Annotation et hiérarchisation de variants non-codants dans le contexte de maladies humaines

Lambert Moyon

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École normale supérieure

Annotation et hiérarchisation de variants non-codants dans le contexte de maladies humaines

Soutenue par
Lambert MOYON
Le 25 septembre 2019

École doctorale n°515
Complexité du Vivant

Spécialité
Génomique fonctionnelle

Composition du jury :

Thomas Bourgeron Professeur, Institut Pasteur, Paris	*Président*
Anaïs Baudot Dr, MMG, Marseille	*Rapporteur*
Pascal Rihet Pr, TAGC, Marseille	*Rapporteur*
Alena Shkumatava Dr, Institut Curie	*Examinateur*
Anaïs Bardet Dr, Université de Strasbourg	*Examinateur*
Chunlong Chen Dr, Institut Curie	*Examinateur*
Hugues Roest Crollius Pr, IBENS	*Directeur de thèse*

1794
ENS | PSL

Remerciements

Cette aventure du doctorat a été l'occasion de nombreuses remises en questions, tant sur le plan scientifique que sur le plan personnel. Et si les jours ont passé, tantôt avec difficulté, tantôt avec légèreté, ce chemin intellectuel et émotionnel n'a pas été parcouru en solitaire : j'ai eu la chance de partager cette expérience avec de nombreuses personnes, qui ont chacune contribué un petit peu à l'aboutissement de ce cycle. Ce manuscrit, matérialisation de ces quatre années passées (si vite !) à grandir et à mûrir, est l'occasion d'un instant revenir sur mon parcours, et de remercier toutes les personnes qui ont partagé avec moi un peu de leur temps, de quelques heures à plusieurs années, collègues et amis.

Tout d'abord mes remerciements vont aux rapporteurs Dr Anaïs Baudot et Pr Pascal Rihet, qui ont pris le temps d'évaluer mon travail, et de me considérer digne de le défendre devant le jury de thèse. Également merci aux autres membres du jury, qui ont accepté de joindre cette évaluation par leur participation à ma défense : Pr Thomas Bourgeron, Dr Alena Shkumatava, Dr Anaïs Bardet, et Dr Chunlong Chen.

A Hugues Roest Crollius, un grand merci pour m'avoir donné l'opportunité de travailler avec toi sur un projet difficile, mais extrêmement stimulant. Cela a été l'occasion pour moi d'apprendre énormément, tant du côté biologie que du côté informatique. Ce projet a évolué dans un contexte scientifique particulièrement dynamique ; merci de m'avoir guidé dans les moments d'hésitations, et d'avoir su me pousser à explorer les pistes malgré mon caractère parfois (un peu) pessimiste.

Je souhaite également remercier les membres du comité de suivi de ma thèse : Dr Lucie Bittner, Dr Valentina Boeva, et Dr Morgane Thomas-Chollier. Notamment, je te remercie Morgane de m'avoir écouté, guidé dans mes réflexions, et poussé à me poser les bonnes questions, en particulier sur mes aspirations scientifiques.

Je remercie aussi toute l'équipe DYOGEN, dont la bonne humeur et la bienveillance ont garanti un plaisir quotidien à les retrouver au travail : les membres passés (Céline, Amélie, Joseph, Christine), et actuels : Alexandra, Camille, Elise, Nga, François, Guillaume, et Yves. Un merci tout particulier à Alexandra, mère de substitution de tous les doctorants dans l'équipe. Je remercie également Camille et Yves, qui ont, comme Morgane, écouté avec bienveillance mes interrogations existentielles, et m'ont aidé à avancer. Plus particulièrement, merci Camille pour ton support scientifique, et ta patience face à mes questions et à mes explications quelquefois bancales. J'en profite pour remercier Clovis et Lisa, avec lesquels l'expérience de tuteur de stage a été très agréable à découvrir. Et je

remercie aussi nos voisins bioinformaticiens : l'équipe CSB, avec qui j'ai partagé avec plaisir les déjeuners-débats (merci donc à Céline, Swann, Hatim, Nathalie, Aurélien, Laura, Morgane, et Denis), ainsi que Stéphane.

J'ai aussi eu le plaisir de découvrir l'expérience de l'enseignement, grâce au monitorat ; aussi je remercie les membres de l'équipe enseignante, en particulier Anne, Morgane, et Pierre. Je remercie au passage le travail phénoménal de Pierre ainsi que de l'ensemble de l'équipe du service informatique : Catherine, Nolwenn, Bilel, et Phi-Phong ; ce projet de thèse n'aurait pas pu être complété sans vous ! Merci aussi à Brigitte et Abdoul pour leur efficacité et leur patience concernant mes demandes administratives, parfois de dernière minutes !

J'ai eu l'occasion de croiser énormément de visages amicaux dans les couloirs du deuxième étage, je souhaite donc remercier l'ensemble de la section génomique fonctionnelle. A plus large échelle, j'ai pu rencontrer beaucoup de personnes très agréables ; je vais évidemment en oublier, je vous prie de m'excuser si votre nom n'est pas dans ces listes. Je souhaite d'abord remercier les membres du bureau SPIBENS, qui ont contribué à animer la vie étudiante au sein de l'institut ; je remercie en particulier Joanne, Yves, et Patrick. Je souhaite également remercier toutes les personnes qui ont contribué à l'expérience YRLS ; ça a été éprouvant, mais très formateur. Un merci en particulier à Claire pour avoir cru en moi, et un énorme merci à Caroline et Harold pour leur aide et soutien (et pour les pièces de monnaie).

Ces quatre années de thèse ont été l'occasion de former une jolie bande d'amis ; les soirées Mölkky-bières resteront un souvenir particulièrement agréable. Merci donc à Kasia, Baptiste, Yves, Camille, France, Benoît, Nikita, Tony, Tiphaine, Rémi, Guillaume, Vinko, Walter, Elisabetta, Swann, Hatim, Felipe (et merci à la K-fêt !) Et un merci tout particulier à Caroline pour sa générosité, et à Virginia pour toutes nos discussions de soutien mutuel (et pour le risotto !). J'ai eu aussi l'occasion de profiter d'amitiés en dehors du laboratoire : je remercie donc Guilhem, Guillaume, François, et Anca, pour avoir animé notre collocation, Rebecca, ainsi que Mathilde, Benoit, Louis, Alexandre, et Emeline.

Je souhaite enfin adresser un immense merci à mes proches, qui m'ont apporté un support sans faille durant cette aventure ; je remercie mes parents et mes soeurs, ainsi que ma grande tante Lucienne. Je remercie chaleureusement mes amis de longue date : Florian et Marine ; nos entrevues étaient sporadiques, mais ont été des bouffées d'oxygène pour moi. J'espère ne pas vous avoir étouffé en retour avec mes histoires de thèse ! Et

bien sûr, Charlotte : avec le temps qui passe et nos chemins respectifs, nos discussions et retrouvailles sont devenus de précieux instants. Merci d'avoir continué à répondre présente après toutes ces années, et merci du fond du coeur pour ces moments d'ailleurs.

Table des matières

Table des figures

Liste des tableaux

Première partie

Introduction

Chapitre 1

Le génome humain : gènes, expression, et régions régulatrices

1.1 Le défi de l'annotation du génome humain

Déchiffrer le code de la vie : tel était l'objectif du projet "Génome Humain" (Human Genome project), collaboration mondiale visant à décoder la séquence nucléotidique composant notre ADN. Ce projet, né à la fin des années 80 (Dulbecco, 1986, Sinsheimer, 1989) devait permettre de résoudre deux problématiques :

- avoir la séquence la plus complète possible du génome humain, représentant plus de 3 milliards de positions, réparties dans 22 paires de chromosomes "autosomaux" ainsi qu'une paire de chromosomes sexuels ;
- localiser et définir les unités fonctionnelles de ce génome, et plus particulièrement les gènes.

Les gènes étaient définis comme des séquences nucléotidiques portant les instructions qui, après transcription de la séquence ADN en un intermédiaire ARN, permettaient la synthèse d'une protéine, produit fonctionnel pouvant intervenir dans différents cycles biochimiques nécessaires au maintien de la cellule (Epp, 1997). Les deux points du projet apparaissaient donc comme des défis majeurs, dont la résolution permettrait de grandes avancées sur la compréhension et l'exploration de la diversité des produits protéiques encodés par les gènes, mais aussi sur l'évaluation de leur implication dans les maladies génétiques.

C'est en 2001 que les premiers résultats du projet "Génome Humain" ont été publiés

(« Initial sequencing and analysis of the human genome » 2001) ; en 2003, le séquençage du génome humain était annoncé comme achevé (International Human Genome Sequencing Consortium, 2004), bien qu'il ait fallu attendre 2006 pour avoir la séquence complète du chromosome 1 (chromosome le plus large, Gregory et al., 2006).

Cette séquence du génome a permis de caractériser plus finement des propriétés précédemment observées : par exemple, la confirmation de variations le long des chromosomes des taux de compositions en Guanine et Cytosine (nucléotides G et C), la confirmation des variations de taux de recombinaison de l'ADN en fonction de différentes régions génomiques, ou encore l'exploration des propriétés de différentes familles de séquences répétées.

C'est également grâce à ces travaux que la composition génique du génome humain a pu être évaluée en détails. Une partie des analyses a concerné certains gènes non-codants, dont le rôle est bien établi (par exemples : les ARN de transfert ou les ARN ribosomaux) ; leurs nombres et localisations ont pu être déterminés plus précisément. Cependant la question des gènes codants était la plus importante à résoudre, avec notamment le défi de donner un nombre exact pour ces gènes, et ainsi conclure sur les différentes estimations proposées auparavant. En effet, certaines estimations proposaient des nombres entre 50 000 et 100 000 gènes (Pertea et al., 2010). Ces estimations ont été largement diminuées par une comparaison des premières ébauches de la séquence du génome humain avec celle du Tétraodon, et ramenées à un intervalle de 28 000 à 34 000 gènes (Crollius et al., 2000). Et finalement, à l'issue du projet Génome humain, le nombre de gènes codants détectés dans les séquences a été réduit à un intervalle entre 20 000 et 24 000 (International Human Genome Sequencing Consortium, 2004). En plus de cette réduction importante par rapport aux précédentes observations, ces analyses ont permis d'observer que tous les gènes présentent des transcrits alternatifs (environs 3 par gène), et que ces gènes présentent de très larges introns, tandis que la séquence codante représente une très faible partie du génome (environs 1.5%).

Le séquençage du génome complet de la souris, obtenu peu de temps après (« Initial sequencing and comparative analysis of the mouse genome » 2002), a été l'occasion d'explorer la conservation de l'architecture et de la séquence du génome humain avec un autre représentant des mammifères. Cette comparaison a permis d'identifier qu'au total, 5% de la séquence du génome humain est conservée chez la souris (Guénet, 2005). Ainsi, le génome non-codant contient certaines régions qui sont sous pression de sélection négative, indiquant que leur séquence joue un rôle potentiellement important dans la cellule. Etant

donnée la complexité observée dans le nombre de transcrits identifiables, ainsi que dans la diversité des types cellulaires composant l'organisme, cette identification de régions non-codantes conservées a invité à l'exploration d'une hypothèse sur le génome non-codant : des régions régulatrices de l'expression des gènes y sont potentiellement localisées, leur importance conduisant à une pression de sélection négative sur leur séquence.

1.2 La régulation de l'expression des gènes

A la suite des résultats obtenus sur l'étude de la séquence du génome humain, de larges consortiums se sont constitués pour explorer les propriétés biochimiques associées au génome, et notamment au génome non-codant. Des projets comme le projet ENCODE (The ENCODE Project Consortium, 2012), le projet Roadmap Epigenomics (Kundaje et al., 2015), ou encore le projet FANTOM (Andersson et al., 2014) ont ainsi permis d'ouvrir la voie vers une meilleure compréhension des signaux biochimiques associés au potentiel régulateur du génome.

Dans cette section, je présente tout d'abord quelques définitions concernant la régulation de l'expression des gènes et ses différents acteurs. Par la suite, je présenterai les signaux principaux identifiés et associés à un potentiel régulateur ; j'aborderai également la question de la prédiction des régions régulatrices. Les associations entre ces régions et les gènes cibles sous leur contrôle seront abordées dans une troisième sous-section.

1.2.1 Les acteurs de la régulation de l'expression des gènes

L'expression d'un gène correspond à la génération depuis sa séquence d'ADN d'un transcrit ARN, composé de différentes sous-régions suivant la nature du gène. Dans le cadre d'un gène codant pour une protéine, on peut identifier les exons et les introns ; une étape d'épissage conduit à l'inclusion d'une partie des exons dans un produit final appelé ARN messager (ou ARNm). Cet ARN messager sera transporté depuis le noyau de la cellule vers le cytoplasme, pour potentiellement conduire à sa traduction en une séquence protéique. D'autres gènes voient également leur séquence transcrite en ARN, mais les étapes intermédiaires de génération et maturation du transcrit ne sont pas toutes partagées avec les transcrits des gènes codants (par exemples : les long non-coding RNA, les ARN de transfert, etc).

L'ensemble des cellules composant les tissus du corps humain partagent le même gé-

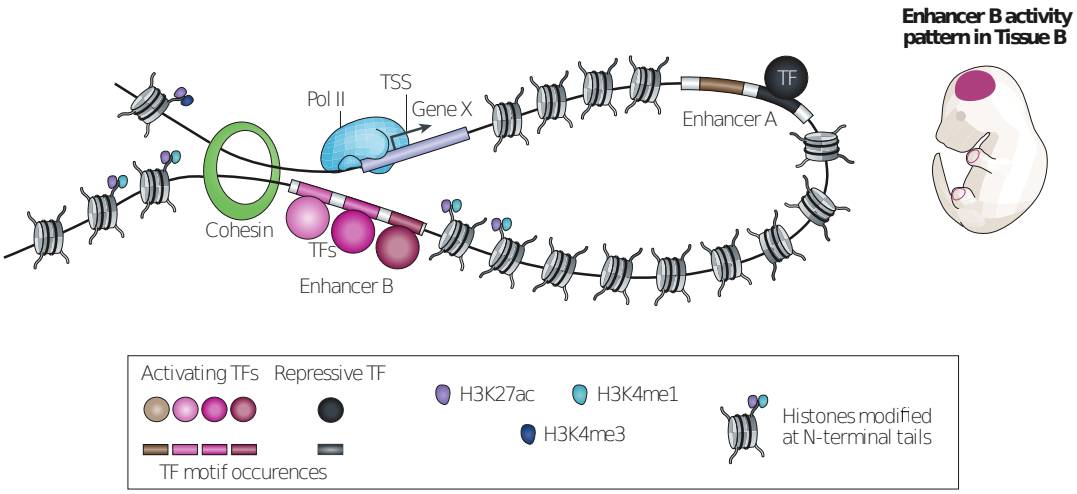


FIGURE 1.1 – Illustration des différents acteurs de la régulation de l'expression des gènes. La machinerie de transcription, résumée à la polymérase II, s'est fixée proche du TSS du gène X par reconnaissance de la séquence promotrice, et va produire un transcrit ARN depuis la séquence ADN du gène. La transcription du gène est faite suivant un schéma spatio-temporel précis, illustré sur l'embryon à droite. Ce schéma est associé à l'activation de cette transcription par la région régulatrice notée "enhancer B". Sa séquence contient les motifs de facteurs de transcription qui définissent ce schéma d'expression spécifique ; en parallèle, l'enhancer A est ici réprimé, et ne peut donc pas appliquer son schéma d'expression au gène X. L'action de la région régulatrice B est possible par sa proximité physique en trois dimensions avec la région promotrice du gène (et donc avec la machinerie de transcription) ; la formation de la boucle permettant cette proximité est notamment médiée par la cohésine. Le promoteur et la région régulatrice sont par ailleurs identifiables par une accessibilité plus marquée de la chromatine, en comparaison avec les régions adjacentes, denses en nucléosomes. Ces nucléosomes présentent enfin des modifications protéiques particulières au niveau de leurs histones, marquant différentes régions ; ici l'enhancer B est entouré de marques d'histones H3K27ac et H3K4me1 (associées aux éléments régulateurs actifs), tandis que le promoteur a en amont un nucléosome portant les modifications H3K27ac et H3K4me3 (associées aux promoteurs actifs). Figure adaptée de SHLYUEVA et al., 2014

nome ; l'expression spatio-temporelle spécifique des gènes, qui permet de définir l'identité de chaque cellule, est donc contrôlée par différentes régions : les régions régulatrices de l'expression des gènes. Je présente ci-dessous quelques définitions concernant ces régions régulatrices et leur lien avec les gènes ; la majorité des éléments présentés sont représentés sur la figure 1.1. A noter que ces régions correspondent à des annotations possibles de la séquence nucléotidique ; l'organisation structurelle de cette séquence est appelée chromatine et désigne l'ensemble formé par la séquence et les protéines de structures, notamment

les protéines d'histones (qui définissent les unités de structure appelées nucléosomes).

Les régions promotrices. Les promoteurs, ou régions promotrices, correspondent aux séquences situées en amont des sites d'initiation de la transcription des gènes. Ces régions permettent l'assemblage de la machinerie de transcription (dont l'ARN polymérase II, qui permet la génération du transcrit). Cet assemblage implique la reconnaissance de motifs particuliers dans la séquence, dont des motifs reconnus par des facteurs de transcription (voir ci-dessous), qui permettent d'activer ou d'inhiber la transcription.

Les régions cis-régulatrices. Les régions cis-régulatrices regroupent plusieurs types de régions, principalement composés des "enhancers" et "silencers". Ces régions sont associées à un contrôle positif ou négatif de l'expression de gènes situés à distance de ces régions (en comparaison avec les promoteurs), mais dans un contexte local (au maximum à quelques mégabases des gènes dont ils impactent l'expression). Les enhancers sont plutôt associés à une activation de l'expression des gènes, tandis que les silencers sont associés à une répression ; ces derniers sont plus difficiles à identifier clairement et à valider, ce qui fait que le terme enhancer est parfois étendu à l'ensemble des régions cis-régulatrices. Ce sont les acteurs principaux de la spécificité spatio-temporelle de l'expression des gènes, par la fixation de facteurs de transcription qui vont contribuer au recrutement et à l'activation de la machinerie de transcription.

Les éléments trans-régulateurs. Deux types d'éléments sont désignés comme "trans-régulateurs" :

- les régions régulatrices localisés à de très longues distances, voire sur des chromosomes différents, des gènes auxquels elles sont associées. Ces régions sont par exemple identifiées par des méthodes de capture de la chromatine (voir section suivante), mais peu sont validées ; je n'en parlerai donc pas plus.
- les facteurs de transcription, protéines régulatrices capables de reconnaître des régions spécifiques dans le génome, de s'y fixer, et d'agir potentiellement sur la machinerie de transcription.

Les facteurs de transcription (ou TF pour "Transcription Factors") sont des acteurs déterminants de la régulation de l'expression des gènes. Certains de ces facteurs sont capables de reconnaître des motifs nucléotidiques dans la séquence du génome, favorisant leur liaison à l'ADN : ce sont les sites de fixation de facteurs de transcription (ou TFBS

pour Transcription Factor Binding Sites), dont la composition en nucléotides est plus ou moins spécifique à chaque facteur. D'autres facteurs, appelés co-facteurs, se lient indirectement par la reconnaissance des facteurs de transcription, plutôt que la reconnaissance de la séquence d'ADN. Ces facteurs de transcriptions sont l'acteur intermédiaire entre les régions régulatrices (évoquées ci-dessus) et la machinerie de transcription. Leur expression (depuis les gènes codants associés) va conduire à une cascade de régulation : la reconnaissance de leurs sites de fixations dans le génome va déterminer l'activation ou l'inhibition de plusieurs gènes (ainsi identifiés comme "régulés" par ces facteurs).

Domaines de régulation. Toutes ces régions génomiques évoquées sont identifiées de manière linéaire sur les chromosomes. Cependant les actions des régions cis-régulatrices et des facteurs de transcription peuvent se réaliser sur des gènes situés à des distances très longues vis-à-vis des régions régulatrices. Ainsi le génome est en réalité une structure en trois dimensions, dont les éléments lointains linéairement peuvent se retrouver proches dans le noyau, par la formation de boucles, faisant intervenir différentes protéines de structure, comme la cohésine (schématisée sur la figure (1.1). Des domaines d'organisation de la chromatine, structure du génome comprenant la séquence d'ADN et les protéines d'histones autour desquelles s'organise l'ADN en nucléosome, peuvent ainsi être identifiés, à de plus ou moins larges échelles. En particulier, des domaines d'associations topologiques (Topological Associations Domain, ou TAD) peuvent être définis à une échelle d'environs 1 million de paires de bases. Ces domaines sont associées à des propriétés de régulation de l'expression des gènes plus ou moins coordonnées (Bemmel et al., 2019), et plus ou moins séparées entre domaines.

1.2.2 Méthodes expérimentales d'identification des régions régulatrices

Comme évoqué précédemment, différents projets ont proposés des approches pour une identification systématiques des régions régulatrices dans le génome. Par exemple les projets ENCODE et Roadmap Epigenomics ont conduit à l'identification de plusieurs signaux biochimiques à travers de multiples types cellulaires ; chacun de ces signaux peut être exploité pour en dériver des prédictions d'éléments régulateurs, mais il est également possible d'avoir une approche intégrative pour identifier plus finement des combinaisons de ces signaux. Je présente dans cette section les différentes approches d'identification de régions régulatrices, avec les signaux fonctionnels associés.

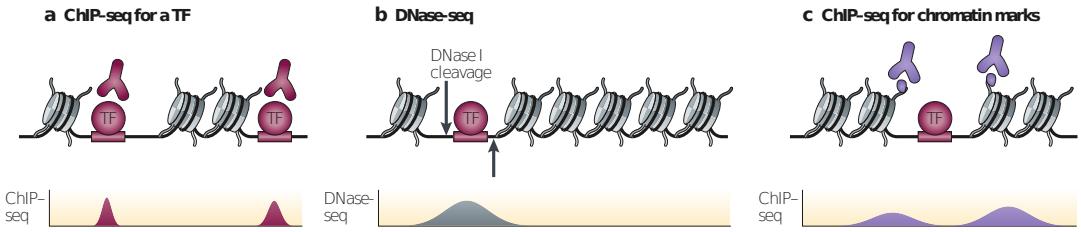


FIGURE 1.2 – Schéma de méthodes de détections de signaux associés aux régions régulatrices. (a) L'identification des sites de fixation pour un facteur de transcription par ChIP-seq permet d'identifier des pics correspondant aux localisations génomiques avec lesquelles le facteur était en contact au moment de l'immunoprécipitation. (b) L'évaluation de l'ouverture de la chromatine par DNase-seq permet de localiser les régions où l'absence de nucléosomes a permis à l'enzyme d'avoir son action de nucléase sur la séquence d'ADN. (c) Les expériences de ChIP-seq appliquées aux modifications d'histones permettent de détecter des régions où les nucléosomes portent une marque d'histone d'intérêt. Dans le cadre de l'identification de marques associées aux éléments régulateurs, les marques d'histones comme la marque H3K27ac vont se trouver sur les nucléosomes autour d'une région accessible de la chromatine (et donc offrant la possibilité d'une fixation d'un facteur de transcription). Figure adaptée de SHLYUEVA et al., 2014

Génomique comparative

Le séquençage de génomes complets de différentes espèces a permis d'identifier dans le génome humain des régions génomiques présentant une conservation en séquence nucléotidique plus ou moins importante (HARDISON, 2003). Cette conservation observée pour certaines positions est associée à l'hypothèse que la sélection négative impose une contrainte sur une séquence, parce que sa composition particulière est fonctionnellement importante. Pour les séquences codantes, la conservation observée correspond au maintien de la séquence protéique, tandis que pour les régions non-codantes, ce maintien est associé à un caractère régulateur, potentiellement médié par des facteurs de transcription.

Comme évoqué plus haut, le génome complet de la souris a été la première occasion d'une comparaison du génome humain avec celui d'un autre mammifère, conduisant à l'identification de régions conservées non-codantes représentant 3.5% du génome humain. D'autres génomes complets d'espèces ont été obtenus, contribuant à une mesure de plus en plus fine de la contrainte de sélection imposée sur les différentes régions du génome. Ainsi, plusieurs approches ont été développées pour identifier les éléments régulateurs dans le génome non-codant, sur la base de la conservation en séquence de régions génomiques.

Par exemple, le projet VISTA (Visel et al., 2007) a proposé une méthode de détections de régions conservés à partir de l'identité exacte en séquences de régions de 200 paires de bases, alignées entre les génomes de l'humain, de la souris et du rat. Le caractère régulateur de ces régions a été testé par construction de gènes rapporteurs (basée sur l'injection dans des embryons de souris d'une construction génétique comprenant la région régulatrice à tester, une région promotrice faible, et un gène dont on peut détecter l'expression), ce qui a permis de valider la capacité de ces régions à activer l'expression des gènes, de manière spécifique à certains tissus de l'embryon.

Pour aller au delà de la simple évaluation d'identité en séquence de segments, des scores plus élaborés ont également été proposés. Le score PhastCons par exemple (Siepel et al., 2005), permet à partir d'un alignement de séquence d'identifier des segments conservés, puis de calculer a posteriori pour chaque base du génome une probabilité d'appartenir à un segment conservé. Les scores PhyloP (Pollard et al., 2010), et GERP (Davydov et al., 2010) permettent quant à eux de calculer un degré de divergence par rapport à un modèle de référence d'évolution de la séquence, et donc d'identifier des positions dont la séquence est plus ou moins conservée qu'attendu.

Régions de chromatine ouverte

Comme représenté sur la figure 1.1, les régions présentant des caractéristiques de régions régulatrices sont généralement associées à une diminution de la présence de nucléosomes, et donc à une accessibilité de la séquence nucléotidique plus importante. Cette accessibilité peut être mesurée par différentes techniques : par exemple, l'enzyme nucléase DNase 1 a été exploitée pour identifier les régions présentant une sensibilité accrue à l'action de cette enzyme, marquant les régions accessibles (Sabo et al., 2006, Song et al., 2010). La figure 1.2 représente sur le panneau b cette approche. Une autre technique (Formaldehyde Assisted Isolation of Regulatory Elements, Giresi et al., 2007) permet de séparer en différentes phases les régions associées aux nucléosomes, de celles qui sont dans un état accessibles, permettant le séquençage de ces dernières. Plus récemment, la méthode d'ATAC-seq s'est démocratisée pour explorer les régions de chromatine ouverte, notamment à l'échelle de chaque cellule ; cette technique implique l'utilisation d'une transposase hyperactive, capable de fragmenter et de marquer les régions ouvertes du génome, permettant leur séquençage.

Des enrichissements en régions promotrices ont été observées pour les régions ouvertes

de la chromatine, conduisant à l'hypothèse que les régions dont l'accessibilité de la séquence est plus importante présentent un potentiel régulateur plus élevé, ce qui permet l'identification de régions régulatrices putatives. Par exemple, dans le cadre du projet ENCODE, ces régions ouvertes de la chromatines ont été identifiées de manière systématiques dans 125 types cellulaires différents (Thurman et al., 2012), conduisant à un catalogue de 2.9 millions de régions accessibles.

Modifications d'histone

Le relâchement de la chromatine peut notamment être expliqué par le dépôt de modifications sur les protéines d'histones qui composent les nucléosomes, ce qui conduit à des modifications du contexte biochimique et favorise ou diminue le compactage de cette chromatine, suivant les modifications d'histones. Il est possible de séquencer à l'échelle du génome l'ensemble des régions qui présentent dans leurs nucléosomes une marque d'histone d'intérêt, par l'application de la méthode d'immuno-précipitation de la chromatine, suivie par une étape de séquençage (comme schématisé sur la figure 1.2 c). Les projets ENCODE et Roadmap Epigenomics ont ainsi proposé une identification systématiques de plusieurs marques d'histones, chacune associée à une conséquence fonctionnelle particulière, à travers de multiples types cellulaires. Quelques marques notables :

- H3K4me1 : cette marque est généralement associée aux régions potentiellement cis-régulatrices ;
- H3K4me3 : cette marque est généralement associée aux régions promoteurs ;
- H3K27ac : cette marque est généralement associée aux régions actives, en particulier enhancers ;
- H3K27me3 : cette marque est généralement associée à une répression des régions régulatrices ;
- H3K9me3 : cette marque est généralement associée à un état hétérochromatinien, inactif ;
- H3K36me3 : cette marque est généralement associée à une transcription active, détectée dans le corps des gènes.

Ces expériences permettent donc d'identifier des régions présentant ces modifications d'histones, et donc potentiellement associées à des propriétés régulatrices particulières.

Dans le cadre des projets ENCODE et Roadmap Epigenomics, la disponibilité de multiples marques d'histones pour les mêmes types cellulaires a conduit à la définition d'états

chromatiniens, par segmentation du génome en régions et assignation pour chacune d'un état correspondant à une combinaison particulière de marques. Par exemple pour le projet Roadmap Epigenomics (Kundaje et al., 2015), les localisations de régions associées à différentes marques d'histones ont été obtenues pour 111 types cellulaires, dont 98 types cellulaires pour lesquels les 6 marques mentionnées ci-dessus ont été obtenues. Par l'utilisation de méthodes de segmentation du génome (Ernst et al., 2010, Ernst et al., 2012), le génome a été divisé en régions de 200 paires de bases, pour chacune desquelles un état chromatinien est assigné, dérivé de la combinaison des marques d'histones identifiées pour cette région.

Comme présenté sur la figure 1.3, les états sont associés à des combinaisons particulières et distinctes de modifications d'histones. Par exemple les états transcrits ("5_Tx" et "6_TxWk") sont associés à des probabilités élevées d'y trouver un signal H3K36me3, tandis que les états enhancers ("7_EnhA1" et "8_EnhA2") sont associés à des probabilités élevées d'y trouver des signaux H3K27ac et H3K4me1.

Prédictions de sites de fixation de facteurs de transcription

Les facteurs de transcription sont capables de reconnaître des contextes biochimiques particuliers, notamment défini par un motif nucléotidique particulier et spécifique à chacun, favorisant leur interaction avec la séquence (Dror et al., 2016). Comme représenté sur la figure 1.2 a, la méthode de ChIP-seq peut être appliquée à l'identification des régions du génome avec lesquelles les facteurs de transcriptions sont capables d'interagir ; à nouveau, plusieurs de ces expériences ont été réalisées dans le cadre du projet ENCODE. Ces régions présentent donc un potentiel à être reconnues par des facteurs de transcriptions, et donc à jouer un rôle dans la régulation des gènes.

Régions régulatrices transcrites

Le projet FANTOM5 (Andersson et al., 2014) a proposé une identification de l'ensemble des événements d'initiation de la transcription le long du génome, par application de la méthode de CAGE (Cap Analysis Gene Expression), permettant de capturer des transcrits naissants. Dans le cadre de ce projet, des événements de transcription bi-directionnels ont été identifiés à différents endroits du génome. Ces événements ont formellement été distingués d'erreurs de la machinerie de transcription, et ont été associé à une activité spécifique de l'ARN polymérase dans ces régions.

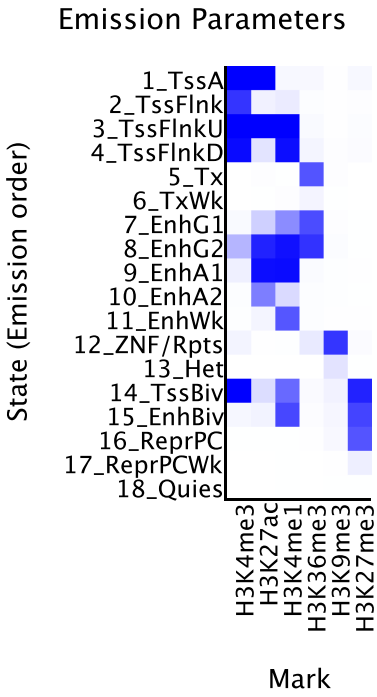


FIGURE 1.3 – Probabilités d'émission des marques d'histones pour chacun des états identifiés par ChromHMM, pour le projet Roadmap Epigenomics. Plus une case est colorée, plus la probabilité de trouver un signal pour la marque donnée est élevée, dans l'état considéré. Source : `https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final/emissions_18_core_K27ac.svg`

Ces événements ont été associés à un potentiel régulateur, dont l'activité est directement mesurable par la quantité d'ARN produit (définissant d'ailleurs une classe de régions régulatrices appelées enhancer RNA). Bien que la fonction de ces eRNA soit floue, les 44 000 régions identifiées dans le cadre de ce projet sont exploitables pour définir des régions régulatrices d'intérêt.

1.2.3 Prédictions d'association entre régions régulatrices et gènes cibles

Le problème de l'identification de régions régulatrices n'est pas évident ; à cette identification, la question des gènes cibles s'ajoute. Comme évoqué dans les définitions, les régions régulatrices ne vont pas agir uniquement de manière linéaire sur les gènes qui leur sont le plus proche, mais vont être potentiellement rapprochées de leurs gènes cibles par des boucles se formant dans la chromatine.

L'identification d'associations non-linéaires est donc un point important pour pouvoir associer fonctionnellement une région identifiée comme régulatrice avec une action sur un ou plusieurs gènes cibles. Ces associations peuvent être mesurées expérimentalement, par détection des boucles de chromatine, soit inférées. Je présente les différentes approches ci-dessous.

Inférence par la conservation de synténie. La conservation plus ou moins marquée de l'organisation des gènes peut aussi être exploitée pour identifier un maintien plus important que prévu de paires de régions entre elles, appelé Gene Regulatory Blocks (GRBs, Kikuta et al., 2007). Ce principe d'étude de la conservation en synténie d'élements conservés et de gènes environnants a été exploité récemment pour prédire des relations d'associations régulatrices entre régions régulatrices et gènes cibles (Naville et al., 2015, Clément et al., 2018).

Inférence par corrélation de signaux fonctionnels. Comme présenté à la section précédente, l'identification de régions potentiellement régulatrices peut être faite à travers différents types cellulaires, par la détection de signaux fonctionnels associés à un potentiel régulateur (expression, ou ouverture de la chromatine). Ces informations peuvent être exploitées pour établir des corrélations statistiques entre régions cis-régulatrices non-codantes et régions promotrices, avec l'hypothèse que l'action médiée par la région cis-régulatrice va conduire à des propriétés similaires entre cette région et le promoteur sur

lequel elle agit.

Ainsi, dans le cadre du projet FANTOM5 (Andersson et al., 2014), des prédictions d'associations entre régions régulatrices et gènes cibles ont été établies sur la base de la corrélation entre les niveaux d'expression mesurés aux régions identifiées comme cis-régulatrices et aux régions promotrices, au travers des 41 tissus étudiés. Cette approche a aussi été proposée par le projet FOCS (Hait et al., 2018), dans le cadre d'une inférence systématique d'associations entre éléments régulateurs et gènes-cibles dans la base de données FOCS, qui propose ces prédictions basées sur des données d'expression, ainsi que sur des données de mesure d'ouverture de la chromatine, provenant des projets ENCODE et Roadmap Epigenomics.

Capture de la conformation chromatinienne. Des méthodes ont été développées récemment pour permettre de lier chimiquement les régions chromatiniennes physiquement proches dans le noyau, pour pouvoir les séquencer et établir des prédictions de proximité physique, indiquant de potentielles relations de régulation (Dekker et al., 2013, Rao et al., 2014). Une méthode en particulier appelée capture Hi-C (Belton et al., 2012), permet d'interroger à l'échelle du génome complet l'ensemble des régions qui sont en interactions avec les régions promotrices. Cette méthode a été utilisée pour identifier des régions régulatrices dans différents types cellulaires (Mifsud et al., 2015, Denker et al., 2016) Cependant, la résolution des régions identifiées est limitée, ce qui conduit notamment à prédire des interactions entre régions très larges (plusieurs milliers de paires de bases) ; ces régions impliquées dans une interaction sont peu probablement régulatrices dans leur entièreté, et il est plus probable que des éléments plus restreints localisés dans ces régions soient à l'origine d'une régulation potentielle.

Associations statistiques avec des différences d'expression. Dans le cadre de l'analyse de variations nucléotidiques dans le génome, le projet GTEX (The GTEx Consortium et al., 2015) a proposé une étude de variants identifiés chez plus de 400 patients, avec une analyse de l'expression des gènes chez ces patients à travers différents tissus. Cette expérience a permis l'association statistique de plus de deux millions de variants avec des variations du niveau d'expression de gènes ; je présente plus en détails ces variants au chapitre suivant. Ces associations peuvent être utilisées pour associer des régions potentiellement régulatrices à des gènes cibles, par leur chevauchement de variants eQTLs.

Ainsi, différentes approches permettent d'établir des associations entre des régions régulatrices et des gènes cibles. Ces méthodes ont récemment été exploitées et centralisées dans un jeu de données appelé GeneHancer (Fishilevich et al., 2017, qui propose l'identification d'un répertoire d'éléments régulateurs, établi sur les bases de différentes prédictions (en prenant par exemple les régions régulatrices du projet FANTOM, ou encore les éléments conservés identifiés par VISTA), et en établissant des prédictions d'associations par leur chevauchement avec des régions obtenues par capture de conformation chromatinienne, ainsi que par les mesures de co-expression du projet FANTOM ; d'autres associations ont été prédites par l'utilisation de variants eQTLs, identifiés pour leur association statistique avec des variations d'expression de gènes.

1.2.4 Résumé

L'identification du répertoire des éléments régulateurs dans le génome non-codant reste un problème complexe, qui n'est pas clairement résolu. La mise en place de larges projets d'études des propriétés biochimiques du génome non-codant ont permis de mettre en lumière des signaux mesurables, dont les corrélations avec quelques régions régulatrices caractérisées ont permis une inférence de la localisation des régions régulatrices dans les génomes. Il est important de garder à l'esprit que ces inférences indiquent des régions avec un potentiel régulateur, plutôt que des régions clairement régulatrices. Par ailleurs ces méthodes sont complémentaires ; l'intégration de ces signaux permet d'identifier des degrés de potentialité pour certaines régions du génome, et donc d'assigner des profils généraux à chaque région du génome. Enfin, ces identifications de régions nécessitent d'être couplées à des prédictions d'interactions entre régions régulatrices et gènes cibles, pour appuyer le rôle des régions identifiées, et potentiellement mesurer leur action. Toutes ces prédictions peuvent être validées expérimentalement, notamment pour vérifier le potentiel fonctionnel des régions régulatrices identifiées (Santiago-Algarra et al., 2017). Des expériences comme les MPRA (Massively Parallel Reporter Assays), les expériences de STARR-seq (self-transcribing active regulatory region sequencing), ou encore l'utilisation de CRISPR-Cas9 pour évaluer l'impact de changement nucléotidiques, permettent de confirmer l'activité régulatrice de régions régulatrices candidates, et sont nécessaires pour formellement valider (à une échelle limitée) les identifications proposées *in silico*.

Nous avons donc à notre disposition de nombreux signaux d'informations pour explorer le génome non-codant, et identifier dans celui-ci quelles sont les positions qui sont

effectivement associées à un potentiel régulateur. Cette disponibilité d'information va être primordiale pour le problème que je souhaite aborder au chapitre suivant : la caractérisation des variations génétiques entre individus, lorsqu'elles se produisent hors des séquences géniques. Avant d'aborder cette question, je souhaite souligner l'importance de la mise à disposition des données qui ont été générées dans ces projets. En effet, les consortiums mentionnés dans ce chapitre ont contribué à une mise en place d'un socle commun d'informations pour explorer plus en détails les propriétés des régions fonctionnelles du génome humain. La mise à disposition systématique et sans condition des résultats et des données issues de ces larges projets est un élément clé de l'avancée des connaissances en génomique ; l'étude présentée dans ce manuscrit a été possible grâce à cette disponibilité.

Chapitre 2

Le Génome : entre variations et contraintes

2.1 Introduction

Comme présenté dans la partie précédente, la disponibilité de la séquence du génome humain a ouvert la voie notamment à l'identification de régions conservées en séquence, par comparaison avec les génomes d'autres espèces. La séquence du génome humain fournit également une base de référence pour comparer et explorer les différences génétiques entre les individus au sein de la population humaine. Je présente dans ce chapitre les méthodes d'identification de ces variations génétiques chez les individus. Pour chaque méthode, j'aborde également l'utilisation possible des variants identifiés pour établir des associations entre des positions nucléotidiques et des traits phénotypiques (différences morphologiques, ou maladies génétiques). Je terminerai ce chapitre par une illustration de ces méthodes appliquées à l'étude des troubles du spectre autistique, et je conclurai sur le défi que représente l'étude des variants identifiés dans le génome non-codant.

2.2 Détection et associations fonctionnelles des variations

2.2.1 Puce de génotypage

Détection des génotypes

Les positions nucléotidiques dans le génome ne sont pas indépendantes les unes des autres : le génome peut être décrit par des blocs contenant des positions nucléotidiques corrélées du fait de la rareté des événements de recombinaison lors de la méïose. Ces blocs correspondent à des régions de déséquilibre de liaison élevées entre nucléotides (Daly et al., 2001), conséquences de variations de tailles de populations durant l'histoire évolutive humaine. Exploiter cette propriété permet d'explorer le génome à une échelle réduite : en effet, il est possible de considérer directement ces blocs de déséquilibre de liaison pour résumer l'information nucléotidique des régions génomiques.

A l'issue du séquençage du génome humain, des projets ont donc proposé d'identifier des positions nucléotidiques présentant des variations de composition dans différentes populations humaines, appelés SNPs (Single Nucleotide Polymorphism). Le but était d'obtenir une couverture la plus large possible du génome avec un nombre restreint de positions, dont le déséquilibre de liaison avec les positions alentours permettrait d'identifier des régions appelées blocs haplotypiques.

Ainsi les projets comme le projet HapMap (« A haplotype map of the human genome » 2005, « Integrating common and rare genetic variation in diverse human populations » 2010) et le groupe de travail "The International SNP Map Working Group" (« A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms » 2001) ont conduit à l'identification et l'exploitation de millions de SNPs dans le génome, afin de résumer l'information génétique des 3 milliards de positions à la composition génétique de quelques millions de positions. Le principe de l'identification de ces SNPs marqueurs est illustré sur la figure 2.1. Il est donc possible d'exploiter ces marqueurs pour cribler et caractériser l'identité génétique de patients (leur génotype), notamment par l'utilisation de puces de génotypage (Louhelainen, 2016), dans lesquelles des sondes à ADN permettent d'identifier l'hybridation allèle-spécifique de régions génomiques contenant les marqueurs. Ces puces permettent ainsi d'identifier pour un patient sa composition allélique aux différents marqueurs, et par extension les blocs haplotypiques qu'il possède dans son génome.

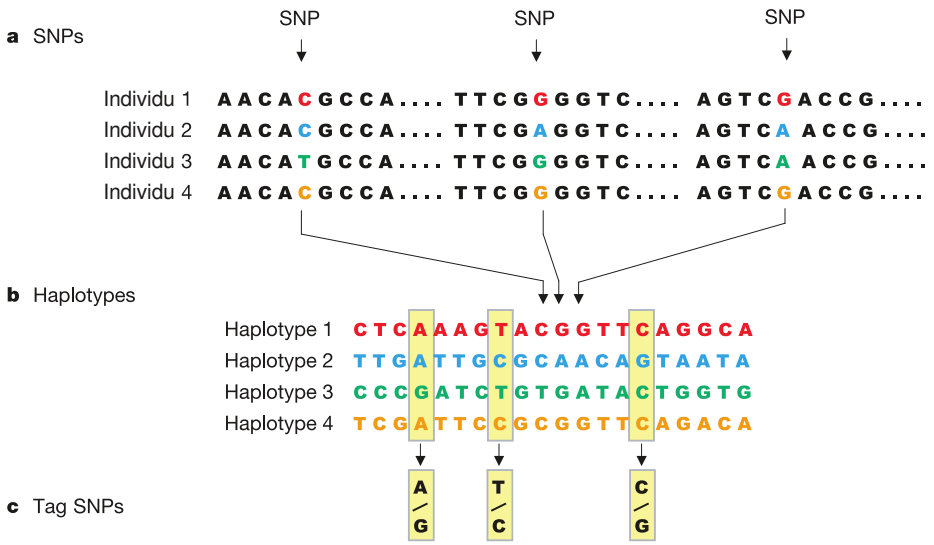


FIGURE 2.1 – Définition des SNPs, blocs haplotypiques, et SNP marqueurs.
(a) L'identification des SNPs (single nucleotide polymorphisms) est faite sur la base de divergence entre individus à certaines positions. Ici, 4 régions génomiques sont comparées ; la majorité des positions sont identiques, cependant trois positions peuvent être identifiés comme présentant une variabilité allélique : ce sont des SNPs.
(b) En agrégeant les positions SNPs chez différents individus, il est possible de définir des combinaisons particulières de ces SNPs. Ces combinaisons peuvent être comparées entre individus, ce qui permet de détecter des co-occurrences systématiques, représentant des blocs de SNPs systématiquement partagés entre individus : ce sont les haplotypes. Ici, quatre haplotypes différents sont détectables à l'échelle d'une population entière.
(c) Les haplotypes peuvent à nouveau être réduits à une combinaison restreinte de SNPs, qui résument la composition de chacun des haplotypes : ce sont les SNP marqueurs (tag SNP). Par exemple, trois marqueurs sont proposés ici ; la combinaison A-T-C pour ces trois positions permet de définir de manière unique l'haplotype 1, lui-même composés de SNPs entre lesquelles les intervalles de nucléotides ont été identifiés comme identiques entre individus.
Ainsi un haplotype permet de couvrir plusieurs milliers de bases dans le génome, en les résumant à leurs SNPs flanquant, et ces haplotypes peuvent être identifiés de manière unique par un sous-ensemble de SNP marqueurs.
Figure adaptée de « A haplotype map of the human genome » 2005.

Associations fonctionnelles des SNP

Par l'utilisation des puces à ADN, il est possible de résumer l'identité génétique de plusieurs milliers d'individus à leurs génotypes. Cette approche peut donc être appliquée pour explorer la diversité génétique de larges populations, et établir des associations sta-

tistiques (par régression logistique, ou par test de chi2) entre les génotypes des individus et leurs phénotypes. C'est le principe à la base des études d'association à l'échelle du génome (ou Genome Wide Association Studies, notées GWAS) : identifier plusieurs individus partageant un trait phénotypique, comme par exemple une maladie, et comparer leurs génotypes à un second groupe contrôle ne présentant pas le trait d'intérêt. Dans cette configuration, l'association statistique entre un SNP et le phénotype se fait par l'évaluation de la présence plus fréquente d'un des allèles par rapport aux autres, chez les individus présentant le trait phénotypique d'intérêt. Ces approches ont ainsi permis d'identifier de nombreuses associations entre SNPs et maladies, permettant ainsi la définition de marqueurs de susceptibilité à ces maladies, (comme par exemple pour le diabète de type 2, Sladek et al., 2007).

Il est également possible de mesurer des associations statistiques entre les variations alléliques et des traits biochimiques mesurables. Par exemple, le projet GTEX a proposée une étude sur les effets de la composition allélique de certains SNP sur les niveaux d'expression de gènes, à travers 44 tissus chez 449 individus (The GTEx Consortium et al., 2015, GTEx Consortium et al., 2017). Par des expériences de séquençage des ARN, les niveaux d'expression des gènes chez ces individus ont été mesurés au travers des 44 tissus ; les associations entre SNPs et niveaux d'expressions ont été statistiquement évaluées, comme schématisé sur la figure 2.2. Les associations statistiquement significatives entre les SNPs et les gènes ont ainsi permis d'identifier des variants eQTL (expression Quantitative Trait Loci).

