609. https://doi.org/10.1101/cshperspect.a012609.

Schneider, Valerie A., Tina Graves-Lindsay, Kerstin Howe, Nathan Bouk, Hsiu-Chuan Chen, Paul A. Kitts, Terence D. Murphy, et al. 2017. « Evaluation of GRCh38 and de Novo Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly ». *Genome Research* 27 (5): 849-64. https://doi.org/10.1101/gr.213611.116.

Schwarze, Katharina, James Buchanan, Jenny C. Taylor, et Sarah Wordsworth. 2018. « Are Whole-Exome and Whole-Genome Sequencing Approaches Cost-Effective? A Systematic Review of the Literature ». *Genetics in Medicine* 20 (10): 1122-30. https://doi.org/10.1038/gim.2017.247.

Shah, Naisha, Ying-Chen Claire Hou, Hung-Chun Yu, Rachana Sainger, C. Thomas Caskey, J. Craig Venter, et Amalio Telenti. 2018. « Identification of Misclassified ClinVar Variants via Disease Population Prevalence ». *American Journal of Human Genetics* 102 (4): 609-19. https://doi.org/10.1016/j.ajhg.2018.02.019.

Shevchenko, Yuriy, et Sherri Bale. 2016. « Clinical Versus Research Sequencing ». *Cold Spring Harbor Perspectives in Medicine* 6 (11): a025809. https://doi.org/10.1101/cshperspect.a025809.

Shi, Xinghua Mindy, éd. 2020. *EQTL Analysis: Methods and Protocols*. Vol. 2082. Methods in Molecular Biology. New York, NY: Springer US. https://doi.org/10.1007/978-1-0716-0026-9.

Shirai, Yuya, et Yukinori Okada. 2021. « Elucidation of disease etiology by trans-layer omics analysis ». *Inflammation and Regeneration* 41 (1): 6. https://doi.org/10.1186/s41232-021-00155-w.

Sims, David, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, et Chris P. Ponting. 2014. « Sequencing Depth and Coverage: Key Considerations in Genomic Analyses ». *Nature Reviews Genetics* 15 (2): 121-32. https://doi.org/10.1038/nrg3642.

Skipper, Paul L., Min Young Kim, H.-L. Patty Sun, Gerald N. Wogan, et Steven R. Tannenbaum. 2010. « Monocyclic aromatic amines as potential human carcinogens: old is new again ». *Carcinogenesis* 31 (1): 50-58. https://doi.org/10.1093/carcin/bgp267.

Smith, L. M., J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, et L. E. Hood. 1986. « Fluorescence Detection in Automated DNA Sequence Analysis ». *Nature* 321 (6071): 674-79. https://doi.org/10.1038/321674a0.

Soon, Wendy Weijia, Manoj Hariharan, et Michael P Snyder. 2013. « High-throughput Sequencing for Biology and Medicine ». *Molecular Systems Biology* 9 (1): 640. https://doi.org/10.1038/msb.2012.61.

Sorek, Rotem, Ron Shamir, et Gil Ast. 2004. « How Prevalent Is Functional Alternative Splicing in the Human Genome? » *Trends in Genetics* 20 (2): 68-71. https://doi.org/10.1016/j.tig.2003.12.004.

Stamm, Stefan, Shani Ben-Ari, Ilona Rafalska, Yesheng Tang, Zhaiyi Zhang, Debra Toiber, T. A. Thanaraj, et Hermona Soreq. 2005. « Function of Alternative Splicing ». *Gene* 344 (janvier): 1-20. https://doi.org/10.1016/j.gene.2004.10.022.

Stenson, Peter D., Matthew Mort, Edward V. Ball, Katy Shaw, Andrew Phillips, et David N. Cooper. 2014. « The Human Gene Mutation Database: Building a Comprehensive Mutation Repository for Clinical and Molecular Genetics, Diagnostic Testing and Personalized Genomic Medicine ». *Human Genetics* 133 (1): 1-9. https://doi.org/10.1007/s00439-013-1358-4.

Stewart, Cheryl, et Michael S. Pepper. 2017. « Cystic Fibrosis in the African Diaspora ». *Annals of the American Thoracic Society* 14 (1): 1-7. https://doi.org/10.1513/AnnalsATS.201606-481FR.

Stone, Eric A., et Arend Sidow. 2005. « Physicochemical Constraint Violation by Missense Substitutions Mediates Impairment of Protein Function and Disease Severity ». *Genome Research* 15 (7): 978-86. https://doi.org/10.1101/gr.3804205.

Sung, Patrick, et Hannah Klein. 2006. « Mechanism of Homologous Recombination: Mediators and Helicases Take on Regulatory Functions ». *Nature Reviews Molecular Cell Biology* 7 (10): 739-50. https://doi.org/10.1038/nrm2008.

Suryamohan, Kushal, Sajesh P. Krishnankutty, Joseph Guillory, Matthew Jevit, Markus S. Schröder, Meng Wu, Boney Kuriakose, et al. 2020. « The Indian Cobra Reference Genome and Transcriptome Enables Comprehensive Identification of Venom Toxins ». *Nature Genetics* 52 (1): 106-17. https://doi.org/10.1038/s41588-019-0559-8.