Ces méthodes d'associations statistiques permettent donc d'établir des liens entre des génotypes et des traits phénotypiques, avec l'idée que le trait mesuré pour un individu est une fonction de son génotype, et donc par la comparaison de nombreux génotypes, il est possible d'identifier et de quantifier ces associations. Cependant, il est important de garder à l'esprit que les variants identifiés par ces méthodes correspondent à des SNP, choisis comme marqueurs représentant de blocs plus larges de variations. L'association entre un SNP et un trait permet donc d'identifier que la région en déséquilibre de liaison avec le SNP marqueur est associée avec le trait, mais il n'est pas possible d'affirmer que le variant est effectivement causal de la variation du trait. Par ailleurs ces positions sont associées à une variation allélique élevée dans la population ; or, les méthodes d'associations statistiques entre SNP et traits dépendent de la taille de l'effet du SNP sur le trait, et de sa fréquence allélique (Guo et al., 2016) : plus un variant est rare dans la population,

plus il sera difficile de l'associer statistiquement à une variation du trait, si son effet ne contrebalance pas sa rareté. Ainsi, la plupart des associations GWAS conduisent à l'identification de SNPs dont la contribution au trait phénotypique est faible ; cela a été par exemple montré pour l'influence des génotypes sur la taille : dans l'article Wood et al., 2014, l'identification de 697 loci ne permet d'expliquer que 16% de la variance de la taille des individus. Les analyses de génotypes sont donc adaptées à l'étude de maladies (ou autres traits) complexes, dont les variations phénotypiques sont associées à de nombreuses positions, et chacune contribuant faiblement à la totalité de l'effet mesuré. Pour tenter d'identifier les variations effectivement causales, et potentiellement compléter les mesures de tailles d'effets, il est possible d'aller explorer l'ensemble des variations génétiques d'un patient grâce aux technologies de séquençage haut-débit. Je présente ces concepts dans la section suivante.

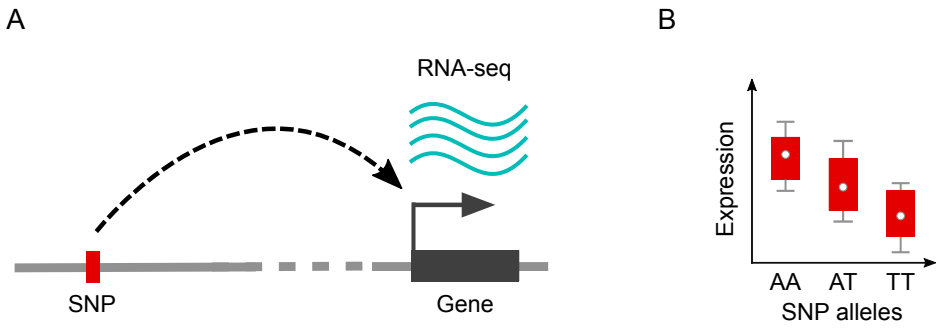


Figure 2.2 – Schéma d'association entre un SNP et un niveau d'expression de gène.
(A) Le design expérimental du projet GTEX a proposé l'évaluation du génotype de 449 individus. Les niveaux d'expressions des gènes chez ces individus ont par ailleurs été mesurés au travers de 44 tissus. Ces deux informations disponibles permettent donc d'évaluer statistiquement le lien entre la composition composition allélique d'un SNP et le niveau d'expression d'un gène aux alentours.
(B) L'exemple schématisé ici suggère que la présence d'un allèle T à l'état homozygote est associé à un niveau d'expression plus faible que la présence d'un allèle A homozygote. Cet effet peut être statistiquement évalué et mesuré, par une régression linéaire par exemple.

2.2.2 Séquençages exome- et génome-complets

Le développement des méthodes de séquençage haut-débit (Shendure et Ji, 2008) a permis de dépasser les limitations des puces de génotypage, en permettant l'identification de l'ensemble des variants d'un patient, soit à l'échelle des régions exoniques (Whole

Exome Sequencing, ou WES), soit à l'échelle du génome complet (Whole Genome Sequencing, ou WGS). Ces méthodes de séquençage haut-débit sont basées sur une fragmentation du génome d'un individu, permettant une amplification des fragments et la caractérisation de leur séquence de manière hautement parallélisée. Les séquences ainsi obtenues sont ensuite alignées sur le génome de référence, et les différences entre la séquence de référence et la séquence de l'individu sont détectées par des algorithmes d'appel de variants (Van der Auwera et al., 2013). Ces algorithmes permettent de détecter des variations de la composition nucléotidiques soit à l'échelle d'une position (variants appelés Single Nucleotide Variants ou SNV), soit à l'échelle de plusieurs positions (correspondant à des insertions ou délétions nucléotidiques, et donc désignés INDEL). Le séquençage haut-débit n'étant pas exempt d'erreurs, ces algorithmes proposent des scores de fiabilité et de qualité pour les variants détectés (Z. Wang et al., 2013).

Ces méthodes de séquençage haut-débit ont été appliquées à différentes échelles, à la fois dans un cadre diagnostique chez un patient, ou bien afin d'explorer les propriétés générales des variations génétiques chez différents individus. Le projet 1 000 Génomes The 1000 Genomes Project Consortium, 2015 a été un pivot important dans la compréhension de la diversité génétique dans les populations humaines, par l'identification des variants génomiques présents chez 2 504 individus provenant de 26 populations différentes. Les résultats obtenus dans le cadre de ce projet ont permis d'établir des ordres de grandeur concernant les variants identifiés ; par exemple, 72% de ces variants (plus de 88 millions identifiés au total) sont considérés comme rares (trouvés chez moins de 0.5% des 2 504 individus). A l'échelle d'un individu, un génome typique contient entre 4 et 5 millions de variants, dont environs 10 000 qui sont localisés dans les séquences codantes ; par ailleurs une minorité de l'ensemble des variants sont privées à celui-ci (entre 6 000 et 20 000).

Dans un cadre diagnostique, les approches de séquençage haut-débit permettent d'identifier des variants rares, potentiellement uniques au patient ou à sa famille, ce qui est particulièrement important dans le cadre d'une maladie héréditaire rare, et dont la cause génétique est peu probablement partagée à l'échelle de la population étant donnée son impact délétère. Il est même possible d'aller identifier spécifiquement les variants qui n'ont pas été transmis par ses parents, et qui sont donc apparus soit dans les cellules gamétiques, soit à la formation du zygote. L'identification de ces variants est particulièrement intéressante lorsqu'une maladie rare est détectée chez un enfant sans trace d'une hérédité chez le reste de la famille (McRae et al., 2017, Acuna-Hidalgo et al., 2016), mais ces mutations

de novo sont également exploitées dans un cadre théorique d'étude des facteurs à l'origine de l'apparition spontanées de mutations (Francioli et al., 2015, Jónsson et al., 2017). Concernant la distinction entre WES et WGS : le choix de se focaliser sur les séquences exoniques peut-être justifié par un coût moins élevé, ainsi qu'une capacité d'interprétation plus importante pour les variants localisées dans les gènes, en comparaison avec les variants non-codants (ce point sera abordé plus en détails dans la section suivante).

2.2.3 Associations fonctionnelles des variants rares

Contrairement aux méthodes de génotypage, l'accès à l'ensemble des variants d'un patient permet d'aller inférer le lien de causalité entre un variant et le phénotype observé. Plusieurs approches d'associations fonctionnelles sont possibles pour les variants rares, suivant le cadre diagnostique dans lequel on se trouve.

Tout d'abord, il est possible d'appliquer une approche similaire aux GWAS, en regroupant plusieurs patients identifiés pour avoir la même maladie génétique, pour tenter d'identifier une association statistique entre des variants à plus haute fréquence chez ces patients, en comparaison avec une population de référence. Cependant la rareté de la majorité des variants identifiés rend cette tâche difficile ; pour augmenter le pouvoir statistique, les variants sont généralement agrégés à l'échelle d'une région fonctionnelle (à l'échelle d'un gène, ou d'une région régulatrice par exemples). Cela permet éventuellement d'identifier des régions présentant une accumulation particulière de variants en comparaison avec des patients contrôles (Auer et al., 2015).

Pour éviter d'avoir à perdre en résolution par cette agrégation, il est aussi possible d'explorer directement chacun des variants chez un patient, dans le cadre d'une maladie rare où un seul patient est considéré (Gilissen et al., 2011). Plusieurs hypothèses sont alors posées : le variant à identifier est rare, avec un effet fonctionnel important, et dont la pénétrance est complète. Cela conduit à chercher parmi les variants d'un patient ceux qui sont extrêmement rares, voire absent de la population générale (Z. Wang et al., 2013). Différents filtres sont ainsi appliqués, à la fois sur la qualité des variants identifiés, ainsi que sur leurs fréquences identifiées dans d'autres projets comme le projet 1 000 génomes (Stitziel et al., 2011). Aussi, l'annotation fonctionnelle de ces variants concernant leur localisation dans un gène ou un élément régulateur connu donne la possibilité d'évaluer leur conséquence potentielle (notamment sur la séquence codante) permet de ne garder que les variants dont l'impact prédit est le plus important (je reviendrai sur les méthodes de

prédiction à la section suivante et au chapitre suivant). L'évaluation de l'impact de variants sur la séquence codante des gènes peut-être focalisée sur des gènes déjà précédemment identifiés et associés au phénotype, mais peut également conduire à une implication d'un nouveau gène pour la maladie considérée si la confiance dans le variant candidat est élevée (Neveling et al., 2013).

Enfin, la pertinence d'un variant candidat peut être confirmée par l'évaluation de sa co-ségrégation avec le phénotype au sein de la famille ; l'identification d'un variant candidat chez d'autres patients atteints de la même maladie vient renforcer le rôle causal d'un variant. Des validations expérimentales permettent aussi de confirmer l'effet prédit d'un variant candidat ; une revue de ces validations est proposée dans l'article de Rodenburg, 2018. Par exemple, dans le cadre de l'évaluation de l'impact d'un variant sur un gène, si la conséquence biochimique est mesurable, alors il est possible d'évaluer à partir d'une culture cellulaire si l'introduction d'un vecteur d'expression contenant le gène non-muté permet un sauvetage du phénotype. Un autre exemple se base sur l'utilisation de morpholino pour bloquer l'expression d'un gène d'intérêt lors du développement du poisson zèbre, afin d'évaluer si les conséquences phénotypiques sont comparables à celles observées chez le patient.

Ainsi, l'analyse de l'ensemble des variants identifiables chez un patient ouvre la possibilité d'un diagnostic génétique, par l'évaluation de leur association fonctionnelle à un gène. Ce diagnostic permet de compléter les connaissances sur les causes génétiques des maladies, et ouvre la voie vers un potentiel traitement ; cette possibilité est illustrée par l'article de Worthey et al., 2011, premier exemple de diagnostic par WES suivi d'un traitement adapté, pour un enfant atteint d'une forme de la maladie de Crohn.

2.3 Le challenge des variants non-codants

Du génotype aux variants rares

Depuis le projet Génome Humain et la production du premier génome de référence, les progrès techniques ont été nombreux, tant pour les puces de génotypage que pour les technologies de séquençage haut débit. La généralisation de l'utilisation de ces technologies a permis de voir émerger des projets dont les nombres de patients sont de plus en plus élevés :

— des analyses de GWAS faites sur des centaines de milliers de patients (par exemples :

GWAS du niveau de scolarité chez plus d'un million d'individus, Lee et al., 2018 ; GWAS des facteurs de risque du diabète de type 2 pour 62 892 patients, Xue et al., 2018) ;

— des analyses de WES pour des dizaines de milliers de patients ; par exemple dans le cadre du projet UK biobank : 49 960 patients séquencés (Hout et al., 2019) ; ou encore le projet ExAC regroupant de données de WES pour 60 706 patients (Lek et al., 2016) ;

— et l'utilisation plus récentes du WGS pour des ordres de grandeurs similaires, que ce soit dans un cadre diagnostique (13 037 individus, dont 9 802 atteints de maladies héréditaires rares Ouwehand, 2019) ou dans le cadre de l'identification de propriétés générales des variations nucléotidiques (2 636 individus de la population islandaise, Gudbjartsson et al., 2015 ; 11 257 génomes pour étudier la diversité en séquence des variants non-codants, Iulio et al., 2018).

Ces méthodes présentent donc un potentiel majeur pour l'élucidation des causes génétiques des maladies. Cependant leur application et les résultats obtenus mettent en lumière l'importance d'une investigation plus poussée des propriétés du génome non-codant ; concernant les analyses de GWAS, il a été montré que la majorité des SNPs candidats identifiés par GWAS sont localisés dans le génome non-codants (Edwards et al., 2013). Le problème de l'interprétation de ces variants non-codants est également évident pour l'analyse des variants obtenus par séquençage complet de génome, étant donné la taille restreinte du génome codant, et le faible nombre de variants qui y sont localisés. L'évaluation des variants non-codants est rendue particulièrement difficile par le nombre réduit d'éléments régulateurs formellement connus, et par l'absence d'une définition claire des propriétés définissant ces éléments régulateurs (voir chapitre 1). S'il est possible de caractériser si un variant est localisé dans une région de chromatine ouverte, dans une région présentant des marques d'histones activatrices, ou dans un pic de ChIP-seq pour un facteur de transcription, aucune de ces informations n'est suffisante pour affirmer que le variant présente un impact fonctionnel sur la régulation d'un gène.

Il est donc nécessaire de faire appel à des méthodes d'intégration de données, pour identifier des propriétés générales associées aux variants non-codants. Des méthodes ont également été développées pour les variants codants, mais elles peuvent exploiter l'information du code génétique pour inférer les conséquences d'un variant sur la séquence codante ; cela n'est pas possible pour les variants non-codants. Ces méthodes présentent

plusieurs intérêts majeurs : elles évitent l'emploi de filtres ad hoc, et permettent d'évaluer le caractère fonctionnel de millions de variants très rapidement pour identifier les meilleurs candidats. Ces méthodes d'intégrations seront présentées au chapitre suivant.

2.3.1 Exemple d'application : étude des causes génétiques de l'autisme

Dans le cadre de mes travaux de thèse, je me suis appuyé sur des résultats récents d'analyses de mutations *de novo* obtenues chez des patients atteints d'autisme (An et al., 2018) ; j'ai donc souhaité présenter quelques points concernant cette maladie ici.

L'autisme, ou maladies du spectre autistiques (Autism Spectrum Disorders, ou ASD), est une maladie dont les causes génétiques sont progressivement décodées par l'utilisation des différentes technologies d'identification de variants fonctionnels.

Les études de cette maladie proposent depuis un certain temps l'idée que les variations *de novo* sont à l'origine du phénotype. Des travaux sur les variations en nombres de copies identifiaient déjà une sur-représentations de ces variations chez les patients affecté en comparaison à des patients contrôles (Sebat et al., 2007, Pinto et al., 2010).

Plus récemment, avec l'essort des technologies de séquençage, différents projets ont abordé le rôle des mutations *de novo* détectées par séquençage d'exome (Iossifov et al., 2012,O'Roak et al., 2011, Sanders et al., 2012), ainsi que par séquençage de génome complet (Michaelson et al., 2012 , Yuen et al., 2016, Turner, Coe et al., 2017, An et al., 2018).

Les travaux sur les données de séquençage d'exomes ont permis de mettre en lumière un enrichissement significatif en variants "perte-de-fonctions" chez les patients atteints du spectre autistiques, en comparaison avec des patients contrôles. Ces analyses ont permis l'élaboration de premiers modèles pour expliquer les causes de cette maladie (Ronemus et al., 2014). Des travaux récents sur des populations spécifiques permettent d'enrichir ces modèles, notamment par l'identification de nouveaux gènes candidats (Leblond et al., 2019). Cependant, malgré la contribution attendue des variants non-codants à cette maladie, l'identification de variants candidats dans les régions régulatrices reste un défi majeur (An et al., 2018).

Chapitre 3

Génome non-codant et intégration de données

3.1 Méthodes de prédiction de fonctionnalité des variants

Nous avons vu au chapitre précédent que l'application des technologies de séquençage haut-débit à un cadre diagnostic permet d'accéder aux millions de variants présents dans le génome d'un patient, pour identifier les causes génétiques de sa maladie. J'ai évoqué différentes méthodes d'associations fonctionnelles permettant d'identifier des variants candidats. L'accent était donné à l'identification de variants fonctionnels dans le génome codant : il est en effet plus simple d'inférer pour un variant codant un impact sur la séquence protéique, en comparaison avec un variant non-codants. Cependant l'augmentation du nombre de patients par projet, et l'application de plus en plus systématique du séquençage génome-complet impose l'utilisation de méthodes automatiques d'annotation et de caractérisation des variants. Je propose dans cette section un aperçu des méthodes existant pour évaluer l'impact de variants codants et non-codants. Ces méthodes sont basées en majorité sur des modèles d'apprentissage machine, dont le but est d'intégrer différents signaux pour fournir un score de prédiction ; ces méthodes peuvent être supervisées (dans le cas où on leur fournit des exemples de variants à identifier comme fonctionnels, et d'autres à identifier comme non-fonctionnels), ou non-supervisées. Dans le cadre de mon travail de thèse, j'ai utilisé un de ces algorithmes d'apprentissage supervisé : l'algorithme de forêt aléatoire ; il sera présenté à la section suivante.

3.1.1 Méthodes pour les variants codants

L'évaluation des variants identifiés comme codants se fait sur la base de leur localisation par rapport à la séquence des transcrits potentiels associés au gène dans lequel ces variants sont localisés. Cette localisation permet d'inférer plus précisément si le variant peut avoir un impact sur la formation d'un transcrit mature, par exemple par sa localisation dans un site d'epissage, ou s'il peut impacter la séquence protéique, par sa localisation dans le segment traduit de l'ARN mature.

Pour ce second cas, la connaissance du cadre de lecture permet d'identifier la localisation du variant par rapport aux codons, et donc de prédire la conséquence sur la séquence protéique. Un variant conduisant à une délétion ou une insertion peut potentiellement décaler le cadre de lecture ; dans ce cas il est possible d'évaluer si la protéine associée voit sa séquence modifiée, ou tronquée. Dans le cadre d'un variant ponctuel (SNV), le variant peut être sans impact sur la séquence protéique (variant silencieux, dû à la dégénérescence du code), ou bien il peut conduire à un changement de l'acide aminé encodé, ou à l'introduction d'un codon stop prématuré.

Mais il est possible d'aller plus loin pour évaluer la conséquence fonctionnelle d'un variant. En effet, plusieurs méthodes d'apprentissage machine, entraînées à reconnaître des variants identifiés comme fonctionnels (sélectionnés depuis la littérature) ont été proposées pour exploitent plus efficacement les propriétés associées aux séquences codantes. Deux méthodes historiques sont classiquement utilisées : SIFT (Kumar et al., 2009) se base sur l'exploitation d'alignements multiples de séquences homogues à un gène, pour évaluer la contrainte imposée sur la conservation de l'identité des acides aminés ; Polyphen (Adzhubei et al., 2010 propose un score prédit par une méthode de classification entraînée sur des propriétés notamment de structures des protéines. Plus récemment des méthodes comme CADD (Kircher et al., 2014) ou REVEL (Ioannidis, Rothstein et al., 2016) ont intégré des annotations comme les marques d'histones, ou les signaux de conservations.

Ces méthodes bénéficient donc d'un ensemble de mesures associées à la séquence codante, ainsi que d'un grand nombre de variants codants identifiés comme fonctionnels dans la littérature, pour prédire efficacement le potentiel fonctionnel de n'importe quel variant localisé dans la séquence d'un gène. Le problème de l'inférence de l'impact des variants codants n'est cependant pas complètement résolu, et il y a encore du potentiel à développer de nouvelles méthodes de prédiction pour les variants codants. En effet, toutes les méthodes déjà disponibles ne sont pas parfaitement corrélées ; leur complémentarité sem-

blerait même être bénéfique pour avoir une vision plus complète de l'impact fonctionnel des variants codants (Riera et al., 2016, Campa et al., 2017). Par ailleurs les projets de séquençage à large échelle récents ont permis d'évaluer plus concrètement les contraintes imposées sur la séquence des gènes au sein de la population, ce qui a conduit à l'identification de gènes présentant des séquences particulièrement peu mutées (Lek et al., 2016) ; ces résultats ont notamment été intégré dans certaines méthodes de prédiction pour raffiner leurs scores d'impacts en fonction des gènes considérés (Velde et al., 2017).

3.1.2 Méthodes pour les variants non-codants

Contrairement aux méthodes dédiées aux variants codants, les méthodes de prédiction de fonctionnalité pour les variants non-codants n'ont commencé à être proposées que récemment. Ces méthodes souffrent en effet de trois difficultés :

— il existe très peu d'exemples de variants fonctionnels non-codants dont on a clairement identifié le rôle régulateur ;
— les localisations des éléments régulateurs ne sont pas aussi bien établies que les localisations des gènes ;
— les signaux de fonctionnalités associés aux éléments régulateurs et leurs combinaisons ne sont pas exactement établis.

Néanmoins plusieurs approches ont été proposées pour définir des scores, résumant les signaux de fonctionnalités associés au génome non-codants (présentés au chapitre 1). L'article de Rojano et al., 2018 propose une revue des méthodes publiées ces dernières années. Je reprends ci-dessous quelques-unes de ces méthodes, qui ont été utilisées dans le cadre de mon travail de thèse ; j'y ai ajouté une brève description de la méthode NCBoost, ainsi que de la méthode FIRE.

CADD. Le score CADD (Combined Annotation Dependent Depletion, Kircher et al., 2014, Rentzsch et al., 2019) propose une intégration de multiples informations de signaux fonctionnels (conservation en séquence, signaux de modifications d'histone, prédictions de changement d'acide aminé, ilôts CpG, etc.), au sein d'un modèle de régression logistique pénalisée entraîné à séparer des variants simulés dans le génome (considérés comme potentiellement fonctionnels) de variants contrôles (pris à partir de variants fixés dans la population humaine, différents de leur état ancestral primate).

FATHMM-MKL. Le score FATHMM-MKL (Shihab et al., 2015) est basé sur une méthode d'apprentissage supervisé appelée Multi-Kernel Learning, qui permet une séparation de variants fonctionnels et non-fonctionnels selon différents groupes d'annotations (annotations de conservation en séquence, de sites de fixation de facteurs de transcriptions, de marques d'histones, et d'accessibilité de la chromatine), utilisé au sein d'une méthode de SVM (Support Vector Machine). Cette approche permet d'établir des séparations linéaires selon chacun des groupes d'annotations, qui sont ensuite pondérées pour la séparation finale du modèle. Le modèle est ici entraîné sur des variants non-codants de la base de données HGMD (Stenson et al., 2017) pour les exemples fonctionnels, tandis que les variants contrôles proviennent du projet 1 000 génomes.

ReMM. Le score de prédiction ReMM (Regulatory Mendelian Mutation) correspond au score de fonctionnalité utilisé dans l'outil Genomiser (Smedley et al., 2016). Ce score est basé sur un entraînement supervisé d'un modèle de forêts aléatoires, entraîné sur 453 variants non-codants fonctionnels identifiés depuis la littérature (qui, après vérification, sont inclus dans la base de données HGMD-DM), et de variants contrôles pris de la même manière que pour CADD. On y retrouve les annotations de conservation en séquence, de marques d'histones, de sites de fixation de facteurs de transcriptions, ainsi que de prédictions FANTOM pour les régions régulatrices.

NCBoost. Le score NCBoost (Caron et al., 2018) est basé sur une méthode d'apprentissage supervisée appelée XGBoost (Chen et al., 2016), dont le principe se rapproche de celui des forêts aléatoires, mais incluant une étape de pondération des entités pendant l'apprentissage qui lui permet forcer une meilleure classification des variants ambiguës. Le score est entraîné à distinguer 737 variants non-codants fonctionnels identifiés depuis la littérature (provenant en partie des variants ReMM et HGMD), de variants contrôles ; à nouveau un ensemble d'annotations comme les propriétés de contrainte de la séquence, ou les marques d'histones, sont utilisés pour décrire les variants.

Eigen. Le score Eigen (Ionita-Laza et al., 2016) proviennent d'une méthode d'apprentissage non-supervisé, appelé apprentissage spectral. Cette méthode de prédiction apprend à séparer deux classes par l'identification de différences entre des combinaisons d'annotations. Les annotations considérées ici correspondent aux scores de conservations, aux

fréquences des variants dans la population (provenant du projet 1 000 génomes), et d'annotations de signaux fonctionnels provenant du projet ENCODE.

FitCons. Le score FitCons (Gulko et al., 2015) correspond à une évaluation de la pression de sélection négative à laquelle sont soumises différents types de régions du génome. Les régions du génomes sont d'abord regroupées selon leur profil de fonctionnalité, établi à partir de données telles que les états chromatiniens, les localisations de séquences codantes, ou encore les régions de chromatines ouvertes. Pour 624 profils uniques, les auteurs ont évalué la fraction de positions sous pression de sélection par la méthode INSIGHT (Gronau et al., 2013), qui leur permet ensuite de propager ce score de contrainte à l'ensemble des régions du génome.

LINSIGHT. Le score LINSIGHT (Huang et al., 2017) est une généralisation de la méthode de prédiction FitCons, permettant une résolution par base de la prédiction d'un score de contrainte de sélection négative. Dans cette méthode, un modèle linéaire est utilisé pour intégrer les données de fonctionnalité (précédemment utilisées pour définir les profils), afin de prédire le score de contrainte mesuré par INSIGHT.

FIRE. La méthode FIRE (Functional Inference of Regulators of Expression, Ioannidis, Davis et al., 2017) se distingue des autres méthodes par les variants considérés : ici la méthode vise à distinguer des variants eQTLs de variants contrôles. Pour cette méthode, un modèle de forêt aléatoire est donc entraîné à identifier des variants potentiellement eQTLs, en se basant sur des annotations similaires à celles utilisées pour les méthodes décrites ci-dessus.

Comparaison et limites des méthodes de prédictions pour les variants non-codants

Un article récent (Liu et al., 2017) a proposé une étude comparative des performances de certaines de ces méthodes, par une évaluation de leur capacité à distinguer des variants fonctionnels issus de la base de données HGMD (Stenson et al., 2017), de variants contrôles identifiés comme variants privés depuis le projet UK10K ; la meilleure méthode rapportée est FATHMM-MKL.

Trois groupes peuvent être distingués depuis les méthodes décrites ci-dessus :

— les méthodes d'apprentissage supervisé dédiées à la distinction de variants fonctionnels potentiellement délétères : CADD, FATHMM-MKL, NCBoost, ReMM ;
— les méthodes sans apprentissage ou avec apprentissage non-supervisé : Eigen, ainsi que LINSIGHT et FitCons (ces deux dernières ayant également pour but l'identification de variants sous pression de sélection négative) ;
— la méthode FIRE, qui est une approche supervisée, mais dédiée à l'identification de variants eQTLs.

Parmi les méthodes d'apprentissage supervisé utilisées, le principe des forêts aléatoires est appliqué pour 3 de ces méthodes (ReMM et Fire, ainsi que NCBoost, qui est basé sur un principe similaire). Les méthodes de forêts aléatoires présentent en effet beaucoup d'avantages pour l'apprentissage depuis des annotations hétérogènes, en intégrant de manière non-linéaire les valeurs des différentes distributions ; les avantages de cette méthode ont motivé mon choix pour mes travaux de thèse, et je présente donc l'algorithme en détails dans la section suivante.

Ces scores de prédictions permettent donc d'avoir, pour une position donnée dans le génome, une information intégrée sur les faisceaux d'indices fournis par les données expérimentales concernant les propriétés régulatrices du génome. Un problème important n'est cependant pas résolu par ces méthodes : l'association à un gène cible. En effet, les régions régulatrices du génome (potentiellement identifiables par ces scores) exercent leur action régulatrice sur des gènes, qui peuvent être situés à de longues distances de la région régulatrices. Comme noté dans la revue de Rojano et al., 2018, l'exploitation de données de capture de la chromatine peut permettre de résoudre ce manque. D'autres méthodes d'associations ont été présentées au chapitre 1 (co-expression de régions régulatrices, co-associations d'états chromatiniens, conservation de synténie) ; ce sont également des sources d'annotations qu'il est possible de prendre en compte pour évaluer l'impact d'un variant potentiellement régulateur sur un gène cible

Enfin, un problème de capacité d'interprétation des prédictions faites par ces modèles se pose. Contrairement aux variants codants, pour lesquels une prédiction de fonctionnalité peut toujours être évaluée au regard de la séquence codante, le manque de connaissances sur les variants non-codants limite l'évaluation et l'interprétation d'un score. Je discute ces points dans le prochain chapitre, dédié à la problématique de mon travail de thèse.

3.2 L'algorithme des forêts aléatoires

L'algorithme des forêts aléatoires correspond à un algorithme d'apprentissage supervisé, au sein duquel des entités composées de valeurs descriptives et d'une valeur cible vont être utilisées pour entraîner un ensemble d'arbres de décisions, qui serviront ensuite à prédire la valeur cible pour n'importe quelle nouvelle entité. Cet algorithme a été décrit pour la première fois en 2001 (Cutler et al., 2001).

Suivant le type de valeur cible, l'entraînement peut correspondre à une tâche de régression (où les arbres de décision cherchent à prédire une valeur continue depuis les valeurs descriptives), ou à une tâche de classification (où les arbres de décision cherchent à discriminer les entités des différentes classes). Dans le cadre de l'identification de variants fonctionnels non-codants, leur identification peut être représentée par une tâche de classification à deux classes : variants fonctionnels et non-fonctionnels. Par simplification je propose donc une explication de l'algorithme pour une tâche de classification à deux classes.

3.2.1 Arbre de décision

L'unité de base de la forêt aléatoire est donc l'arbre de décision. Celui-ci est entraîné à séparer du mieux possible les deux classes d'entités, en appliquant une série de seuils sur les descripteurs de ces entités, pour séparer successivement les entités des deux classes. Un arbre est donc composé par ces successions de seuils, représentées sous forme de noeuds.

Un noeud dans l'arbre correspond à l'unité de décision où un certain nombre d'entités du jeu d'entraînement vont se retrouver décrite par un seul descripteur, choisi de manière optimisée, et pour lequel un seuil est appliqué, afin de définir deux noeuds fils contenant les entités dont les valeurs du descripteur sont inférieures ou supérieures au seuil respectivement. Le premier noeud de l'arbre correspond à la racine (où l'ensemble des entités du jeu d'entraînement se retrouvent), tandis que le dernier noeud d'une succession de seuils est appelé noeud terminal, ou feuille. Dans cette feuille, une majorité a théoriquement été atteinte pour l'une des classes ; c'est cette classe majoritaire qui détermine la classe prédite dans cette feuille, ce qui signifie que toute nouvelle entité qui suivra la succession de noeuds conduisant à cette feuille se verra assignée cette classe majoritaire comme prédiction.

Un arbre sera donc composé d'un ensemble de chemins de décisions, chacun correspondant à une série de seuils successifs appliqués sur les différents descripteurs, et conduisant à

une séparation progressive des entités dans leurs classes respectives, jusqu'à aboutir à une feuille, où la classe majoritaire déterminée au moment de l'entraînement permet la prédiction. Il est à noter que différents critères d'arrêts sont possibles pour moduler la forme d'un arbre : profondeur maximale des chemins de décision, nombre minimum d'entité par noeud, etc.

3.2.2 Critère de choix des seuils

Les choix des descripteurs et des seuils associés à chacun des noeuds est fait de manière à séparer du mieux possible les entités en leurs deux classes. Cette optimisation est donc basée sur une mesure de la diversité en classe à chacun des noeud, et vise à maximiser sa diminution lorsqu'un seuil est appliqué à un noeud parent pour créer les noeuds fils.

La mesure de la diversité en classe correspond à l'indice de diversité de Gini, qui permet à partir des proportions de chaque classe (notées p_i) dans un noeud de quantifier à quel point ce noeud est hétérogène (équation 3.1).

$$I_G(noeud) = \sum_{i=1}^{m} p_i(1 - p_i) \tag{3.1}$$

L'algorithme définit donc le choix d'un seuil et d'un descripteur à un noeud par l'exploration heuristique d'un ensemble de seuils possibles à appliquer pour chacun des descripteurs, avec l'objectif de maximiser la diminution de l'indice de diversité de Gini (équation 3.2)

$$\Delta I_G(noeud) = I_G(noeud) - I_G(noeud_{fils1}) - I_G(noeud_{fils2}) \tag{3.2}$$

Il est important de noter que cet algorithme est particulièrement intéressant pour exploiter des descripteurs sans avoir besoin de les transformer : les distributions sont considérées indépendamment les unes des autres, et ne nécessitent donc pas d'être normalisées avant utilisation de cet algorithme. En revanche, il existe un biais de sélection des descripteurs : l'algorithme aura plus de facilité à exploiter des descripteurs présentant beaucoup de valeurs, en comparaison avec un descripteur catégoriel avec un nombre limité de catégories (avec le cas extrême des descripteurs binaires) ; ce biais est d'autant plus vrai que les annotations de catégories sont creuses.

3.2.3 Construire la forêt aléatoire depuis les arbres

Un arbre de décision représente donc un ensemble de chemins de décisions, générés de manière optimisée pour un jeu d'entités ayant servi à l'entraînement. Cela conduit généralement à un modèle de décision dont la généralisation est limitée : il a été construit de manière optimisée pour le jeu d'entraînement (en particulier si aucune limite n'est imposée sur sa structure), ce qui fait que son application à des entités différentes est limitée.

Pour éviter ce sur-apprentissage, il est possible de combiner plusieurs arbres de décisions au sein d'une forêt ; les variations dans les chemins appris par chacun de ces arbres vont permettre de renforcer la généralisations des règles identifiées. Pour augmenter un peu plus la variation au sein des arbres, une étape d'agrégation "bootstrap" des entités est introduite en amont de la génération de chacun des arbres. Cette étape consiste à échantillonner avec remise parmi les entités du jeu d'entraînement le même nombre d'entités ; ainsi chaque arbre va apprendre des règles sur un sous-jeu d'entités (où certaines se retrouve en plusieurs exemplaires) de l'ensemble initial, ce qui augmente un peu plus les variations de choix de seuils et de descripteurs dans les noeuds.

L'algorithme des forêts aléatoires est un algorithme très puissant et versatile, capable d'exploiter des descripteurs hétérogènes, en établissant des combinaisons non-linéaires de leurs valeurs pour aboutir à une prédiction de classe. On peut noter que la prédiction finale de la forêt aléatoire correspond à un vote majoritaire depuis les décisions émises par chacun des arbres. Il est donc possible de calculer un pourcentage d'arbre votant pour une classe donnée, ce qui peut servir comme une mesure de confiance sur la décision émise par la forêt.

Chapitre 4

Problématique

La régulation de l'expression des gènes est un processus biologique important pour définir clairement les schémas d'expression spatio-temporels précis de ces gènes, que ce soit pour le développement correct de l'embryon, ou pour le maintien des fonctions cellulaires tissus-spécifiques. Cette régulation est guidée par l'action de régions génomiques portant des fonctions activatrices ou inhibitrice de l'expression des gènes. Bien que nous ne connaissons pas clairement les localisations de ces régions et leurs relations d'associations avec les gènes, plusieurs signaux fonctionnels sont disponibles pour inférer leur positions, et prédire des associations fonctionnelles.

La caractérisation de ces régions régulatrices est d'une importance majeure pour l'évaluation et l'interprétation des variations nucléotidiques se produisant hors des séquences codantes des gènes. En effet, le développement et la démocratisation de l'utilisation des méthodes d'identification des variants à l'échelle du génome chez des patients conduit à une explosion du nombre de variants identifiés dans les régions non-codantes du génome. Or, les résultats d'expérience GWAS et les limitations diagnostiques des approches focalisées sur l'exome nous conduisent à nous interroger sur les méthodes à employer pour caractériser et hiérarchiser ces variants non-codants.

Plusieurs méthodes ont été proposées ces dernières années pour répondre au problème de l'identification de variants non-codants potentiellement fonctionnels. Ces méthodes exploitent notamment les propriétés identifiables du génome non-codant (présentées au chapitre 1), pour proposer des scores de fonctionnalité, résumant à l'échelle d'une position l'ensemble des signaux fonctionnels disponibles pour décrire cette position.

Cependant, ces méthodes dédiées au génome non-codant présentent des limites no-

tables :

— les scores de fonctionnalité des variants ne permettent pas de les associer à des régions régulatrices ;

— contrairement aux variants codants, la capacité d'interprétation des scores de fonctionnalité est bien plus limitée.

Par ce projet de thèse, j'ai souhaité proposer une approche automatisée de hiérarchisation des variants non-codants identifiés dans un cadre diagnostique, pour permettre d'identifier des variants candidats pertinents. La méthode que j'ai développée se distingue sur trois points par rapport aux méthodes existantes :

— plusieurs jeux de variants fonctionnels ont été considérés, pour explorer un espace du génome régulateur le plus divers possible ;

— des jeux de données de prédictions de régions régulatrices avec leur gènes cibles ont été intégrés à ma méthode, pour permettre de hiérarchiser des variants associés à des gènes d'intérêt ;

— une approche d'évaluation et de compréhension des décisions du modèle de prédiction a été mise en place, pour permettre d'avoir un regard plus informé sur les scores de prédiction associés aux variants non-codants évalués.

Cette méthode, appelée FINSURF (Functional Interpretation of Non-coding Sequences Using Random Forests) est décrite dans la partie de ce manuscrit dédiées aux résultats. Ces résultats sont organisés en quatre parties. Tout d'abord, je propose quelques analyses préliminaires concernant les jeux de données utilisés dans le cadre de ce projet, avec notamment une comparaison des différents ensembles de variants fonctionnels identifiés et choisis pour entraîner mes modèles de prédiction. Le second chapitre des résultats est dédié à la description et l'analyse des différentes étapes de l'entraînement de ces modèles. Dans le chapitre suivant, deux approches sont proposées pour caractériser les modèles, et augmenter la capacité d'interprétation sur leur fonctionnement. Enfin, le dernier chapitre propose un exemple d'application de l'un des modèles de prédictions, pour illustrer l'intérêt de la méthode FINSURF.

Deuxième partie

Matériel et Méthodes

Chapitre 5

Origine des données

Sauf mention spéciale, toutes les données et analyses ont été faites sur la version hg19 / GRCh37 du génome humain (Pruitt et al., 2007).

Lorsque ce n'est pas le cas, comme par exemple pour l'utilisation de données générées pour la version GRCh38 du génome (Church et al., 2015), une précision est apportée sur la conversion des données, faisant appel à un/des outil(s) présenté(s) dans le chapitre suivant.

5.1 Annotations génomiques

5.1.1 Scores de conservation

Les scores de conservation permettent d'évaluer le degré de contrainte de la composition nucléotidique d'une position, soit à partir d'un alignement multiple de séquences obtenues chez différentes espèces, soit par évaluation du degré de polymorphisme au sein de la population humaine.

PhastCons et PhyloP. Deux premiers jeux d'annotations ont été identifiés, issus de deux méthodes différentes : PhastCons et PhyloP. Ces scores mesurent le degré de contrainte sur la composition nucléotidique à une position génomique au cours de l'évolution. PhastCons correspond à la probabilité a posteriori pour une position d'appartenir à un élément conservé ; à 0, la position est considérée comme non-conservée (c'est à dire que la composition nucléotidique n'est pas contrainte au cours de l'évolution), et à 1 la position est considérée comme conservée. PhyloP mesure la probabilité pour une position

de suivre un modèle neutre d'évolution de la séquence ; les valeurs négatives indiquent les positions où la composition nucléotidique évolue plus rapidement qu'attendu, et au contraire les valeurs positives indiquent une contrainte sur cette composition.

Les scores de conservation pour la version hg19 du génome humain ont été téléchargés à partir du site FTP de l'UCSC (Kent, Sugnet et al., 2002, Karolchik et al., 2004, Raney et al., 2014). Un premier groupe de scores a été téléchargé en décembre 2015 ; ce sont des scores calculés depuis les alignements multiples de séquence des génomes de 100 espèces de vertébrés :

— PhyloP100w

— PhastCons100w

Des scores de conservation plus récents, ont été téléchargés en février 2019, pour la version hg38 du génome humain :

— PhyloP20w

— PhastCons20w

Ces scores ont été générés à partir d'alignements multiples de 20 génomes de primates ; ils ont donc été inclus pour une meilleure évaluation potentielle des contraintes sélectives spécifiques à la lignée primate. Afin de les utiliser avec la version hg19 du génome, j'ai appliqué plusieurs commandes de conversion :

— une conversion des fichiers BigWig au format BED ;

— une transformation des positions génomiques de la version hg38 à la version hg19 du génome ;

— enfin, les régions sont réordonnées, fusionnées si chevauchantes, et reconverties vers un format BigWig, avec un score par base.

GERP. Ce score a été calculé à partir d'alignements multiples effectués sur 33 séquences de mammifères. Il évalue le taux de conservation par mesure du déficit en substitution à chaque position ; plus le taux est positif, moins il y a de substitutions identifiées dans l'alignement. Le fichier de scores (au format BigWig) a été téléchargé depuis la page du projet. Par ailleurs, un fichier BED d'identification de régions génomiques conservées (appelées éléments GERP) a également été téléchargé.

CDTS. Le projet gnomAD (Karczewski et al., 2019) a conduit à la collection de données de séquençage de génomes complets pour plus de quinze mille patients, ce qui a permis d'évaluer la répartition le long du génome des variations nucléotidiques entre

individus. Le score CDTS ("context-dependent tolerance score") a été calculé à partir de 15 491 génomes, et correspond à la différence entre la proportion observées de mutations et la proportion attendue, évaluée à l'échelle d'une région de 10 paires de bases. Plus le score est faible, plus la région est intolérante aux variations. Le fichier a été téléchargé au format BED le 19/12/2018 depuis le site web du projet (`http://www.hli-opendata.com/noncoding/`), et converti au format BigWig.

5.1.2 Données experimentales de fonctionalité

Roadmap Epigenomics : états chromatiniens

Le projet Roadmap Epigenomics (Kundaje et al., 2015) a conduit à la localisation génomique de plusieurs modifications d'histone à travers 111 types cellulaires. Les auteurs ont réalisé une intégration de ces signaux biochimiques, et ont défini des états chromatiniens, découverts par application de chaînes de Markov cachées (Ernst et al., 2010, Ernst et al., 2012). Depuis le site internet dédié au projet, j'ai téléchargé les jeux de données correspondant à la découverte de 18 états chromatiniens différents, à travers 98 types cellulaires. Ces états ont été identifiés pour des régions de 200 paires de bases, à partir des signaux biochimiques suivants :

— signaux de ChIP-seq pour la modification d'histone H3K4me3 ;
— signaux de ChIP-seq pour la modification d'histone H3K4me1 ;
— signaux de ChIP-seq pour la modification d'histone H3K36me3 ;
— signaux de ChIP-seq pour la modification d'histone H3K27me3 ;
— signaux de ChIP-seq pour la modification d'histone H3K9me3 ;
— signaux de ChIP-seq pour la modification d'histone H3K27ac.

Comme présenté sur la figure 1.3 (voir chapitre 1), les états sont associés à des combinaisons particulières et distinctes de modifications d'histones. Par exemple les états transcrits ("5_Tx" et "6_TxWk") sont associés à des probabilités élevées d'y trouver un signal H3K36me3, tandis que les états enhancers ("7_EnhA1" et "8_EnhA2") sont associés à des probabilités élevées d'y trouver des signaux H3K27ac et H3K4me1.

J'avais initialement extrait pour chacun des 98 types cellulaires les régions pour lesquels l'état chromatinien correspond à un état régulateur (par exemples : "flankingTSS", "BivalentEnhancer", ou encore "Active Enhancer"). Dans le cadre de l'annotation de variants, ces régions étaient utilisées pour définir si un variant était situé ou non dans un état régulateur pour un type cellulaire donné ; cela conduisait donc à 98 annotations binaires

et creuses (pour chaque tissu, une minorité de variants étaient localisés dans une région régulatrice). La combinaison des caractères "creux" et "binaire" conduisait à une sous-exploitation de ces annotations par les modèles d'apprentissages utilisés dans ce projet ; j'ai donc choisi d'exploiter autrement ces annotations, en calculant pour chaque région du génome le nombre de types cellulaires où un état donné est identifié, et ce pour les différents états.

Par simplification, les 18 états ont été réduits à 7 états principaux :

— promoteur : regroupe les états "Active TSS", "Flanking TSS", "Flanking TSS Upstream", "Flanking TSS Downstream", et "Bivalent/Poised TSS" ;
— transcrit : regroupe les états "Strong transcription" et "Weak transcription" ;
— enhancer : regroupe les états "Genic enhancer" 1 et 2, "Active enhancer" 1 et 2, "Weak enhancer" et "Bivalent enhancer" ;
— ZNFRpts : correspond à l'état "ZNF genes and repeats", décrivant des régions génomiques avec des propriétés d'hétérochromatines (marque H3K9me3), mais aussi marqué par des signaux associés à la transcription (H3K36me3), provenant d'une concentration élevés en gènes à doigt de zinc ("ZNF genes") ;
— hétérochromatine : correspond à l'état "Heterochromatin" ;
— réprimé : regroupe les états "Repressed PolyComb" et "Weak Repressed PolyComb" ;
— quiescent : correspond à l'état "Quiescent/Low".

Les états chromatiniens des différents types cellulaires ont donc été agglomérés dans un seul fichier, et une étape de re-segmentation des régions génomiques a été appliquée, pour définir des régions non-chevauchantes. Chaque région a été réannotée avec les comptages de chacun des états principaux, évalués sur les 98 types cellulaires ; chaque région est ainsi associée à 7 valeurs discrètes. Ces opérations ont été réalisée avec les outils Bedops et Bedtools.

Roadmap Epigenomics : marques d'histones

En plus des états chromatiniens, j'ai téléchargé les signaux de Fold Change obtenus sur les 98 types cellulaires du projet Roadmap Epigenomics, pour trois modifications d'histones d'intérêt, identifiant les régions régulatrices :

— H3K4me1 : plutôt associées aux enhancers ;
— H3K4me3 : plutôt associées aux séquences des promoteurs ;

— H3K27ac : plutôt associées aux enhancers actifs.

Les fichiers ont été téléchargés le 25/11/2018. Pour chaque modification, et afin de résumer les informations provenant des 98 types cellulaires, les valeurs médianes à chaque position du génome ont été extraites avec l'outil WiggleTool. Les fichiers BED obtenus ont ensuite été filtrés (par exemples : calcul des scores moyens pour les régions chevauchantes, retrait de régions aberrantes), et convertis en fichiers BigWig.

Données de sites de fixation de facteurs de transcription

Trois jeux de données ont été sélectionnés pour explorer l'impact des variants sur des sites potentiels de fixation de facteur de transcription ; ils sont présentés ci-dessous. Ces fichiers ont été téléchargés le 26/11/2018. Dans le cadre de l'utilisation de ces jeux de données pour annoter des positions génomiques, deux informations peuvent être extraites de ces identifications de sites de fixation : le nombre de sites chevauchant une position donnée, et le score maximal parmi les scores d'identification des sites ; ce sont ces deux informations que j'utiliserai pour annoter les variants non-codants pour mes modèles de classification.

Sites de fixation dans des pics de ChIP-seq. Un ensemble de jeux de données dédiés aux annotations de régions régulatrices a été compilé dans la base de données d'Ensembl (Zerbino, Wilder et al., 2015) ; les données proviennent des projets ENCODE et Roadmap Epigenomics, et ont été réutilisées pour définir cet ensemble d'annotations appelé "Ensembl Regulatory Build". Parmi les annotations disponibles, j'ai téléchargé une collection de régions correspondant à des sites de fixation de facteurs de transcription (48 facteurs au total), spécifiquement localisés dans des régions de pics de ChIP-seq associés à ces facteurs. Ces prédictions permettent donc d'avoir une identification de sites dans le génome, dont la pertinence biologique est appuyée par une observation expérimentale de fixation.

Sites de fixation conservés. Depuis le navigateur de génomes de l'UCSC, j'ai téléchargé un jeu de prédictions de sites de fixation de facteurs de transcription nommé "HMR Conserved Transcription Factor Binding Sites". Les motifs des facteurs de transcription de la base de données Transfac (v7) ont été utilisé pour prédire des sites de fixation dans les séquences alignées de trois espèces : l'humain, le rat, et la souris. Ces prédictions per-

mettent de calculer un score mesurant le degré de qualité d'identification et de conservation de cette identification chez les trois espèces ; ces sites correspondent donc à des sites de fixation potentiels qui ont été vraisemblablement conservés au cours de l'évolution.

Sites de fixation regroupés. Depuis le navigateur de génomes de l'UCSC, j'ai téléchargé un jeu de prédictions de sites de fixation de facteurs de transcription nommé "Transcription Factor ChIP-seq Clusters" (`http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeRegTfbsClusteredV3`). Ce jeu de données a été généré depuis l'identification de sites de fixation pour 161 facteurs de transcriptions dans 91 types cellulaires, issus du projet ENCODE. Les différentes prédictions ont été agrégées depuis les types cellulaires, permettant pour chaque facteur le calcul d'un score de "clustering" ou regroupement, mesurant le degré d'identification systématique d'un site dans une région génomique donnée.

Régions chromatiniennes ouvertes agglomérées

Depuis le navigateur de génomes de l'UCSC, j'ai téléchargé un jeu de régions identifiés par des expériences de sensibilité à la DNase 1, appelé "DNaseI Hypersensitivity Clusters". Ces régions correspondent à une agglomération de régions sensibles à la DNase 1, identifiées dans le cadre du projet ENCODE sur 125 types cellulaires. Chaque région est associée à un score entre 100 et 1000, mesurant le niveau d'accessibilité de la chromatine, agrégé depuis les types cellulaires.

REMAP : nouvelles identifications de pics de ChIP-seq

La base de données REMAP (Chèneby et al., 2018) propose un catalogue de localisations de pics de ChIP-seq pour des facteurs de transcriptions, par une ré-analyse de différents jeux de données publics (dont les données du projet ENCODE). Un total de 35.5 millions de sites de fixations non-redondants de facteurs de transcription sont disponibles dans le catalogue, pour un total de 485 facteurs de transcriptions.

5.1.3 Séquences et caractérisations des régions

Annotations GENCODE

Les données GENCODE d'annotations de gènes sur le génome humain (Frankish et al., 2019) ont été téléchargées le 22/11/2018, à partir de l'URL `https://www.gencodegenes.org/human/release_29lift37.html`.

La version à cette date là était la version "v29lifthg19" ; le document utilisé correspond au jeu de données "Comprehensive gene annotation", qui contient la liste complète des annotations de gènes obtenues initialement pour les chromosomes de la référence GRCh38, mappés sur l'assemblage GRCh37 du génome humain.

Différentes étapes de conversions et de filtres ont été appliquées :

— la manipulation des gènes se fait par les noms de gènes ; quand c'était nécessaire, une mise à jour de ces noms a été faite, avec la liste des noms proposés par le comité de nomenclature des gènes HUGO (HGNC) ;

— les gènes dont les noms sont associés à plusieurs localisations génomiques ont été réduit au gène dont l'importance biologique est la plus grande (la majorité des cas correspondent à des situations de gènes antisens, lncRNA, ou pseudo-gènes, partageant le nom d'un gène codant pour une protéine ; c'est donc ce dernier qui est gardé dans tous les cas) ;

— l'ajout des régions introniques pour les gènes codants (définies comme toute région située entre deux exons d'un même gène) et la distinction des 5'UTR et 3'UTR (à partir des régions UTRs définies, et de l'orientation des transcrits) ;

— l'annotation des régions promotrices.

Un fichier BED est ainsi généré, permettant d'identifier les gènes (codants ou non-codants) et leurs biotypes (régions génomiques plus précises, comme les UTRs ou les CDS) dans lesquels sont localisés les variants que je souhaite analyser dans ce projet. Puisque les annotations sont potentiellement chevauchantes (transcrits multiples d'un même gène, ou de gènes différents), j'extrais depuis les annotations associée à une position le biotype le plus important. Ce degré d'importance est établi de telle manière que les biotypes concernant la séquence codante (CDS, codon start, codon stop, site de splicing) sont plus importants que les régions non-codantes de transcrits de gènes codants (site de splicing, UTRs, promoteur, intron), eux-mêmes plus importants que des régions annotées comme transcrits non-codants (lncRNA, snRNA, etc.), qui sont finalement plus importants que les pseudo-gènes ; toute région non-annotée est considérée comme intergénique.

Séquences génomiques et composition

Séquences du génome, assemblage hg19. Les séquences génomiques ont été téléchargées le 4 février 2014 à partir du FTP de la base de données Ensembl (Zerbino, Achuthan et al., 2018) : `ftp://ftp.ensembl.org/pub/release-75/fasta/homo_sapiens/dna/`. Les séquences utilisées correspondent aux fichiers "rm", dans lesquels les séquences répétées sont remplacées par des "N".

Les fichiers ont été téléchargés au format Fasta, et indexés avec l'outil Samtools faidx (Li et al., 2009).

Dinucléotides CG. Les positions des dinucléotides CG ont été extraites pour chaque chromosome, et converties en un score (1 pour les dinucléotides CG, 0 pour les autres positions). Les fichiers (au format BED) ont été réassemblés dans un fichier au format BigWig.

Ilôts CpG. Les îlots CpG (régions riches en dinucléotides CG et associées aux régions promotrices de certains gènes), ont été téléchargés le 31/05/2015 depuis le navigateur de tables de l'UCSC, au format BED.

Éléments répétés. Les annotations d'éléments répétés ont été téléchargées le 10/09/2018, à partir du navigateur de tables du site de l'UCSC. Ces annotations correspondent aux régions identifiées par l'outil "repeat masker" (*RepeatMasker Open-4.0* 2013 - 2015).

Bandes cytogénétiques. Les annotations de bandes cytogénétiques (larges régions chromosomiques, associées à des propriétés de compaction de la chromatine plus ou moins importante) ont été téléchargées le 20 novembre 2018, à partir du navigateur de tables du site de l'UCSC.

Régions non-assemblées, et régions problématiques. Deux types de régions à exclure des analyses ont été identifiés :

— les régions non-assemblées du génome (correspondant à des régions de composition nucléotidiques particulière, comme par exemple les centromères, et pour lesquelles la détection de variants ou de signaux fonctionnels est peu fiable) ;

— les régions problématiques identifiées dans le projet ENCODE (The ENCODE Project Consortium, 2012), correspondant à des régions d'alignements anormaux de fragments de séquences obtenus notamment dans des expériences d'étude de l'ouverture de la chromatine.

Les régions non-assemblées du génome (trous) ont été téléchargées depuis le répertoire de la base de données de l'UCSC : `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/gap.txt.gz`.

Une compilation des régions problématique a été téléchargé sur le site personnel d'Anshul Kundaje (voir la documentation associée : `http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg19-human/hg19-blacklist-README.pdf`).

5.1.4 Prédictions d'associations entre régions régulatrices et gènes cibles

Les différents jeux de données ci-dessous proposent une identification de régions génomiques ayant un potentiel régulateur ; ces identifications sont basées sur des hypothèses et propriétés biologiques différentes, et complémentaires. Chaque région est également associée à un ou plusieurs gènes cibles ; à nouveau ces associations sont basées sur des propriétés biologiques différentes.

Dans le cadre de mes annotations de variants, un variant peut être localisé dans un ou plusieurs éléments éléments régulateurs prédits par les différentes sources. J'extrais donc pour chaque source le meilleure des scores d'association entre l'élément régulateur et les gènes cibles, ainsi que le nombre de gènes cibles. Par ailleurs, je calcule un score de partage des cibles entre différentes sources (dans le cas où un variant chevauche plusieurs éléments régulateurs) : ce score est le ratio du nombre de gènes-cibles communs aux différentes sources d'annotations, sur le nombre total de gènes-cibles. Ce score représente donc un degré d'accord entre les sources de prédictions d'associations.

PEGASUS - conservation de synténie. Le jeu de données PEGASUS (Clément et al., 2018) a été obtenu par communication interne au laboratoire ; ces données correspondent à un format tabulaire de régions régulatrices prédites avec leurs gènes cibles prédits, ainsi que différentes mesures de qualité. Ces prédictions sont basées sur la conservation en séquence des régions, et sur la conservation de la synténie entre les régions et leurs gènes cibles.

FANTOM5 - association par co-expression. Dans le cadre du projet FANTOM5 (Andersson et al., 2014), différents jeux de données de prédictions de régions régulatrices ont été générés. Ces jeux de données sont disponibles en téléchargement direct depuis l'URL : `http://enhancer.binf.ku.dk/presets/`. Le jeu de données correspondant aux prédictions d'associations entre régions régulatrices et gènes-cibles est celui nommé "enhancer_tss_associations.bed.gz" ; il a été téléchargé le 19/11/2018. Ces prédictions d'associations sont basées sur une corrélation de l'expression entre les régions non-codantes et les gènes cibles avoisinants.

GeneHancer. Le jeu de données GeneHancer (Fishilevich et al., 2017) propose une prédiction de régions régulatrices et d'associations aux gènes-cibles, par une compilation de différents types de données. Les ressources pour identifier les régions régulatrices proviennent des projets FANTOM, Ensembl Regulatory Build, VISTA, et ENCODE. Ces ressources permettent dans un premier temps de définir les régions régulatrices, avec une quantification de ce caractère régulateur suivant le nombre de sources associées à une région donnée. L'association aux gènes cibles est faite par l'utilisation des variants eQTLs du projet GTEX (voir section suivante), aux co-expressions identifiées dans le projet FANTOM, à la co-expression entre facteurs de transcriptions et gènes cibles (après identification de sites de fixation dans les régions régulatrices), et grâce à des données de capture de la conformation chromatinienne.

FOCS - redéfinition statistique d'associations. La base de données FOCS (Hait et al., 2018) propose des prédictions d'associations entre régions régulatrices et gènes-cibles, recalculées à partir des différents projets d'étude de régions fonctionnelles (FANTOM, ENCODE, Roadmap Epigenomics). Trois jeux de prédictions sont téléchargés : des prédictions basées sur des données d'expressions (des données de CAGE, pour le projet FANTOM, et des données de Gro-seq), et des prédictions basées sur une corrélation d'états chromatiniens (Roadmap Epigenomics).

5.1.5 Divers

"Chain files". Pour convertir les coordonnées génomiques associées à une version d'assemblage du génome à une autre, les outils utilisés requièrent des "chain files". L'ensemble du projet ayant été réalisé sur la base du génome de référence GRCh37/hg19, les fichiers

de conversion suivant ont été téléchargés :

— hg38 vers hg19 : `http://hgdownload.cse.ucsc.edu/goldenPath/hg38/liftOver/hg38ToHg19.over.chain.gz`

— hg18 vers hg19 : `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver/hg18ToHg19.over.chain.gz`

Taille des chromosomes. Différents outils de conversion de formats de fichiers ont besoin des tailles totales des chromosomes pour opérer. Un fichier "hg19.chrom.sizes" (associant le nom d'un chromosome à sa taille totale) a été téléchargé depuis le serveur FTP de l'UCSC : `http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes`.

5.2 Jeux de variants

5.2.1 Variants pour l'entraînement et évaluation des modèles de prédiction

"Damaging mutations - Genomiser." Dans le cadre de la création de l'outil "Genomiser" (Smedley et al., 2016), les auteurs ont compilé manuellement un jeu de données de 448 variants, associés à des maladies Mendéliennes et dont la pathogénicité a été jugée plausible sur la base de leur co-ségrégation, des validations expérimentales associées, ou de considérations similaires.

Nous avons extraits ces variants au format VCF le 9 janvier 2018, par téléchargement direct de la table S6 associée à la publication.

Ce jeu de données a initialement servi pour l'évaluation des modèles de prédiction proposés dans mon projet. Mais le chevauchement quasi-complet de ces variants avec ceux de la base de données HGMD a conduit à les écarter et à proposer une méthode alternative d'évaluation.

"Damaging mutations - HGMD." La base de données HGMD (Stenson et al., 2017) propose une compilation manuellement validée de variants identifiés dans la littérature scientifique par leur association à des maladies génétiques. Sont ainsi compilés plusieurs dizaines de milliers de variants, pour lesquels un degré de validation expérimental minimum doit être rapporté dans les articles associés. Un fichier VCF correspondant à la version

professionnelle de la base de données (version 2017.2) a été obtenu par communication personnelle dans le cadre d'une collaboration avec le consortium "NIHR bioresources for rare diseases".

eQTLs. Les variants eQTLs, correspondant à des polymorphismes associés statistiquement à des changements d'expression de gènes, ont été obtenus depuis le projet GTEx (Lonsdale et al., 2013, The GTEx Consortium et al., 2015, GTEx Consortium et al., 2017). La version 7 du projet a été téléchargée ; cela correspond à un total de plus de 36 millions d'associations SNP-gène cible à travers 48 tissus, pour un total de 3 124 345 positions eQTLs.

ClinVar - variants non-pathogéniques. Depuis le site FTP de la base de données ClinVar (Landrum et al., 2014), j'ai téléchargé le jeu de données correspondant à des mutations fréquentes, sans impact connu, qui a été assemblé le 5 septembre 2017. Ces variants correspondent à des variants communs (défini comme des variants dont la fréquence de l'allèle mineur est supérieure ou égale à 0.01), et pour lesquels aucune conséquence associée à une maladie n'a été identifiée.

5.2.2 Variations pour l'application du modèle

Dans le cadre d'une illustration de l'utilisation de mon modèle de prédiction, j'ai identifié une publication scientifique (An et al., 2018) concernant des variants *de novo* potentiellement causaux du syndrome autistique chez les enfants de la cohorte de patients analysée. Ces variants ont été identifiés par séquençage complet des génomes, pour des familles où seul un enfant parmi la fratrie est malade. Le fichier tabulaire (fourni en matériel supplémentaire) contenant les 255 106 mutations identifiés a été téléchargé depuis la version web de l'article.

5.3 Jeux de gènes

5.3.1 Standardisation des noms de gènes

Le comité HUGO de nomclature des gènes (Yates et al., 2017), ou HGNC, est dédié à la définition de standards pour normaliser la nomenclature associée aux gènes. En effets différentes ressources d'annotations utilisent des identifiants de gènes qui ne sont

pas toujours inter-compatibles ; par ailleurs un même locus fonctionnel peut être associé à plusieurs noms de gènes différents, pour des raisons historiques. L'objectif de ce comité est d'établir une standardisation de ces identifiants. J'ai donc téléchargé la table d'identification proposée par ce comité depuis le site web associé, le 22/11/2018. Cette table a servi de base pour la standardisation et la mise à jour des identifiants des gènes au travers des différents jeux de données utilisés dans ce projet.

5.3.2 Relations de régulation entre gènes

Dans le cadre de l'identification de relations de régulation pertinentes entre des prédictions de sites de fixations de facteurs de transcription, et des gènes cibles potentiels, j'ai utilisé la base de données TRRUST (« TRRUST v2 » p.d.).

Cette base de données compile des prédictions de relations de régulation entre des facteurs de transcription (795) et des gènes cibles (pour un total de 9396 relations prédites), sur la base d'une extraction automatique de ces relations depuis la littérature ; ainsi dans ce fichier, un facteur de transcription donné est associé à une relation d'activation ou de répression de un ou plusieurs gènes cibles.

5.3.3 Annotations de gènes

Pour caractériser différents groupes de gènes d'intérêts, j'ai utilisé différentes ressources, détaillées ici. Toutes les fichiers sont au format tabulaire, et compilent des noms de gènes, associés ou non à un score suivant la ressource considérée.

Gènes dosage-sensitifs. Certains gènes ont une contrainte d'expression importante, qui leur impose d'avoir une régulation maintenue constante, et surtout de ne pas être dupliqués (au cours de l'évolution, ou dans le cadre de mutations segmentales dans la population humaine). Ces gènes sont nommés "dosage-sensitifs" ; ce caractère peut s'expliquer par un besoin de niveau minimal d'expression pour être actif, d'une balance stoechiométrique avec leurs partenaires, ou encore par une nocivité lorsqu'ils sont exprimés en trop forte quantité. Un ensemble de 200 gènes dosage-sensitifs a été compilé (Rice et al., 2017), par identification de gènes enrichis dans des variants CNV trouvés chez des patients malades, par rapport à des CNVs identifiés chez des patients sains.

Gènes intolérants aux mutations : pLI. Dans le cadre du projet Exac (Lek et al., 2016), une mesure nommée "pLI" a été définie pour caractériser l'intolérance des gènes aux mutations type perte de fonction (mutations "Loss of Function" ou LoF, introduisant des codons stop dans la séquence protéique). Ce score a été calculé à partir de variations identifiés par séquençage d'exomes pour 60 706 patients ; il mesure une déviation du nombre de mutations observées par rapport au nombre attendu. Ansi chaque gène est associé à un score de tolérance : plus il est élevé, plus le gène est intolérant aux mutations (avec un seuil proposé de 0.9 pour identifier les gènes contraints).

Gènes intolérants aux mutations : RVIS. Dans le cadre du projet ESP (Exome Sequencing Project), un score nommé RVIS a été défini pour identifier les gènes intolérants aux mutations (Petrovski et al., 2013). Ce score est basé sur une comparaison par gène entre le nombre total de mutations observées, et le nombre de variations potentiellement fonctionnelles (mutations introduisant un codon stop, ou changeant la séquence protéique), calculés à partir des variations obtenus par séquençage d'exomes de 6 503 patients. Plus ce score est faible, plus le gène est intolérant aux mutations impactant la séquence protéique (avec un seuil proposé à 20, pour identifier les gènes contraints).

Gènes associées aux maladies : OMIM La base de données OMIM (Amberger et al., 2015) propose une compilation d'associations entre gènes et phénotypes, identifiées par extraction automatisée d'informations depuis la littérature. Un sous-jeu de ce catalogue a été téléchargé depuis l'outil Biomart de la base de données Ensembl, et correspond à une liste de 3 310 gènes associés à des phénotypes de maladie.

Gènes associés aux maladies : OpenTargets La base de données OpenTargets propose un répertoire d'associations entre phénotypes et gènes, basés sur un ensemble de sources diverses (Koscielny et al., 2017, Carvalho-Silva et al., 2019). Les associations sont établis sur la base d'associations génétiques (GWAS), de validations expérimentales (validées manuellement), ou encore par associations automatiques par extraction d'information automatisée depuis la littérature. Dans le cadre des analyses des variants associés au syndrome du spectre autistique, une liste de 2 257 gènes associés à l'autisme (identifiant EFO_0003758) a été téléchargée depuis le site web d'OpenTargets.

Gènes associés au syndrome du spectre autistique - SFARI La fondation "Simons Foundation Autism Research Initiative", dédiée à la recherche sur l'autisme, propose au téléchargement un tableau de 1 055 gènes, qui ont été identifiés comme associés au syndrome du spectre autistique dans le cadre de leurs recherches.

5.4 Scores de fonctionnalité

Au chapitre 3, des méthodes de prédiction dédiés à l'évaluation des variants non-codants ont été présentées. Une sous-partie de ces méthode a été sélectionnée, à la fois par soucis de facilité d'utilisation, et par soucis de diversité des scores.

Je présente ici les différents scores de fonctionnalités, avec les étapes de modifications appliquées.

CADD

Le score CADD (Kircher et al., 2014, Rentzsch et al., 2019), dans sa version 1.4, est disponible pour chaque substitution possible pour l'ensemble des nucléotides du génome. Le fichier est au format tabulaire, et a été téléchargé depuis le site web du projet (`https://cadd.gs.washington.edu/download`) ; un index a été généré pour accélérer son utilisation par mes outils.

Eigen

Les scores Eigen et Eigen-PC (Ionita-Laza et al., 2016) ont été téléchargés depuis le site web du projet (`https://xioniti01.u.hpc.mssm.edu/v1.1/`). Comme pour CADD, chaque position du génome est associée à 3 valeurs, pour chacune des substitutions possibles. Le fichier initial a été réduit pour ne garder que les deux scores Eigen et Eigen-PC, puis indexé pour accélérer son utilisation.

FATHMM-MKL

Le score FATHMM-MKL (Shihab et al., 2015) a été téléchargé depuis le site web du projet (`http://fathmm.biocompute.org.uk/database/fathmm-MKL_Current_zerobased.tab.gz`). Comme pour CADD, chaque position du génome est associée à 3 valeurs, pour chacune des substitutions possibles. Deux scores sont disponibles : un score pour les va-

riants codants, et un score pour les variants non-codants. C'est ce dernier qui sera utilisé pour annoter les variants non-codants et comparer les différents méthodes avec la mienne.

FIRE

Les scores FIRE (Ioannidis, Davis et al., 2017) ont été téléchargés pour chaque chromosome depuis le site web du projet. Comme pour CADD, chaque position du génome est associée à 3 valeurs, pour chacune des substitutions possibles. Les fichiers par chromosomes ont été agrégés en un seul fichier, qui a ensuite été indexé.

B-score

Le score de pression de sélection ("Background selection score", McVicker et al., 2009) a été téléchargé depuis la page du laboratoire associés (`http://www.phrap.org/othersoftware.html`). Les fichiers initiaux (un par chromosomes) correspondent à des fichiers tabulaires dans lesquel le B-score est fixé pour une taille de région génomique donnée, pour le génome humain dans sa version hg18. J'ai donc tout d'abord converti ce format en un format BED, en utilisant les coordonnées génomiques des chromosomes depuis l'assemblage hg18, puis transformé ces positions vers la version hg19 du génome. Enfin, puisque chaque position est associée à un seul score, le fichier BED est converti en un format BigWig, pour accélérer la manipulation de cette ressource.

NCBoost

Les scores NCBoost (Caron et al., 2018) ont été téléchargés depuis le site web du projet (`https://github.com/RausellLab/NCBoost`). Le fichier correspond à un format BED, et propose pour chaque position du génome un ou plusieurs scores, étant donné l'annotation génomique et le biotype considéré (si une position chevauche plusieurs transcrits, un score est proposé pour chacun des transcrits). Dans le cadre de l'annotation des variants, je garderai le meilleur des scores associés à une position.

ReMM - Genomiser

Les scores ReMM (Smedley et al., 2016), associés à l'outil Genomiser, ont été téléchargés depuis la page du projet (`https://charite.github.io/software-remm-score.html`). Le fichier est dans un format tabulaire, avec une seule coordonnée par position.

Puisqu'ici chaque position est associée à un seul score, le fichier est transformé dans un format BigWig pour une manipulation plus rapide.

FitCons

Les scores FitCons (Gulko et al., 2015) ont été téléchargés depuis la page du projet `http://compgen.cshl.edu/fitCons/` ; plusieurs fichiers, au format BigWig, sont disponibles. Cependant pour comparer leur méthode à d'autres approches, les auteurs invitent les utilisateurs à n'utiliser que le fichier "i6", qui correspond à une version de leur score intégrée sur tous les types cellulaires qu'ils ont considérés dans leur approche.

Linsight

Le score Linsight (Huang et al., 2017) a été téléchargé depuis la page du projet (`https://github.com/CshlSiepelLab/LINSIGHT/blob/master/README.md`), au format BigWig.

Chapitre 6

Méthodes

6.1 Manipulations des fichiers de données

6.1.1 Fichiers de variants : VCF

Les fichiers de variants sont au format VCF, ou "Variant Call Format" (Danecek et al., 2011). Ce format est composé d'un ensemble de lignes de commentaires décrivant les différents champs disponibles, suivies de l'ensemble des variants appelés, dans le cadre d'une expérience de séquençage appliqué chez un ou plusieurs patients. Ces variants sont identifiés par leur position génomique, et annotés avec différentes caractéristiques, comme des scores de qualité, ou encore des métadonnées relatives aux patients chez qui ces variants sont trouvés. Ces fichiers ont principalement été utilisés de manière similaires aux fichiers d'intervalles génomiques, et convertis en leur équivalent "BED" pour les différentes manipulations effectuées (annotations de variants, sélections, etc.). Pour faciliter et accélérer leur manipulation, il est possible de les indexer grâce à l'outil Tabix (Li, 2011).

6.1.2 BED et TSV

Les formats BED et TSV ("tab separated values") correspondent à des fichiers de données tabulaires), et peuvent donc être manipulé avec la plupart des outils de manipulation de fichiers textes sous GNU/Linux (awk, sed, less, ou encore Excel).

Les fichiers au format BED correspondent plus spécifiquement à des fichiers d'intervalles génomiques, permettant d'associer une région génomique (définie par un chromosome, une position de début incluse et une position de fin exclue) à différentes annotations

et caractéristiques (score, label, identifiant, etc).

Deux outils ont été utilisés pour manipuler ces fichiers BED :

— BedTools, en version 2.27.1 (Quinlan et al., 2010) : qui permet notamment de réaliser des intersections entre intervalles, ainsi que des sélections (par exemple : obtenir les régions les plus proches d'une région d'intérêt, à une certaine distance). L'interface Python pyBedTools (Dale et al., 2011) a été utilisée dans les différents modules de mon outil FINSURF.

— BedOps, en version 2.4.15 (Neph et al., 2012) : qui permet de réaliser des opérations plus complexes sur les intervalles. Notamment, j'ai employé cette suite d'outil pour isoler des régions non-chevauchantes après avoir regroupé des intervalles de différentes sources (outil bedtools –partition), et pour réannoter ces régions isolées avec les annotations des sources initiales (outil bedmap).

Pour faciliter et accélérer leur manipulation, il est possible de les indexer grâce à l'outil Tabix (Li, 2011).

6.1.3 Fichiers de séquences Fasta

Le format Fasta permet de stocker des séquences (nucléotidiques ou protéiques), associées à un identifiant. Pour faciliter la manipulation de ces fichiers, différents modules de la suite d'outils Samtools ont été utilisés (Li et al., 2009), à la fois pour indexer et pour interroger les fichiers.

6.1.4 Fichiers WIG et BigWig

Les fichiers WIG et BigWig permettent d'associer de manière optimisée des valeurs continues ou discrètes à un ensemble de positions dans le génome. Le fichier WIG correspond à un fichier texte, dans lequel sont stockés des blocs, identifiés par une ligne définissant notamment la manière dont sont représentés les intervalles génomiques (taille fixe, ou variable ; taille de chaque intervalle ; début de l'intervalle), qui est suivie par un ensemble de valeurs représentant par exemple un score, et qui sont associées aux intervalles du bloc. Les fichiers BigWig (Kent, Zweig et al., 2010) correspondent à une version binaire, indexée des fichiers WIG. Ils sont idéaux pour interroger et obtenir rapidement des scores associés à un ensemble de positions d'intérêt.

Concernant les outils :

— Les fichiers WIG ont représenté des intermédiaires lors des manipulations des fichiers d'annotations ; leur conversion en fichiers BigWig a été réalisée grâce à l'outil "wigToBigWig" de la suite d'outils de l'UCSC (Kuhn et al., 2013).
— Les fichiers BigWig ont été manipulés grâce aux outils "bwtool" (version 1.0, Pohl et al., 2014), "pyBigWig" pour l'intégrations à mes modules Python (version 0.3.12, Ramírez et al., 2014), et WiggleTools (Zerbino, Johnson et al., 2014).

6.1.5 Autres outils

Certains des outils utilisés dans le cadre de la manipulation des fichiers proviennent de ressources proposées par le navigateur de génomes de l'UCSC (Kent, Sugnet et al., 2002). Notamment l'outil LiftOver (Hinrichs et al., 2006), qui a permis la conversion de coordonnées génomiques d'une version d'assemblage à une autre.

Par ailleurs depuis le serveur web du navigateur de génome de l'UCSC, le navigateur de tables (Karolchik et al., 2004) a permis de télécharger certaines des ressources d'annotations utilisées dans ce projet.

L'équivalent proposé par la base de donnée Ensembl (Zerbino, Achuthan et al., 2018), appelé Biomart (Kinsella et al., 2011), a également servi pour le téléchargement de certains jeux de données.

6.2 Mesures et tests statistiques

6.2.1 Mesure de taille d'effet

Pour quantifier l'écart entre deux groupes pour une distribution donnée de valeurs, et pouvoir le comparer à un autre écart, j'ai recourt à des mesures de taille d'effet standardisées.

Coefficient d de Cohen. Pour une distribution continue ou discrète, ce coefficient permet de comparer deux populations sur la base de leurs moyennes, par le calcul d'une différence standardisée :

$$d_{cohen} = \frac{(\bar{x}_A - \bar{x}_B)}{s} \tag{6.1}$$

où s représente l'écart-type cumulé :

$$s = \sqrt{\frac{(N_A - 1)s_A^2 + (N_B - 1)s_B^2}{N_A + N_B - 2}} \tag{6.2}$$

Coefficient h de Cohen. Pour une distribution catégorielle ou binaire, ce coefficient permet de comparer deux populations sur la base de leurs proportions dans les différentes catégories. Dans mon cas, je transformerai systématiquement toute distribution catégorielle en plusieurs distributions binaires ; dans ce cas il est possible de calculer une proportion p d'entités ayant une valeur non-nulle. Le coefficient h de Cohen est basé sur la différence entre deux proportions :

$$h_{cohen} = 2(\arcsin\sqrt{p_A} - \arcsin\sqrt{p_B}) \tag{6.3}$$

où p_A et p_B sont les proportions respectives des deux populations, et $arcsin$ est la fonction arc sinus. Ce calcul permet d'atténuer l'amplitude des différences pour des proportions très faibles.

6.3 Méthodes et mesures spécifiques à l'apprentissage machine

Dans le cadre d'une tâche de classification, on cherche à développer un modèle de prédiction capable d'assigner à une entité un label, correspondant à la classe prédite par le modèle ; ce label est idéalement le même que la classe réelle de l'entité, autrement le modèle s'est trompé. Lorsque l'entité n'a jamais été vue auparavant, on ne peut pas savoir si le modèle se trompe ou non. En revanche au moment de l'entraînement du modèle il est possible d'utiliser une partie du jeu d'entités dont on connaît les vraies classes pour évaluer et valider le modèle.

Différentes mesures sont disponibles pour évaluer la qualité d'un modèle de classification, à partir d'un vecteur de classes vraies associées aux entités, et d'un second vecteur de labels prédits, obtenu par application du modèle aux entités. Je présente ces mesures ci-dessous. A noter que je me place dans un problème de classification binaire, où seules deux classes (positive et négative) sont à déterminer par le modèle de classification : en effet, dans le cadre de ce projet de thèse, nous cherchons à entraîner des modèles pour distinguer des variants fonctionnels (classe contrôle-positif) de variants non-fonctionnels (classe contrôle-négatif).

6.3.1 Matrice de confusion et catégories

Tout d'abord, dans le cadre d'une classification binaire, il est possible d'identifier les entités correctement classées ou non, en fonction de leur vrai label :

— Les **vrais-positifs** (ou "True positives" ou "TP") sont les entités de la classe des contrôles-positifs dont le label a correctement été prédit par le modèle ;

— Les **vrais-négatifs** (ou "True negatives" ou "TN") sont les entités de la classe des contrôles-négatifs dont le label a correctement été prédit par le modèle ;

— Les **faux-positifs** ou ("False positives" ou "FP") sont les entités de la classe des contrôles-négatifs qui sont incorrectement labellisés "positifs" ;

— Les **faux-négatifs** ou ("False negatives" ou "FN") sont les entités de la classe des contrôles-positifs qui sont incorrectement labellisés "négatifs".

Toutes ces valeurs sont regroupées dans un tableau nommé **matrice de confusion**, dont les lignes correspondent aux vrais labels, et les colonnes correspondent aux labels prédits. Ainsi, de haut en bas et de gauche à droites se trouvent dans l'ordre les vrais-négatifs, les faux-positifs, les faux-négatifs, et les vrais-positifs. Les valeurs de cette matrice seront exprimées soit en comptages, soit en proportions pour chacun des "vrais" labels.

6.3.2 Scores de classification

À partir de ces valeurs, il est possible de calculer les différents scores de qualités évoqués plus haut.

L'**accuracy**, que je traduis dans ce manuscript par "exactitude", mesure la fraction totale d'entités dont la classe a correctement été prédite par le modèle.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6.4}$$

La **balanced accuracy**, que je traduis dans ce manuscript par "exactitude balancée", mesure pour chaque classe la fraction d'entités correctement labellisée par le modèle, et calcule la moyenne des deux valeurs (dans le cadre d'une classification binaire). Cette mesure permet d'avoir une évaluation de la qualité générale du modèle en prenant en compte le déséquilibre de représentation des deux classes.

$$Balanced\ Accuracy = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \tag{6.5}$$

Le **True Positive Rate** (abrégé "TPR") ou "taux de vrais positifs", également appelé "sensitivity" (sensibilité) ou "recall", mesure la proportion d'entités de la classe d'intérêts qui sont correctement labellisés en "positifs", sur le total du nombre d'entités de cette classe qui étaient à identifier.

$$True\ Positive\ Rate = \frac{TP}{TP + FN} \tag{6.6}$$

Le **True Negative Rate** (abrégé "TNR") ou "taux de vrais négatifs", également appelé "specificity" (spécificité), mesure la proportion d'entités de la classe contrôle qui sont correctement labellisés en "négatifs", sur le total du nombre d'entités de cette classe qui étaient à identifier.

$$True\ Negative\ Rate = \frac{TN}{TN + FP} \tag{6.7}$$

Le **False Positive Rate** (abrégé "FPR") ou "taux de faux positifs", mesure la proportion d'entités de la classe contrôle qui sont incorrectement labellisés en "positifs", sur le total du nombre d'entités de la classe contrôle qui étaient à identifier.

$$\begin{aligned} False\ Positive\ Rate &= \frac{FP}{TN + FP} \\ &= 1 - True\ Negative\ Rate \end{aligned} \tag{6.8}$$

La **Precision** ou "précision", mesure la proportion d'entités de la classe d'intérêt correctement labellisées parmi l'ensemble des entités labellisées "positifs".

$$Precision = \frac{TP}{TP + FP} \tag{6.9}$$

Le **False Discovery Rate** (abrégé "FDR") ou "Taux de fausses découvertes", mesure la proportion d'entités de la classe contrôle qui sont incorrectement labellisés en "positifs", sur le total du nombre d'entités labellisées "positifs".

$$\begin{aligned} False\ Discovery\ Rate &= \frac{FP}{TP + FP} \\ &= 1 - Precision \end{aligned} \tag{6.10}$$

Le **score F1** correspond à la moyenne harmonique de la précision et de la sensibilité (ou "recall") :

$$F1\ score = \frac{2 \cdot Precision \cdot TPR}{Precision + TPR} \tag{6.11}$$

6.3.3 Courbes de qualité

Généralement, les modèles de classifications fournissent pour des entités à classifier un score, plutôt qu'une classe prédite. C'est grâce à un seuil sur ce score qu'il est possible d'obtenir les prédictions de classes. Une illustration pour une tâche de classification binaire est proposée dans la figure 6.1 : deux populations d'entités (positives en rouge et négatives en bleu) ont eu un score d'assigné par un modèle de prédiction. Il est possible d'appliquer un seuil sur ce score, pour définir les classes prédites associés à ces entités.

Si l'on utilise directement les scores de prédictions plutôt que les classes prédites, il est possible de définir deux courbes explorant la qualité du modèle de prédictions. Ces deux courbes sont construites par exploration de l'espace des seuils possibles, et consistent à calculer pour chacun des seuils les différentes valeurs composant la matrice de confusion.

La première courbe est appelée **Receiver Operating Characteristics curve** (abrégée "ROC curve"), et correspond à l'évolution du taux de vrais positifs détectés en fonction du taux de faux positifs introduits, mesurés aux différents seuils sur le score. A partir de la représentation de cette courbe, une mesure est généralement calculée : l'aire sous la courbe. Un classificateur parfait est associé à une aire sous la courbe de 1 (100% de vrais positifs obtenus pour 0% de faux positifs introduits dès le premier seuil), tandis qu'un classificateur aléatoire correspond à une diagonale sur cette courbe, et donc à une aire sous la courbe de 0.5. Un modèle peut être évalué comme bon classificateur s'il permet d'obtenir une courbe ROC s'éloignant de la diagonale, et donc dont la valeur de l'aire sous la courbe s'approche de 1.

La seconde courbe est appelée **Precision-Recall curve** (abrégée "PrecRec curve" ou "PR curve"), et correspond à l'évolution de la précision des positifs prédits en fonction du taux de vrais positifs détectés. Dans cette représentation, un classificateur aléatoire introduit autant de vrais positifs que de faux positifs à n'importe quel seuil fixé. Cette prédiction aléatoire conduit donc à des taux de vrais et faux positifs qui sont déterminés par les proportions initiales de chacune des classes. Ainsi, un classificateur aléatoire apparaîtra dans cette représentation sous la forme d'une droite horizontale, dont l'intersection avec l'axe des y (la précision) se fait à la proportion de la classe minoritaire sur l'ensemble du jeu de données. L'aire sous la courbe associée est égale à cette proportion. Un classificateur

parfait en revanche apparaîtra comme une constante de valeur 1 pour toute valeur du taux de vrais-positifs, et sera donc associé à une aire sous la courbe de 1. Un modèle peut être évalué comme bon classificateur s'il permet d'obtenir une courbe PR s'éloignant de cette constante horizontale, et donc dont la valeur de l'aire sous la courbe s'approche de 1.

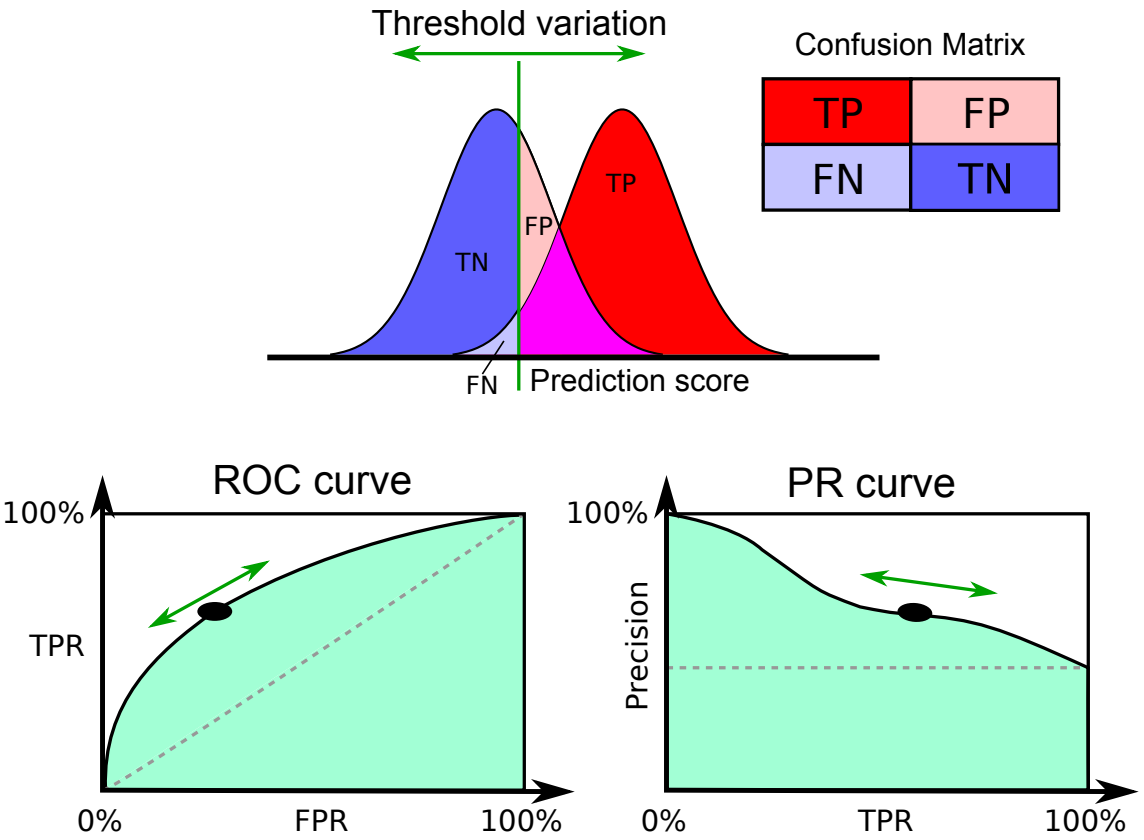


FIGURE 6.1 – Illustration des courbes ROC et PR. Pour une tâche de classification binaires, les deux populations (positifs et négatifs) sont associées à des distributions de scores de prédictions plus ou moins distinctes, comme présenté sur le panel du haut. Afin d'assigner le label prédit, un seuil (en vert) est appliqué sur le score de prédiction : toutes les valeurs au dessus sont labellisés "positives" et celles en dessous sont labellisées "négatives". Cela conduit les positifs (ensemble rouge) à être séparés en vrais positifs et faux négatifs, et les négatifs (ensemble bleu) à être séparés en vrais négatifs et faux positifs ; ces valeurs sont répertoriées dans la matrice de confusion. En faisant varier le seuil, il est possible de générer la courbe ROC (panel bas gauche), et la courbe Precision-Recall (panel bas droit) ; les aires sous les courbes (surfaces colorées) peuvent être rapportées pour évaluer la qualité globale d'un modèle.
Adapté de `https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic`.

6.3.4 Algorithme des K-means et mesures

L'algorithme des K-means consiste en la découverte et définition de K centroïdes auxquels sont assignés l'ensemble des entités que l'on cherche à regrouper en communautés homogènes.

Distance Euclidienne. Pour une entité, son assignation à un des groupes est déterminée par la plus courte des distances entre cette entité et chacun des centroïdes définissant les groupes. Cette distance correspond à la distance euclidienne mesurée entre deux vecteurs :

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{\sum_{m=1}^{M} (\boldsymbol{p}_m - \boldsymbol{q}_m)^2} \tag{6.12}$$

où $\boldsymbol{p}$ et $\boldsymbol{q}$ sont deux vecteurs composés chacun de M valeurs.

Inertie. Pour évaluer la qualité d'identification des groupes obtenus après application de l'algorithme, il est possible de mesurer l'inertie :

$$inertie = \sum_{i=0}^{n} \underset{\mu_j \in C}{min}(d(\boldsymbol{x}_i, \boldsymbol{\mu}_j)^2) \tag{6.13}$$

où C représente l'ensemble des groupes, μ_j étant le centroïde du groupe C_j, et x_i l'une des entités. La distance au centroïde le plus proche de chaque entité est utilisée pour calculer une somme sur l'ensemble N des entités.

Cette mesure peut être interprétée comme une quantification de la cohérence interne des groupes définis par les centroïdes : plus la valeur est faible, plus les groupes sont homogènes.

Mesure de silhouette. Une seconde mesure utilisée est le coefficient de silhouette (Rousseeuw, 1987). Pour une entité $\boldsymbol{x_i}$, assignée au groupe C_j, on peut mesurer une première valeur a_i, correspondant à la distance entre cette entité et l'ensemble des entités du même groupe :

$$a_i = \frac{\sum_{k \in C_j, k \neq i} d(\boldsymbol{x}_i, \boldsymbol{x}_k)}{|C_j|} \tag{6.14}$$

Une seconde valeur, b_i, correspond à la distance moyenne entre l'entité $\boldsymbol{x}_i$ et l'ensemble des entités du second groupe le plus proche de $\boldsymbol{x}_i$ (après C_j), que l'on note C_h :

$$b_i = \frac{\sum_{k \in C_h} d(\boldsymbol{x}_i, \boldsymbol{x}_h)}{|C_h|} \tag{6.15}$$

La silhouette associée à $\boldsymbol{x_i}$ est calculée à partir de ces deux valeurs :

$$Silhouette_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{6.16}$$

Plus cette valeur est élevée, plus l'entité $\boldsymbol{x_i}$ est à la fois proche des entités du groupe auquel elle appartient, et éloignée des autres entités. On peut alors calculer une silhouette moyenne sur l'ensemble N des entités : plus elle est élevée, plus les groupes identifiés sont homogènes et distincts les uns des autres.

6.4 Analyse de motifs et détection de site de fixations de facteurs de transcription

Dans le cadre de mes analyses de variants, j'ai procédé à la prédiction de sites de fixation de facteurs de transcription, et à l'évaluation de l'impact de certains variants sur ces sites. L'ensemble des analyses a été réalisé avec la suite d'outils RSAT (Nguyen et al., 2018). Je précise ci-dessous les différents modules et ressources utilisés.

6.4.1 Motifs de facteurs de transcriptions

Les facteurs de transcriptions sont des effecteurs protéiques capables de reconnaître des compositions nucléotidiques particulière dans le génome, et de s'y lier de manière transitoire, pour favoriser ou inhiber la transcription d'un gène (voir chapitre 1). Ces sites de fixations sont spécifiques à chaque facteur, et peuvent être modélisé sous la forme de matrices de poids, à partir d'expériences de détection d'affinités. Ces matrices de poids contiennent à chaque position associée à un motif donné les proportions de chacun des nucléotides.

Dans le cadre des analyses effectuées avec l'outil RSAT, les motifs de facteurs de transcriptions évalués proviennent de la base de motifs JASPAR (Khan et al., 2018).

6.4.2 Détection des sites de fixation et analyse d'impact des variants

La détection d'un site de fixation d'un facteur de transcription dans une séquence peut être faite par l'outil *matrix-scan* disponible dans la suite RSAT. Pour un facteur de trans-

cription représenté par un motif, cet outil permet de scanner une séquence nucléotidique donnée pour calculer un score pour chaque segment de la taille du motif. Ce score mesure la correspondance entre la matrice de poids du motif et le segment nucléotidique considéré : le score associé est d'autant plus élevé que la séquence correspond à la composition préférentielle mesurée par la matrice de poids. Un modèle de référence de la composition nucléotidique du génome permet d'évaluer la probabilité pour un segment nucléotidique donné de présenter un vrai site de fixation. Ainsi chaque segment nucléotidique dans une région évaluée est associé à un score de prédiction et à une probabilité.

Il est alors possible d'évaluer la conséquence d'un variant nucléotidique dans un segment sur le score et la probabilité d'un site de fixation prédit. La diminution de ces valeurs indiquerait que le variant réduit potentiellement la capacité du segment à être reconnu par un facteur de transcription, et inversement. Cette évaluation peut être automatisée par l'outil *variation-scan*, qui permet de scanner avec une base de motifs un ensemble de séquences et de variants associées, pour identifier les sites de fixations potentiels, et les conséquences des variants sur la prédiction de ces sites.

6.5 Développement des programmes

Les analyses et outils développés dans le cadre de ce projet ont été réalisés dans le langage de programmation Python (version 3.6), dont la gestion des bibliothèques additionnelle a été permis grâce à l'outil Conda (qui permet également une portabilité de l'environnement, et un suivi des versions des bibliothèques installées). Par ailleurs les analyses et figure générées ont été développées dans des notebooks Jupyter, pour permettre de rendre ces analyses reproductibles.