Swaney, Danielle L., Dana J. Ramms, Zhiyong Wang, Jisoo Park, Yusuke Goto, Margaret Soucheray, Neil Bhola, et al. 2021. « A Protein Network Map of Head and Neck Cancer Reveals PIK3CA Mutant Drug Sensitivity ». *Science (New York, N.Y.)* 374 (6563): eabf2911. https://doi.org/10.1126/science.abf2911.

Wetterstrand. 2020. « The Cost of Sequencing a Human Genome ». Genome.Gov. 2020. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

Tian, Yuan, Tina Pesaran, Adam Chamberlin, R. Bryn Fenwick, Shuwei Li, Chia-Ling Gau, Elizabeth C. Chao, Hsiao-Mei Lu, Mary Helen Black, et Dajun Qian. 2019. « REVEL and BayesDel Outperform Other in Silico Meta-Predictors for Clinical Variant Classification ». *Scientific Reports* 9 (1): 12752. https://doi.org/10.1038/s41598-019-49224-8.

Toro, J R, M-H Wei, G M Glenn, M Weinreich, O Toure, C Vocke, M Turner, et al. 2008. « BHD Mutations, Clinical and Molecular Genetic Investigations of Birt-Hogg-Dube Syndrome: A New Series of 50 Families and a Review of Published Reports ». *Journal of Medical Genetics* 45 (6): 321-31. https://doi.org/10.1136/jmg.2007.054304.

Tweedie, Susan, Bryony Braschi, Kristian Gray, Tamsin E M Jones, Ruth L Seal, Bethan Yates, et Elspeth A Bruford. 2021. « Genenames.org: the HGNC and VGNC resources in 2021 ». *Nucleic Acids Research* 49 (D1): D939-46. https://doi.org/10.1093/nar/gkaa980.

Uffelmann, Emil, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, et Danielle Posthuma. 2021. « Genome-Wide Association Studies ». *Nature Reviews Methods Primers* 1 (1): 1-21. https://doi.org/10.1038/s43586-021-00056-9.

Uhlén, Mathias, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, et al. 2015. « Tissue-based map of the human proteome ». *Science* 347 (6220): 1260419. https://doi.org/10.1126/science.1260419.

Uhlén, Mathias, Björn M Hallström, Cecilia Lindskog, Adil Mardinoglu, Fredrik Pontén, et Jens Nielsen. 2016. « Transcriptomics resources of human tissues and organs ». *Molecular Systems Biology* 12 (4): 862. https://doi.org/10.15252/msb.20155865.

Urbanowicz, Ryan J., et Jason H. Moore. 2009. « Learning Classifier Systems: A Complete Introduction, Review, and Roadmap ». *Journal of Artificial Evolution and Applications* 2009 (septembre): 1-25. https://doi.org/10.1155/2009/736398.

Vandeweyer, Geert, Lut Van Laer, Bart Loeys, Tim Van den Bulcke, et R. Frank Kooy. 2014. « VariantDB: a flexible annotation and filtering portal for next generation sequencing data ». *Genome Medicine*, octobre. https://doi.org/10.1186/s13073-014-0074-6.

Velde, K. Joeri van der, Eddy N. de Boer, Cleo C. van Diemen, Birgit Sikkema-Raddatz, Kristin M. Abbott, Alain Knopperts, Lude Franke, et al. 2017. « GAVIN: Gene-Aware Variant INterpretation for medical sequencing ». *Genome Biology* 18 (1): 6. https://doi.org/10.1186/s13059-016-1141-7.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, et Christopher B. Burge. 2008. « Alternative Isoform Regulation in Human Tissue Transcriptomes ». *Nature* 456 (7221): 470-76. https://doi.org/10.1038/nature07509.

Watson, J. D., et F. H. C. Crick. 1953. « Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid ». *Nature* 171 (4356): 737-38. https://doi.org/10.1038/171737a0.

Wei, Chih-Hsuan, Alexis Allot, Robert Leaman, et Zhiyong Lu. 2019. « PubTator central: automated concept annotation for biomedical full text articles ». *Nucleic Acids Research* 47 (W1): W587-93. https://doi.org/10.1093/nar/gkz389.

Weyn-Vanhentenryck, Sebastien M., Huijuan Feng, Dmytro Ustianenko, Rachel Duffié, Qinghong Yan, Martin Jacko, Jose C. Martinez, et al. 2018. « Precise Temporal Regulation of Alternative Splicing during Neural Development ». *Nature Communications* 9 (1): 2189. https://doi.org/10.1038/s41467-018-04559-0.

Xin, Jiwen, Adam Mark, Cyrus Afrasiabi, Ginger Tsueng, Moritz Juchler, Nikhil Gopal, Gregory S. Stupp, et al. 2016. « High-performance web services for querying gene and variant annotation ». *Genome Biology* 17 (1): 91. https://doi.org/10.1186/s13059-016-0953-9.