Troisième partie

Résultats

Chapitre 7

Analyses préliminaires

7.1 Récapitulatif des annotations utilisées

L'objectif de ce projet est de définir des modèles de classification dédiés à l'identification de variants génomiques non-codants potentiellement fonctionnels. Ces modèles de classification vont être entraînés à distinguer des variants identifiés comme fonctionnels (contrôles positifs), de variants identifiés comme non-fonctionnels (contrôles négatifs). Cette distinction nécessite d'avoir des valeurs permettant de décrire les variants, pour y identifier des différences entre les deux classes. Ces descripteurs sont issus de différents jeux d'annotations génomiques, décrits dans le chapitre 5. Ils sont résumés dans le tableau ci-dessous. Chacune des annotations est associée à une courte description, ainsi qu'à une information concernant le type de distribution. Quatre classes d'annotations y sont répertoriées :

— des annotations issues d'information d'annotation de la séquence ;
— des annotations concernant la conservation nucléotidique ;
— des annotations de marques fonctionnelles biochimiques ;
— des prédictions d'associations entre éléments régulateurs et gènes cibles ;

7.2 Caractérisation des jeux de données de variants

Les modèles de classification que j'ai souhaité développer dans ce projet doivent être capables de prédire pour un variant d'intérêt son potentiel fonctionnel. Nous avons vu à la section précédente les caractéristiques utilisées pour décrire des variants, et qui se-

Classe	Identifiant	Description	Distribution
Séquence	vartrans.ord	Type de mutation.	Discrète
	Cgdinit	Localisation dans un dinucléotide CG.	Binaire
	CpGisland	Localisation dans un îlot CpG annoté.	Binaire
Conservation	phyl100w	Conservation à la position mesurée par le score PhyloP sur les alignements de séquences de 100 espèces vertébrés.	Continue
	phyl20w	Conservation à la position mesurée par le score PhyloP sur les alignements de séquences de 20 espèces primates	Continue
	phast100w	Conservation à la position mesurée par le score PhastCons sur les alignements de séquences de 100 espèces vertébrés	Continue
	phast20w	Conservation à la position mesurée par le score PhastCons sur les alignements de séquences de 20 espèces primates	Continue
	gerpScore	Conservation à la position mesurée par le score GERP sur les alignements de séquences de 33 espèces de mammifères	Continue
	gerpElem	Présence dans un élément conservé identifié par accumulation de positions nucléotidiques conservées selon le score GERP.	Binaire
	CDTS	Score de tolérance aux mutations dans la population (projet gnomAD) calculé sur des régions de 10bp, et rapporté pour la position.	Continue
Propriétés biochimiques	dnaseClust	Clusters de pics de sensibilité à la DNase 1 identifiés dans 125 types cellulaires, dans le cadre du projet ENCODE.	Discrète
	roadmapState.promoter	Comptage du nombre de types cellulaires dans lesquels la position est associée à une région évaluée comme “promoter”.	Discrète
	roadmapState.transcribed	Comptage du nombre de types cellulaires dans lesquels la position est associée à une région évaluée comme “transcrit”.	Discrète
	roadmapState.enhancer	Comptage du nombre de types cellulaires dans lesquels la position est associée à une région évaluée comme “enhancer”.	Discrète
	roadmapState.ZNFRpts	Comptage du nombre de types cellulaires dans lesquels la position est associée à une région évaluée comme “transcrit de gène à doigt de zinc.”.	Discrète
	roadmapState.heterochrom	Comptage du nombre de types cellulaires dans lesquels la position est associée à une région évaluée comme “hétérochromatinienne”.	Discrète
	roadmapState.repressed	Comptage du nombre de types cellulaires dans lesquels la position est associée à une région évaluée comme “réprimée”.	Discrète
	roadmapState.quies	Comptage du nombre de types cellulaires dans lesquels la position est associée à une région évaluée comme “quiescente”.	Discrète
	H3K4me1_medFC	Valeur médiane du fold-change associée à la marque d’histone H3K4me1 (enhancer) calculée depuis 98 types cellulaires, et mesurée à la position.	Continue
	H3K4me3_medFC	Valeur médiane du fold-change associée à la marque d’histone H3K4me3 (promoter) calculée depuis 98 types cellulaires, et mesurée à la position.	Continue
	H3K27ac_medFC	Valeur médiane du fold-change associée à la marque d’histone H3K27ac (enhancer) calculée depuis 98 types cellulaires, et mesurée à la position.	Continue
	tfbsEnsembl.count	A une position : comptage du nombre de facteurs de transcription pour lesquels le motif a été trouvé dans un pic de ChIP-seq correspondant.	Discrète
	tfbsEnsembl.scoremax	A une position : score maximal de détection d’un motif parmi les sites identifiés.	Discrète
	tfbsCons.count	A une position : comptage du nombre de facteurs de transcription pour lesquels le motif a été trouvé comme conservé chez l’humain, le rat, et la souris.	Discrète
	tfbsCons.scoremax	A une position : score maximal de détection d’un motif parmi les sites identifiés.	Discrète
	tfbsClust.count	A une position : comptage du nombre de facteurs de transcription pour lesquels le motif a été trouvé, à travers 161 types cellulaires (projet ENCODE).	Discrète
	tfbsClust.scoremax	A une position : score maximal de détection d’un motif parmi les sites identifiés.	Discrète
Prédictions de regions régulatrices et d’associations	pegasus.score	A une position : score maximal d’association entre une région régulatrice prédite (méthode PEGASUS) et un ou plusieurs gènes cibles.	Continue
	pegasus.count_targs	A une position : nombre d’associations entre une région régulatrice prédite (méthode PEGASUS) et un ou plusieurs gènes cibles.	Discrète
	fantom.bestScore	A une position : score maximal d’association entre une région régulatrice prédite (méthode FANTOM) et un ou plusieurs gènes cibles.	Continue
	fantom.count_targs	A une position : nombre d’associations entre une région régulatrice prédite (méthode FANTOM) et un ou plusieurs gènes cibles.	Continue
	genhancer.bestScore	A une position : score maximal d’identification de région régulatrice, par la méthode GENEHANCER.	Continue
	genhancer.countTargs	A une position : nombre d’associations entre une région régulatrice prédite (méthode GENEHANCER) et un ou plusieurs gènes cibles.	Discrète
	genhancer.bestScore_targs	A une position : score maximal d’association entre une région régulatrice prédite (méthode GENEHANCER) et un ou plusieurs gènes cibles.	Continue
	focs_fantom.bestScore	A une position : score maximal d’association entre une région régulatrice prédite (méthode FANTOM) et un ou plusieurs gènes cibles.	Continue
	focs_fantom.count_targs	A une position : nombre d’associations entre une région régulatrice prédite (méthode FANTOM) et un ou plusieurs gènes cibles.	Discrète
	focs_groseq.bestScore	A une position : score maximal d’association entre une région régulatrice prédite (méthode FANTOM) et un ou plusieurs gènes cibles.	Continue
	focs_groseq.count_targs	A une position : nombre d’associations entre une région régulatrice prédite (méthode FANTOM) et un ou plusieurs gènes cibles.	Discrète
	focs_roadmap.bestScore	A une position : score maximal d’association entre une région régulatrice prédite (méthode FANTOM) et un ou plusieurs gènes cibles.	Continue
	focs_roadmap.count_targs	A une position : nombre d’associations entre une région régulatrice prédite (méthode FANTOM) et un ou plusieurs gènes cibles.	Discrète
	ratio_shared_targets	A une position : calcul du nombre de gènes partagés entre les prédictions de différentes sources, divisé par le nombre de gènes totals associés.	Discrète
	targets_associations	Degré d’association entre une position et un ou plusieurs gènes (fournis par l’utilisateur), calculé depuis la localisation génomique et les données de prédictions d’associations. (0 : non associé ; 1 : gène adjacent ; 2 : cible d’enhancer ou dans le gène ; 3 : enhancer génique)	Discrète

TABLE 7.1 – Tableau résumé des annotations

ront exploitées par les modèles de prédiction. Je présente dans cette partie les choix faits concernant les ensembles de variants identifiés et sélectionnés depuis la littérature.

J'aborde tout d'abord la question des outils développés et utilisés pour l'annotation des variants avec les ressources sélectionnées. Puis j'évoque les ensembles de variants fonctionnels identifiés, qui seront des contrôles positifs pour les modèles d'apprentissage machine. J'aborde également la génération de différents ensembles de variants non-fonctionnels, pris comme contrôles négatifs pour les modèles. L'évaluation de différences mesurables entre les contrôles positifs et négatifs me conduira à justifier l'utilisation de modèles d'apprentissage machine, qui seront développés au chapitre suivant.

7.2.1 Annotation des variants

Afin d'annoter le plus rapidement possible de larges jeux de variations (au format BED ou VCF), à partir de données hétérogènes, j'ai développé plusieurs modules d'annotation et de conversion de ces annotations, présentés dans la figure 7.1, et que je détaille ci-dessous.

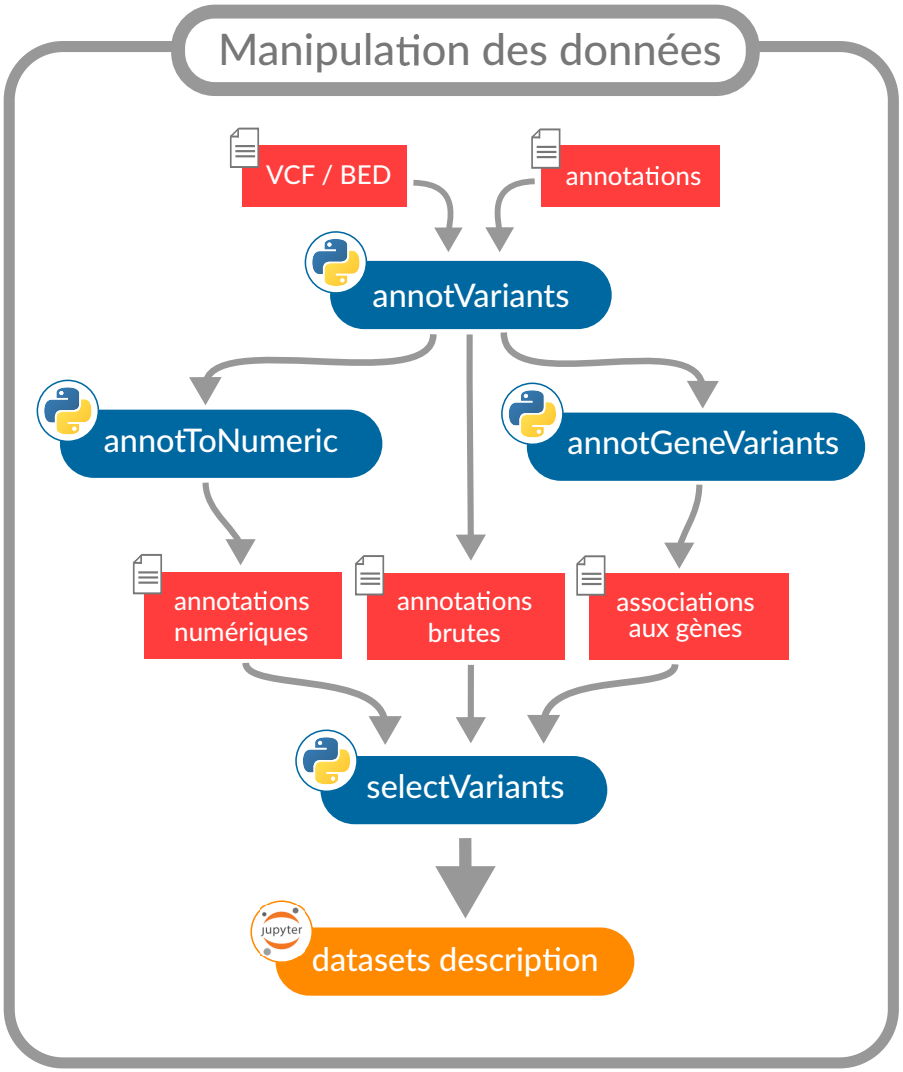


FIGURE 7.1 – Schéma des modules pour l'annotation et la sélection des variants.

Un premier module pour l'outil FINSURF, appelé **"annotVariants"**, permet d'interroger les différents fichiers d'annotations utilisés pour en extraire les informations associées aux variants évalués. Ce module a été développé en Python, et fait appel à plusieurs bibliothèques basées sur des outils permettant de manipuler les différents formats de fichiers des données d'annotations : pyBigWig (pour les fichier bigwig), pyBedTools (pour les fichiers BED). Un fichier de configuration permet de définir les jeux d'annotations à utiliser et les informations à lire depuis les sources d'annotations. Pour limiter l'impact sur la mémoire vive, le fichier de variations est annoté par bloc. En revanche pour maximiser la vitesse de

cette étape, chaque fichier d'annotation peut être lu et exploité en parallèle.

Un deuxième module, nommé **"annotToNumeric"**, est dédié à la conversion des annotations précédentes d'un format texte à un format numérique. En effet certaines des annotations sont obtenues sous forme de chaînes de caractères, afin de préserver une lecture intelligible. Par exemple pour les données de prédictions d'associations entre régions régulatrices et gènes, l'annotation est composée d'un score d'association, ainsi que des noms des gènes cibles associés, afin de pouvoir interpréter et sélectionner des variants d'intérêt grâce à cette information. Cependant, les modèles d'apprentissage machine utilisés ne permettent pas une utilisation directe de ce type de format d'annotations. Ce module "annotToNumeric" permet donc d'appliquer les conversions nécessaires, définies par un second fichier de configuration. Pour reprendre l'exemple des données de prédictions d'associations : ce deuxième module est capable d'extraire le score d'association depuis l'annotation formatée. Il permet également de calculer un degré d'association entre chaque variant et un ou plusieurs gènes, en évaluant la localisation de ces variants relatives aux gènes (fournis par l'utilisateur) depuis les annotations de gènes et les prédictions d'associations (voir le tableau 7.1, et l'annotation targets_associations).

Ces deux modules ont été appliqués sur les trois sources de variants décrites précédemment :

- des variants non-codants associées à des maladies héréditaires, et provenant de la base de données HGMD (nommés ci-après "HGMD-DM") ;
- les variants eQTLs provenant de la base de données GTEx, correspondant à des positions polymorphiques associées à des variations du niveau d'expression de gènes (nommés ci-après "eQTLs") ;
- les variants de la base de données ClinVar, identifiés comme n'ayant pas d'impact fonctionnel, et pris comme contrôles négatifs pour les entraînements de modèles (nommés ci-après "ClinVar").

Enfin, le module **"selectVariants"**, utilisé avec le notebook Jupyter "datasets description", permet de sélectionner les différents jeux de variants, qui seront à la base de l'entraînement des modèles de prédiction de FINSURF, que je présente au chapitre suivant.

7.2.2 Analyse des annotations associées aux variants

Comme mentionné dans la section précédente, j'ai identifié trois jeux de variants différents : deux qui correspondent à des ensembles de variants fonctionnels, et serviront

de variants contrôles positifs (les variants HGMD-DM et les variants eQTLs), et un ensemble de variants contrôles négatifs (ClinVar). Tous ces variants ont été annotés avec les différentes ressources d'annotations, qui peuvent être exploitées par les modèles d'apprentissage machine pour séparer les contrôles positifs des contrôles négatifs. Je présente dans cette section les différences identifiables entre les ensembles.

Composition et localisation des variants

J'ai tout d'abord remarqué que les variants HGMD-DM sont principalement localisés dans les séquences promoteurs et 5'UTR des gènes auxquels ils sont fonctionnellement associés. J'ai donc voulu confirmer ce biais de localisation à partir des annotations génomiques GENCODE, et évaluer la répartition génomique des variants provenant des deux autres ressources.

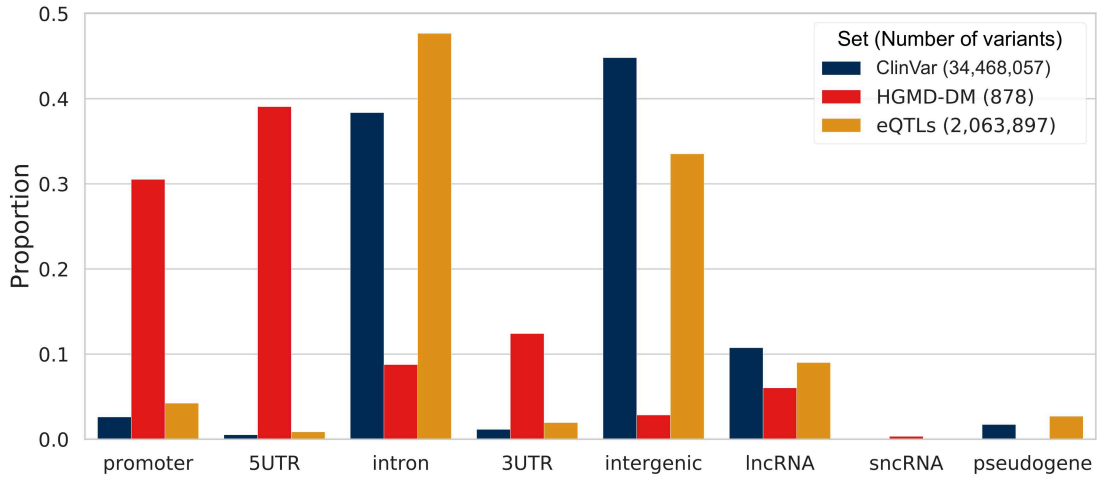


FIGURE 7.2 – Proportions des jeux de variants par annotation génomique. Les annotations génomiques correspondent aux annotations provenant de GENCODE (version 29). Les 3 jeux de données sont pris bruts (sans filtre), et une seule annotation est prise par variant (certains variants pouvant couvrir plusieurs annotations) selon leur degrés d'importance (voir méthodes).

Dans la figure 7.2, nous pouvons constater que les variants HGMD-DM sont effectivement concentrés dans les régions promotrices et UTR5' des gènes. En revanche, les jeux de variants eQTLs et ClinVar présentent des proportions beaucoup plus importantes dans les introns et dans les régions intergéniques. Les eQTLs constituent donc un ensemble de variants "fonctionnels" qui va nous permettre d'explorer des relations de régulation de

l'expression des gènes sur de plus longues distances que pour les variants HGMD-DM. Ces deux jeux de variants fonctionnels sont donc complémentaires pour mon objectif d'analyse des propriétés des variants non-codants régulateurs.

Définition de sous-ensembles de variants

Dans notre analyse, les variants ClinVar nous servent de variants contrôles négatifs. Comme nous l'avons vu, leur répartition génomique diffère de celles des variants contrôles positifs (HGMD-DM ou eQTLs). Or, nous souhaitons que la méthode d'apprentissage machine détecte des différences d'annotations fonctionnelles qui sont dues au caractère fonctionnel ou non-fonctionnel intrinsèque des variants contrôles, plutôt qu'elle ne détecte des différence dues au différences de localisations génomiques. Nous proposons donc trois méthodes d'échantillonnage des variants contrôles, pour ajuster les répartitions des contrôles négatifs à celles de chacun des ensembles de contrôles positifs :

— un échantillonnage des variants ClinVar à l'échelle du génome complet, prenant en compte la répartition des variants fonctionnels dans les régions génomiques définies par GENCODE (ci-après désignée GENCODE-match) ;
— un échantillonnage des variants ClinVar prenant en compte la répartition du jeu de variants fonctionnels dans les régions génomiques, mais fait exclusivement dans les cytobandes contenant les variants contrôles positifs (ci-après désignée Cytoband-match) ;
— un échantillonnage local des variants contrôles, pris à une distance maximale de mille paires de bases d'un variant fonctionnel (ci-après désignée Distance-match).

La différence entre l'approche GENCODE-match et l'approche Cytoband-match est essentiellement justifiée par le fait que le jeu de variants HGMD-DM est petit (878 variants au total) et restreint à certaines parties du génome (par exemple, le chromosome 11 concentre 144 des variants, tandis que le chromosome 18 n'en contient que 4). L'approche Cytoband-match vise donc à respecter cette restriction en échantillonnant les variants contrôles négatifs dans espace génomique plus proche de celui des variants contrôles positifs.

J'ai donc appliqué ces méthodes d'échantillonnage pour définir des jeux d'entraînements basés sur les variants HGMD-DM, et des jeux d'entraînement basés sur les variants eQTLs. Concernant ces derniers : puisque les modèles de prédiction ont vocation à être appliqués dans un cadre diagnostique, j'ai choisi de réduire le jeu de variants eQTLs à un

sous-jeu d'eQTLs affectant spécifiquement l'expression de gènes associés à des maladies ; ces gènes provenant de la base de données OMIM, les modèles basés sur cette sélection seront nommés par la suite "eQTLs-OMIM".

Le tableau 7.2 résume l'ensemble des comptages et jeux de données générés par ces sélections.

		Sans sélection	
Modèle	**Jeu de variants**	**Nombre initial de variants**	**Nombres de variants après filtres (codants, régions exclues)**
HGMD-DM	HGMD-DM	878	878
	ClinVar	38 056 330	34 468 057
eQTLs-OMIM	eQTLs	695,454 (Total = 3,124,345)	465,134 (Total = 2,063,897)
	ClinVar	38 056 330	34 468 057

		Sélection contrôles : GENCODE-MATCH		
Modèle	**Jeu de variants**	**Nombres de variants gardées**	**Nombres exclus (INDEL ; chevauchement de positifs)**	**Nombre final (dont X% de positifs)**
HGMD-DM	HGMD-DM	878	(0 ; 0)	**1 780 981 (0.0493%)**
	ClinVar	3 061 561	(1,281,417 ; 41)	
eQTLs-OMIM	eQTLs	872 344	(0 ; 0)	**36 292 018 (1.61%)**
	ClinVar	39 630 252	(0 ; 538 927)	

		Sélection contrôles : CYTOBAND-MATCH		
Modèle	**Jeu de variants**	**Nombres de variants gardées**	**Nombres exclus (INDEL ; chevauchement de positifs)**	**Nombre final (dont X% de positifs)**
HGMD-DM	HGMD-DM	878	(0 ; 0)	**682 081 (0.129%)**
	ClinVar	1 179 844	(498 598 ; 43)	
eQTLs-OMIM	eQTLs	872 344	(0 ; 0)	**35 426 175 (1.65%)**
	ClinVar	38 509 978	(0 ; 543 548)	

		Sélection contrôles : DISTANCE-MATCH		
Modèle	**Jeu de variants**	**Nombres de variants gardées**	**Nombres exclus (INDEL ; chevauchement de positifs)**	**Nombre final (dont X% de positifs)**
HGMD-DM	HGMD-DM	877	(0 ; 0)	**6 985 (12.6%)**
	ClinVar	10 806	(4 652 ; 46)	
eQTLs-OMIM	eQTLs	872 344	(0 ; 0)	**7 477 028 (7.8%)**
	ClinVar	8 207 726	(0 ; 617 385)	

TABLE 7.2 – Comptages de variants sélectionnés pour entraîner les modèles

A noter que pour les étapes d'entraînement, un filtre supplémentaire est appliqué sur les jeux de variants sélectionnés : il consiste à enlever les variants affectant plusieurs bases (Insertions et délétions, ou INDEL), et les variants du jeu contrôle dont les positions chevauchent des positions de variants fonctionnels. Concernant le filtre des INDEL : celui-ci n'est appliqué que pour les modèles HGMD-DM. En effet pour ces derniers, les contrôles positifs ne contenant pas d'INDEL, le fait d'en inclure parmi les contrôles négatifs risquerait d'indiquer au modèle d'apprentissage que le caractère négatif est associé au fait d'être un INDEL.

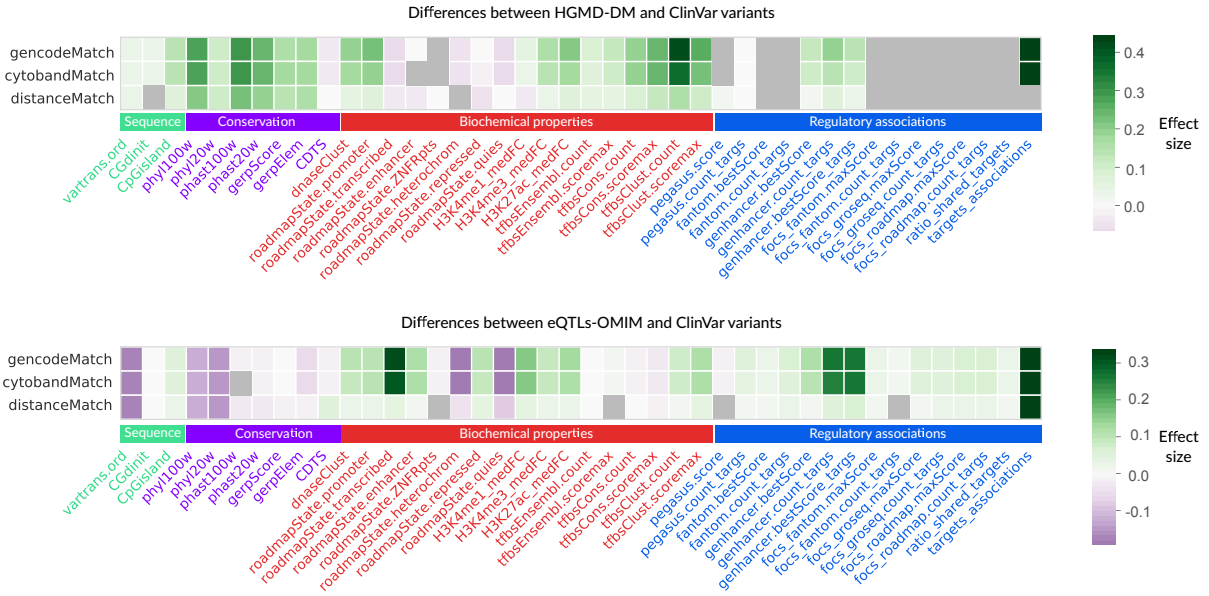


FIGURE 7.3 – Heatmap des différences entre variants contrôles positifs et négatifs. Les différences sont représentées par leur taille d'effet mesurée entre les contrôles positifs et négatifs (voir méthodes), et indiquent si la valeur d'une annotation est plus élevée dans les positifs (couleur verte), ou plus élevée dans les négatifs (couleur violette). Les cases grisées correspondent aux annotations pour lesquelles la différence n'est pas significative (tests : Mann–Whitney U test ou Chi2 de contingence, en fonction du type de distribution associée à l'annotation ; correction multiple par la méthode de Holm-Bonferroni).

Vision générale des annotations et différences

Une fois annotés, j'ai souhaité évaluer les différences mesurables entre les variants identifiés comme contrôles positifs (HGMD-DM et eQTLs-OMIM) et les échantillons de variants contrôles négatifs associés. Ces différences sont la base sur laquelle les modèles de prédiction vont pouvoir être entraînés ; les variants ayant été sélectionnés pour leur fonctionnalité a priori, je souhaite vérifier que les annotations reflètent cette fonctionnalité, en comparaison avec les contrôles négatifs issus de la base de données ClinVar.

Les différences entre contrôles positifs et négatifs sont mesurées par la taille d'effet, équivalent à une différence de proportions pour les variables binaires (coefficient h de Cohen) ou à une différence de moyennes normalisée pour les autres types de distributions (coefficient d de Cohen).

On peut constater de manière générale que les variants fonctionnels (HGMD-DM non-codant, ou eQTLs) présentent effectivement des différences marquées en leur faveur, qui confirment nos attentes : les annotations associées à des propriétés régulatrices et/ou

fonctionnelles ont des valeurs plus élevées pour ces variants par rapport aux variants contrôles :

- des annotations comme les marques d'histones (H3K27ac, H3K4me3, H3K4me1), ou les états chromatiniens prédits (notamment promoteur), ou encore les annotations de site de fixations de facteurs de transcription, présentent des valeurs plus élevées pour les contrôles positifs en comparaison avec les contrôles négatifs (ce qui est particulièrement visible pour les pistes "gencodeMatch" et "cytobandMatch" ;
- en revanche pour les états chromatiniens "réprimés" et "quiescents", les tailles d'effet indiquent que les contrôles négatifs ont des valeurs plus élevées que les contrôles positifs pour ces annotations ;
- les annotations correspondant aux associations entre régions régulatrices et gènes cibles montrent également une différence en faveur des contrôles positifs, plus prononcée dans le cas des variants eQTLs que dans le cas des variants HGMD-DM. On peut remarquer pour les ensembles basés sur les eQTLs qu'une différence est détectable et significative pour toutes les annotations de cette catégorie, excepté pour l'annotation du nombre de gènes cibles associés aux enhancers transcrits identifiés par FOCS (focs_fantom.count_targs) pour le schéma d'échantillonnage Distance-Match). En revanche pour les ensembles basés sur les variants HGMD-DM, seules les annotations associées à la ressource Genehancer sont détectées comme significatives. On détecte par ailleurs une forte association génomique des contrôles positifs à leur gène cible (colonne "targets_associations").

Concernant les annotations de conservation, on peut observer que les variants HGMD-DM sont plus conservés en moyenne que les variants ClinVar. Cela reste vrai pour tous les schémas de sélection des contrôles négatifs, même si la différence s'atténue légèrement pour les variants ClinVar pris dans un rayon de mille pairs de bases autour des variants fonctionnels (Distance-match). En revanche pour les variants eQTLs, il apparaît que ces variants sont moins conservés que les contrôles négatifs (ceci est particulièrement visible pour les annotations phyl100w et phyl20w).

C'est une différence qui est potentiellement surprenante, mais qui peut s'expliquer par la nature des eQTLs : ces positions sont principalement identifiées comme eQTLs grâce à une association statistique entre des combinaisons alléliques et des niveaux d'expression de gènes. Or il existe une corrélation négative entre la fréquence allélique d'un variant alternatif et son association statistique : plus il est rare, plus il faudra un effet prononcé

sur l'expression pour qu'il soit détecté. Par conséquent, les eQTLs sont enrichis en variants montrant une forte variabilité dans la population humaine, ce qui est cohérent avec leur faible conservation évolutive entre espèces.

Enfin, je souhaite souligner les conséquences de l'échantillonnage Distance-match sur les mesures de différences entre contrôles positifs et négatifs, pour les annotations correspondant à des intervalles génomiques, comme les prédictions de régions régulatrices, les prédictions d'états chromatiniens, ou encore les sites de fixation de facteurs de transcription. On constate en effet que les différences détectables pour les schéma d'échantillonnage GENCODE-match et Cytoband-match s'atténuent fortement pour le troisième schéma d'échantillonnage, lorsque les variants ClinVar ont été pris dans un rayon de mille paires de bases des variants fonctionnels. Cette diminution est flagrante par exemple pour les annotations "dnaseClust" (identifiant les régions de chromatine ouverte, potentiellement accessible pour des facteurs de trancription pour pour la machinerie de transcription), "roadmapState.promoter" (identifiant les régions dont les marques d'histones indiquent un état promoteur), ou encore "tfbsClust.count" (indiquant une concentration de sites de fixation de facteurs de transcription), que l'on regarde la heatmap des variants HGMD-DM, ou celle des variants eQTLs-OMIM. Cette diminution est également vraie pour les annotations correspondant aux associations entre régions régulatrices et gènes cibles. Ce schéma d'échantillonnage est le plus strict : il vise à tester la résolution des modèles de prédiction, et notamment à évaluer si ces derniers se contentent d'utiliser certaines annotations d'intervalles (ce qui ne permettrait pas de discriminer deux positions dans ces intervalles).

En conclusion, les annotations choisies pour caractériser la fonctionnalité de variants non-codants sont effectivement associées à des différences détectables entre les contrôles positifs et négatifs qui serviront à entraîner nos modèles de prédiction. Les deux jeux de variants fonctionnels ne sont pas associés aux mêmes propriétés biologiques, ce qui se traduit par des distinctions entre les profils de différences avec les contrôles négatifs. Ces deux jeux conduiront donc probablement à des modèles de prédiction de variants fonctionnels non-codants qui ne détecteront pas les mêmes signaux dans les annotations.

Pouvoir discriminant individuel des annotations

Pour justifier l'intérêt de la création d'un modèle d'apprentissage machine basé sur un ensemble d'annotations, il est intéressant d'évaluer le pouvoir discriminant associé

à chacune de ces annotations : en effet, si une annotation individuelle apparaît comme aussi bonne ou meilleure que le modèle complet pour discriminer les variants fonctionnels des non-fonctionnels, alors il est suffisant d'utiliser cette annotation comme prédicteur de fonctionnalité.

Pour montrer la capacité de chaque annotation à séparer les variants contrôles positifs (HGMD-DM ou eQTLs-OMIM) des contrôles négatifs, j'utilise pour chaque annotation une courbe ROC et une courbe PR, dont sont calculées les aires sous les courbes (voir la figure 6.1). Ces dernières servent de coordonnées pour visualiser le pouvoir discriminant des annotations, sachant qu'un prédicteur idéal correspondrait à un point de coordonnées (1, 1), tandis qu'un prédicteur aléatoire aurait pour aire sous la courbe ROC une valeur de 0.5, et une valeur d'aire sous la courbe PR correspondant à la proportion de variants fonctionnels sur le jeu total (voir table 7.2).

La figure 7.4 présente le pouvoir discriminant individuel pour l'ensemble des annotations présentées précédemment (groupées par classes fonctionnelles), pour les trois stratégies de contrôles négatifs. On remarque que certaines annotations présentent des pouvoirs discriminants élevés, en particulier pour les annotations de conservation en séquence : les meilleurs correspondent aux annotations gerpElem (évaluant la présence des variants dans des éléments conservés), et phast100w (mesurant la probabilité d'un variant d'appartenir à un élément conservé). On peut également noter le bon pouvoir discriminant associé aux annotations de sites de facteurs de transcriptions : tfbsClusts et tfbsCons, ainsi que les états promoteurs. Parmi les annotations de prédictions de régulation, seules les annotations provenant des prédictions d'association Genehancer permettent de détecter des variants fonctionnels (valeur d'aire sous la courbe ROC élevée), mais avec une très mauvaise précision (valeur sous la courbe PR faible). L'annotation "targets_associations", qui quantifie le degré d'association entre un variant et son gène cible, est également un bon prédicteur de fonctionnalité. Ce commentaire est vrai pour les schémas de sélection "GENCODE-match" et "Cytoband-match", mais pas pour le schéma d'échantillonnage Distance-match.

Ces bons prédicteurs, identifiés par les aires sous les courbes ROC et PR, correspondent effectivement à des annotations dont les différences mesurées entre contrôles positifs et négatifs étaient importantes (voir la figure 7.3).

Une figure similaire est proposée pour les variants eQTLs-OMIM (figure 7.5). Les annotations apparaissent comme de moins bons prédicteurs de manière générale (en comparai-

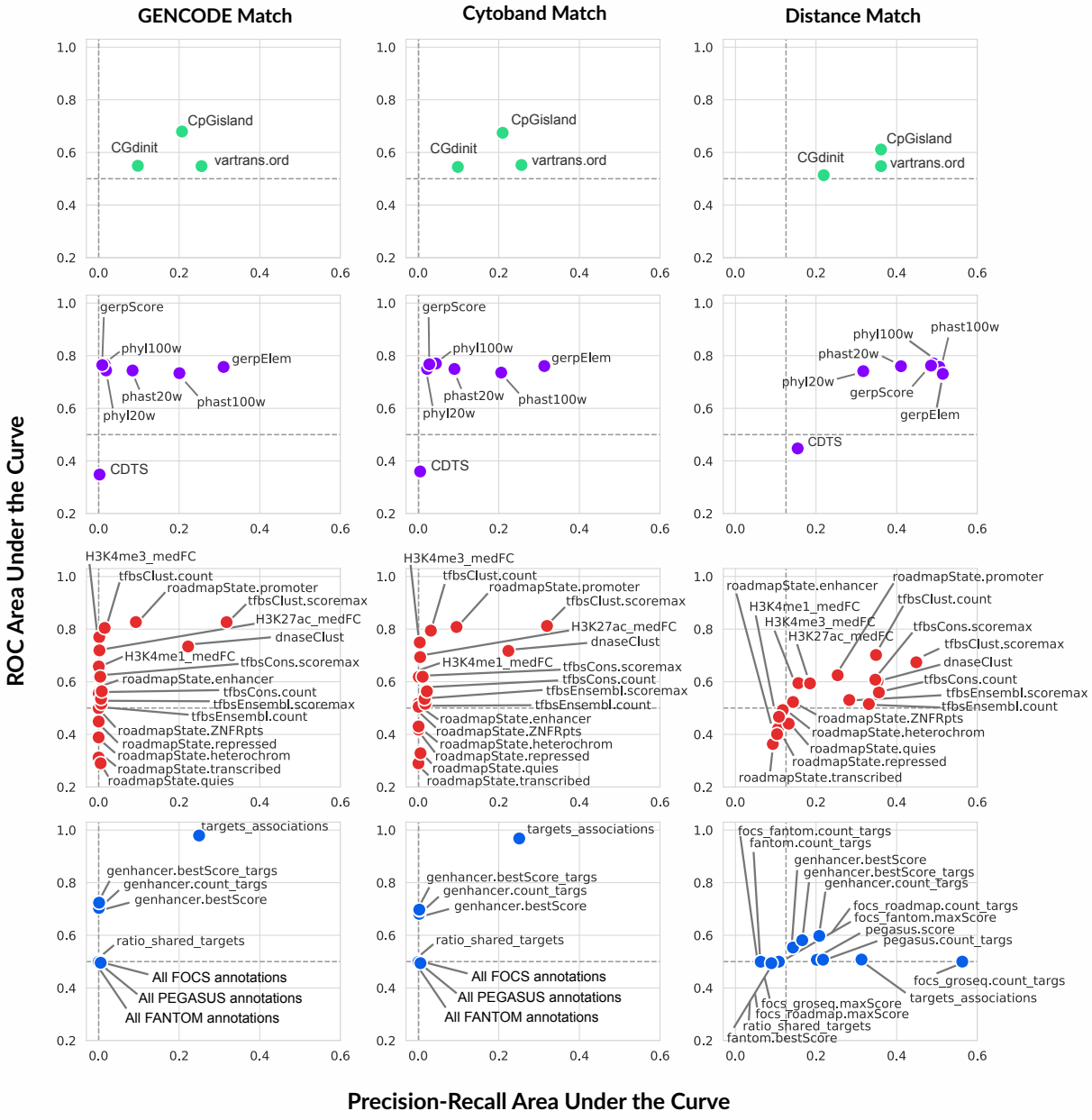


FIGURE 7.4 – Pouvoir discriminant des annotations prises indépendemment pour les variants HGMD-DM. La capacité de discrimination d'une annotation est résumée à son aire sous la courbe ROC (ROC AUC, axe vertical) et à son aire sous la courbe "Precision-Recall" (Precision-Recall Area Under the Curve, axe horizontal). Le pouvoir discriminant de chaque annotation est présenté pour les 3 jeux de variants contrôles. Les lignes pointillées correspondent aux valeurs d'aires sous les courbes pour un modèle aléatoire.

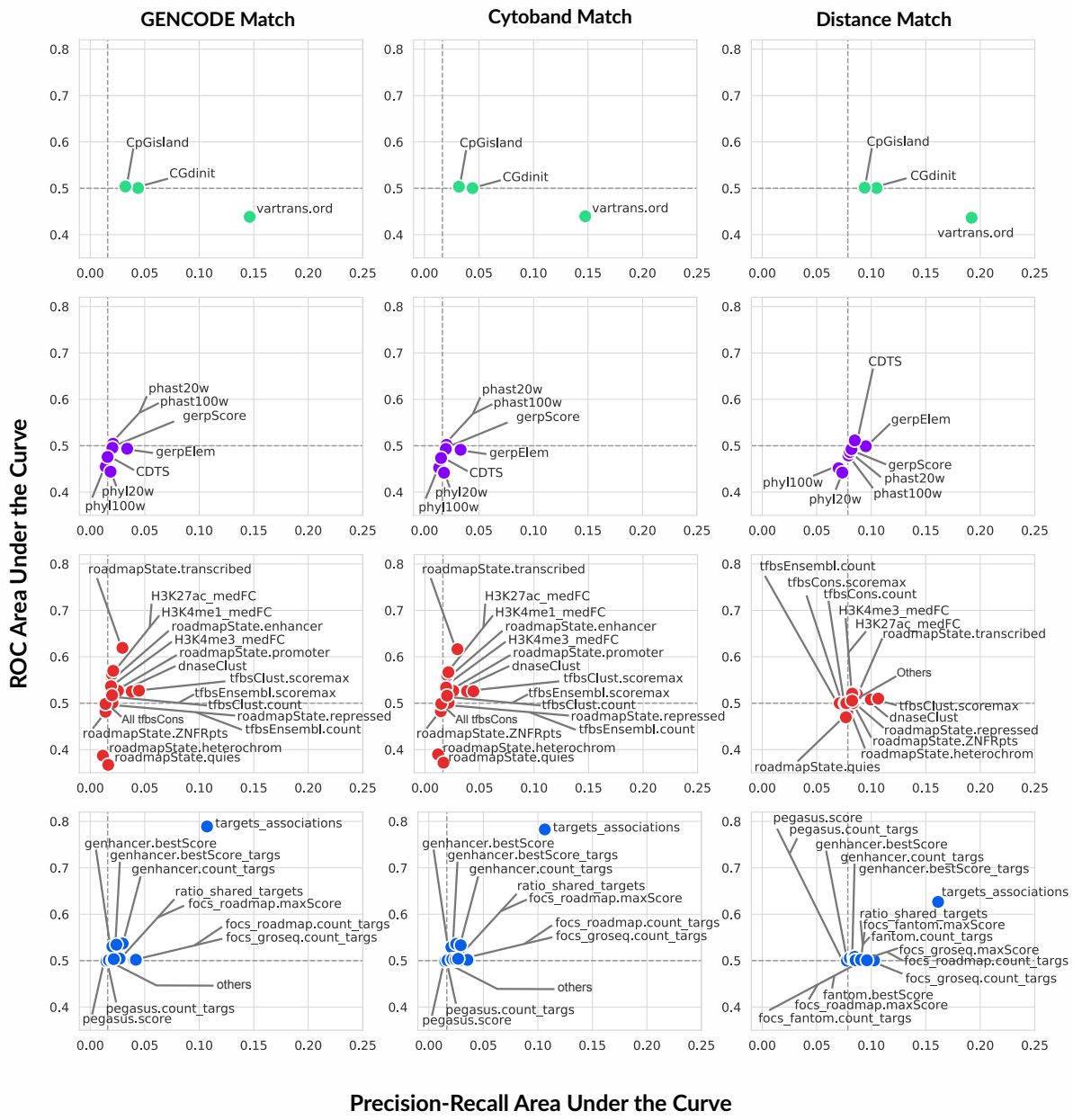


FIGURE 7.5 – Pouvoir discriminant des annotations prises individuellement pour les variants eQTLs-OMIM. La capacité de discrimination d'une annotation est résumée à son aire sous la courbe ROC (ROC AUC, axe vertical) et à son aire sous la courbe "Precision-Recall" (Precision-Recall Area Under the Curve, axe horizontal). Le pouvoir discriminant de chaque annotation est présenté pour les 3 jeux de variants contrôles. Les lignes pointillées correspondent aux valeurs d'aires sous les courbes pour un modèle aléatoire.

son avec les valeurs observées pour les jeux HGMD-DM). Les annotations de conservation séparent très mal les contrôles positifs des contrôles négatifs pour ces jeux de variants, et apparaissent comme de très mauvais prédicteurs. Les seules annotations qui présentent des valeurs d'aires sous les courbes notables sont les annotations d'état chromatinien transcrit (roadmapState.transcribed), probablement du fait de la localisation des eQTLs en majorité dans les introns, et les marques d'histones H3K4me1 et H3K27ac, qui sont effectivement associées à des propriétés de régions régulatrices. Par ailleurs l'annotation "targets_associations" apparaît comme un bon prédicteur, confirmant que les eQTLs tendent à être génomiquement associés à leurs gènes cibles (soit par leur présence dans le gène, soit par leur présence dans un élément de régulation prédit pour cibler le gène).

7.3 Conclusions

Nous avons confirmé l'intérêt des jeux de variants identifiés comme exemples de variants non-codants régulateurs : ces variants présentent effectivement des propriétés d'annotations distinctes de variants contrôles négatifs.

J'ai défini différents schémas d'échantillonnage de ces variants contrôles négatifs, pour évaluer l'influence de ces sélections sur la capacité des différentes annotations à séparer les contrôles positifs des négatifs, pour des sélections de plus en plus strictes. Ces sélections permettent d'explorer différentes questions biologiques : notamment, prendre des variants contrôles négatifs à une faible distance des variants contrôles positifs permet potentiellement d'entraîner un modèle dédié à l'identification des positions fonctionnelles au sein d'une région régulatrice par exemple ; en revanche les sélections GENCODE-match et Cytoband-match, plus relâchées, permettent d'explorer des variants à l'échelle du génome.

Les annotations prises individuellement présentent des capacités de discriminations variables, assez faibles, et cela justifie de chercher à les intégrer au sein d'un modèle d'apprentissage machine, qui pourra les combiner pour atteindre un pouvoir discriminant plus élevé. Certaines annotations présentent des pouvoirs discriminants très faibles (proche de classificateurs aléatoires), et sont associés à des tailles d'effets très faibles (voir la figure 7.3). Par exemple les annotations relatives aux dinucléotides CG, ou encore certains des états chromatiniens (par exemple l'état enhancer) présentent des différences faibles entre contrôles positifs et négatifs, et sont associés à des pouvoirs discriminants faibles. Mais la capacité des modèles d'apprentissage utilisés dans ce projet (les forêts aléatoires) à combiner les annotations de nombreuses manières permettra potentiellement d'exploiter

plus efficacement ces annotations.

Nous verrons donc dans les chapitres suivants les étapes d'entraînement de ces modèles de prédiction, avec l'influence de ces différents schémas d'échantillonnage des contrôles négatifs, ainsi que les différences entre les modèles basés sur les variants HGMD-DM et ceux basés sur les variants eQTLs-OMIM.

Chapitre 8

Conception des modèles de prédiction

Durant les étapes initiales de mon projet de thèse, j'avais tenté de créer empiriquement un score unique par position génomique, combinant les différentes annotations présentées dans le chapitre précédent. Ce score correspondait à une simple somme de valeurs dérivées depuis les quatre classes fonctionnelles d'annotations (annotations de séquence, conservation en séquence, annotations biochimiques, et prédictions d'associations régulatrices). Si cette méthode conduisait effectivement à donner des scores plus élevés pour des variants fonctionnels en comparaison avec des variants contrôles, il était difficile d'évaluer de manière correcte les contributions de chaque annotation. Par ailleurs, une évaluation plus en détails de la littérature m'avait conduit à identifier la base de données "RegulomeDB" (Boyle et al., 2012), qui proposait une séparation similaire des variations en plusieurs catégories, chacune associée avec une combinaison spécifique d'annotations de régulation. J'ai donc souhaité mettre en place une approche statistique plus formelle, basée une méthode d'apprentissage machine de forêts aléatoires, et que je présente dans ce chapitre.

8.1 Définition du pipeline et présentation des modèles explorés

L'outil FINSURF s'articule autour de modules dédiés aux différentes étapes de la création des modèles de classification, entraînés à séparer des variants fonctionnels, de

variants non-fonctionnels.

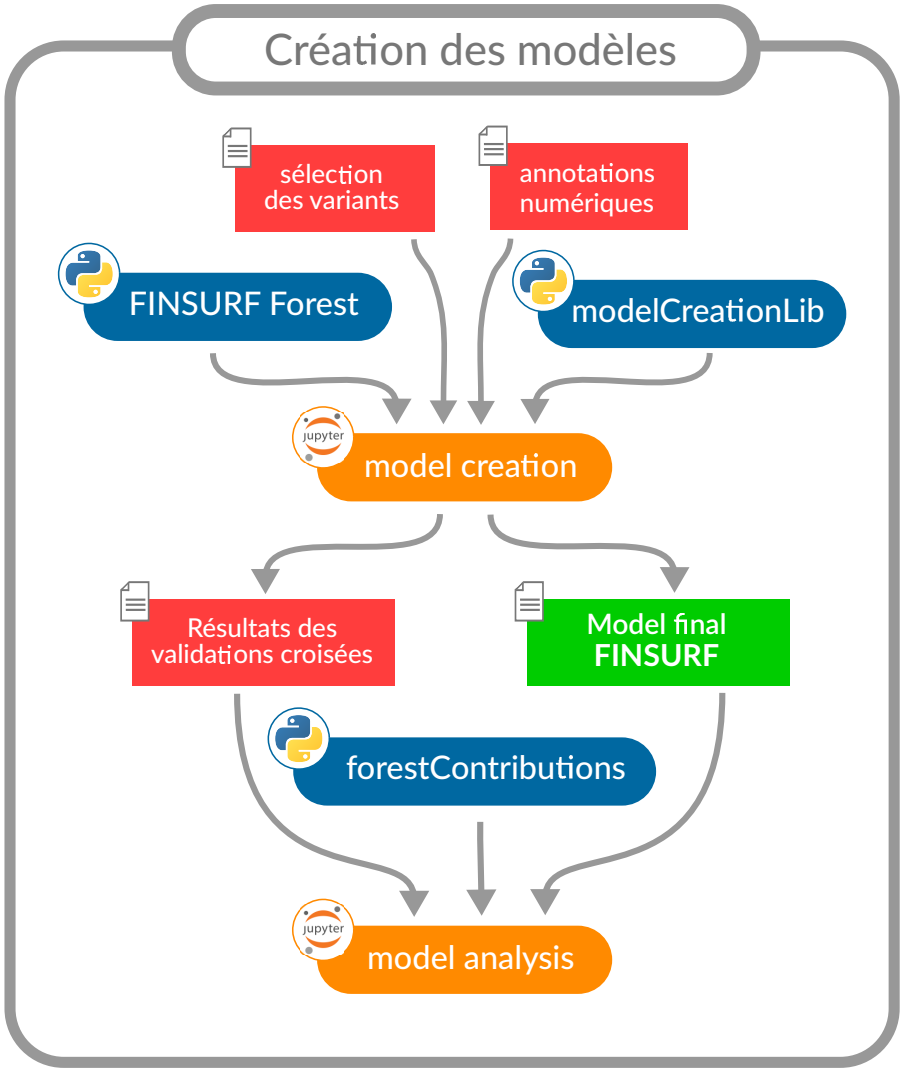


FIGURE 8.1 – Schéma des modules pour la création des modèles FINSURF. La première étape est dédiée à la création des différents modèles, ainsi que la génération des résultats de la procédure de validation croisée des modèles. Ces résultats de ces validations sont compilés et évalués dans une seconde étape, grâce à un notebook Jupyter dédié à l'analyse des modèles.

Les fichiers d'annotations des variants sélectionnés (présentés au chapitre précédent) sont donc utilisés ici pour l'entraînement des modèles de prédiction. Comme montré sur la figure 8.1, deux modules Python interviennent à cette étape.

FINSURF Forest. J'ai modifié l'implémentation des forêts aléatoires proposée dans la bibliothèque Scikit-learn, afin d'obtenir une version de l'algorithme capable de gérer les différences de proportions des classes. En effet, comme présenté au chapitre d'introduction aux outils d'intégration de données, les noeuds dans les arbres d'une forêt aléatoire sont définis par optimisation de la séparation des différentes classes présentes. Or si une classe est sur-représentée par rapport à une autre, le taux d'erreurs sera biaisé par ce déséquilibre, et l'algorithme tendra à ignorer la classe minoritaire (cette dernière ne contribuant que faiblement au calcul du taux d'erreurs de classification). L'implémentation par défaut n'est pas capable de gérer ces différences.

Différentes méthodes existent pour résoudre ce problème au niveau du jeu de données, soit par réduction de la classe majoritaire, soit par génération de nouvelles entités pour la classe minoritaire (Chawla et al., 2002, Maldonado et al., 2019). Il est cependant possible d'inclure une étape de ré-équilibrage des classes au sein de l'algorithme des forêts aléatoires : plutôt que d'équilibrer avant d'entraîner le modèle, l'équilibrage est fait pendant sa génération. A chacun des arbres, l'étape de "Bootstrap Aggregation" est appliquée de manière à ce que le nombre d'entités prises au hasard dans la classe majoritaire soit équivalent au nombre d'entités de la classe minoritaire. La modification que j'ai apportée au code source de l'algorithme des forêts aléatoires n'affecte qu'une seule fonction de la bibliothèque Scikit-Learn, permettant ainsi un impact minime sur l'utilisation de cet algorithme.

modelCreationLib. Durant la création des modèles de prédiction, j'ai souhaité explorer et corriger les biais qui peuvent être introduits par inadvertance durant les différentes étapes de l'entraînement d'un modèle de prédiction, et qui conduisent à fausser un modèle. Ce deuxième module contient donc les différentes fonctions et méthodes utilisées pour la génération des modèles de prédiction, pour leur validation interne au moment des entraînements, et pour l'évaluation des biais, que je présente ci-dessous.

Une première source de biais, qui n'est pas spécifique à l'analyse de séquences génomiques, provient de l'inférence des valeurs manquantes. Il est tentant de pré-traiter un ensemble d'annotations en amont de l'entraînement, en remplissant les valeurs manquantes à partir de l'entièreté du jeu de données (par exemple, en utilisant la valeur moyenne d'une annotation sur le jeu total, ou encore en utilisant la moyenne par classe). Cela n'est pas une approche correcte : en effet, au moment de l'entraînement du modèle, on cherche à évaluer ce dernier par séparation du jeu d'entraînement en sous-jeux (voir le chapitre des

méthodes). Or, si l'imputation est effectuée sur l'entièreté du jeu de données, cela signifie que les valeurs imputées pour les entités du jeu de validation le sont en partie grâce aux valeurs des entités du jeu d'entraînement : les deux jeux ne sont donc plus indépendants, faussant cette étape de validation (ce problème est appelé "data leakage", ou "fuite d'information"). Le module "modelCreationLib" et l'utilisation du notebook Jupyter "model creation" permettent de gérer correctement cette situation par l'établissement d'un pipeline combinant une étape d'inférence (basée sur la bibliothèque Python "skearn_pandas" et une étape d'entraînement du modèle de prédiction ; ce pipeline, par son implémentation, permet que l'inférence des valeurs manquantes soit faite indépendemment du jeu d'entraînement pour le jeu de validation, prévenant ainsi une fuite d'information depuis le premier.

Une seconde source de biais provient de la façon dont est effectuée la séparation des entités du jeu d'entraînement de celles du jeu de validation. Ce biais est spécifique à l'analyse de séquences génomiques (et des problèmes faisant intervenir un lien séquentiel entre entités) dans le sens où les entités manipulées (les variants) ne sont pas indépendantes les unes des autres : les variants sont définis par des positions génomiques, et leur plus ou moins grande proximité est corrélée avec un partage potentiel d'annotations. Cela signifie que lors de l'assignation aléatoire des entités aux jeux d'entraînement et de validation, il est possible que deux entités proches soient séparés chacune dans un de ces groupes ; dans ce cas le modèle aura potentiellement plus de facilités à assigner l'entité du jeu de validation à sa classe correcte, faussant l'évaluation. Certaines des méthodes décrites au chapitre d'introduction proposent des évaluations prenant en compte ce lien : par exemple FATHMM-MKL et FIRE proposent une séparation "entraînement/validation" basée sur les chromosomes ; le score ReMM utilisé dans Genomiser propose une séparation sur les bandes cytogénétiques. J'ai choisi une séparation similaire à cette dernière, en implémentant une fonction d'évaluation des modèles de prédiction qui isole une partie des bandes cytogénétiques pour validation, tandis que les variants présents dans les autres bandes cytogénétiques sont utilisés pour l'entraînement.

Les résultats de ces explorations sont présentés à la fin de ce chapitre, pour des modèles de classification entraînés sur les différents jeux de variants fonctionnels et non-fonctionnels définis au chapitre précédent. Avant cela, je présente dans la section suivante le choix des hyper-paramètres utilisés pour ces modèles.

8.2 Optimisation et choix des hyper-paramètres

Le modèle des forêts aléatoires a été décrit en détail dans le chapitre d'introduction dédié aux méthodes d'intégration de données. Différents paramètres ont été évoqués, permettant de décrire la structure du modèle lors de l'apprentissage. Certains de ces paramètres sont définis et optimisés pendant l'apprentissage, et dépendent des données : ils correspondent à l'ensemble des choix d'annotations et des seuils associés à chacun des noeuds des arbres composant la forêt. Un autre groupe de paramètres correspond à la définition de la structure du modèle d'apprentissage, et ils sont définis avant l'étape d'entraînement : ce sont les "hyper-paramètres".

Dans la bibliothèque Python Scikit-learn, chaque modèle de prédiction est proposé avec des valeurs par défaut pour ces hyper-paramètres. Bien que ces modèles permettent généralement d'obtenir des résultats corrects, il est intéressant de pouvoir explorer un ensemble de combinaisons plus élaborées, et d'évaluer les améliorations potentielles pour les modèles associés. En effet il n'existe pas vraiment de règles permettant d'établir a priori les paramètres optimaux ; seule l'exploration empirique permet d'ajuster les paramètres adaptés à la stratégie expérimentale d'intérêt. Cette exploration de paramètres est rendue possible par l'objet "GridSearchCV" dans la bibliothèque Scikit-learn. Sont définis dans cet objet les listes de valeurs à tester pour chacun des paramètres du modèle. L'objet "GridSearchCV" va alors entraîner un modèle pour chacune des combinaisons, et l'évaluer sur une validation croisée. Un score de qualité de classification est obtenu depuis cette validation croisée, et permet ainsi d'ordonner l'ensemble des modèles testés selon ce score. L'objet "GridSearchCV" permet finalement de sélectionner le meilleur modèle identifié grâce à ce score, et donc la meilleure combinaison de paramètres.

Cependant, il est nécessaire de vérifier que l'optimisation a bien conduit à une amélioration notable de la qualité de classification, en comparaison avec un modèle "par défaut". Il est important que cette comparaison soit faite sur un jeu de données indépendant, plutôt que d'utiliser directement le score obtenu lors de l'optimisation : si l'on utilise directement ce dernier, il est possible que l'optimisation soit spécifique au jeu de données utilisé, mais ne corresponde pas à un modèle généralisable avec une bonne qualité de classification sur de nouvelles entités à classifier.

La documentation officielle de la bibliothèque Scikit-learn fournit une démonstration de l'objet "GridSearchCV", avec une proposition de validation indépendante, faite en validation croisée (`https://scikit-learn.org/stable/auto_examples/model_selection/plot_`

`nested_cross_validation_iris.html`). Cependant, dans l'exemple donné, une nouvelle validation croisée est faite pour évaluer le meilleur modèle identifié, mais elle opère directement sur le jeu de données qui a servi à l'optimisation. Dans ce cas, cette évaluation ne permet pas d'évaluer la généralisation de la qualité du classificateur, et peut conduire à une situation de sur-apprentissage.

J'ai donc réalisé une optimisation et validation de paramètres par validations-croisées imbriquées, garantissant une indépendance entre le jeu de données servant à l'optimisation des paramètres, et le jeu de données "test" servant à évaluer la qualité réelle du classificateur optimisé.

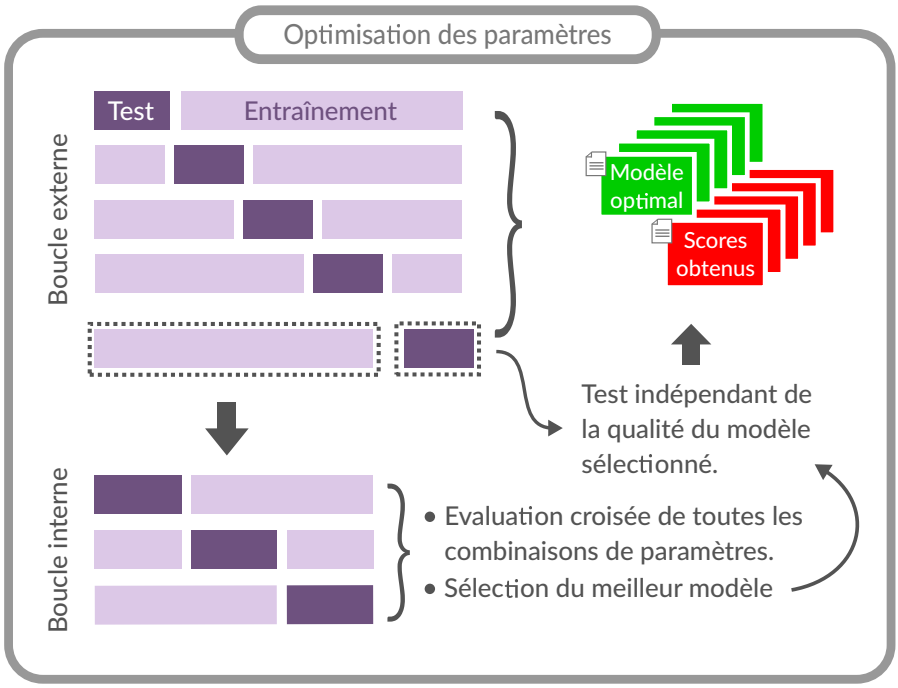


FIGURE 8.2 – Schéma des validations croisées imbriquées pour l'optimisation des paramètres d'un modèle d'apprentissage machine, avec évaluation indépendante. Pour un ensemble fini de valeurs de paramètres à explorer, les différentes combinaisons de ces valeurs sont utilisées pour entraîner des modèles, sur la boucle interne. La validation croisée dans cette boucle interne permet de sélectionner automatiquement la meilleure combinaison parmi toutes celles explorées. La validation externe permet un test indépendant de la qualité du modèle sélectionné. Ici la validation externe correspond à une séparation en K=5 groupes, tandis que la séparation interne est faite en 3 groupes.

Comme présenté dans la figure 8.2, le jeu de données initial est séparé en plusieurs groupes, en respectant les proportions de chacune des classes : c'est la boucle externe.

Paramètre	Valeurs explorées	Question
Nombre d'arbres	100 à 2000, par 200	Le nombre d'arbre augmente théoriquement l'exploration de combinaisons de descripteurs : est-ce associé à une amélioration de la qualité du modèle ?
Profondeur des arbres	1, puis 5 à 35, de 5 en 5	Par arbre, la profondeur des branches est associée à une combinaison de séparations dichotomiques : augmenter la profondeur améliore-t-il la qualité du modèle ?
Nombre d'entités par feuille	1 à 9, par 2	Plus il y a d'entités par feuille, plus le chemin de décision de la racine à la feuille est "générale" ; quel est l'impact sur la qualité du modèle ?

TABLE 8.1 – Paramètres choisis pour optimisation.

Chaque sous-groupe est alors mis de côté en tant que jeu de "test", tandis qu'une évaluation et optimisation de paramètres est faite avec l'objet "GridSearchCV" sur les groupes restants. Cette évaluation est faite par validation croisée, interne au sous-groupe ("boucle interne"). Elle permet d'obtenir pour l'ensemble des combinaisons de paramètres un score moyen de qualité de classification pour chacun des modèles, et donc de leur assigner un rang. Il est alors possible d'évaluer le meilleur modèle sélectionné, en le testant de manière indépendante sur le sous-groupe initialement mis de côté. L'optimisation est répétée K fois, ce qui permet d'évaluer une convergence indépendante des paramètres optimisés.

J'ai ainsi exploré un ensemble de 400 combinaisons de paramètres, dont l'optimisation a été faite par calcul de qualité de classification sur une validation croisée interne en 3 groupes. Cette qualité de classification est évaluée sur le score F1, permettant de maximiser à la fois la sensibilité et la précision du modèle. L'optimisation a été répétée 5 fois, correspondant à une séparation externe en 5 sous-groupes utilisés pour évaluation indépendante. Ce sont donc 600,000 entraînements qui ont été effectués durant cette étape. Les paramètres qui ont été évalués sont résumés dans le tableau 8.1 ci-dessous.

La figure 8.3 présente les résultats obtenus à partir des validations croisées internes, sur l'ensemble des combinaisons de paramètres. J'ai souhaité évaluer les tendances générales d'association entre la valeur d'un paramètre donné et la qualité de classification. On peut constater que les deux paramètres les plus critiques sont la profondeur des arbres et le nombre d'entités par feuille. Sur cette figure, il est assez net que plus les arbres sont profonds (donc plus le nombre de combinaisons explorées est grand), meilleure est la qualité. Cette amélioration semble converger assez rapidement cependant, et à partir d'une profondeur de 10 (donc un maximum de 10 décisions combinées au sein des chemins de décision), la qualité du modèle ne semble plus affectée par ce paramètre. En revanche, on peut observer que forcer un nombre d'entités important par feuille conduit à une dégradation de la qualité du classificateur. Enfin, il semble que le nombre d'arbre dans la forêt ne semble pas être un facteur extrêmement décisif sur la qualité du classificateur.

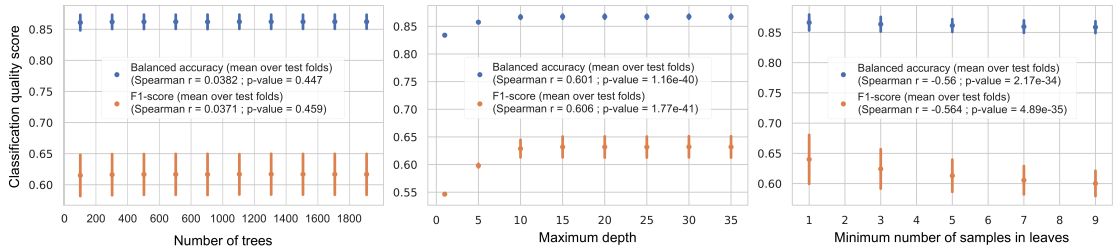


FIGURE 8.3 – Influence de la variation de trois paramètres sur la qualité de la classification par un modèle de forêt aléatoire. L'évaluation est faite à partir des résultats obtenus pour 5 sous-jeux tests, exclus de l'étape d'optimisation. Chaque point représente la moyenne sur l'ensemble des modèles entraînés avec la valeur du paramètre fixée, et les traits représentent la déviation standard de la valeur de qualité de classification. Deux mesures sont présentées : l'exactitude de classification de chaque classe moyennée (balanced accuracy) et le score F1 (avec lequel l'optimisation des paramètres est faite).

Les valeurs les plus faibles sont effectivement associées à des scores de classification légèrement plus faibles que pour les modèles ayant un nombre d'arbres plus élevés, mais cette augmentation semble stabilisée dès un nombre d'arbres autour de 300.

Par ailleurs, j'ai extrait le meilleur modèle identifié pour chacune des 5 séparations initiales du jeu de données. Chaque modèle est alors testé sur le sous-groupe qui a été gardé indépendant, afin d'évaluer sa qualité, dans le cadre d'une généralisation de son application. Les résultats sont présentés dans le tableau 8.2. On peut remarquer que le meilleur modèle n'est pas toujours le même dans les 5 séparations ; pour chaque modèle identifié dans les sous-groupes, son rang dans les autres sous-groupes est rapporté. Par ailleurs, les scores de qualité de classification depuis validation croisée interne (ayant conduit à la sélection des hyper-paramètres) sont également rapportés, pour comparer avec les scores obtenus sur le jeu indépendant. Cette comparaison permet de vérifier qu'il n'y a pas de diminution trop importante de la qualité du modèle lorsqu'il est testé sur des entités indépendantes ; les meilleurs modèles sont bien généralisables. La comparaison avec le modèle avec des paramètres par défaut permet par ailleurs de constater que l'optimisation conduit à une légère augmentation des mesures de qualité, confirmant l'intérêt d'identifier des valeurs d'hyper-paramètres plus adaptées à la tâche de classification à laquelle je suis confronté.

Les paramètres finaux pour mes modèles sont : un nombre d'entité par feuille de 1 (clairement associé à une meilleure qualité de prédiction des modèles, voir figure 8.3), une profondeur maximale de 10 (seuil à partir duquel la qualité ne semble plus augmenter, voir

	Inner loop test scores		Outer loop test scores		Default model performance		Rank of the selected model in the outer splits				
Selected model	**Balanced accuracy**	**F1 score**	**Balanced accuracy**	**F1 score**	**Balanced accuracy**	**F1 score**	**rank_0**	**rank_1**	**rank_2**	**rank_3**	**rank_4**
N_tree=300 Min_samples_leaf=1 Max_depth=15	0.8831	0.682	0.871	0.674	0.8424	0.662	1	49	49	19	17
N_tree=1300 Min_samples_leaf=1 Max_depth=35	0.87	0.666	0.869	0.667	0.852	0.648	18	1	35	2	29
N_tree=1500 Min_samples_leaf=1 Max_depth=20	0.875	0.665	0.879	0.664	0.877	0.676	32	42	1	16	41
N_tree=1100 Min_samples_leaf=1 Max_depth=20	0.87	0.661	0.89	0.694	0.88	0.699	37	36	21	1	9
N_tree=100 Min_samples_leaf=1 Max_depth=30	0.876	0.667	0.88	0.658	0.873	0.676	11	25	4	8	1

TABLE 8.2 – Résultats des 5 modèles identifiés comme les meilleurs par optimisation de paramètres. Un modèle est identifié comme le meilleur pour chaque groupe de la boucle externe ; ses performances dans la boucle interne et dans la boucle externe sont présentées ; un modèle par défaut (100 arbres, 1 entité minimum par feuille, pas de limite sur la profondeur) est entraîné et évalué sur le jeu test de la boucle externe pour comparaison. Le rang de chacun des modèles est également identifié depuis chacune des 5 optimisations ; chaque modèle est le premier dans la boucle interne où il a été sélectionné, et son rang dans les autres groupes est récupéré. Par exemple pour le 2e modèle, il est le meilleur modèle identifié dans la boucle 2 (son rang est de 1), et il est au second rang pour la boucle 4.

figure 8.3), et un nombre d'arbres de 1 000. Ce dernier paramètre est moins critique que les premiers, mais son augmentation semble légèrement associée à de meilleures performances ; le nombre de 1 000 permettra une interprétation simple des scores de prédictions dans les prochaines étapes.

8.3 Entraînement et validation des modèles

8.3.1 Définition des modèles explorés

J'ai donc défini le choix de la structure des forêts aléatoires ; les variants fonctionnels et non-fonctionnels ont été présentés au chapitre précédent, avec les différences détectables dans les annotations qui servent à les décrire. Nous avons donc tous les éléments nécessaires à l'entraînement de modèles de prédictions. J'ai souhaité profiter de ces entraînements pour évaluer les conséquences de certains points sur les qualités de prédiction des modèles ; ces points sont :

— les différents schémas d'échantillonnage des contrôles négatifs ;
— l'inclusion de l'annotation "targets_associations" ;
— la méthode de séparation des jeux d'entraînements en sous-jeux "entraînement/validation",

pour les validations croisées.

Influence de l'échantillonnage des contrôles négatifs. Comme présenté dans le chapitre précédent, plusieurs échantillonnages de variants contrôles négatifs ont été effectuées, chacun plus ou moins strict :

— l'échantillonnage GENCODE-match (visant à ajuster les répartitions génomiques des contrôles négatifs à celle des contrôles positifs) est le plus général ;

— l'échantillonnage Cytoband-match est similaire au GENCODE-match, mais s'en distingue surtout pour les contrôles positifs du jeu HGMD-DM : leur faible nombre a justifié de choisir des contrôles négatifs dans les régions génomiques contenant spécifiquement ces contrôles positifs ;

— l'échantillonnage Distance-match est le plus strict, et sera l'occasion d'évaluer la capacité des classificateurs à distinguer des variants très proches en coordonnées génomiques.

Inclusion de l'annotation "targets_association". Cette annotation correspond à une mesure du degré d'association génomique d'un variant contrôle positif au gène auquel il est fonctionnellement associé (association expérimentalement validée pour les HGMD-DM, ou association statistique pour les eQTLs, depuis des mesures d'expression des gènes). L'annotation est à 4 niveaux : le plus élevé correspond à une situation où le variant est dans le gène auquel il est fonctionnellement associé, ainsi que dans une région régulatrice prédite ; le plus faible correspond à une situation où le variant n'est pas génomiquement associé à son gène (il n'est pas ni dans le corps du gène, n'est pas dans une région régulatrice ciblant ce gène, et aucun des gènes flanquants ne lui correspond).

L'analyse des représentations heatmap (7.3) des différences entre les contrôles positifs et négatifs, ainsi que des pouvoirs discriminants des annotations, avait indiqué que cette annotation d'association au gène-cible correspond à un très bon prédicteur. J'ai souhaité comparer des modèles où l'annotation est exclue, avec des modèles où l'annotation est incluse, afin de déterminer si les niveaux d'associations élevés des variants contrôles positifs conduisait les modèles à ne considérer que cette information (conduisant à un modèle trivial, incapable d'identifier des relations plus complexes dans les annotations).

Validation croisée et séparations des sous-jeux Enfin, un dernier point exploré dans cette partie correspond au schéma de séparation appliqué durant la validation croisée.

Par défaut, la bibliothèque Python Scikit-Learn propose différentes méthodes de séparation automatique pour obtenir K groupes à partir d'un jeu d'entraînement. La plus adaptée à notre situation est appelé "StratifiedKFold" : il permet de séparer le jeu de données (ici les variants) en K groupes tout en respectant au mieux les proportions des différentes classes.

Cependant, cette séparation ne permet pas de tenir compte de la nature des entités présentes dans le jeu de données, et en particulier des liens existants entre ces entités. Dans notre cas cela pose un problème puisque les entités ne sont pas indépendantes les unes des autres : les variants sont localisés à des distances plus ou moins faibles, et peuvent parfois être proches. Cette proximité génomique peut conduire ces entités à avoir des annotations communes, et/ou corrélées. Si une séparation aléatoire est appliquée, il est possible que deux variants proches se retrouvent respectivement dans le jeu d'entraînement et dans le jeu de validation : cela facilite le modèle d'apprentissage, et conduit potentiellement à une évaluation optimiste de sa qualité.

Il est donc important de proposer une validation croisée prenant en compte cette proximité des variants. Pour le modèle FINSURF, j'ai implémenté une séparation en K groupes prenant en compte les localisations des variants, par une séparation des variants basée sur les bandes cytogénétiques (similaire à la séparation décrite dans la publication de Genomiser, Smedley et al., 2016). J'évaluerai les conséquences de cette séparation, nommée "Location-aware", sur la qualité mesurée des modèles.

8.3.2 Résultats des entraînements et validation des modèles

Ainsi, ce sont 24 modèles dont je présente l'exploration dans cette partie : 12 pour le jeu "HGMD-DM" et 12 pour le jeu "eQTLs-OMIM". Pour mesurer la qualité des modèles entraînés, j'ai appliqué une validation croisée basée sur la génération de 10 sous-groupes : pour chaque sous-groupe, un modèle est entraîné sur les 9 autres, et le modèle est utilisé pour prédire les classes du sous-groupe restant ; la courbe ROC et la courbe PR sont générées pour évaluation de la qualité. Les courbes moyennes, calculées à partir des modèles des 10 sous-groupes, sont obtenues, et les valeurs d'aires sous ces courbes sont rapportées dans le tableau 8.3, pour évaluer l'impact des différents paramètres sur la qualité des modèles. Pour aider à l'évaluation de ces valeurs, je propose la visualisation des différentes courbes, obtenues sur les validations croisées basées sur les séparations Location-aware : la figure 8.4 présente les courbes obtenues pour les modèles entraînés sur les variants fonctionnels HGMD-DM, et la figure 8.5 présente les courbes obtenues pour les modèles

entraînés sur les variants fonctionnels eQTLs-OMIM.

		Stratified K-fold split				Location-aware			
		ROC AUC		PR AUC		ROC AUC		PR AUC	
		with Targs	without Targs	with Targs	without Targs	with Targs	without Targs	with Targs	without Targs
HGMD-DM	GENCODE Match	0.9927	0.9424	0.9807	0.8257	0.9918	0.9169	0.9604	0.6562
	Cytoband Match	0.9891	0.9368	0.9591	0.8111	0.9823	0.9068	0.9033	0.6448
	Distance Match	0.8836	0.8817	0.7143	0.713	0.8343	0.843	0.5292	0.5348
eQTLs-OMIM	GENCODE Match	0.8413	0.7182	0.332	0.1791	0.8354	0.7043	0.3133	0.157
	Cytoband Match	0.8365	0.7156	0.3238	0.1769	0.8307	0.7029	0.3098	0.1567
	Distance Match	0.7052	0.6462	0.1959	0.1559	0.6915	0.6304	0.1738	0.135

TABLE 8.3 – Valeurs des aires sous les courbes ROC et PR pour l'ensemble des modèles FINSURF explorés. Les valeurs sont groupées par jeux de variants pour les lignes. Deux groupes de colonnes séparent les méthodes de séparations pour la validation croisée. Au sein de ces groupes, les valeurs d'aires sous les courbes sont groupées par courbe (ROC ou PR), et les valeurs des modèles avec ou sans l'annotation "targets_associations" ("with Targs" et "without Targs" respectivement) sont juxtaposées pour simplifier la lecture.

Influence des échantillonnages. Le premier point que je souhaitais évaluer concernait les différents échantillons de variants contrôles négatifs. On voit une diminution systématique de la qualité des modèles : les modèles basés sur l'échantillonnage Distance-match sont systématiquement plus faibles, que les modèles basés sur les échantillonnages Cytoband-match et GENCODE-match. Ceci est vrai pour les contrôles positifs HGMD-DM et les contrôles positifs eQTLs-OMIM, et pour les deux mesures (ROC AUC et PR AUC).

Conceptuellement, l'échantillonnage Cytoband-match est plus rigoureux que l'échantillonnage GENCODE-match ; la légère différence de qualité confirme également que le premier conduit à une sélection de contrôles négatifs plus stricte. L'échantillonnage Distance-match quant à lui teste à un niveau local la capacité des modèles à discriminer des variants fonctionnels de variants non-fonctionnels à une échelle locale, permettant potentiellement de discriminer deux positions au sein d'un élément régulateur par exemple ; la diminution de qualité des modèles (qui restent bons cependant) est donc attendue, et une exploration plus détaillée des conséquences de cet échantillonnage sera présentée par la suite. Ainsi dans la section et le chapitre suivants, je me focaliserai sur les modèles basés sur les échantillonnages Cytoband-match et Distance-match, dont les performances sont élevées, et qui explorent les différences entre variants fonctionnels et des variants non-fonctionnels

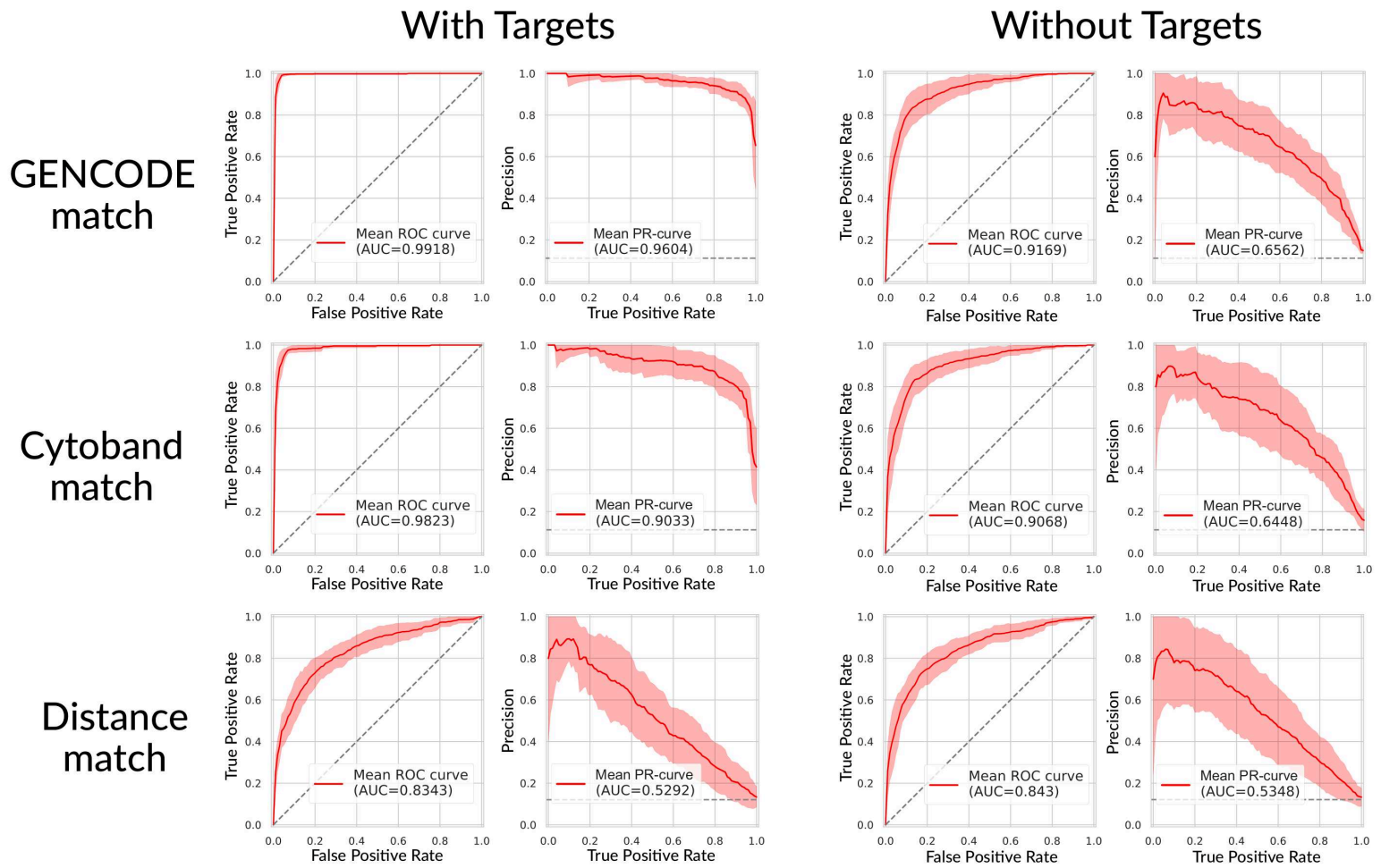


FIGURE 8.4 – Courbes ROC et PR pour les modèles HGMD-DM évalués par validation croisée Location-aware. Chaque graphique correspond à 10 évaluations, à partir desquelles sont calculées une courbe moyenne, et une aire correspondant à la déviation standard. Une paire de graphiques "courbe ROC" et "courbe PR" correspond à un modèle, entraîné avec le jeu de variant contrôle négative identifié par ligne, et avec ou sans l'annotation "targets_associations", identifiée par groupe de colonne.

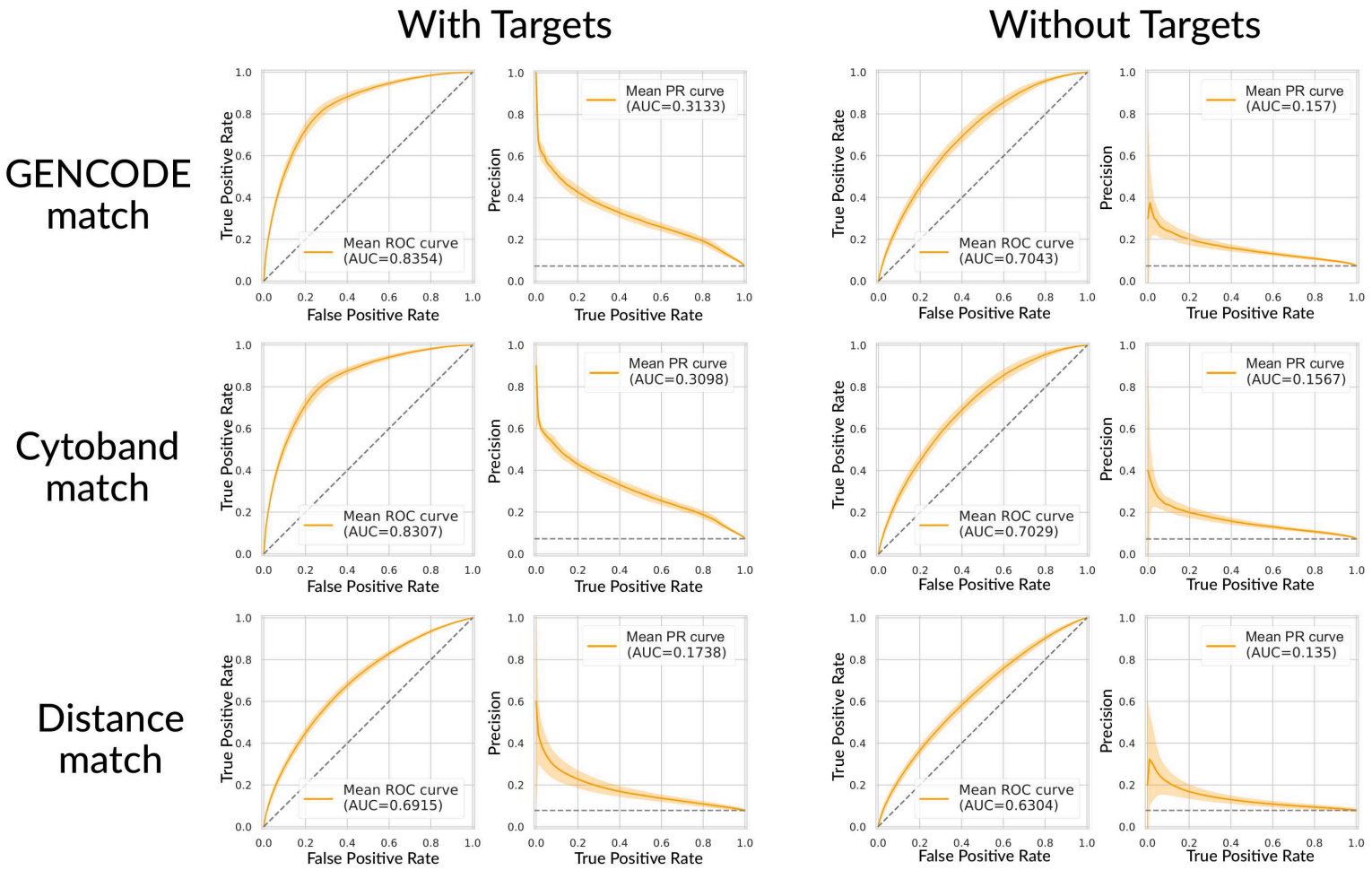


FIGURE 8.5 – Courbes ROC et PR pour les modèles eQTLs-OMIM évalués par validation croisée Location-aware. Chaque graphique correspond à 10 évaluations, à partir desquelles sont calculées une courbe moyenne, et une aire correspondant à la déviation standard. Une paire de graphiques "courbe ROC" et "courbe PR" correspond à un modèle, entraîné avec le jeu de variant contrôle négative identifié par ligne, et avec ou sans l'annotation "targets_associations", identifiée par groupe de colonne.

sélectionnés de manière rigoureuse.

Influence de l'annotation "targets_associations". En comparant les colonnes "with Targs" et "without Targs" du tableau 8.3, on peut voir que l'absence de cette annotation est systématiquement associée à une diminution des aires sous les courbes, plus ou moins importante, pour les schémas d'échantillonnage Cytoband-match et GENCODE-match, et ce pour les deux jeux fonctionnels (HGMD-DM et eQTLs-OMIM).

En revanche lorsque l'on considère les modèles HGMD-DM Distance-match, on peut voir que les différences de valeurs d'aires entre les modèles avec l'annotation et sans l'annotation sont très faibles (par exemple : les valeurs d'aire sous la courbe ROC sont de 0.8836 et 0.8817 avec et sans l'annotation respectivement, pour la validation croisée "Stratified K-fold Split"). Cela s'explique par le fait que les contrôles négatifs ont été pris à proximité des contrôles positifs, et sont donc localisé potentiellement dans les mêmes gènes, ou les mêmes éléments régulateurs ; l'annotation n'est donc plus discriminante. Ce n'est pas le cas pour les modèles eQTLs-OMIM Distance-match, dont les scores pour le modèle avec l'annotation "targets_associations" sont meilleurs que sans. Cela rejoint l'observation du chapitre précédent (7.3), qui montrait que ce schéma d'échantillonnage n'abolissait pas totalement la différence forte entre contrôles positifs et négatifs pour cette annotation.

Il est intéressant de noter que pour les classificateurs HGMD-DM GENCODE-Match et Cytoband-match incluant cette annotation, les valeurs d'aires sous les courbes sont extrêmement proches de celles d'un classificateur parfait (ROC AUC = 0.9927 et PR AUC = 0.9807 pour GENCODE-Match par exemple). Cela indique que l'annotation est trop corrélée à la classe des variants (fonctionnel/non-fonctionnel), et que le classificateur n'a pas besoin d'apprendre à partir des autres informations : il est suffisant de savoir qu'un variant est associé génomiquement à un gène d'intérêt pour prédire que ce variant est fonctionnel. Cette situation conduit donc à un modèle trivial, qui se contente de prédire un variant comme fonctionnel dès lors qu'il est situé dans le corps ou dans un élément régulateur associé à un gène d'intérêt. Si l'on cherche effectivement à définir des modèles de prédiction de qualité, qui apprennent des relations complexes entre les différentes annotation, il n'est pas souhaitable de garder cette annotation ; cela justifie de ne pas continuer à travailler avec les modèles incluant l'annotation "targets_associations".

Différences dues aux méthodes de validations croisées On peut constater que pour chaque modèle (combinaison d'un jeu de contrôles positifs, d'un schéma d'échan-

tillonnage des négatifs, et de l'utilisation ou non de l'annotation "targets_associations"), les valeurs d'aires sous les courbes sont systématiquement plus faibles lorsque calculées depuis une validation croisée basée sur les localisations des variants dans les bandes cytogénétiques (Location-aware), en comparaison avec une validation croisée aléatoire (stratified K-fold split).

Les différences sont particulièrement flagrantes pour les modèles HGMD-DM Distance-match : par exemple pour les modèles n'incluant pas les association aux gènes cibles, on passe d'une PR AUC de 0.713 (validation croisée simple) à une PR AUC de 0.538 (validation croisée Location-aware). Cette différence souligne que les modèles entraînés par validation croisée simple arrivaient à reconnaître des variants contrôles positifs et négatifs dans les jeux de validation par leur proximité génomiques aux variants contrôles négatifs ; tandis que dans le cadre de la validation croisée Location-aware, cette reconnaissance par proximité génomique n'est plus possible (les variants pour validation étant localisés dans des bandes cytogénétiques différentes). Les diminutions d'aires sous les courbes sont moins élevées pour les autres modèles, mais tout de même présentes.

Ainsi, je confirme qu'une séparation simple aléatoire du jeu d'entraînement conduit à sur-estimer la qualité d'un modèle de classification ; au moment de l'évaluation finale des modèles, il est donc préférable de rapporter les valeurs associées à une validation croisée basée sur les bandes cytogénétiques dans notre cas.

Conclusions J'ai exploré dans cette partie les performances de différents modèles de prédictions, définis pour tester les limites associées à plusieurs facteurs pouvant biaiser leur apprentissage.

J'ai tout d'abord montré que les performances de ces modèles varient selon la nature des variants contrôles négatifs, et la manière dont ils ont été sélectionnés ; cela m'a conduit à identifier les modèles basés sur les échantillonnages Cytoband-Match et Distance-match comme étant les plus rigoureux, chacun permettant d'évaluer la fonctionnalité de variants à une échelle globale et à une échelle locale respectivement.

J'ai également montré que l'utilisation de l'annotation targets_associations, qui avait été identifiée comme fortement corrélée avec les variants fonctionnels (chapitre 8), conduit à la génération de modèles de prédictions triviaux, ce qui m'a poussé à ne pas considérer ces modèles comme pertinents.

Enfin, la comparaison des méthodes de séparations des jeux d'entraînement pendant la validation croisée a permis de mettre en évidence que ne pas prendre en compte la

proximité génomique des variants conduisait à une évaluation optimiste de la qualité des modèles, en particulier pour les modèles avec l'échantillonnage Distance-match. Ces étapes de validation-croisées sont importantes puisqu'elles servent à évaluer la qualité potentielle d'un modèle à être généralisable et applicable en routine. Il est également important d'en tenir compte pour comparer de manière rigoureuse un modèle de prédiction à un autre (voir section suivante).

Les modèles finaux qui seront explorés et utilisés dans les chapitres suivant sont donc les modèles HGMD-DM Cytoband-match, HGMD-DM Distance-match, eQTLs-OMIM Cytoband-match, et eQTLs-OMIM Distance-match, tous entraîné sans considérer l'annotation targets_associations ; à noter que pour leur utilisation aux chapitres suivants, les modèles sont entraînés sur l'entièreté de leurs jeux d'entraînement respectifs.

8.4 Comparaison à la littérature

Pour comparer différents scores de prédictions existants, il est nécessaire d'avoir une base de variants indépendants du jeu d'entraînement. J'avais initialement proposé une comparaison des modèles FINSURF avec d'autres méthodes publié, en les applicant au jeu de variants fonctionnels non-codants rapportés dans la publication associée au score Genomiser (Smedley et al., 2016). Cependant, après vérifications, les variants de ce jeu de données correspondent à des variants présents dans la base de données HGMD que j'ai utilisée. Cela ne permettait donc pas d'avoir une comparaison équitable entre différents scores : mon modèle FINSURF, le modèle Genomiser, ou encore le modèle NCBoost (entraîné sur une partie des variants de Genomiser, Caron et al., 2018) pourraient être avantagés si on utilise des variants utilisés pour leur entraînement, par rapport à d'autres méthodes qui ne considèrent pas ces variants (par exemple CADD). Malheureusement, je n'ai pas trouvé de bases de données alternatives de variants fonctionnels non-codants impliqués dans des maladies, et complètement indépendants de la base de données HGMD.

Je propose donc une comparaison de mes modèles FINSURF (HGMD-DM et eQTLs-OMIM), en ré-utilisant les sous-jeux définis lors des validations croisées Location-aware, présentés dans la section précédente. Ainsi j'ai la garantie de ne pas avantager les modèles FINSURF par rapport aux autres scores : comme précédemment décrit, les modèles FINSURF seront entraînés sur un sous-jeu de variants qui seront distincts des variants du jeu de validation (par la séparation Location-aware). FINSURF sera même légèrement désavantagé : les méthodes de prédiction entraînées avec ces mêmes jeux d'entraînement vont

avoir une facilité supérieure à reconnaître les variants fonctionnels, et donc à les classer correctement.

Pour résumer l'approche de comparaison, en considérant les modèles HGMD-DM : les 10 sous-jeux précédemment définis par la validation croisée Location-aware sont ré-utilisés pour comparer chacun des modèles FINSURF (Cytoband-match ou Distance-match) avec les modèles provenant de la littérature, dont les scores de prédictions sont obtenus pour les variants des 10 sous-jeux. Pour chacun des modèles (FINSURF, ou modèles comparés), une courbe ROC et une courbe PR sont calculées à partir de la moyenne de leurs performances sur les 10 sous-jeux. Le même principe est appliqué pour les modèles eQTLs-OMIM.