Xu, Qiang, Barmak Modrek, et Christopher Lee. 2002. « Genome-wide detection of tissue-specific alternative splicing in the human transcriptome ». *Nucleic Acids Research* 30 (17): 3754-66.

Yeo, Gene, et Christopher B. Burge. 2004. « Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals ». *Journal of Computational Biology* 11 (2-3): 377-94. https://doi.org/10.1089/1066527041410418.

Yeo, Gene, Dirk Holste, Gabriel Kreiman, et Christopher B. Burge. 2004. « Variation in Alternative Splicing across Human Tissues ». *Genome Biology* 5 (10): R74. https://doi.org/10.1186/gb-2004-5-10-r74.

Yeo, Giles S. H., Chiao-Chien Connie Hung, Justin Rochford, Julia Keogh, Juliette Gray, Shoba Sivaramakrishnan, Stephen O'Rahilly, et I. Sadaf Farooqi. 2004. « A de Novo Mutation Affecting Human TrkB Associated with Severe Obesity and Developmental Delay ». *Nature Neuroscience* 7 (11): 1187-89. https://doi.org/10.1038/nn1336.

Zednik, Carlos. 2021. « Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence ». *Philosophy & Technology* 34 (2): 265-88. https://doi.org/10.1007/s13347-019-00382-7.

Zeng, Jiaming, Berk Ustun, et Cynthia Rudin. 2017. « Interpretable Classification Models for Recidivism Prediction ». *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180 (3): 689-722. https://doi.org/10.1111/rssa.12227.

Zhao, Bailin, Eli Rothenberg, Dale A. Ramsden, et Michael R. Lieber. 2020. « The Molecular Basis and Disease Relevance of Non-Homologous DNA End Joining ». *Nature Reviews Molecular Cell Biology* 21 (12): 765-81. https://doi.org/10.1038/s41580-020-00297-8.

Zhao, Nan, Jing Ginger Han, Chi-Ren Shyu, et Dmitry Korkin. 2014. « Determining Effects of Non-synonymous SNPs on Protein-Protein Interactions using Supervised and Semi-supervised Learning ». *PLoS Computational Biology* 10 (5): e1003592. https://doi.org/10.1371/journal.pcbi.1003592.

Zheng, Baoying, et Ling Cao. 2017. « Differences in gene mutations between Chinese and Caucasian cystic fibrosis patients ». *Pediatric Pulmonology* 52 (3): E11-14. https://doi.org/10.1002/ppul.23539.

Zhou, Jian, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, et Olga G. Troyanskaya. 2018. « Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk ». *Nature genetics* 50 (8): 1171-79. https://doi.org/10.1038/s41588-018-0160-6.

Université de Strasbourg

Thomas WEBER

École Doctorale des Sciences de la Vie et de la Santé STRASBOURG

Nouvelles méthodes d'évaluation des variations génétiques *via* une approche bioinformatique : application aux maladies humaines.

Résumé :

Le séquençage du génome humain a bouleversé la biologie et ouvert la voie à une meilleure identification et interprétation des variations génétiques, reflet de notre diversité, mais pouvant entrainer des maladies génétiques rares. L'objectif de ma thèse était de développer des outils pour mieux caractériser des variations génétiques impliquées dans les maladies génétiques rares. Mes travaux se sont organisés autour de deux axes majeurs : Premièrement, le développement de MISTIC (*MISsense deleTeriousness predICtor*), nouvel outil basé sur de l'intelligence artificielle, visant à prédire l'impact des variations faux-sens. Les performances élevées de MISTIC découlent d'une architecture originale et d'un choix minutieux des descripteurs intégrés. Deuxièmement, la création de duxt (*differential usage across tissues*), une métrique pour mieux caractériser les variations situées dans les exons alternatifs. L'application de duxt a permis d'identifier des exons fortement/faiblement utilisés dans certains tissus et d'explorer leurs relations avec des variations impliquées dans certains phénotypes atypiques de maladies génétiques rares.

Mots-clés : variations génétiques, génome humain, faux-sens, intelligence artificielle, maladies génétiques rares, *Big Data*

Summary :

The sequencing of the human genome has dramatically changed biology and opened the way to a better identification and interpretation of genetic variations, which reflect our diversity but can lead to rare genetic diseases. The objective of my thesis was to develop tools to better characterise genetic variations involved in rare genetic diseases. My work was organised around two major axes: First, the development of MISTIC (MISsense deleTeriousness predICtor), a new tool based on artificial intelligence, aimed at predicting the impact of missense variations. The high performance of MISTIC is the result of an original architecture and a careful choice of embedded descriptors. Secondly, the creation of duxt (differential usage across tissues), a metric to better characterise variations located in alternative exons. The application of duxt has made it possible to identify exons with high/low usage in specific tissues and to explore their relationship with variations involved in certain atypical phenotypes of rare genetic diseases.

Keywords : genetic variations, human genome, missense, artificial intelligence, rare genetic diseases, Big Data