Concernant le choix des autres méthodes à comparer avec FINSURF, je me suis basé sur la disponibilité (en ligne) de scores précalculés soit pour chaque position du génome (un score par position), soit pour des variations alléliques spécifiques (un score pour chacune des 3 mutations possibles pour l'allèle de référence à chaque position. Ces modèles (présentés au chapitre 3) couvrent également des approches variées (tant sur les méthodes de calcul, que sur les hypothèses biologiques couvertes). Les scores pré-calculés, étaient disponibles dans des formats similaires à ceux des annotations utilisées pour mes modèles, ce m'a permis de réutiliser l'outil annotVariants développé pour FINSURF (voir chapitre 7), afin d'annoter avec ces scores l'ensemble des variants utilisés pour les entraînements des modèles FINSURF. A noter que toutes les positions ne sont pas annotées par l'ensemble des scores. Notamment certains scores ne proposent pas de valeurs pour les variants sur les chromosomes sexuels. Par ailleurs certains proposent une valeur par variation allélique : cela ne permet pas d'annoter les variants INDELs. J'ai choisi d'écarter toutes les positions pour lesquelles au moins un des scores était manquant ; cela conduit à écarter entre 35% et 55% des variants des jeux de validations pour le modèle HGMD-DM, et entre 73% et 77% des variants des jeux de validations pour le modèle eQTLs-OMIM.

Observation :

— FINSURF HGMD-DM : le modèle FINSURF HGMD-DM (trait rouge plein) apparaît meilleur que la totalité des autres méthodes ; pour comparaison, je présente le modèle entraîné en incluant les "targets_associations" (trait rouge pointillé) : celui-ci se comporte comme un modèle quasi-parfait pour l'échantillonnage Cytoband-match, et revient à un comportement similaire au modèle excluant l'annotation pour l'échantillonnage Distance-match, ce qui était déjà observable précédemment et conforte le choix de l'exclure pour les analyses à suivre.

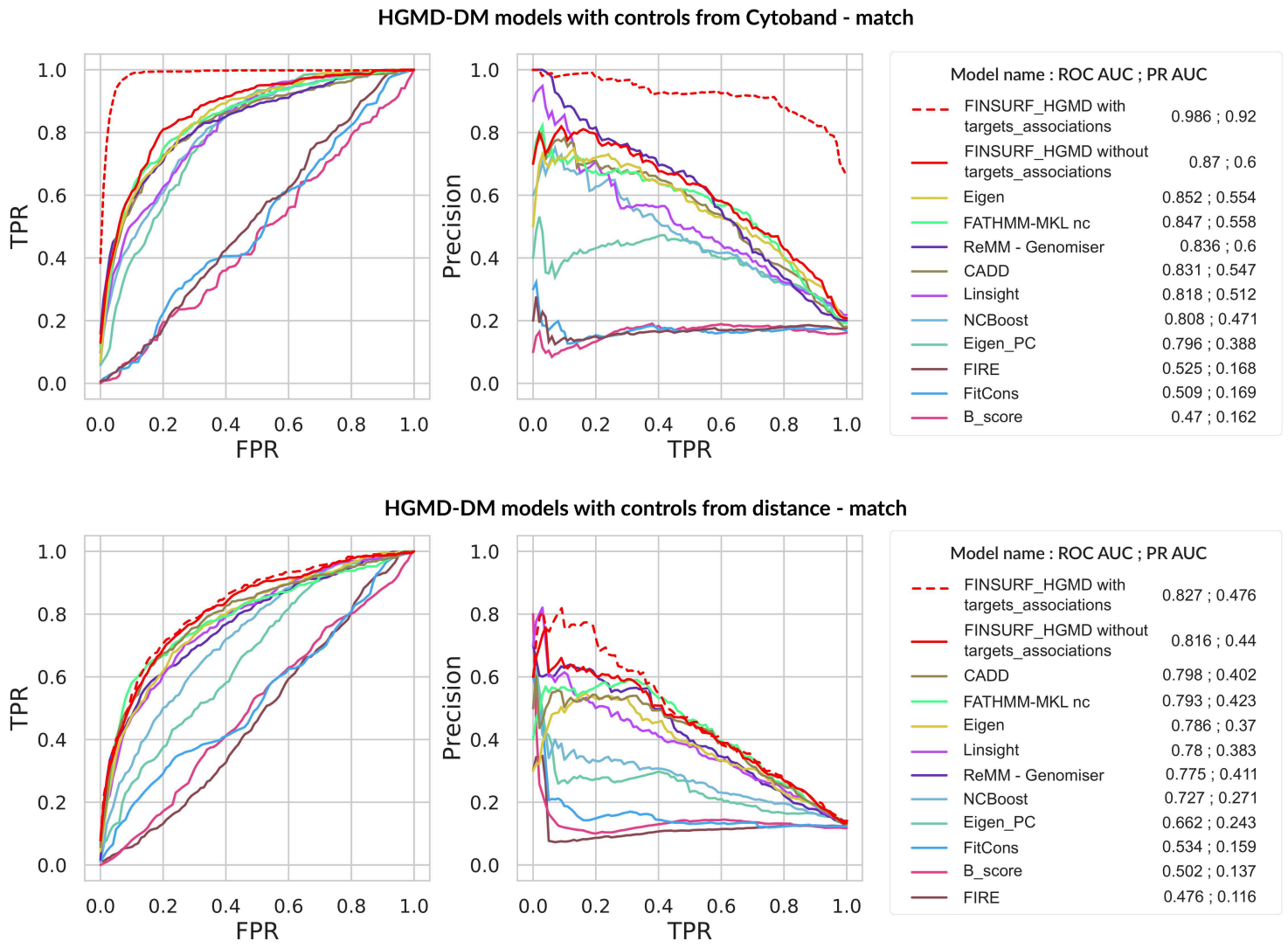


FIGURE 8.6 – Comparaison des modèles FINSURF HGMD-DM à d'autres méthodes de prédiction de variants fonctionnels non-codants. Les modèles FINSURF HGMD-DM Cytoband-match (première ligne), et Distance-match (seconde ligne) sont comparés avec 10 autres méthodes (décrites au chapitre 3 de l'introduction). La comparaison se fait sur les sous-jeux définis par validation croisée Location-aware, qui ont servi à évaluer les performances des modèles FINSURF (voir section précédente).

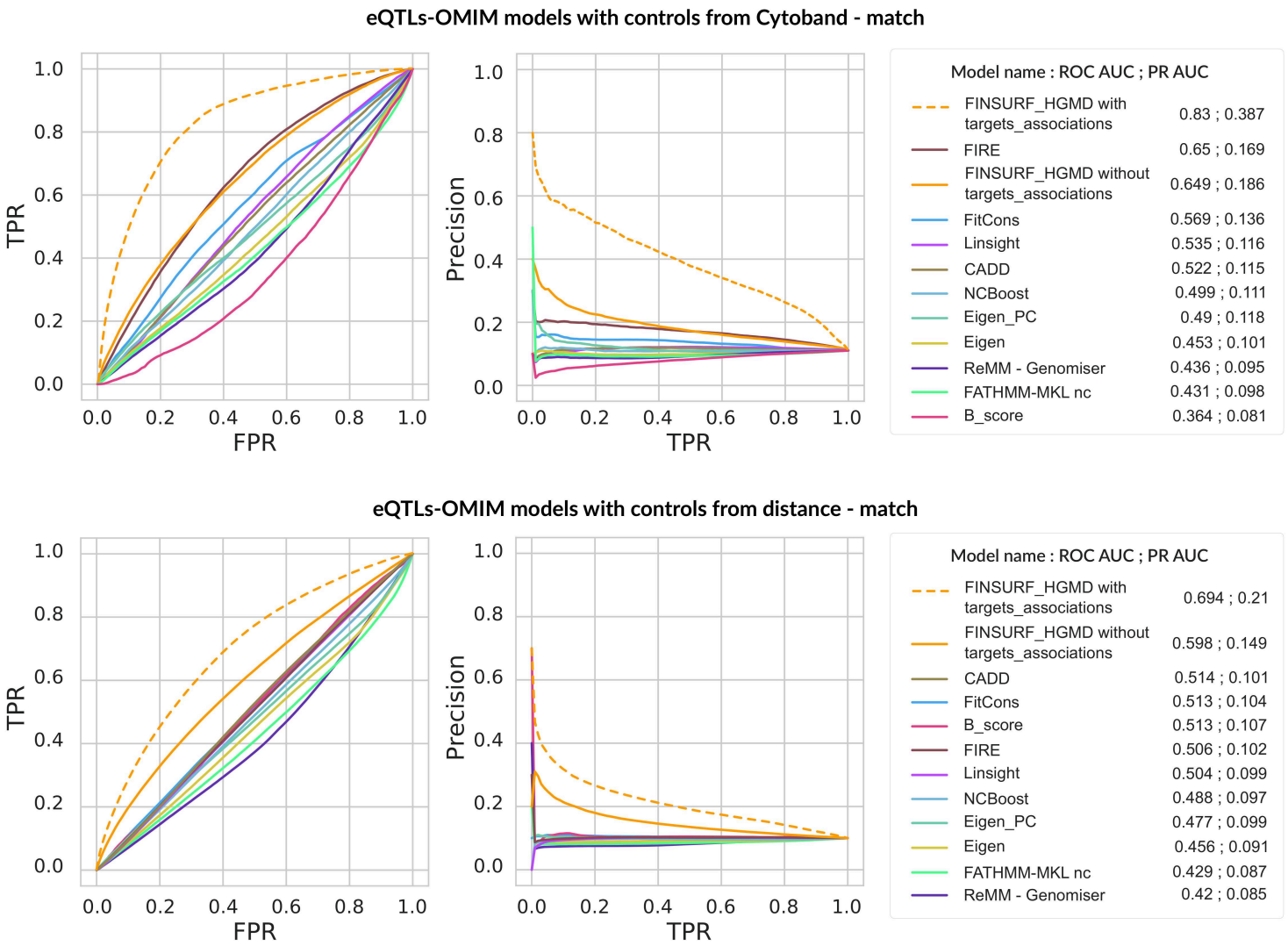


FIGURE 8.7 – Comparaison des modèles FINSURF eQTLs-OMIM à d'autres méthodes de prédiction de variants fonctionnels non-codants. Les modèles FINSURF eQTLs-OMIM Cytoband-match (première ligne), et Distance-match (seconde ligne) sont comparés avec 10 autres méthodes (décrites au chapitre 3 de l'introduction). La comparaison se fait sur les sous-jeux définis par validation croisée Location-aware, qui ont servi à évaluer les performances des modèles FINSURF (voir section précédente).

Il n'empêche que le modèle excluant cette annotation reste le meilleur parmi les méthodes évalués, et ce pour les deux types d'échantillonnage. Il est intéressant de noter que malgré leur entraînement sur des variants similaires, les modèles ReMM et NCBoost ne font pas partie des meilleurs modèles.
Enfin on peut noter que les moins bons modèles sont FIRE, FitCons et le score de sélection Background (B score).

— FINSURF eQTLs-OMIM : A nouveau, le modèle incluant l'annotation "targets_associations" (trait jaune pointillé) se distingue nettement des autres modèles, indiquant une corrélation trop importante entre cette annotation et la classe des variants, confirmant à nouveau la nécessité de l'exclure des analyses à suivre. Pour le modèle eQTLs-OMIM excluant cette annotation (trait jaune plein), on constate qu'il présente une qualité comparable à celle du modèle FIRE (avec toutefois une meilleure aire sous la courbe PR pour FINSURF).
Cela peut s'expliquer par le fait que FINSURF eQTLs-OMIM et FIRE sont tous deux entraînés à reconnaître des eQTLs : bien que le modèle eQTLs-OMIM soit focalisé sur un sous-jeu d'eQTLs associés à des gènes de maladie, il est finalement probable que les règles apprises par le modèle correspondent à des propriétés générales associées au caractère fonctionnel "eQTL", conduisant à des détections similaire à FIRE (dont l'apprentissage n'est pas biaisé sur une classe particulière d'eQTLs).
Il est intéressant de noter que le modèle FINSURF eQTLs-OMIM présente des meilleurs scores que FIRE et que les autres méthodes dans le cadre des contrôles négatifs "Distance-match" : c'est d'ailleurs le seul modèle qui semble arriver à distinguer une partie des contrôles positifs et négatifs (bien que ses scores de qualité soit assez faibles).

8.5 Conclusions

Ce chapitre m'a permis de développer en détails différents points concernant l'entraînement des modèles de prédiction FINSURF, allant de l'évaluation des conséquences de différents biais sur les mesures de performances, jusqu'à la comparaison des perfomances des modèles FINSURF comparés à d'autres méthodes de la littérature.

Les conclusions sont les suivantes

— une séparation rigoureuse des variants d'entraînement et de validation est nécessaire

pour évaluer de manière objective les performances des modèles, afin de ne pas sur-estimer sa capacité de généralisation : la proposition de la séparation selon les bandes cytogénétiques permet effectivement une évaluation rigoureuse de mes modèles ;

— les échantillonnages de contrôles négatifs conduisent à des différences de performances notables ; les modèles associés répondent à des hypothèses différentes (évaluation d'une fonctionnalité à l'échelle du génome ou à l'échelle locale), et les qualités de prédictions associées montrent que dans un contexte local la discrimination entre variants fonctionnels et non-fonctionnels est plus délicate (mais les performances restent bonnes) ;

— j'ai souligné l'importance d'identifier des annotations qui conduisent à des résultats trop parfaits, et donc des modèles dont les prédictions sont triviales ;

— la comparaison avec d'autres méthodes de prédiction montre que les modèles FINSURF sont aussi bons voire meilleurs que ces modèles, confirmant donc l'intérêt d'utiliser FINSURF pour évaluer la fonctionnalité de variants non-codants.

Tout cela me conduit à continuer mes analyses avec les modèles HGMD-DM Cytoband-match et Distance-match, ainsi que les modèles eQTLs-OMIM Cytoband-match et Distance-match. Dans la partie suivante, je présente des approches permettant d'explorer plus en détails les propriétés des modèles entraînés, afin d'interpréter les règles biologiques identifiées par les modèles.

Chapitre 9

Interprétation des modèles

9.1 Evaluation de l'importance globale des annotations

9.1.1 Mesure de l'importance des annotations pour la classification

Pour un modèle d'apprentissage donné, il est intéressant d'avoir une information sur les annotations qui ont été effectivement utiles pour l'apprentissage. Cela permet par exemple de voir si des annotations précédemment identifiées comme pertinentes sont effectivement utilisées pour la classification. Cela permet également d'opérer une sélection parmi ces annotations, afin de définir un modèle minimal basé uniquement sur les annotations importantes.

Pour évaluer l'importance d'une annotation au sein d'un modèle, il nous faut donc une mesure. La bibliothèque Scikit-Learn propose une méthode de calcul pour les modèles de forêts aléatoires. Celle-ci est basée sur la façon dont sont construites les forêts aléatoires (présentées en détails au chapitre 3), et dont je reprends un point crucial pour le calcul ici. Au moment de l'entraînement d'un modèle de forêt aléatoire, la génération d'un arbre de décision de cette forêt est basée sur une séparation progressive des entités de chaque classe, à travers un succession de noeuds où des seuils sont appliqués aux valeurs des annotations. Pour quantifier cette séparation des classes, l'indice de diversité de Gini (équation 9.1) est calculé à chacun des noeuds, à partir des proportions de chaque classe (notées p_i) dans le noeud. Le choix des annotations et des seuils à ces noeuds est fait par l'optimisation de la diminution de l'indice de diversité de Gini (équation 9.2) :

$$I_G(noeud) = \sum_{i=1}^{m} p_i(1 - p_i) \tag{9.1}$$

$$\Delta I_G(noeud) = I_G(noeud) - I_G(noeud_{fils1}) - I_G(noeud_{fils2}) \tag{9.2}$$

La mesure de cette diminution peut être utilisée pour calculer l'importance de l'annotation au sein du modèle. Chaque annotation est ainsi associée à la somme des ΔI_G (pondérés par le nombre d'entités par noeud) de l'ensemble des noeuds où cette annotation a été utilisée. À l'échelle de l'arbre, la valeur finale d'importance d'une annotation est normalisée par le nombre d'entités à la racine, et par la somme totale des valeurs d'importance des annotations. Une moyenne est calculée depuis l'ensemble des arbres, pour obtenir la valeur d'importance d'une annotation dans un modèle de forêt aléatoire. Cette valeur peut donc être interprétée comme la proportion moyenne de diminution de l'indice de diversité de Gini associée à l'utilisation de l'annotation : plus la valeur d'importance est élevée, plus la séparation des entités des différentes classes a été importante grâce à cette annotation, et donc plus la diminution de l'indice de Gini a été élevée, au moment de l'entraînement du modèle. Cette mesure d'importance nous permet donc d'évaluer l'utilité des annotations pour l'entraînement des modèles, et de les comparer par modèle et entre modèles.

J'ai appliqué cette méthode aux différents modèles présentés au chapitre précédent. Pour chaque jeu de variants contrôles positifs (HGMD-DM et eQTLs-OMIM), j'ai souhaité évaluer l'effet des échantillonnages de variants négatifs sur les valeurs d'importance des annotations. En effet, certaines des annotations couvrent des régions génomiques, dans lesquelles chaque position va être associée à la même valeur. Or nous avons vu que les modèles basés sur des contrôles négatives Distance-match ont des qualités de classification moindre comparées aux autres modèles ; la comparaison des valeurs d'importance permet donc potentiellement de voir si les annotations de régions contribuent moins efficacement pour ces modèles. Je compare donc ci-dessous les modèles entraînés sur les contrôles négatifs Cytoband-match, avec les modèles entraînés sur les variants contrôles négatifs Distance-match.

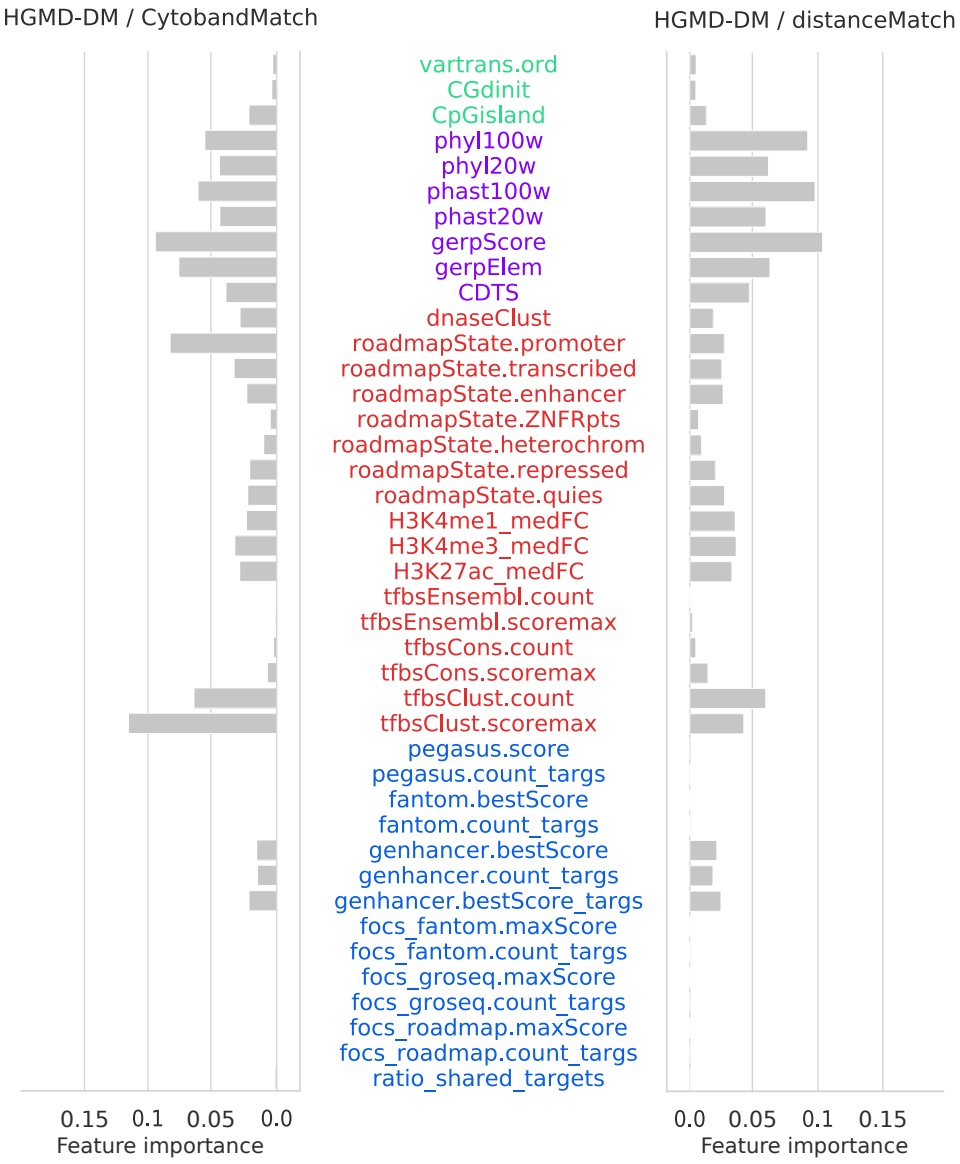


FIGURE 9.1 – Mesure d'importance des annotations pour les modèles HGMD-DM Cytoband-match et Distance-match. Les noms des annotations sont colorés selon leur classe d'annotation : vert pour les annotations de séquence, violet pour les données de conservations ; rouge pour les marques biochimiques ; bleu pour les prédictions d'associations régulatrices.

9.1.2 Mesures d'importances des annotations pour les modèles HGMD-DM

La figure 9.1 nous permet de comparer les valeurs d'importance des annotations calculées depuis les modèles HGDM-DM Cytoband-match et HGDM-DM Distance-match. On peut tout d'abord constater que pour le modèle Cytoband-match, l'annotation la plus importante est l'annotation tfbsClust.scoremax, correspondant à la concentration locale de prédiction de sites de fixation de facteurs de transcription (TFBS). Cette annotation était effectivement l'une de celles qui présentait une différence particulièrement marquée lors de l'évaluation des tailles d'effets entre contrôles positifs et négatifs (figure 7.3) ; il est donc intéressant de voir qu'elle a effectivement été exploitée par le modèle de prédiction. Au contraire, les autres annotations concernant des TFBS n'ont visiblement pas contribué au modèle, malgré des différences marquées lors des mesures de taille d'effet. Parmi les autres annotations identifiées comme importantes, on retrouve les annotations concernant la conservation en séquence (en particulier le score GERP) ; l'état chromatinien promoteur (roadmapState.promoter) est également une annotation ayant été exploitée par le modèle. Enfin, on peut constater que des annotations comme les états chromatiniens quiescent (roadmapState.quies) et réprimé (roadmapState.repressed) ont une valeur d'importance non-nulle ; ces annotations étaient associées à des valeurs plus importantes pour les variants contrôles négatifs, ce que le modèle semble avoir réussi à exploiter.

La comparaison avec le modèle Distance-match permet de constater une baisse marquée de l'importance des annotations tfbsClust.scoremax et roadmapState.promoter, tandis que les scores de conservation gagnent en importance. Cela peut s'expliquer par un point déjà évoqué au chapitre précédent : l'échantillonnage Distance-match est plus strict, et conduit à sélectionner des variants contrôles négatifs à proximité des variants contrôles positifs, et donc potentiellement dans les mêmes annotations d'intervalles génomiques que ces contrôles positifs. Le modèle a donc potentiellement plus de mal à exploiter ces annotations, où la différence entre positifs et négatifs est moins marquée ; cela semble le conduire en revanche à utiliser de manière plus importante les annotations de conservations, qui elles sont disponibles par position, et permettent donc une résolution plus fine. La diminution de l'utilisation des annotations d'intervalles pouvaient être attendue, puisque l'analyse des tailles d'effet permettaient déjà d'observer la diminution importante de la différence entre contrôles positifs et négatifs pour les annotations comme les TFBS.

Les modèles basés sur les variants fonctionnels HGMD-DM semblent donc baser leurs

prédictions sur une identification de variants soumis à une contrainte évolutive, affectant des sites de fixations de facteurs de transcription, et présentant des signaux de fonctionnalités associés à des régions actives.

9.1.3 Mesures d'importances des annotations pour les modèles eQTLs-OMIM

La figure 9.2 nous permet de comparer les valeurs d'importance des annotations calculées depuis les modèles eQTLs-OMIM Cytoband-match et eQTLs-OMIM Distance-match. On peut voir que les profils d'importances des annotations sont très différents des profils obtenus pour les modèles HGMD-DM. Tout d'abord pour le modèle eQTLs-OMIM Cytoband-match, on identifie distinctement quatre annotations dont l'importance est élevée : vartrans.ord (dont les valeurs ordonnés indiquent si un variant correspond à une transition, une transversion, ou un INDEL), roadmapState.transcribed (indiquant les états chromatiniens transcrits), roadmapState.heterochrom (états d'hétérochromatine) et roadmapState.quies (états quiescents). Ces quatre annotations correspondent effectivement à des annotations dont on pouvait mesurer des différences élevées lors des comparaisons de taille d'effet (7.3), avec des directions différentes : les contrôles négatifs semblent enrichis en variants INDEL ou transversion, dans des états hétérochromatiniens ou quiescents en comparaison avec les eQTLs-OMIM, tandis que ces derniers présentent des valeurs plus élevés que les contrôles négatifs pour les annotations d'états transcrits. D'autres annotations présentent des valeurs d'importances moindres, mais non-nulles, comme les annotations de conservation (dont les valeurs étaient moindres pour les eQTLs-OMIM par rapport aux contrôles négatifs), ainsi que des signaux de marques d'histones (associées à des régions à potentiel régulateur), et les prédictions de régions régulatrices Genehancer.

La comparaison avec le modèle Distance-match permet de voir que les annotations d'intervalles voient leur importance diminuer fortement (roadmapState.transcribed, roadmapState.quies, roadmapState.heterochrom). La même explication que pour les modèles HGMD-DM s'applique : les variants contrôles négatifs se trouvent plus probablement dans les mêmes états chromatiniens que les variants contrôles positifs, ce qui ne permet plus au modèle de les discriminer sur cette base. L'annotation qui présente la plus forte importance est l'annotation vartrans.ord ; cela indique probablement une sur-représentation de variants INDEL ou transversion dans l'un des jeux contrôles. On peut aussi constater que les annotations de conservation voient leur importance augmentée pour ce modèle : en

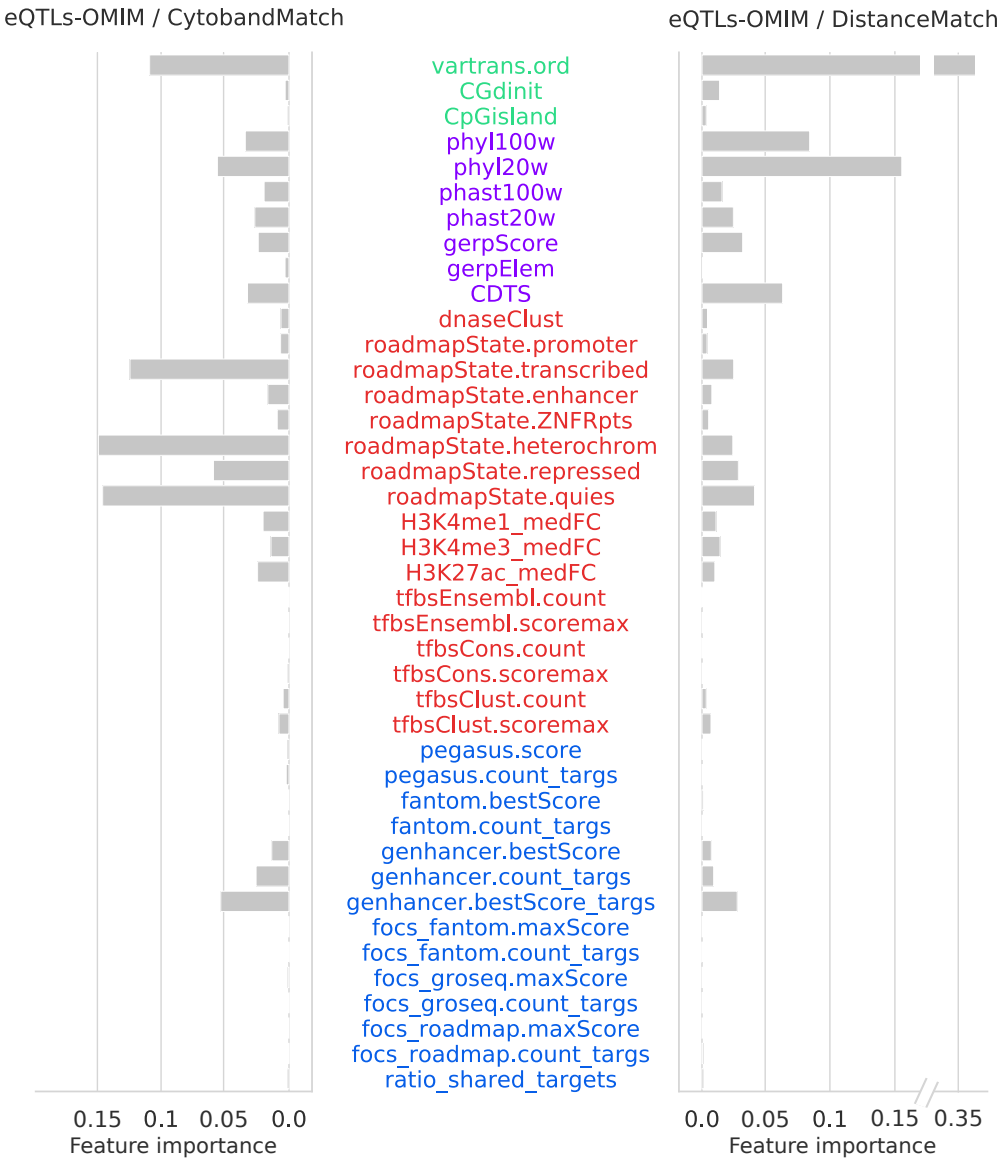


FIGURE 9.2 – Mesure d'importance des annotations pour les modèles eQTLs-OMIM Cytoband-match et Distance-match. Les noms des annotations sont colorés selon leur classe d'annotation : vert pour les annotations de séquence, violet pour les données de conservations ; rouge pour les marques biochimiques ; bleu pour les prédictions d'associations régulatrices.

particulier pour les annotations phyl100w et phyl20w. En observant les tailles d'effets mesurées, il est probable que le modèle exploite les valeurs plus faibles associées aux eQTLs, et associe donc à l'inverse le caractère non-fonctionnel à des valeurs de conservation élevées.

Dans l'ensemble ces analyses permettent de mieux comprendre la façon dont les modèles exploitent les annotations. Cela permet d'identifier des annotations qui pourraient probablement être retirées sans diminuer la qualité des modèles (par exemple l'ensemble des prédictions de régions régulatrices provenant de la base de données FOCS) ; ces annotations étaient déjà précédemment identifiées comme peu discriminantes (figure 7.4, et figure 7.5). D'autres annotations pourraient également être enlevées malgré des différences qui étaient tout de même détectables : par exemple les annotations de TFBS "tfbsEnsembl" et "tfbsCons" ont des valeurs d'importance suffisamment faibles dans mes modèles pour pouvoir être enlevées. Néanmoins, l'outil FINSURF vise non seulement à fournir une prédiction de fonctionnalité pour un variant, mais également un profil d'annotations associées à cette fonctionnalité ; c'est pourquoi je garderai l'ensemble des annotations sélectionnées. Par ailleurs, ces valeurs d'importance sont une vision globale de l'utilisation des annotations par le modèle. Il serait intéressant de pouvoir accéder à une mesure de l'importance des annotations à l'échelle de chaque variant : cela conduirait à une meilleure compréhension de ce que le modèle est capable d'utiliser pour un variant donné, et permettrait potentiellement d'identifier des groupes de variants différents, représentant les multiples chemins de décisions dans les arbres. C'est cette approche que j'ai souhaité suivre, et que je développe dans la section suivante.

9.2 Contributions des annotations par variant

9.2.1 Objectifs et méthode de calcul

La mesure d'importance des annotations présentée dans la section précédente permet d'avoir une vision globale de l'apport d'une annotation donnée à un modèle de forêt aléatoire. Cependant cela ne permet pas d'explorer en détails les structures existant dans cette forêt. En effet les forêts aléatoires conduisent, au moment de l'entraînement, à la génération de nombreux chemins de décisions (suites de seuils appliqués sur les annotations) dans les arbres. Ces modèles sont parfois désignés comme des modèles "boîtes noires" à cause de cette complexité. Une approche a cependant été proposée pour explorer les chemins de décisions, et ce pour chaque entité évaluée par le modèle : ce sont les valeurs de contributions des annotations (Kuz'min et al., 2011, Palczewska et al., 2014), dont je présente le principe ci-dessous.

Pour un arbre de la forêt, une entité va passer par un chemin de décisions (séquence de noeuds de décisions), depuis la racine de l'arbre jusqu'à une feuille (noeud terminal du chemin), où la prédiction de la classe de l'entité est faite. Lors de l'entraînement du modèle, les seuils de décisions successifs appliqués aux noeuds ont conduit à une séparation progressive des entités en fonction de leurs classes, jusqu'à un critère d'arrêt (par exemples : un paramètre de profondeur maximale, ou un nombre minimum d'entités à séparer). A chaque noeud de ce chemin, les proportions d'entités de chaque classe ont donc changé progressivement, toujours avec l'objectif de diminuer l'indice de diversité de Gini (voir équation 3.2) ; à la feuille, l'indice de Gini est théoriquement minimal étant donnés les critères d'arrêt. Dans une feuille, ce sont les proportions d'entités pour chaque classe qui déterminent quelle prédiction est faite : la classe prédite est celle dont la proportion est la plus élevée. Ainsi, toutes les entités à classer qui emprunteront le même chemin, et termineront dans la même feuille, seront affectées à la même prédiction.

Dans le cadre d'une tâche de classification à deux classes, désignées C_0 et C_1, on peut donc écrire qu'à chaque noeud n, deux proportions existent :

$$Y_0^n = \frac{|x_n \in C_0|}{|n|}$$

et

$$Y_1^n = \frac{|x_n \in C_1|}{|n|}$$

où

$$Y_0^n + Y_1^n = 1$$

Et au noeud terminal d'un chemin, l'assignation à la classe C_1 est déterminé par $Y_1^{feuille} > 0.5$. On définit alors l'incrémentation locale de la proportion de la classe C_1 entre un noeud parent p et un noeud fils f, associée à l'annotation k :

$$IL^n_{C_1,k} = \begin{cases} Y^f_{C_1} - Y^p_{C_1} & \text{si l'annotation } k \text{ est utilisée à ce noeud} \\ 0, & \text{autrement} \end{cases} \tag{9.3}$$

Par simplification, et par la relation $IL^n_{C_1,k} = 1 - IL^n_{C_0,k}$, je note dans les équations suivantes $IL^n = IL^n_{C_1,k}$ et $Y^n = Y^n_1$, C_1 représentant notre classe d'intérêt.

Pour une entité x qui emprunte un chemin de décisions dans un arbre, on peut calculer la contribution de l'annotation k :

$$contrib(x,k) = \sum_{n=1}^{N} \mathbb{1}_{x \in n} IL^n_{C_1,k} \tag{9.4}$$

La fonction de décision notée $f(x)$ associée au chemin emprunté par x, déterminant l'assignation de x à une classe, est ainsi notée :

$$\begin{aligned} f(x) &= Y^{n=0} + \sum_{k=1}^{K} contrib(x,k) \\ &= Y_f^{feuille} \end{aligned} \tag{9.5}$$

On généralise alors cette fonction de décision à l'échelle de la forêt :

$$\begin{aligned} F(x) &= \frac{1}{T}\sum_{t=1}^{T} Y_{f_t(x)}^{feuille} \\ &= \frac{1}{T}\sum_{t=1}^{T} f_t(x) \\ &= \frac{1}{T}\sum_{t=1}^{T} Y_t^{n=0} + \sum_{k=1}^{K}(\frac{1}{T}\sum_{t=1}^{T} contrib_t(x,k)) \end{aligned} \tag{9.6}$$

Ainsi dans l'équation 9.6, le premier terme de l'addition correspond au biais, qui est la proportion moyenne d'entités de la classe C_1 à la racine des arbres ; le second terme correspond à la somme sur les annotations K de la contribution moyenne de chaque annotation k, calculée pour l'entité x. C'est ce second terme que l'on extrait, pour établir pour chaque entité les valeurs de contributions de ses annotations à la classe prédite. Dans le cadre de mes modèles de prédictions de variants fonctionnels, chaque variant sera donc associé à un vecteur de valeurs de contributions (une par annotation). Ce vecteur permettra de déterminer quelles sont les annotations qui contribuent à sa prédiction en tant que variant fonctionnel (valeurs de contributions positives, augmentant la valeur $F(x)$), celles qui contribuent à sa prédiction en tant que variant non-fonctionnel (valeurs de contributions négatives, diminuant la valeur $F(x)$), et celles qui ne contribuent pas à sa prédiction (valeurs nulles).

9.2.2 Implémentation

La bibliothèque Python Treeinterpreter (Saabas, 2015) propose une implémentation du calcul des valeurs de contributions, pour les modèles de forêts aléatoires disponibles dans la bibliothèque Scikit-learn. Cette bibliothèque permet de calculer les vecteurs de contributions d'annotations pour un ensemble d'entités, étant donné un modèle de prédictions de forêt aléatoire entraîné. Mon module "forestContributions" (8.1) est donc basé sur l'implémentation disponible dans cette bibliothèque, dont j'ai modifié la structure pour pouvoir apporter deux modifications :

— J'ai optimisé le calcul des contributions : initialement le calcul est effectué de manière indépendante pour chaque entité. Puisque plusieurs entités peuvent partager la même feuille dans un arbre, j'ai modifié l'implémentation pour que le calcul des contributions soit d'abord fait pour chaque chemin dans l'ensemble des arbres.

Cela permet pour une entité d'aller directement récupérer depuis les vecteurs précalculés celui correspondant à la feuille où cette entité est assignée, pour chacun des arbres, ce qui accélère la procédure.

— Pour appliquer ce calcul de contributions de manière non-biaisée sur les jeux d'entraînement de mes modèles, j'ai introduit une étape d'association entre les entités et les arbres où ces entités n'ont pas été utilisées pour l'entraînement, du fait de la sélection aléatoire des entités par bootstrap aggregating au moment de la génération d'un nouvel arbre (voir introduction). Cela conduit à un calcul des contributions pour une entité qui n'est effectué en moyenne que sur 2/3 de la forêt, mais garantit une vision non-biaisée de ces contributions (inclure les arbres où l'entité a servi à la génération conduirait à sur-estimer les contributions des annotations de cette entité).

9.2.3 Profils des variants correctement et incorrectement classés

Les modèles de classification permettent d'obtenir un score pour toute entité à classer, que l'on peut convertir en la classe prédite par application d'un seuil. Cela conduit à l'identification de 4 types d'entités, au moment des étapes d'entraînement et validation : les vrais positifs et faux négatifs (entités de la classe positive correctement et incorrectement classées respectivement, notés TP et FN), et les vrais négatifs et faux positifs (entités de la classe négative correctement et incorrectement prédites respectivement, notés TN et FP).

Grâce au calcul des valeurs de contributions des annotations, il est possible d'aller explorer spécifiquement les profils associés à chacun de ces groupes, pour identifier potentiellement les raisons pour lesquels certaines entités sont mal classées.

J'ai choisi d'utiliser ici deux des modèles précédemment décrits : le modèle HGMD-DM avec les contrôles négatifs Cytoband-match, et le modèle eQTLs-OMIM avec les contrôles négatifs Cytoband-match également. Pour chaque modèle, les vecteurs de contributions des annotations sont calculés sur le jeu d'entraînement, comme décrit dans la section précédente. Les prédictions associées aux variants sont converties en prédictions de classes par l'utilisation de seuils identifiés pour chaque modèle. Par défaut le seuil est à 0.5 ; puisque nous avons entraîné les modèles en utilisant pour leur optimisation le score F1, j'ai à nouveau utilisé ce score pour déterminer le meilleur seuil, à partir des validations croisées effectuées sur les cytobandes. Cela m'a conduit à choisir un seuil de 0.55 pour le modèle HGMD-DM, et un seuil de 0.6 pour le modèle eQTLs-OMIM.

Pour chaque modèle, les variants sont groupés dans les 4 catégories mentionnées (TP, FN, TN, FP), et le profil moyen des contributions des annotations est calculé.

Profils pour les variants HGMD-DM

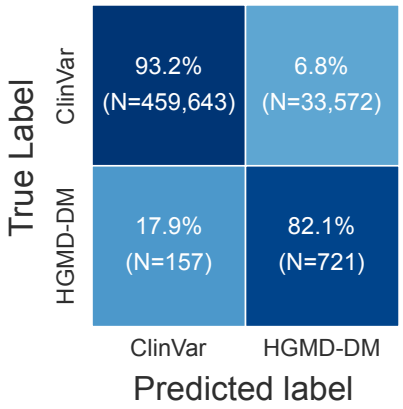
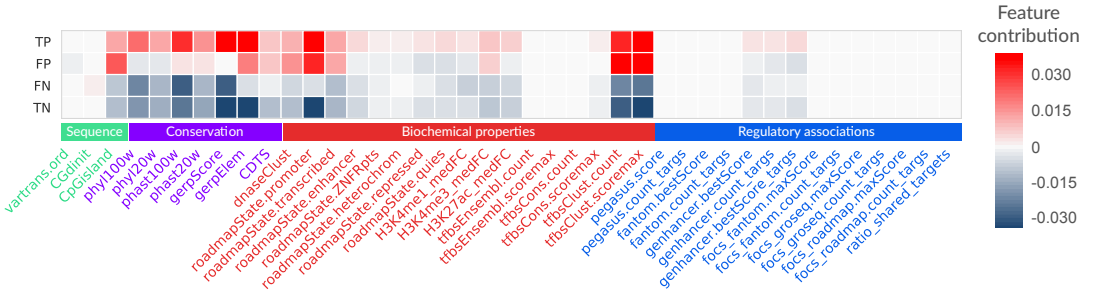


FIGURE 9.3 – Matrice de confusion pour le modèle HGMD-DM Cytoband-Match, appliqué sur son jeu d'entraînement. Le seuil sur le score de prédiction est de 0.55. Pour chaque classe réelle, les pourcentages de variants par classe prédite sont rapportés, et utilisés pour colorer les cases.

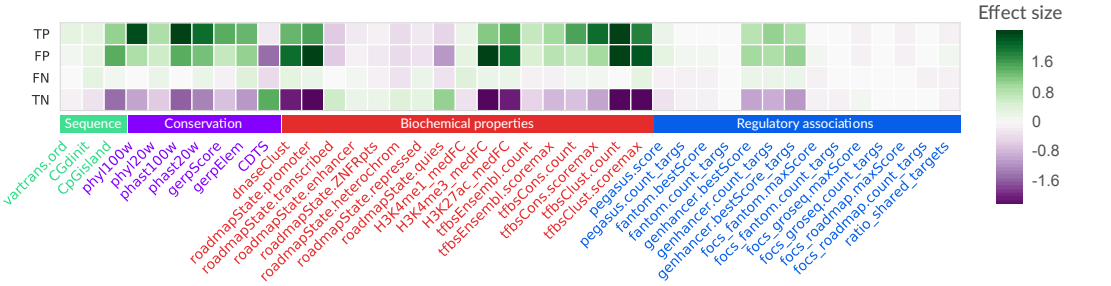
La figure 9.3 présente pour le modèle HGMD-DM Cytoband-match les nombres de variants dans chacune des 4 catégories, au seuil de 0.55 sur le score de prédiction. On peut constater qu'à ce seuil, une large proportion des variants HGMD-DM sont correctement prédits comme HGMD-DM (82.1%), de même qu'une large proportion de variants ClinVar sont bien identifiés comme contrôles négatifs (93.2%). Je souhaite donc évaluer les valeurs de contributions associées à ces catégories de variants, ainsi qu'aux variants qui sont incorrectement classés : cela permettra d'identifier les annotations ayant correctement et incorrectement contribués à ces classifications.

Les vecteurs de valeurs de contributions sont donc calculés pour chaque entité, et moyennés par catégorie ; les résultats sont présentés sur la figure 9.4a.

Variants correctement classés. On peut constater que les vrais positifs (TP) présentent effectivement des contributions avec valeurs positives (donc contribuant à un score de prédiction élevé) sur la plupart de leurs annotations En particulier les annotations de conservation en séquences phast100w, gerpScore, et gerpElem, l'annotation d'état promoteur roadmapState.promoter, et les annotations de TFBS semblent contribuer fortement



(a) Visualisation des contributions moyennes des annotations pour les quatre groupes de résultats de classification.



(b) Visualisation des tailles d'effet moyennes des annotations pour chaque groupe comparé aux autres groupes.

FIGURE 9.4 – Évaluation des profils d'annotations pour les variants HGMD-DM, suivant leur classe prédite par le modèle. Les variants sont assignés à une classe, et par comparaison à leur classe réelle, les groupes de vrais positifs (TP), faux positifs (FP), faux négatifs (FN) et vrais négatifs (TN) sont établis.

à leur identification correcte. Ce sont des annotations que l'on avait déjà mentionné lors de l'analyse des valeurs d'importances au début de ce chapitre. De manière générale, les annotations de conservation ont des contributions élevés ; les annotations concernant les états chromatiniens (autres que promoteurs) ont des valeurs plus faibles, mais positives, de même pour les prédictions d'associations régulatrices du jeu de données Genehancer. Les vrais négatifs (TN) présentent un profil de contributions en miroir de celui des TP : les annotations citées contribuent négativement aux scores de prédictions, ce qui les conduit effectivement à être détectés comme variants non-fonctionnels.

Variants incorrectement classés. Il est intéressant de regarder les profils des variants mal classés (FP et FN).

— Concernant les FP, on peut voir que leur profil de contributions est proche de celui

des TP, avec toutefois des valeurs de contributions plus faibles pour les annotations de conservation (voire nulles, comme pour les annotations phyl100w et phyl20w). Les annotations comme les états chromatiniens enhancer, quiescent, ou réprimé, qui contribuaient légèrement positivement pour les TP, sont ici des contributions négatives (mais faibles), et qui donc ont plutôt tendance à diminuer les scores de prédictions de ces variants.

— Concernant les FN, on voit que leur profil est clairement proche de celui des variants négatifs : aucune des annotations (à part l'annotation concernant la localisation dans un dinucléotide CG) n'a contribué positivement à leur score de prédiction, ce qui les conduit à être identifiés comme des variants négatifs.

Comparaison avec les tailles d'effet des annotations mesurées entre groupes. Ces valeurs de contributions des annotations ne nous permettent pas de connaître les valeurs réelles des annotations associées à ces variants : par exemple, si les variants contrôles positifs sont systématiquement moins conservés que les variants contrôles négatifs, un score de conservation phyl100w très faible pourrait être associé à une contribution au score très élevée. Je propose donc dans la figure 9.4b une visualisation des tailles d'effet entre chaque groupe, comparé au reste du jeu de données, pour évaluer ces différences réelles dans les annotations (la visualisation reprend celle proposée au chapitre des analyses préliminaires).

On peut voir que les variants des groupes TP et FP sont effectivement plus conservés que le reste du jeu de variants, ce qui est en adéquation avec les valeurs de contributions positives observées dans la première figure ; on observe aussi des valeurs de taille d'effet importantes pour les annotations correspondant au potentiel régulateur (chromatine ouverte avec le signal dnaseClust, état promoteur, marques d'histones H3K4me3 et H3K27ac), pour les annotations de TFBS, et pour les prédictions d'associations régulatrices Genehancer. Des différences légères entre ces deux groupes sont intéressantes à remarquer : les TP sont visiblement plus conservés que les FP, tandis que ces derniers ont des signaux de marques biochimiques plus importants.

Concernant les TN : on peut voir que ces variants sont beaucoup moins conservés que le reste des variants, moins présents dans des TFBS et dans les régions régulatrices Genehancer, et présentent des valeurs moins élevées pour les marques d'histones H3K4me3 et H3K27ac. En revanche ils présentent des valeurs plus élevées pour les états transcrits et quiescents. Ces variants sont donc clairement identifiables comme non-fonctionnels, ce qui est bien détecté par le modèle de prédiction. Les faux négatifs sont en revanche des

variants qui présentent des profils très peu marqués : les tailles d'effets sont très proches de 0 pour la majorité de leurs annotation. Cela signifie que ce sont des variants fonctionnels qui ne sont pas clairement distinguables des variants non-fonctionnels, ce qui explique que le modèle de prédiction ne soit pas capable de les classer correctement.

L'analyse de ces profils de contributions a donc permis de comprendre plus clairement les propriétés utilisées par le modèle de prédiction ; les profils de contributions des annotations sont en adéquation avec les différences mesurables par taille d'effet, et permettent d'expliquer plus clairement pourquoi certains variants sont incorrectement prédits. Par ailleurs, l'analyse des profils de contributions permet de voir que certaines annotations ont des valeurs de contributions nulles, comme par exemple les annotations de TFBS d'Ensembl, ou les prédictions d'associations régulatrices de la base de données FOCS. Ce sont des annotations que nous avions déjà mentionné dans la section précédente, concernant les valeurs d'importances des annotations. Si les annotations FOCS ne présentent pas non plus de tailles d'effet détectables (et pourraient donc probablement être ignorées), les annotations de TFBS ont des valeurs non-nulles de taille d'effet, qui vont dans le même sens que celles des TFBS groupés (tfbsClust). Cela signifie probablement que cette dernière ressource a pris le dessus en terme d'utilité au moment de l'entraînement. Il est toutefois intéressant pour l'interprétation d'un variant donné de garder les annotations de TFBS conservés, et de TFBS expérimentalement confirmés, malgré leur contribution nulle. Dans l'ensemble ce sont des profils de contributions cohérents avec les propriétés des variants HGMD-DM. Ils indiquent que le modèle identifie comme fonctionnels des variants sous forte pression de sélection, présents dans des régions génomiques à haut potentiel régulateur, et modifiant des sites de fixation de facteur de transcription.

Profils pour les variants eQTLs-OMIM

Je propose ici une analyse similaire pour le modèle eQTLs-OMIM CytobandMatch, avec un seuil sur le score de 0.6. La figure 9.5 présente les nombres de variants dans chacune des 4 catégories. A noter qu'à ce seuil, la proportion de variants eQTLs-OMIM correctement identifiée est faible (36.6%) ; il a néanmoins été identifié comme le seuil optimisant le score F1, ce qui signifie qu'un seuil plus faible, même s'il permettrait d'augmenter la proportion d'eQTLs correctement identifiés, diminuerait probablement la précision en introduisant beaucoup de faux positifs (déjà nombreux).

Je propose néanmoins d'analyser les profils de contributions pour chaque catégorie,

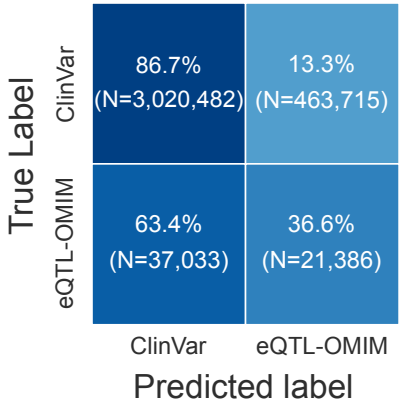


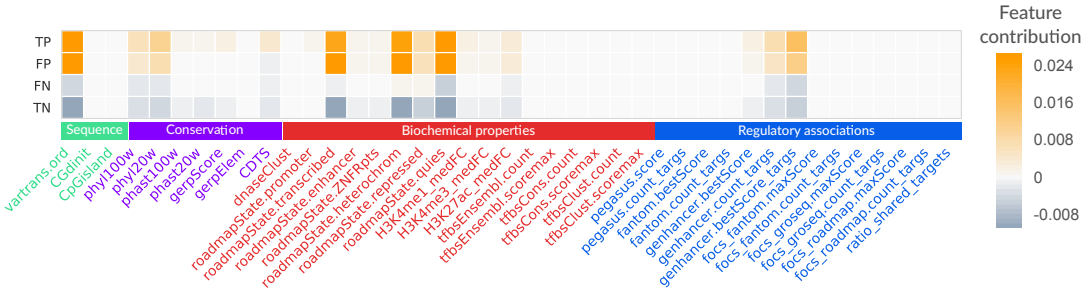
FIGURE 9.5 – Matrice de confusion pour le modèle eQTLs-OMIM Cytoband-Match, appliqué sur son jeu d'entraînement. Le seuil sur le score de prédiction est de 0.6. Pour chaque classe réelle, les pourcentages de variants par classe prédite sont rapportés, et utilisés pour colorer les cases.

comme pour le modèle HGMD-DM Cytoband-match ; les résultats sont présentés sur la figure 9.6a.

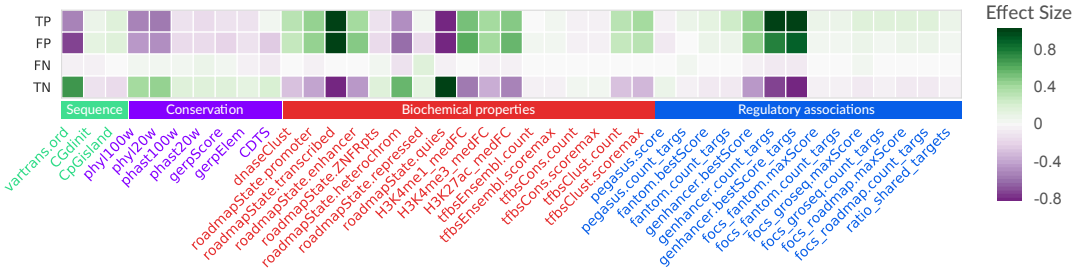
Variants correctement classés. Quatre annotations contribuent clairement à l'identification correcte des TP : vartrans.ord (dont les valeurs ordonnés indiquent si un variant correspond à une transition, une transversion, ou un INDEL), roadmapState.transcribed (indiquant les états chromatiniens transcrits), roadmapState.heterochrom (états d'hétérochromatine) et roadmapState.quies (états quiescents). Quelques autres annotations présentent des contributions positives plus faibles : les annotations de conservation, de modifications d'histones, et de prédictions d'association régulatrices GeneHancer. Comme pour les groupes du modèle HGMD-DM Cytoband-match, on voit que le groupe des TN est en miroir de celui des TP.

Variants incorrectement classés. Pour ce modèle, le groupe des faux positifs présente un profil de contributions extrêmement similaire à celui des vrais positifs, différant essentiellement pour les annotations de conservations (mais celles-ci ne présentent pas de valeurs très élevées pour les TP). Cela indique que le modèle n'est pas capable de distinguer ce variants faux positifs des vrais positifs, et que les valeurs de leurs annotations sont probablement très proches de celles de ces derniers.

Pour les variants faux négatifs (FN), la ressemblance avec les vrais négatifs est moins



(a) Visualisation des contributions moyennes des annotations pour les quatre groupes de résultats de classification.



(b) Visualisation des tailles d'effet moyennes des annotations pour chaque groupe comparé aux autres groupes.

FIGURE 9.6 – Évaluation des profils d'annotations pour les variants eQTLs-OMIM, suivant leur classe prédite par le modèle. Les variants sont assignés à une classe, et par comparaison à leur classe réelle, les groupes de vrais positifs (TP), faux positifs (FP), faux négatifs (FN) et vrais négatifs (TN) sont établis.

marquée : effectivement la plupart des annotations présentent des contributions négatives au score de prédiction, mais ces contributions sont moins notablement négatives que lorsque l'on considère le profil des TN. Ce groupe doit donc correspondre à des variants pour lesquels le score de prédiction final doit être moyen, légèrement en dessous de 0.5, indiquant que le modèle ne les identifie pas clairement comme des négatifs, mais n'est pas non plus capable d'exploiter leurs annotations pour les détecter comme de vrais positifs.

Comparaison avec les tailles d'effet des annotations mesurées entre groupes. La figure 9.6b montre les différences mesurées par taille d'effet entre chacune des catégories de variants prédits, et le reste du jeu de variants. Pour les variants TP et FP, on peut constater que certaines annotations contribuant fortement à leur classification correcte sont en adéquation avec les différences détectables depuis les valeurs de ces annotations.

Par exemples : l'annotation d'état chromatinien transcrit, et les annotations de prédictions de régions régulatrices Genehancer, ont des taille d'effet importantes, indiquant que les variants positifs sont plus présents dans ce type de régions que les autres variants. En revanche, d'autres annotations sont en oppositions avec leur taille d'effet : vartrans.ord contribue positivement à la classification, mais les tailles d'effets associées indiquent que les variants TP et FP ont des valeurs plus faibles que le reste du jeu de données (donc que ces variants sont plutôt des transitions). Une observation similaire peut être faite pour les annotations d'états "hétérochromatine" et "quiescent", ce qui indique que les variants détectés comme fonctionnels par le modèle sont effectivement moins présents dans ces états que les variants détectés comme non-fonctionnels.

Pour les faux négatifs : comme pour le modèle HGMD-DM, ces variants présentent des profils de taille d'effet très peu marqués : ce sont des variants fonctionnels qui ne sont pas clairement distinguables des variants non-fonctionnels, et ils ne sont pas non plus identifiables comme des variants fonctionnels. Le modèle a donc effectivement des difficultés pour les classifier correctement. Les vrais négatifs en revanche présentent des différences marquées :

— ils sont moins présents dans des régions potentiellement fonctionnelles (annotations de région ouverte de chromatine, d'état promoteur, transcrit, enhancer, ainsi que les TFBS et les annotations Genehancers).
— ils sont en revanche plus présents dans des états associés à un faible potentiel fonctionnel (hétérochromatine, état réprimé, et surtout état quiescent).

On peut enfin noter que pour les annotations de conservation, nous avions identifié une différence associée à une plus forte conservation des variants ClinVar en comparaison avec les eQTLs. Nous la retrouvons ici dans les contributions de ces annotations, ainsi que dans les tailles d'effets mesurés : les vrais négatifs sont clairement plus conservés que les autres groupes, ce qui a bien été identifié par le modèle comme une propriété de variant non-fonctionnel ; je discute plus en détails ce point à la sous-section de conclusion.

Ainsi, comme pour le modèle HGMD-DM Cytoband-match, l'analyse de ces profils permet de comprendre plus clairement les propriétés utilisées par le modèle de prédiction. En revanche ici, toutes les contributions ne sont pas en adéquation avec les différences mesurées par taille d'effet : par exemple, des valeurs faibles pour l'annotation "état hétérochromatinien" sont associées à des valeurs élevées de contributions. Dans l'ensemble ce sont des profils de contributions cohérents avec les propriétés des variants eQTLs. Ils

indiquent que le modèle basé les eQTLs identifie comme fonctionnels des variants plutôt présents dans des régions génomiques à haut potentiel régulateur, mais qui sont moins conservés que les variants contrôles négatifs.

Conclusions sur les profils des catégories de prédiction

Le calcul de vecteurs de contributions des annotations, pour chacun des variants des différents jeux de données, a permis d'aller explorer plus en détails les règles de décisions apprises par nos modèles. Les valeurs d'importances nous donnaient déjà un premier aperçu de l'utilisation des annotations, mais ne permettaient pas de clairement établir la façon dont étaient utilisées ces annotations. Grâce aux valeurs de contributions, nous avons pu aller plus loin dans la compréhension de nos modèles. Nous avons ainsi établi que les groupes de variants correctement et incorrectement identifiés pouvaient effectivement se confirmer par des utilisations distinctes des annotations : les variants faux positifs sont effectivement détectés par les modèles comme similaires aux vrais positifs, et la même observation peut être faite pour les faux négatifs et vrais négatifs. Les comparaisons avec les tailles d'effets mesurables entre catégories a également permis de comprendre comment le lien entre les différences observables dans les distributions de valeurs, et les valeurs de contributions que les modèles y associent. Par exemple pour le modèle HGMD-DM Cytoband-match, les contributions positives des annotations de conservation sont associées à des valeurs plus élevées des scores. En revanche pour le modèle eQTLs-OMIM Cytoband-match, les contributions des annotations de conservation sont en opposition avec les différences mesurées par taille d'effet : les variants vrais positifs ont des valeurs plus faibles pour ces scores que les autres variants. Cela peut s'expliquer par la nature des eQTLs : ces variants correspondent en majorité à des SNPs marqueurs, identifiés comme eQTLs par association statistique à des niveaux d'expression de gènes. Comme évoqué dans le chapitre 2 de l'introduction, la détection d'une association statistique est facilité par la fréquence élevée du variant (plus il y a de possibilités de voir le variant dans chacun des états homozygotes et héterozygote, plus l'association statistique sera aisée). Ces positions sont donc peu probablement soumises à une pression de sélection évolutive.

La séparation des variants effectuée dans cette partie a été limitée aux quatre catégories définies par les classes réelles et prédites par les modèles. Il est cependant possible d'explorer encore plus en détails ces profils de contributions : c'est ce que je présente dans la section suivante.

9.2.4 Identification de structures dans les contributions des annotations

Les valeurs de contributions des annotations permettent pour chaque variant de comprendre comment ses annotations ont été exploitées par le modèle de prédiction, et combien elles ont contribué au score final de fonctionnalité fourni par le modèle. La partie précédente a permis de distinguer les propriétés spécifiques aux catégories de classification obtenues après un seuil sur le score. Cependant, nous avons pris les variants contrôles positifs et contrôles négatifs comme des ensembles chacun homogène ; les modèles ont été entraîné à distinguer ces deux classes. Mais il est probable que tous les variants d'un groupe contrôle ne soient pas identiques ; c'est d'ailleurs ce qui semble conduire les modèles à se tromper, générant les faux positifs et faux négatifs.

Dans cette partie, je cherche donc à identifier des structures particulières parmi les variants contrôles, en me basant sur les différents profils de contributions de leurs annotations. Pour cette analyse, je me focalise sur le modèle HGMD-DM Cytoband-match, et sur le jeu de variants contrôles positifs de ce modèle : nous avons vu dans la section précédente qu'au seuil fixé, le modèle se trompait sur une minorité des variants HGMD-DM, et que les profils associés étaient clairement définis. Par ailleurs ce modèle sera le seul appliqué dans le chapitre suivant, en partie vis-à-vis du contexte de l'application, et également parce que ce modèle présente les meilleures performances ; j'ai donc préféré limiter la présentation de ces analyses à ce modèle.

Je souhaite donc explorer deux points :

— est-ce que la forêt aléatoire a détecté différentes communautés de variants au sein de l'ensemble des variants contrôles positifs utilisés pour l'entraînement ?

— est-ce que ces communautés sont associées à des qualités de prédiction différentes ? Retrouve-t-on par exemple la séparation entre vrais positifs et faux négatifs ?

Détection et prédiction de communautés

La première étape consiste à détecter des communautés de variants dont les vecteurs de contributions des annotations sont similaires, par l'algorithme de K-means. L'algorithme des K-means est une forme d'apprentissage non-supervisé, qui permet de définir des groupes à partir d'une mesure de distance entre des entités ; le paramètre K correspond au nombre de communautés attendues. N'ayant pas d'a priori sur ce nombre, j'ai choisi d'explorer un nombre croissant de communautés à prédire. L'optimisation de ce nombre s'est faite sur la base du score d'inertie, qui mesure la cohérence interne et dont

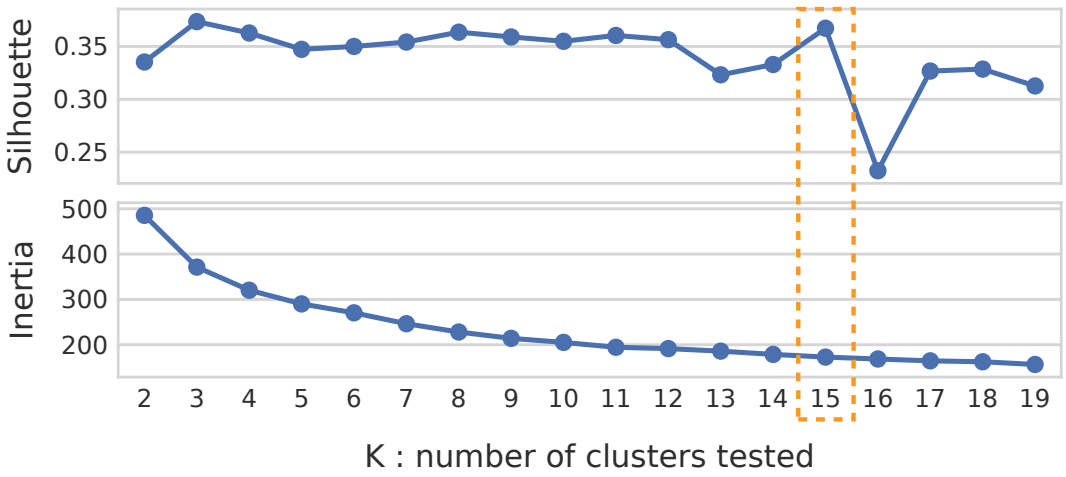
on cherche à minimiser la valeur, et du score de silhouette, qui mesure la différence entre les distances inter- et intra-groupes, et dont on cherche à maximiser la valeur (voir chapitre des méthodes).

J'ai donc utilisé l'algorithme des K-means sur les vecteurs de contributions des 878 variants HGMD-DM, avec un nombre de clusters allant de 2 à 19 ; les valeurs d'inertie et de silhouette ont été mesurées pour évaluer la qualité d'identification de ces groupes. Les résultats sont présentés sur la figure 9.7. Pour identifier le nombre de groupes optimal, j'ai cherché le nombre K associés à une inertie minimale (les communautés prédites sont homogènes) et à une silhouette maximale (les communautés prédites sont homogènes ainsi que distinctes les unes des autres) ; cela m'a conduit à sélectionner un nombre de 15 groupes (figure 9.7a).

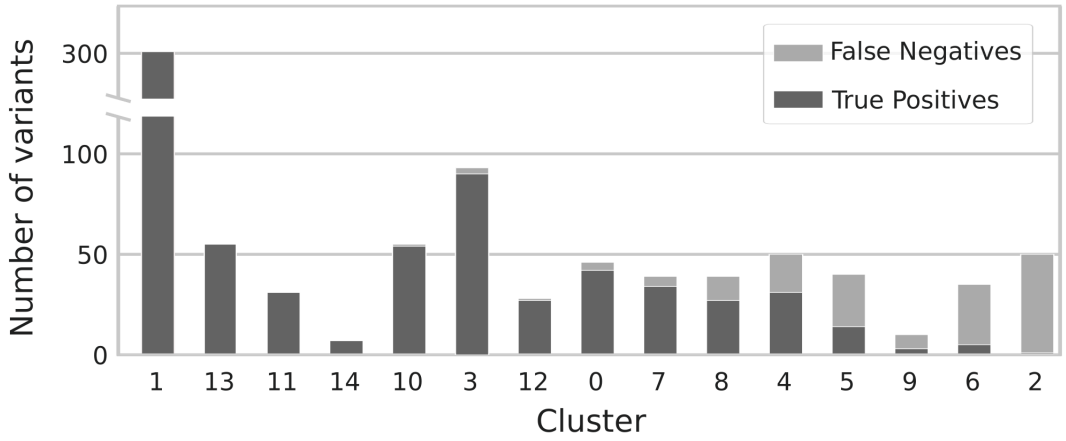
En regardant la composition en variants correctement et incorrectement prédits pour ces groupes, on peut voir sur la figure 9.7b) que les groupes 2, 6, 9 et 5 concentrent les faux négatifs, tandis que les groupes 4, 8 et 7 ont des compositions en vrais positifs et faux positifs plus partagées ; les autres groupes sont principalement composés de vrais positifs. Cela signifie donc que parmi les vrais négatifs, la forêt a appris à identifier 4 profils de variants différents, tandis que pour les vrais positifs, 8 profils différents sont identifiables, dont un (le 1) qui semble concentrer la majorité de ces variants. J'explore ces profils ci-dessous.

Analyses des profils des communautés

La figure 9.8a présente les profils résumés des contributions pour chacune des 15 communautés prédites. A la section précédente, nous avions identifié que les vrais positifs étaient identifiables par des contributions élevées provenant des annotations de conservations, d'état promoteur, et de TFBS groupés. Les faux négatifs quant à eux ressemblaient à des vrais négatifs : les annotations de conservations, d'état promoteurs, et de localisation dans les TFBS, étaient négatives, et conduisaient donc ces variants à avoir des scores de prédictions faibles. On peut voir dans cette représentation que les communautés identifiées présentent des profils de contributions beaucoup plus divers ; on peut par exemple remarquer que tous les groupes à majorité de vrais positifs n'ont pas des contributions systématiquement positives provenant des annotations de conservations. Je décris ci-dessous plus en détails les différences identifiables entre ces communautés prédites. Les annotations des variants de chaque groupe sont comparées à celles de l'ensemble du jeu de données



(a) Mesures de qualité d'identification de groupes par K-means en fonction du nombre de groupes découverts.



(b) Composition en vrais positifs et faux négatifs pour un nombre K=15 de groupes découverts.

FIGURE 9.7 – Découverte de groupes par l'algorithme des K-means sur les vecteurs de contributions des variants fonctionnels HGMD-DM. 9.7a Un nombre croissant de groupes a été exploré, et deux mesures ont permis d'évaluer la qualité des groupes identifiés. Le choix s'est porté sur un nombre K=15 de groupes. 9.7b Le nombre et la composition des 15 groupes sont rapportés. Les groupes ont été ordonnés par le pourcentage décroissant de vrais positifs, et par la taille.

d'entraînement, et représentées par leur taille d'effet dans la figure 9.8b. Les deux sources d'informations (contribution moyenne d'une annotation, et taille d'effet mesurable entre le groupe et le reste du jeu de données) sont présentées dans les descriptions :

— Le groupe 1, concentrant la majorité de vrais positifs, correspond clairement au profil identifié précédemment : les variants sont très conservés, localisés dans des régions à haut potentiel régulateur (état promoteur, région Genehancer, avec des signaux de marques d'histones relatifs aux régions régulatrices), et impactent potentiellement des TFBS. Ce sont tous ces signaux qui sont effectivement détectés par la forêt aléatoire, ce qui se traduit par des valeurs de contributions élevées pour les annotations correspondantes, et conduit ces variants à des scores de prédiction élevés.
— Le groupe 12 est similaire au premier, à l'exception des annotations de concentration de TFBS (tfbsClust) et prédictions Genehancer ; l'absence de signal pour tfbsClust a même été associé à une contribution négative de cette annotation au score de prédiction. Il est à noter cependant que ces variants présentent une taille d'effet importante pour les annotations tfbsCons ; cependant la contribution associée est faiblement positive.
— Le groupe 11 ressemble également au premier, excepté pour les annotations phyl100w et phast100w, qui présentent des contributions négatives. En regardant les tailles d'effet, les valeurs de conservations pour ces variants sont en moyenne plus élevées que le reste du jeu de variants, mais semblent effectivement moins conservés que celles du groupe 1. Le caractère fonctionnel des variants a donc été associé par le modèle à des valeurs de conservation particulièrement élevées pour les annotations phyl100w et phast100w. Ces variants restent néanmoins associés à des scores de prédictions élevés.
— Le groupe 14 présente aussi des contributions négatives provenant des annotations de conservation en séquence, ainsi que des contributions négatives provenant des annotations d'états chromatiniens promoteur et transcribed. Les tailles d'effets pour ces deux états confirment que ces variants sont situés dans des régions particulièrement bien identifiées comme transcrites. Il est donc probable que leur caractère fonctionnel soit associé à un impact sur le transcrit lui-même, plutôt que sur le contrôle de la transcription.
— Le groupe 10 concentre des variants qui ne sont pas présents dans des éléments conservés GERP, mais apparaissent néanmoins conservées selon ce score, ce qui est associé à une contribution positive importante de cette annotation.
— Le groupe 3 est un groupe intéressant : il concentre un nombre élevé de variants

fonctionnels qui ont des scores de conservation légèrement plus faibles que le reste du jeu de variants, ce qui se traduit par des valeurs de contributions négatives de ces annotations. Cependant ils semblent localisés dans des ilôts CpG, des éléments régulateurs (régions ouvertes, états promoteurs) et dans des TFBS, et cela a contribué positivement à leur scores de prédiction finaux, et donc à une concentration de vrais positifs.

— Les groupes 12 et 0 diffèrent essentiellement par les contributions des annotations phast100w et phast20w : elles sont positives pour le groupe 12, et négative et nulle (respectivement) pour le groupe 0. Cela se retrouve dans les tailles d'effets mesurés : ce sont des groupes qui apparaissent comme conservés par rapport au reste du jeu de variants, mais la conservation est plus notable pour le groupe 12. Par ailleurs ces deux groupes sont définis par une forte contribution de l'annotation gerpElem, des annotations de TFBS groupés, et dans une moindre mesure, des annotations associées aux régions régulatrices (état promoteur, chromatine ouverte).

— Le groupe 7 est défini par une très forte contribution des annotations de conservations, que l'on retrouve dans les tailles d'effets. Les annotations corrélées avec cette annotation contribuent également positivement à leur score de prédiction : ce sont des variants qui sont localisés dans des TFBS conservés, et dans des éléments régulateurs prédits par PEGASUS (basé sur la conservation).

— Les groupes 8 et 4 font partie des groupes où la composition en vrais positifs est plus faible. Cela signifie que pour certains des variants, la somme des contributions positives n'a pas suffit à contrebalancer les contributions négatives, et le score final de prédiction était en dessous du seuil. Les contributions négatives proviennent des annotations de conservation : ces groupes ont effectivement des tailles d'effets moyennes nulles ou légèrement négatives en comparaison avec le reste du jeu de données. Ils sont cependant associés à un potentiel régulateur important, identifiable par une forte contribution positive des annotations tfbsClust, ainsi qu'une contribution positive des annotations de chromatine ouverte (plutôt pour le groupe 8), d'état promoteur (très marqué pour le groupe 4), ainsi que d'état transcrit.

— Le groupe 5 est le premier des groupes à concentrer une majorité de variants faux négatifs. Les variants de ce groupe partagent le profil moyen de contributions qui est présenté, mais seule une minorité a obtenu une somme de contributions positives suffisamment élevée pour dépasser le seuil de 0.55 fixé. On peut voir que ces variants

ont des contributions négatives provenant de leurs annotations de conservation en séquence, excepté pour l'annotation gerpElem ; l'observation des tailles d'effets confirme que malgré leur présence dans des éléments identifiés comme conservés, ces variants correspondent à des positions faiblement conservées, qui ont conduit le modèle à assigner une contribution négative pour ces annotations. Par ailleurs, l'absence de TFBS associés à ces positions a également contribué fortement négativement à leur score de prédiction.

— Les groupes 9, 6, et 2 sont les autres groupes qui concentrent les variants faux négatifs. Contribution négative des annotations de conservation : ils sont effectivement peu ou pas conservés par rapport au reste du jeu de données. Contribution négative de l'annotation état promoteur (excepté pour le groupe 6), et des annotations TFBS : ces variants, malgré leur label "fonctionnel" ne présentent pas d'annotations ayant permis au classifieur de les détecter correctement.

Ainsi, cette approche d'identification de communautés parmi les variants fonctionnels du modèle HGMD-DM Cytoband-match permet de voir que les vrais positifs et les faux négatifs ne sont pas des catégories complètement homogènes. Le modèle de prédiction est en effet capable d'identifier des combinaisons d'annotations différentes, probablement associées à des propriétés biologiques différentes pour ces variants.

La description et l'analyse succincte des différentes communautés confirme qu'une grande partie des variants fonctionnels ont effectivement le profil précédemment associé à la catégorie des vrais positifs : le groupe 1 concentre la majorité de ces vrais positifs, et on y retrouve le profil de contributions associé à une forte conservation, et des signaux de potentiel régulateur et d'impact de TFBS marqués. Cependant d'autres communautés de vrais positifs sont nettement identifiables, avec des profils particuliers. Par exemple le groupe 3 (deuxième groupe de variants vrais positifs le plus large) contient des variants plutôt peu ou pas conservés, mais associés à des signaux de régions fonctionnelles particulièrement forts (les signaux d'état promoteur, d'ilôts CpG, et de localisation dans des TFBS). Il est intéressant de noter que les annotations de conservations, qui peuvent paraître redondantes, sont utilisées de manières différentes par le modèle de prédiction, ce qui conduit à des combinaisons de contributions distinctes dans les communautés ; étant données que ces valeurs de conservation sont mesurées à différentes échelles de temps, il est possible que les variants associés présentent des niveaux de conservation variables suivant les espèces comparées.

La diversité de populations de variants, a priori ignorée (puisque tous ces variants ont été classé simplement comme ”fonctionnels”), a pu être détectée par la forêt aléatoire, par la génération de chemins de décisions différents, qui se retrouvent dans les vecteurs de contributions. Le modèle n'est cependant pas parfait, et l'analyse des groupes de variants concentrant les faux négatifs confirme que ces variants ne présentent pas de valeurs d'annotations suffisamment distinctes des variants non-fonctionnels, ce qui conduit le modèle de prédiction à se tromper.

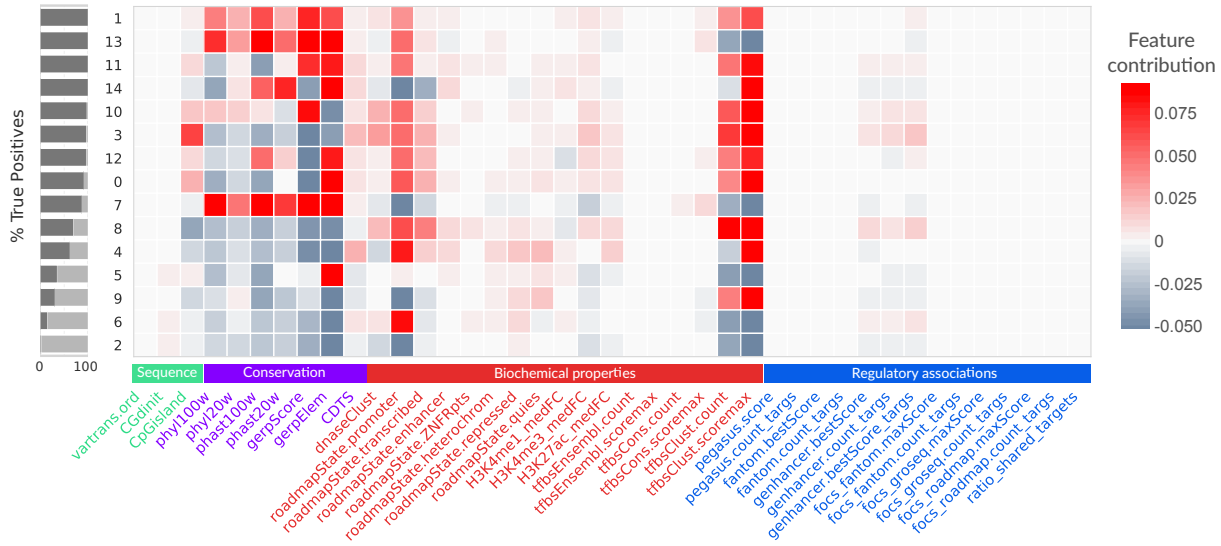
9.3 Conclusions

J'ai proposé dans ce chapitre deux approches pour explorer les modèles de prédiction entraînés dans le cadre de ce projet. La première approche permet de calculer une valeur d'importance de chaque annotation dans un modèle. Ce calcul donne une vision générale de l'apport de chaque annotation à la séparation des contrôles positifs et négatifs dans chacun des modèles. Les comparaisons de ces valeurs d'importance ont permis d'identifier que certaines annotations étaient probablement superflues, et pourraient être exclues des modèles sans conséquence sur la qualité de classification. Par ailleurs, les comparaisons de ces valeurs d'importance entre modèles a permis de voir que le schéma d'échantillonnage Distance-match conduit à une diminution de l'importance des annotations d'intervalles en comparaison avec l'échantillonnage Cytoband-match ; ceci est vrai pour les modèles HGMD-DM et pour les modèles eQTLs-OMIM. Cette diminution est due au fait que les contrôles négatifs Distance-match se trouvent localisés plus probablement dans les mêmes intervalles génomiques que les contrôles positifs ; les annotations correspondantes ne sont alors plus discriminantes, et celles dont les valeurs sont disponibles par position individuelle deviennent plus importantes pour le modèle.

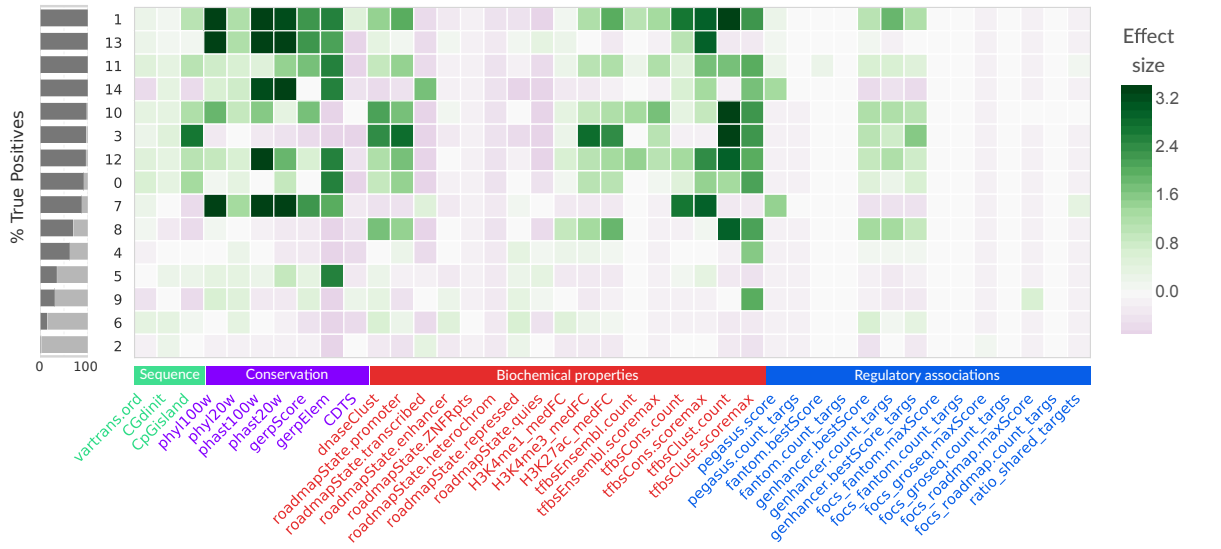
Pour aller plus loin dans la compréhension des modèles de prédiction, j'ai proposé une deuxième approche consistant à calculer des valeurs de contributions depuis les chemins de décisions des arbres des forêts aléatoires. Ces contributions sont calculables pour chaque variant classé par mes modèles, et permettent donc de comprendre plus en détails comment sont utilisées les valeurs particulières de annotations. Pour les modèles HGMD-DM Cytoband-match, et eQTLs-OMIM Cytoband-match, l'analyse des catégories vrais et faux positifs et vrais et faux négatifs a permis de comprendre quelles sont les annotations exploitées par ces modèles pour identifier les vrais positifs et vrais négatifs, et a permis de confirmer que les faux négatifs ont effectivement des profils proches des vrais

négatifs (de même pour les faux et vrais positifs). Concernant le modèle eQTLs-OMIM Cytoband-match, j'ai souligné que les annotations de conservation sont utilisées par le modèle de manière opposée à leur utilisation dans le modèle HGMD-DM Cytoband-match : les variants eQTLs-OMIM sont associés à des valeurs de conservation plus faibles que les variants ClinVar. Enfin, une identification de communautés parmi les contrôles positifs du modèle HGMD-DM Cytoband-match a conduit à identifier des sous-groupes de vrais positifs associés à des contributions de leurs annotations distinctes, confirmant la capacité des forêts aléatoires à reconnaître des propriétés et combinaisons d'annotations particulières dans un ensemble de variants hétérogènes, qui ne sont pas traités tels quels a priori.

Ces analyses sont donc l'occasion de comprendre plus en détails comment les modèles de prédictions entraînés ont exploité les annotations. Cette meilleure compréhension permet de mieux comprendre quelles sont les propriétés que les modèles seront capables de détecter pour de nouveaux variants à classifier, et d'établir les limites d'application des modèles. Par exemple, il apparaît clairement que les propriétés identifiées par les modèles basés sur les variants eQTLs sont en partielle opposition avec celles identifiées par les modèles basés sur les variants HGMD-DM : la forte conservation de ces derniers (pour une majorité) sera probablement associée à au caractère "non-fonctionnel" par les modèle eQTLs. Comme je l'ai évoqué, la nature de ces variants eQTLs est potentiellement la source de cette différence : ils correspondent en majorité à des SNPs marqueurs de blocs de déséquilibre de liaison, dont l'identification est basée sur une forte variabilité dans la population, ce qui contribue également à leur association statistique aux variations d'expression génique ; le choix de les utiliser comme contrôles positifs a été motivé par l'idée que ces positions marquent des régions d'activité fonctionnelle (ce qui semble être détecté par le modèle, notamment pour les signaux de marques d'histones activatrices, et la localisation dans des états plutôt fonctionnels), mais concernant l'absence de conservation de la position exacte, il est probable que cela soit dû à leur nature de variant très variable. Le modèle HGMD-DM Cytoband-match quant à lui dépend des annotations de conservation, d'état promoteur, et de la présence des variants dans des TFBS identifiés par l'annotation tfbsClust ; les combinaisons de ces annotations sont diverses, mais des valeurs faibles dans les trois conduiront probablement un variant à être identifié comme non-fonctionnel, malgré d'autres annotations pertinentes.



(a) Contribution moyenne des annotations pour les variants HGMD-DM, après identification de groupes par K-means.



(b) Taille d'effet moyenne des annotations pour les variants HGMD-DM, après identification de groupes par K-means.

FIGURE 9.8 – Contributions et tailles d'effet moyennes des annotations pour les variants HGMD-DM, après découverte de groupes par K-means. (9.8a) Les vecteurs de contributions sont utilisés pour identifier 15 groupes parmi les variants fonctionnels, après leur classification par le modèle. Chaque groupe est associé à une composition plus ou moins importante de vrais positifs et faux négatifs. (9.8b) Les contributions identifiées par le modèle correspondent plus ou moins directement à des différences d'annotations entre les variants de chaque groupe et le reste du jeu de données, mesurées par leur taille d'effet.

Chapitre 10

Application - mutations *de novo* et Autisme

10.1 Introduction et données

10.1.1 La cohorte "Simons Simplex Collection"

L'organisme SFARI (Simons Foundation Autism Research Initiative, `https://www.sfari.org/`) gère un programme de recherche sur l'autisme, en finançant des projets de recherche sur les causes génétiques de cette maladie, et sur la découverte de potentiels traitements. Dans le cadre de ce programme de recherche, différentes cohortes de patients sont rassemblées, pour faciliter l'étude sur les causes de l'autisme. L'une de ces cohortes est appelée **Simons Simplex Collection** ou SSC (`https://www.sfari.org/resource/simons-simplex-collection/`), et regroupe 2,600 familles chez lesquelles un enfant est atteint d'autisme, tandis que les autres membres de la famille sont sains (parents et frères et soeurs). Ces configurations sont appelées "simplex", et correspondent à une occurrence isolée de la maladie, dont l'apparition peut être expliquée par une mutation *de novo*, un motif particulier d'hérédité, ou une pénétrance variable. L'article de Ronemus et al., 2014 se penche sur la question des différents modèles de transmission de la maladie, et évoque notamment la composition de cette cohorte.

Dans un article récent (An et al., 2018), des chercheurs se sont focalisés sur l'analyse de 1,902 familles provenant de la cohorte SSC. Pour ces familles, les mutations *de novo* ont été obtenues chez les enfants suite au séquençage des génomes complets de chaque

parent, de l'enfant atteint du trouble du spectre de l'autisme, ainsi que d'un second enfant ne présentant pas les symptômes de la maladie, et donc pris comme patient contrôle. Au total, 255 106 mutations *de novo* ont été identifiées parmi tous les enfants de la cohorte, avec en moyenne 61.5 SNV et 5.6 INDEL *de novo* chez les enfants malades. Ces variants ont été identifiés grâce à une combinaison de filtres sur les différents critères de qualité d'appel des variants, qui a été décrite dans une de leurs précédentes publications (Werling et al., 2018) ; le taux de faux positifs, évalué grâce à la validation par séquençage Sanger de plus d'un millier de variants, est extrêmement faible (1.2%).

Les auteurs ont ainsi exploré les enrichissements de ces mutations *de novo*, spécifiquement chez les patients malades ou chez les contrôles, pour 55 143 catégories de régions génomiques, manuellement définies à partir de différents critères d'annotations (score de conservation, listes de gènes, annotations de régions fonctionnelles, etc). Dans leur analyse, il apparaît que seules les mutations *de novo* affectant la séquence codante de gènes présentent un signal d'enrichissement chez les patients malades statistiquement significatif. Or ces variants codants représentent une minorité parmi l'ensemble des mutations *de novo*, avec une à deux mutations *de novo* codantes observées par patient (Acuna-Hidalgo et al., 2016). Les auteurs ont alors défini un modèle de classification pour le statut "malade", par l'utilisation d'un modèle de régression avec sélection d'annotations (régression lasso), entraîné sur les enrichissements identifiés dans les différentes catégories. Finalement, le modèle se base en majorité sur les enrichissements dans les catégories concernant les gènes ; par ailleurs, parmi les 163 catégories "non-codantes" du modèle (sur 238), le pouvoir prédicteur semble provenir principalement de 45 catégories concernant les promoteurs, et en particulier des positions mutées présentant une conservation en séquence importante. Par ces analyses, les auteurs ont proposé une caractérisation des propriétés générales des mutations *de novo* identifiées chez ces patients, plutôt qu'une approche gène-candidat. Cependant leur approche se révèle limitée pour l'association de variants non-codants à la maladie. Par ailleurs, aucune estimation n'est proposée concernant le nombre de patients expliqués par les mutations testées.

J'ai donc souhaité évaluer l'apport de mon outil FINSURF à l'identification de mutations candidates chez ces familles. En effet, l'apport d'un modèle de prédiction entraîné spécifiquement sur des profils de mutations régulatrices peut potentiellement permettre d'augmenter le nombre de mutations fonctionnelles candidates chez les patients malades, et dont le caractère significatif ne dépendrait pas d'une accumulation dans des régions gé-

nomiques données. L'utilisation des prédictions d'associations peut également contribuer à l'identification de mutations candidates dans des régions associées à des gènes d'intérêt, mais en dehors des gènes. Je présente les résultats de cette analyse dans les sections suivantes.

10.1.2 Étapes préliminaires et filtres

Les données des mutations *de novo* étaient disponibles dans les documents supplémentaires de l'article, ce qui m'a permis de les télécharger, les annoter, et les évaluer avec FINSURF.

J'ai appliqué quelques étapes préliminaires de traitement et de sélection des variants avant analyse. Par exemple, certains patients sont considérés comme déjà diagnostiqués ; or nous souhaitons proposer une approche diagnostique d'identification de mutations candidates par patient : il n'est donc pas pertinent d'inclure les familles déjà "résolues". De même pour les patients chez qui une mutation codante à fort impact fonctionnel est trouvée dans un gène pertinent. Je décris ci-dessous un ensemble de filtres, appliqués indépendamment sur l'ensemble des mutations ; un résumé des filtres et des nombres de variants exclus suivra ces descriptions.

Génome de référence. Les variants présentés dans l'article ont initialement été obtenus par appel de variants sur la version hg20 de l'assemblage du génome humain. Mon outil et mes annotations étant basés sur la version hg19, j'ai tout d'abord converti les positions des variants *de novo* d'un assemblage à l'autre, grâce à l'outil liftOver. Cette conversion ne s'effectue pas toujours sans perte, dues aux présences et absences de certaines régions spécifiques à chaque assemblage. Ainsi, le jeu de données initial contient 255 106 mutations *de novo* identifiées sur la version hg38 du génome. Durant l'étape de liftOver, 336 variants sont perdus, tous correspondant à des positions absentes dans la version hg19 du génome.

Séquences nucléotidiques. Dans le cadre d'une analyse de motifs de facteurs de transcription chevauchant des variants d'intérêt (voir section suivante), les séquences nucléotidiques des régions contenant les variants ont été téléchargées grâce à l'outil "fetch-sequence" de la suite RSAT. Cette étape a conduit à la perte de 185 variants dont la séquence de référence identifiée dans le fichier VCF ne correspond pas à celle obtenue.

Variants codants et familles diagnostiquées. Dans l'article présenté plus haut (An et al., 2018), l'analyse de la collection des 1 902 familles a porté sur l'enrichissement en mutations *de novo* dans différentes catégories d'annotations génomiques, en comparant les patients malades et les patients contrôles. Cet enrichissement n'empêchait pas l'analyse des catégories "non-codantes" ; cependant pour un patient dont le phénotype est expliqué par un variant codant, il est peu probable que ses mutations *de novo* non-codantes contribuent également à la maladie. J'ai donc souhaité ici exclure les familles déjà diagnostiquées, ainsi que celles pour lesquelles une mutation *de novo* codante pouvait être identifiée comme pertinente.

Dans un premier temps, j'ai téléchargé depuis les documents supplémentaires de l'article un tableau contenant différentes métadonnées sur les patients, dont l'information sur le diagnostic (établi à partir de l'identification de mutations tronquant des gènes pertinents, de larges délétions, ou de réarrangements chromosomiques). Ce diagnostic est ainsi établi pour 211 familles.

J'ai également effectué une analyse complémentaire des mutations *de novo* avec l'outil VEP (McLaren, Pritchard et al., 2010, McLaren, Gil et al., 2016), qui permet d'obtenir un résumé de conséquences potentielles des variants, notamment sur les séquences codantes. J'ai ainsi identifié tous les variants codants prédits pour avoir un impact sur la séquence protéique de gènes associés à l'autisme (provenant de la liste de gènes de la base de données de SFARI, voir méthodes). Les conséquences rapportées par VEP et identifiées comme importantes sont les suivantes : impact sur des sites d'epissage, impact sur la séquence codante (synonyme, faux-sens, ou non-sens), décalage du cadre de lecture, et impact sur un codon start ou stop. Sont ainsi expliquées 144 familles (dont 80 déjà précédemment diagnostiquées), chacune expliquée par une mutation ; les mutations se répartissent comme suit :

— décalage de cadre de lecture : 64
— gain de codon stop : 50
— impact de site d'epissage : 28
— perte d'un codon stop : 2

L'ensemble des mutations *de novo* provenant de ces familles sera donc exclu.

Identification de faux-positifs parmi les mutations *de novo*. J'ai souhaité identifier de potentiels évènement de positions mutées de manière indépendante chez plusieurs familles, qui pourraient correspondre à des mutations candidates pour expliquer les phé-

notypes. Cette étape conduit à identifier 3 340 variants "dupliqués" sur l'ensemble de la cohorte (correspondant à des variants dont les positions affectées sont retrouvées chez plusieurs patients).

Cependant des étapes supplémentaires de caractérisation de ces mutations multiples conduisent à en identifier 2 032 présentes à la fois chez un patient malade et un patient contrôle d'une même famille. Ces cas pourraient correspondre à une erreur de détection du variant chez l'un des parents, ou bien à une situation de mosaïcisme, conduisant à la présence d'une mutation restreinte à la lignée germinale chez l'un des parents, et donc transmise systématiquement par le parent à ses enfants. Par ailleurs, 346 mutations *de novo* sont identifiées comme dupliquées au sein d'un même patient ; c'est à dire que la même position nucléotidique est affectée par plusieurs mutations différentes chez un même patient. Ces situations correspondent potentiellement à des erreurs de détection des variants à ces positions ; il est également possible que ces variants affectaient différentes positions dans l'assemblage hg38. Enfin, quatre mutations sont identifiées comme dupliquées chez des patients distincts, mais les mutations associées sont exactement inversées entre patients (par exemple : à la position 99695240 du chromosome 9 une mutation C>CTTTCT est détectée chez le patient 14540.p1, tandis que chez le patient 11963.p1, la mutation à cette position est CTTTCT>C) ; ces cas étranges sont écartés. Ainsi, seules 958 mutations sont identifiées au sein de la cohorte comme étant de multiples événements de novo survenus aux mêmes positions de manière indépendante chez les différents patients, et sont donc gardées dans les analyses.

Nombres aberrants de mutations *de novo*. Si en moyenne les patients présentent un nombre attendu de mutations *de novo* (environ 65), j'ai constaté que certaines familles présentaient des nombres particulièrement élevés (ou bas) de mutations *de novo* chez l'un des enfants : le maximum identifié chez un patient est de 229 mutations *de novo*, tandis que le minimum est à 33. Par ailleurs, entre deux enfants d'une même fratrie, les différences attendues en terme de comptages sont assez faibles, dépendant principalement de l'âge du père à la conception. La différence moyenne calculée pour cette cohorte est de 11 mutations *de novo* entre 2 enfants d'une famille ; mais certaines familles présentent des nombres inattendus : notamment, la différence maximale associée à l'une des familles est de 128.

J'ai donc défini des seuils sur les nombres minimum et maximum de variants identifiés par patient, ainsi que la différence maximale acceptée entre deux enfants d'une même

fratrie ; les seuils sont identifiés depuis les distributions de ces valeurs, et fixés à 2 déviations standards de la moyenne (soit une différence maximum de 30 mutations entre deux enfants, et des comptages de mutations *de novo* compris entre 47 et 99). L'ensemble des familles dont les valeurs dépassent ces seuils sont exclues : 121 au total.

Résumé des filtres et nombre final de variants. Au total ce sont 50 666 mutations *de novo* qui sont exclues, soit 19.9% des 254 770 mutations obtenues après changement des coordonnées des mutations. La principale source d'exclusion des mutations correspond au caractère diagnostiqué des familles (51.2% des mutations exclues). La seconde source d'exclusion (32.3%) correspond aux familles dont les nombres de mutations *de novo* identifiées chez les enfants sont anormaux par rapport au reste de la cohorte : 113 familles parmi les 121 sont exclues selon ce critère uniquement ; les 8 autres familles sont aussi exclues sur la base d'un diagnostic établi, et sont donc comptées indépendamment de ces deux premières catégories (correspondant donc à 2.3% des mutations exclues). Les autres critères contribuant à l'exclusion sont le caractère codant des mutations (7.3% des mutations), et les variants identifiés comme faux positifs pour le caractère *de novo* (3.9%).

Ces quatre critères représentent 97% des variants exclus, les autres étant exclus selon différentes combinaisons des critères possibles (par exemple, 610 variants sont exclus à la fois parce qu'ils sont codants, et qu'ils sont trouvés chez des patients déjà expliqués). Les analyses qui suivent sont donc effectuées sur les 204 104 mutations *de novo* restantes, trouvées chez 1 582 familles, pour lesquelles on peut considérer l'hypothèse qu'une ou plusieurs mutations non-codantes affectant un gène d'intérêt sont à l'origine de la maladie chez le patient concerné.

10.2 Analyses des mutations *de novo*

Dans l'ensemble des analyses qui suivent, le score FINSURF utilisé pour évaluer la fonctionnalité des mutations *de novo* provient du modèle HGMD-DM Cytoband-match. Ce choix est basé à la fois sur la qualité des prédictions évaluées par les étapes de validation croisée (voir chapitre 8), et sur les jeux de variants qui ont servi à son entraînement. Les variants contrôles positifs de ce modèle sont des variants fonctionnels associés à des maladies héréditaires ; les contrôles négatifs ont quant à eux été échantillonnés dans les bandes cytogénétiques où se trouvent ces variants fonctionnels, sans imposer de distance minimum aux variants fonctionnels. Ce modèle s'inscrit donc dans un contexte d'identifi-

cation de variants candidats parmi un ensemble de variants à l'échelle du génome, ce qui correspond à la tâche proposée ici.

10.2.1 Analyse des variants par famille

Une première analyse que j'ai souhaité faire a consisté en la réplication d'un résultat évoqué dans l'article de Zhou et al., 2019. Dans cet article, les auteurs présentent les résultats de l'application d'une méthode d'identification de variants fonctionnels non-codants, basé sur des réseaux de neurones profonds. L'un des premiers résultats qu'ils présentent correspond à une comparaison des scores de fonctionnalité moyens des variants pour chaque famille, entre le patient malade et le patient contrôle. Le score moyen des variants de chaque patient est calculé, et normalisé par la distribution de scores depuis l'ensemble des patients contrôles. Ce Z-score ainsi calculé est pairé à celui calculé chez le patient contrôle de la même famille ; les paires de Z-scores sont donc utilisées pour tester statistiquement la différence de scores entre les patients malades et les patients contrôles. Les auteurs identifient ainsi une différence statistiquement significative indiquant des scores plus élevés pour les patients malades ; cette différence est particulièrement importante pour les variants associés aux gènes intolérants aux mutations (score pLI, Lek et al., 2016).

J'ai donc appliqué une procédure similaire de calcul des Z-scores pour les variants de chaque patient, comparés entre les patients malades et les patients contrôles. La figure 10.1 présente les résultats de cette approche. Les valeurs représentées correspondent à la moyenne des Z-scores moyens calculés chez les patients d'un groupe (malade ou contrôle). Les catégories d'associations aux gènes sont définies ici grâce aux annotations des prédictions d'associations utilisées dans le projet (voir méthodes), ainsi qu'aux annotations GENCODE : un variant est identifié comme associé à un variant d'une catégorie donnée si ce variant est dans une région régulatrice avec une association prédite avec l'un des gènes de cette catégorie, ou bien s'il se trouve dans un gène de cette catégorie. Les différentes catégories de gènes d'intérêt regroupent des gènes associés à des propriétés particulières : les gènes associés à l'autisme (ASD genes) provienent de la fusion des bases de données SFARI et OpenTargets ; les gènes de la base de données OMIM (OMIM genes) correspondent à des gènes associés à des maladies en général ; les gènes sensibles au dosage sont des gènes identifiés comme soumis à une sélection négative concernant leur duplication, et donc associés à un niveau d'expression contraint ; et les gènes pLI et RVIS correspondent à des gènes dont la séquence est sous contrainte dans la population humaine (voir méthodes).

On peut observer que les variants des patients malades ont en moyenne un score FINSURF légèrement plus élevé pour différentes catégories de gènes d'intérêt. Cette différence est observable par exemple pour les gènes associés à l'autisme (ASD genes), aux gènes de maladies en général (OMIM genes), ou aux gènes intolérants aux mutations (pLI constrained genes). Cependant aucune de ces différences n'est significative, contrairement à ce que les auteurs de l'article Zhou et al., 2019 ont réussi à détecter. Il est possible que la méthode de prédiction proposée par les auteurs soit plus efficace que FINSURF à détecter des variants fonctionnels avec un score élevé. Il est possible également que les jeux de variants soient très différents entre mon analyse et la leur (ils identifient 127 140 mutations *de novo* chez 1 790 familles ; je n'ai pas pu comparer leur sélection de variants à la mienne). Enfin d'un point de vue biologique, l'observation d'une différence qui n'est pas statistiquement significative ne paraît pas aberrant : en effet, le cadre de cette analyse des mutations *de novo* implique que parmi la soixantaine de mutations identifiées chez un patient malade, seules quelques unes (voire une seule) sont à l'origine de la maladie. Ainsi, on ne s'attend pas à ce que la poignée de variants effectivement fonctionnels conduisent à un déplacement significatif de la moyenne des scores des variants pour un patient donné. La tendance observée dans mes résultats est donc intéressante, mais reste non-significative, ne permettant pas de conclure clairement ; ce n'est cependant pas inattendu en considérant le modèle biologique, et peut-être qu'un test statistique sur les valeurs extrêmes des distributions de scores serait plus adapté à ces évaluations de différences.

10.2.2 Analyse des variants par catégorie d'association aux gènes d'intérêts.

Je propose ici une seconde analyse, consistant à identifier des catégories de mutations *de novo* présentant des scores particulièrement élevés chez les patients malades, en comparaison avec les mutations identifiées chez les patients contrôles ; cette analyse mêle l'évaluation d'enrichissement en mutation *de novo* dans certaines catégories d'annotations génomiques (Werling et al., 2018), et une comparaison des scores de prédictions qui leur sont associés (Zhou et al., 2019). Dans cette approche, les mutations ne sont plus comparées par famille ; on souhaite identifier des propriétés générales distinguant les mutations *de novo* des patients malades de mutations *de novo* détectées chez des patients contrôles.

J'ai donc défini plusieurs catégories de mutations, sur la base de leurs associations aux gènes des différents groupes définis précédemment ; ces catégories sont non-exclusives (un

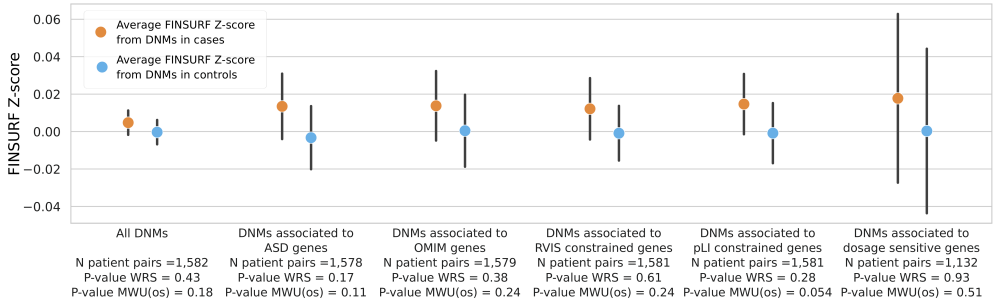


FIGURE 10.1 – Comparaison du Z-score FINSURF moyen des mutations *de novo* chez chaque patient, entre patients malades et patients contrôles. Le Z-score par patient est calculé comme la moyenne des scores FINSURF des variants de ce patient, normalisé par la moyenne des scores des variants de tous les patients contrôles, pour une catégorie considérée. Par catégorie, le score moyen sur l'ensemble des patients malades, ou patients contrôles, est rapporté ; l'intervalle de confiance de cette valeur est identifié par la barre d'erreur. Les variants sont assignés aux catégories par rapport à leur association aux gènes de cette catégorie. L'association d'un variant à un gène est définie à partir de sa localisation dans le gène, ou par sa localisation dans une région régulatrice avec prédiction d'association au gène. Le nombre de paires de patients considérées peut varier, par l'exclusion des paires où l'un des patients ne présente pas de variants associés à la catégorie de gènes considérée. Les tests statistiques appliqués sont le test des rangs signés de Wilcoxon, pour évaluer les différences statistiques entre paires de patients (WRS). Un test de Wilcoxon-Mann-Whitney non-pairé (MWU(os)) est également appliqué. Les P-valeurs ne sont pas corrigées pour les tests multiples.

variant peut être associé à des gènes de différents jeux, et donc être retrouvé dans plusieurs catégories). Pour chacune de ces catégories, le score FINSURF est utilisé pour identifier le nombre de mutations fonctionnelles chez les patients malades et chez les contrôles. Ces mutations prédites comme fonctionnelles sont distinguées des non-fonctionnelles grâce au seuil de 0.55, précédemment identifié pour maximiser le nombre de vrais positifs et minimiser le nombre de faux positifs (voir chapitre 9). En résumé, je souhaite identifier des catégories d'associations entre mutations *de novo* et gènes d'intérêt qui concentrent des mutations hautement fonctionnelles, et ce de manière spécifique aux patients malade.

La figure 10.2 correspond aux résultats de cette analyse. On peut tout d'abord constater qu'aucune des catégories ne présente de différence statistiquement significative si l'on corrige pour les tests multiples. On peut cependant remarquer quelques catégories de variants qui semblent présenter des nombres de variants fonctionnels particulièrement plus élevés pour les patients malades que pour les patients contrôles : la catégorie la plus notable correspond aux variants localisés dans des régions régulatrices associées aux gènes

sensibles au dosage. Dans cette catégorie, la proportion de variants fonctionnels est 1,65 fois supérieure pour les variants des patients malades, comparée aux variants des patients contrôles (sur 207 DNMs chez les patients malades, 15.8% sont identifiées comme fonctionnelles, contre 9.6% des 230 DNMs des patients contrôles). Les deux autres catégories qui ont des P-valeurs non-corrigées significatives correspondent aux variants à proximité de gènes de l'autisme (provenant de la base de données SFARI, avec et sans fusion de la base de données OpenTargets). Cependant les différences de proportions mesurées sont extrêmement faibles. Deux catégories présentent des proportions de variants fonctionnels légèrement plus élevées pour les patients malades : les catégories de variants associés par régions régulatrices géniques aux gènes de l'autisme (ASD genes - self-targeting genic regulatory region) ; les P-valeurs ne sont cependant pas significatives, même sans correction. Dans cette figure, j'ai annoté deux catégories qui sont associées à une proportion de variants fonctionnels plus importante chez les patients contrôles que chez les variants malades, pour des DNMs localisés dans des régions régulatrices associés à des gènes de l'autisme ; cette différence est très faible, mais indique que notre modèle introduit potentiellement des variants faux positifs, ou bien que ces variants ont effectivement un caractère fonctionnel, mais qui n'est pas associé au phénotype malade.

Dans l'ensemble, il est difficile de conclure sur un enrichissement clairement notable et significatif dans aucune des catégories définies ici. Des tendances semblent se dégager en faveur d'une sur-représentation chez les patients malades de variants fonctionnels dans des régions régulatrices associées à des gènes d'intérêt. En particulier l'enrichissement de variants fonctionnels dans des régions régulatrices associées à des gènes sensibles au dosage est très intéressant : cette catégorie correspond à des gènes dont le maintien d'un niveau expression stable est extrêmement important ; l'impact de variants régulateurs sur ces gènes est donc potentiellement associé à une conséquence phénotypique importante. Toutefois ces tendances sont à contre-balancer par une sur-représentation de variants fonctionnels chez les patients contrôles pour des catégories également intéressantes par rapport au phénotype. Dans la section suivante, je propose une approche de sélection de variants candidats pour chaque patient.

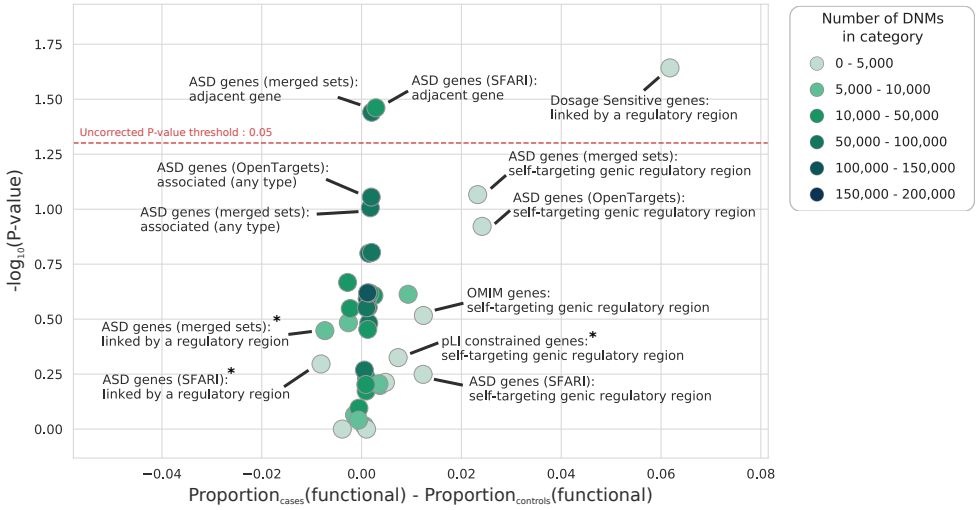


FIGURE 10.2 – Comparaison du nombre de mutations prédites comme fonctionnels entre patients et malades, pour différentes catégories d'associations régulatrices à des gènes d'intérêt. Les catégories de variants sont définies selon leur association aux gènes de différents jeux d'intérêt : gènes de maladie (OMIM), gènes associés à l'autisme (SFARI, OpenTargets, ou les deux combinées), gènes sensibles au dosage. Les associations sont définies selon la localisation des variants par rapport aux gènes considérés : associés par n'importe quelle localisation, associés par gène adjacent, associés par une région régulatrice, ou associés par une région régulatrice dans le même gène. Pour chaque catégorie, les scores FINSURF des variants des patients malades et contrôles sont utilisés pour calculer le nombre de variants fonctionnels, selon le seuil de 0.55. La différence de proportions de variants fonctionnels entre les patients malades et les contrôles est rapportée en abscisses. L'utilisation d'un test de Fisher permet d'évaluer statistiquement cette différence ; la P-valeur de ce test est rapportée en ordonnée. Le seuil de significativité à 0.05 sans correction est identifié sur le graphique ; avec un nombre de tests de 35, la valeur en ordonnée correspondante après correction de Bonferroni serait de 2.85, et n'est donc pas rapportée par soucis de lisibilité. Par soucis de lisibilité également, toutes les catégories ne sont pas identifiées : seules celles présentant une différence de proportions supérieure à 1%, ou avec une P-valeur inférieure à 0.1 sont identifiées ; les catégories marquées d'une astérisque sont cependant identifiées pour l'interprétation.

10.3 Identification de mutations candidates chez les patients malades

10.3.1 Approche et nombre de patients potentiellement expliqués

L'objectif de la méthode FINSURF est de pouvoir aller identifier des variants prédits comme hautement fonctionnels, d'avoir un profil de leurs annotations et des contributions de chacune au score de prédiction, et de pouvoir sélectionner parmi les variants ceux associés à des gènes d'intérêt pour une maladie considérée. J'ai appliqué cette approche diagnostique pour chacun des 1 582 patients malades de la cohorte. Pour chacun, le meilleur variant est sélectionné selon le score FINSURF, provenant du modèle HGMD-DM Cytoband-match. J'avais déterminé pour ce modèle un seuil de 0.55 à appliquer sur le score, pour identifier un maximum de vrais positifs en introduisant un minimum de faux positifs ; ce seuil est donc appliqué aux variants sélectionnés. Seuls 1 206 patients parmi les 1 582 présentent au moins une mutation *de novo* avec un score dépassant ce seuil. Il est possible que pour les 376 patients exclus, plusieurs hypothèses sont possibles :

— la ou les mutations *de novo* contribuant au phénotype ont été perdues lors des différents filtres ;
— elles n'ont pas été détectées lors de l'appel de variants ;
— il est possible également que le score FINSURF assigné à la vraie mutation causale soit trop faible, puisque l'on a vu que le modèle n'était pas capable de détecter comme vrais positifs certains profils de variants (voir chapitre 9) ;
— le patient a hérité d'une mutation causale, à pénétrance variable, qui n'a pas conduit à un syndrome détectable chez les parents.

Parmi les 1 206 variants détectés comme fonctionnels, j'ai identifié ceux associés à des gènes d'intérêt. Ces gènes d'intérêts proviennent de trois listes évoquées dans la section précédente : les gènes associés à l'autisme provenant de la base de données SFARI, les gènes associés à l'autisme provenant de la base de données OpenTargets, et les gènes identifiés comme sensibles au dosage ; le jeu total de gènes d'intérêt est ainsi composé de 2 886 gènes.

Pour comparer l'apport des prédictions d'associations aux gènes cibles, j'ai défini deux méthodes d'association des variants candidats aux gènes d'intérêt :

— la première approche consiste à identifier les variants localisés dans la séquence (non-codante) des gènes d'intérêt, ou bien dans des régions régulatrices prédites (provenant de PEGASUS, FANTOM, FOCS, ou Genehancer), et pour lesquelles

un des gènes flanquants (en amont ou en aval) est un gène d'intérêt. Dans cette approche, les prédictions d'associations ne sont pas considérées, et l'association de la région régulatrice à un gène cible est faite aux gènes flanquants ;

— la seconde approche consiste à identifier les variants localisés dans la séquence des gènes d'intérêt, ou bien dans les régions régulatrices en considérant cette fois l'association prédite pour les gènes d'intérêt.

Avec la première approche, seuls 360 des patients sont identifiés comme ayant une mutation candidate à score élevé et associée à un gène d'intérêt, soit 30% des patients explicables. En revanche avec la seconde approche, ce sont 622 patients qui sont identifiés comme ayant une mutation candidate avec un score élevé et associée à un gène d'intérêt, soit 51.6% des patients explicables. A noter que seuls 147 patients sont partagés entre les deux approches. Cela signifie que la considération des prédictions d'associations entre régions régulatrices et gènes cibles permettent d'identifier une mutation candidate pour 475 patients supplémentaires, tandis que l'approche d'association par défaut d'un variant dans une région régulatrice à un gène flanquant permettait de proposer une mutation candidate pour 213 autres patients, qu'on ne retrouve pas par les associations prédites (provenant des prédictions PEGASUS, FANTOM, FOCS, ou GeneHancer). Il est possible que toutes les associations biologiques ne soient pas exhaustivement identifiées par ces méthodes, et donc que les 213 mutations représentent des variants candidats tout de même pertinents ; je ne les considérerai cependant pas pour la suite.

Il est également intéressant de noter que pour ces 622 mutations candidates, 68.6% correspondent à la mutation à plus haut score parmi toutes les mutations à haut scores de chaque patient malade, tandis que 31.4% des mutations affectant une région associée à un gène d'intérêt ne possèdent pas le score le plus élevé dans leurs génomes respectifs.

Pour évaluer si cette proportion est due à un enrichissement particulier des mutations *de novo* de ces 622 patients dans des régions régulatrices associées à des gènes d'intérêt, j'ai effectué 1 000 tirages aléatoires d'une mutation candidate par patient ; ce tirage est fait soit parmi l'ensemble des mutations de chaque patient soit parmi les mutations localisées dans des gènes ou des éléments régulateurs (cette localisation pouvant contribuer à un score FINSURF plus élevé). La comparaison, présentée sur la figure 10.3 B, confirme que les meilleures mutations identifiées chez les 622 patients ne sont pas associées à des gènes d'intérêt par hasard (17% si l'on considère toutes les mutations, et 27% en considérant le sous-jeu de mutations dans des gènes ou régions régulatrices) : ces patients présentent

effectivement des variants candidats à fort potentiel fonctionnel spécifiquement associés à des gènes d'intérêt.

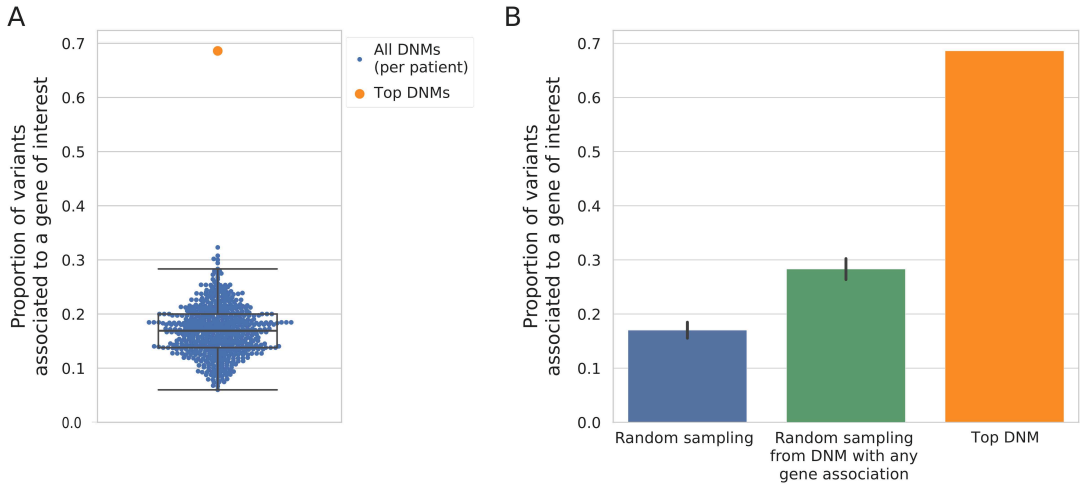


FIGURE 10.3 – Étude des proportions de variants associés à des gènes d'intérêt pour 622 patients potentiellement expliqués.
En A, la proportion de mutations *de novo* associées à des gènes d'intérêts (gènes impliqués dans l'autisme, ou gènes sensibles au dosage) est rapportée pour chaque patient, en prenant l'ensemble de ses mutations *de novo*. La proportion calculée pour les 622 mutations candidates est également présentée.
En B, cette proportion est comparée à la proportion attendue par hasard, calculée à partir de 1 000 tirages aléatoires d'une mutation candidate par patient, pour les 622 patients. Le tirage aléatoire peut être fait sur l'ensemble des mutations *de novo* d'un patient (barre bleu) ou bien uniquement parmi les mutations *de novo* pour lesquelles une association à un gène (d'intérêt ou non) peut être détectée par leur localisation dans un gène ou dans un élément régulateur (barre verte). La barre d'erreur représente l'écart-type.

La question de la récurrence d'association à certains gènes se pose alors. En effet parmi les 622 mutations identifiées, il est possible qu'un des gènes d'intérêt (gènes liés à l'autisme, ou gènes sensibles au dosage) soit associé plusieurs fois à des mutations candidates chez différents patients ; cela permettrait d'identifier des gènes d'autant plus pertinents à étudier. Je rapporte dans le tableau 10.1 les gènes identifiés avec plusieurs associations indépendantes parmi les 622 candidats. On peut voir par exemple que NPAS3 et RBFOX1 sont chacun identifié comme le gène potentiellement impacté pour 6 variants candidats différents. Pour illustration, les 6 variants candidats associés au gène RBFOX1 sont présentés dans le tableau 10.2. A noter qu'en tirant aléatoirement 622 gènes au hasard parmi les 2 886 gènes d'intérêt identifiés, la probabilité d'obtenir 6 fois le même est de 0.0001. Une évaluation statistique plus rigoureuse intégrant la taille des gènes et des régions ré-

gulatrices qui leur sont associées serait cependant nécessaire pour confirmer clairement le caractère exceptionnel de ces observations ; par exemple dans le cas de RBFOX1, le transcrit le plus long identifié dépasse 2 millions de paires de bases, ce qui augmente les chances d'identifier des variants fonctionnels localisés dans sa séquence par rapport à un gène plus court.

Gene name	**Number of associations**	**in ASD genes (OpenTargets)**	**in ASD genes (SFARI)**	**in Dosage Sensitive genes**
NPAS3	6	True	False	False
RBFOX1	6	True	True	True
WWOX	5	False	True	False
ALG6	4	False	True	False
CADM1	4	True	True	False
DIAPH3	4	True	True	False
EBF3	4	True	True	False
LPP	4	True	False	True
MKI67	4	True	False	False
NBEA	4	True	True	False
SOBP	4	True	False	False
VAMP1	4	True	False	False
ZNF521	4	True	False	False

TABLE 10.1 – Tableau des gènes codants d'intérêt associés de manière récurrente aux 622 variants candidats. Seuls les gènes présentant un nombre d'association récurrente supérieur à 3 sont présentés.

chrom	**start**	**end**	**ref**	**alt**	**SampleId**	**Genomic location**	**Gene association**	**FINSURF score**	**Best DNM overall in patient**
chr16	5735463	5735464	T	C	13543.p1	intron	in_gene	0.6996	Yes
chr16	6533628	6533629	G	C	14344.p1	intron	in_gene	0.5786	Yes
chr16	7132804	7132805	GAGGCCACCAG TAGCTTAATCAGT GTTGGTTTC	G	12620.p1	intron	in_gene	0.6079	Yes
chr16	7492374	7492375	G	A	12285.p1	intron	in_gene	0.6674	No
chr16	7763242	7763243	C	T	13780.p1	3UTR	in_gene	0.689	Yes
chr16	8184548	8184549	G	A	14671.p1	intergenic	in_enhancer	0.61	No

TABLE 10.2 – Tableau des variants candidats associés au gène RBFOX1.

Ainsi, par l'utilisation de la méthode FINSURF de hiérarchisation des variants, et en exploitant les associations prédites entre régions régulatrices et gènes cibles, il est possible de trouver une mutation *de novo* candidate pour 51.6% des patients, parmi lesquelles

68.6% sont la meilleure mutation identifiable parmi celles présentes chez chacun des patients. Dans la section suivante, je propose des étapes supplémentaires de sélection de variants candidats, et j'illustre l'utilisation des valeurs de contributions obtenues depuis le modèle FINSURF pour comprendre plus en détails les profils fonctionnels des mutations candidates.

10.3.2 Profils fonctionnels de mutations candidates

Afin d'identifier des variants candidats les plus pertinents possibles, j'ai souhaité évaluer les modifications introduites par les 622 mutations *de novo* candidates sur des sites potentiels de fixation de facteurs de transcription. L'outil variation-scan, disponible dans la suite d'outils RSAT (Nguyen et al., 2018) a été utilisé pour prédire des TFBS potentiels chevauchant ces mutations ; la base de données JASPAR (Khan et al., 2018) a servi de source pour les motifs de facteurs de transcription utilisés.

Différents modèles de fréquences en di-nucléotides et tri-nucléotides ont été utilisés dans le cadre de cette identification de TFBS ; ces fréquences (que j'ai calculées depuis l'ensemble des régions régulatrices du génome, en incluant ou non les régions introniques) sont utilisées par l'outil variation-scan pour calculer des probabilités de découverte par hasard d'un TFBS dans une séquence génomique donnée. Pour garantir la robustesse des TFBS identifiés, je n'ai gardé que les prédictions de sites identifiés systématiquement avec les différents modèles de fréquences nucléotidiques. Par ailleurs, le modèle de variation-scan permet d'évaluer la modification d'une mutation sur un TFBS prédit en comparant les scores d'identité entre le motif et la séquence obtenue pour la version de référence et la version mutée de cette séquence. Seules les mutations entraînant une diminution notable des scores d'identité sont gardées (sur conseils des développeurs, pour minimiser le nombre de faux positifs identifiés, un seuil minimum de $1e-4$ est appliqué sur le TFBS découvert, et un seuil minimum de 100 est appliqué pour le ratio de la P-valeur du site découvert pour la séquence référence sur la P-valeur du site découvert pour la séquence mutée).

J'ai ainsi identifié 58 mutations candidates qui sont prédites par l'outil variation-scan pour diminuer ou détruire un total de 141 TFBS putatifs, correspondant à 115 facteurs de transcriptions différents. La moyenne du nombre de sites affectés par mutation est de 2.43.

J'ai par ailleurs tenté d'appuyer ces prédictions de modifications par des preuves supplémentaires de pertinence :

— en utilisant la base de données TRRUSST (« TRRUST v2 » p.d.) : pour un variant associé à une modification de TFBS, y-a-t'il une relation de régulation déjà connue entre le facteur de transcription et le gène cible d'intérêt associée au variant ; aucun des sites ne présente une telle association cependant.
— en scannant les sites identifiés avec les régions de pics de ChIP-seq provenant de la base de données REMAP (Chèneby et al., 2018). Cette étape permet d'identifier de confirmer si un TFBS prédit est également soutenu par un signal de ChIP-seq correspondant, identifié dans un des types cellulaires étudiés dans le projet ENCODE. Un total de 14 variants candidats, modifiant 18 TFBS, sont ainsi détectés.

Ainsi cet ensemble d'analyses supplémentaires permet d'identifier 14 mutations *de novo* candidates qui sont :

— identifiée comme fonctionnelle par FINSURF ;
— associée à un gène d'intérêt soit en étant dans le corps du gène, soit dans une région régulatrice avec association prédite ;
— et conduisant à une perte d'un ou plusieurs TFBS qui chacun chevauchent un signal de ChIP-seq correspondant.

Ces variants représentent donc des candidats hautement pertinents pour expliquer le phénotype des patients chez qui ils sont trouvés. Pour comprendre plus en détails la prédiction fournie par FINSURF, il est possible d'explorer les valeurs de contributions des annotations au score de prédiction. Je propose ici d'évaluer le variant candidat pour le patient 11187.p1 : chr3 :114866116 :C>G ; ce variant fait partie des 14 candidats mentionnés ci-dessus.

Ce variant est identifié comme la mutation *de novo* la plus fonctionnelle parmi les mutations de ce patient (avec un score FINSURF de 0.884), et est localisé dans l'UTR5' du gène ZBTB20 ; ce gène est un gène associé au spectre de l'autisme dans les bases de données SFARI et OpenTargets. Par l'analyse des modifications sur des TFBS, ce variant est associé à 3 pertes de TFBS : SP1 (avec un score passant de 10.87 à 0.54), SP2 (avec un score passant de 9.70 à 1.03), et KLF5 (avec un score passant de 10.01 à -0.67) ; pour chacun un pic de ChIP-seq chevauche la région. La visualisation d'un profil fonctionnel pour ce variant (figure 10.4) permet d'obtenir plus de détails sur les propriétés de ce variant, et comment ses annotations ont été évaluées par le modèle. On voit ainsi que le variant présente une conservation notable, et que le modèle a pu exploiter cette information, qui a contribué positivement au score ; de même, les signaux de régions fonctionnelles (ilôt

CpG, concentration de régions de chromatine ouverte, état promoteur, valeurs médianes de détection de marques H3K27ac et H3K4me3) présentent des valeurs plus élevées pour ce variant par rapport au reste des mutations *de novo*, et ont contribué positivement au score FINSURF. Ce variant est également associé à une localisation dans des sites de facteurs de transcription identifiés par les différentes ressources (Ensembl, TFBS conservés, TFBS groupés) ; ces valeurs contribuent positivement aussi au score FINSURF, sauf pour l'annotation tfbsEnsembl.scoremax, qui ne semble pas avoir été utilisée par le modèle malgré une valeur extrêmement élevée.

Cette caractérisation détaillée des annotations du variant permet de confirmer que le variant considéré présente bien des caractéristiques d'un variant fonctionnel, et que le score attribué par le FINSURF est bien pertinent ; ce profil fonctionnel rappelle d'ailleurs les profils que nous avions déjà pu observés lors de l'analyses des variants du jeu d'entraînement (voir chapitre 9).

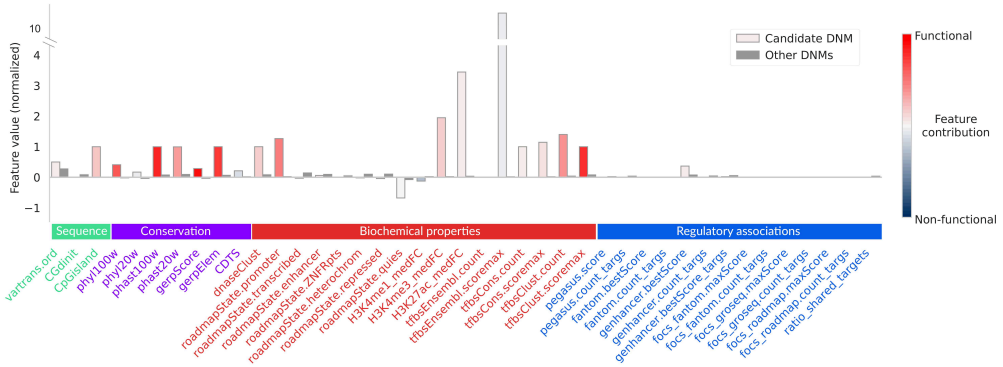


FIGURE 10.4 – Profil fonctionnel du variant candidat identifié chez le patient 11187.p1. Ce profil combine les valeurs mesurées pour les différentes annotations associées au variant (normalisées sur l'ensemble du jeu de DNMs, et représentées sous forme de barres), et les valeurs de contributions des annotations, utilisées pour colorer les barres.

Par ailleurs, la visualisation de ce variant grâce au navigateur de génomes de l'UCSC (figure 10.5) On peut notamment constater que la région dans laquelle est localisée ce variant est bien associée à une ouverture de la chromatine, et à une concentration de marques de chromatine associées à une activité régulatrice ; on peut également voir que cette région est conservée, et que différentes ressources y identifient une concentration de TFBS. Cette figure permet de confirmer visuellement la localisation de ce variant dans une région présentant plusieurs caractéristiques de région régulatrice, et confirme la localisation du variant vis-à-vis du gène ZBTB20.

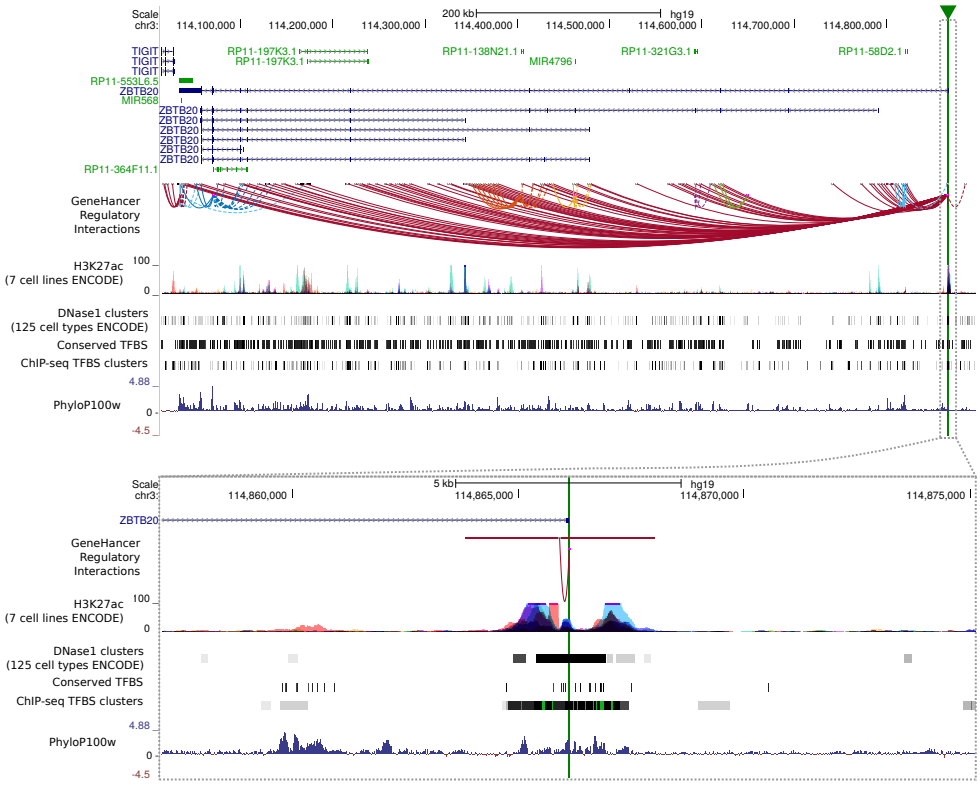


FIGURE 10.5 – Visualisation du contexte génomique du variant candidat pour le patient 11187.p1. Le navigateur de génomes de l'UCSC permet de présenter une vue d'ensemble du variant, dans son contexte génomique. Le panneau du haut présente une visualisation avec la totalité du gène ZBTB20 ; le panneau du bas représente une région plus restreinte, centrée sur le variant. Cette visualisation propose différentes pistes représentant certains signaux de fonctionnalités ; par exemple la piste DNase1 clusters, qui est une des annotations utilisées dans le modèle FINSURF, ou encore la piste GeneHancer, dont les prédictions d'interactions sont visualisées.

Quatrième partie

Discussions et conclusions

Chapitre 11

Discussion

11.1 Résumés des travaux

Mon travail de thèse a eu pour objectif de définir une approche d'évaluation et de hiérarchisation de variants fonctionnels non-codants, pour pouvoir identifier rapidement parmi plusieurs millions de mutations obtenues chez un patient lesquelles sont potentiellement fonctionnelles, et associées à la régulation de l'expression d'un ou plusieurs gènes d'intérêt.

En effet, le séquençage de génome complet est de plus en plus utilisé pour identifier les causes génétiques de maladies rares, par l'étude de l'ensemble des variants nucléotidiques présents dans le génome d'un patient (entre 4 et 5 millions au total). La majorité des variants identifiés chez un patient se trouvant hors des gènes codants, l'évaluation et l'analyse de variants nucléotidiques localisés dans le génome non-codant représentent un enjeu important. Notre capacité à comprendre l'impact des variants non-codants est d'autant plus importante que pour des patients atteints de maladies rares, l'identification de variants codants pertinents n'est possible que pour une partie de ces patients (de 8% à 70%, Wright et al., 2018). De plus, la majorité des variants identifiés comme facteurs de risques dans le cadre de GWAS, sont localisés dans les régions non-codantes du génome. Il est donc nécessaire d'avoir une méthode automatisée de hiérarchisation, afin d'identifier des variants non-codants à fort potentiel fonctionnel, permettant de sélectionner des variants candidats pour des validations expérimentales plus poussées.

C'est cette étape de hiérarchisation à laquelle j'ai souhaité répondre dans le cadre de mon travail de thèse, ce qui a abouti au développement d'une méthode appelée FINSURF

(Functional Interpretation of Non-coding Sequences Using Random Forests), qui se distingue par ses performances et ses caractéristiques. La mise en place de cette méthode a consisté en le développement d'un ensemble d'outils dédiés à l'annotation et l'entraînement de modèles d'apprentissage machine basés sur la distinction entre des variants contrôles positifs et des variants contrôles négatifs, sélectionnés exclusivement dans le génome non-codant. Plusieurs modèles ont été entraînés, par l'utilisation de variants contrôles positifs différents : des variants fonctionnellement associés à des maladies (HGMD-DM), et des variants associés à des changements d'expressions de gènes de maladies (eQTLs-OMIM). Différents échantillons de variants contrôles négatifs (provenant de la base de données ClinVar), ont été établis, pour explorer différentes hypothèses biologiques et étudier les conséquences de ces échantillonnages sur les propriétés des modèles. Les étapes d'entraînement des modèles ont été l'occasion d'évaluer l'importance de la méthodologie appliquée pour valider les modèles, ainsi que l'importance de sélectionner correctement les annotations qui seront utilisées par ces modèles ; ces analyses ont été développées au chapitre 8. J'y ai notamment confirmé que les modèles FINSURF, évalués dans le cadre d'une validation croisée garantissant l'indépendance des phases d'entraînement et de validation, présentent les meilleures performances par rapport à un panel de méthodes identifiées dans la littérature.

La question de l'utilisation de ces annotations a été l'objet d'un ensemble d'analyses (présentées au chapitre 9), dans le cadre desquelles je souhaitais plus clairement comprendre les propriétés biologiques exploitées par mes modèles. Tout d'abord, les mesures globales d'importance des annotations au sein des modèles ont permis un premier regard général sur les différences existant entre mes modèles, en comparant la contribution globale de chaque annotation à la qualité de discrimination des classes pour chaque modèle. Ensuite, la mesure pour chaque variant des contributions des annotations au score de prédiction a permis d'identifier plus en détails des motifs dans l'utilisation de ces annotations par mes modèles. En particulier, l'étude proposée pour le modèle HGMD-DM Cytobandmatch a été l'occasion de voir que le modèle de forêt aléatoire est capable d'identifier différentes sous-populations parmi les variants fonctionnels HGMD-DM ; ces sous-populations sont associées à des propriétés biologiques différentes pour ces variants. Ces profils de contributions permettent ainsi d'inférer des fonctions biologiques plus précises que le seul score de prédiction donné par le modèle pour un variant.

Enfin, dans le chapitre 10, j'ai illustré l'utilité de la méthode FINSURF pour l'identifi-

cation de variants candidats, par l'application du modèle HGMD-DM Cytoband-match à des mutations *de novo* obtenues chez des patients atteints de troubles du spectre autistique. Deux analyses ont été proposées pour étudier ces variants en les groupant dans différentes catégories génomiques d'associations à des gènes d'intérêt (notamment des gènes d'autismes, et des gènes sensibles au dosage). Si des tendances se dégagent en faveur d'une fonctionnalité plus importante des mutations *de novo* provenant des patients malades, aucune des catégories ne présente une différence significative par rapport aux variants identifiés chez les patients contrôles. L'utilisation des scores a cependant permis d'identifier pour une partie des patients malades une mutation à fort potentiel fonctionnel ; ma méthode permet ainsi de proposer une mutation candidate associée à un gène d'intérêt pour 622 patients (soit 51.6% de l'ensemble de patients présentant au moins une mutation fonctionnelle). J'ai proposé une analyse plus poussée de certains de ces variants candidats, associés à des gènes pertinents, et affectant des sites de fixation de facteurs de transcription. Enfin, un de ces variants a été illustré par son profil fonctionnel, établi à partir de ses annotations, et des valeurs de contributions de ces annotations, permettant de comprendre comment le modèle de prédiction a identifié ce variant candidat comme fonctionnel. Cette représentation correspond à une caractéristique importante de la méthode FINSURF, qui contribue à une hiérarchisation intelligible des variants candidats.

La méthode FINSURF permet donc de hiérarchiser des variants non-codants potentiellement fonctionnels, et de comprendre quelles annotations ont contribué au score de prédiction de ces variants. L'utilisation de données de prédictions d'associations entre régions régulatrices et gènes cibles constitue un avantage majeur dans l'identification de variants affectant potentiellement l'expression de gènes d'intérêt.

11.2 Apports de FINSURF à l'état de l'art

11.2.1 Intérêts pratiques et théoriques

Par le développement de la méthode FINSURF, j'ai voulu proposer des modèles de prédiction dédiés aux variants non-codants, en explorant trois points :

— exploiter les prédictions d'associations entre régions régulatrices et gènes-cibles ;
— évaluer les conséquences associées aux méthodes de validation croisée des modèles, étant donné la nature séquentielle du génome ;
— proposer une méthode de calcul qui permette de comprendre plus précisément com-

ment sont établies les décisions fournies par le modèle.

Prédictions d'associations régulatrices. Les prédictions d'associations entre régions régulatrices et gènes-cibles représentent un élément crucial dans l'interprétation d'un variant non-codant. En effet, les variants non-codants régulateurs doivent être associés à un gène pour inférer une conséquence phénotypique. En l'absence de ces prédictions d'associations, il est toujours possible d'assigner l'impact d'un variant régulateur à l'un de ses gènes flanquants ; mais comme on l'a vu au chapitre 10, cela réduit notre capacité à identifier des variants candidats (51.6% des patients sont potentiellement expliqués si l'on considère les prédictions d'interactions, contre 30% sans). Ainsi, l'utilisation des prédictions d'associations est importante pour inférer les conséquences biologiques d'un variant non-codant.

Étapes d'entraînement des modèles. Dans le cadre de l'entraînement de mes modèles, j'ai tenté d'utiliser ces prédictions d'associations comme une annotation pour décrire mes variants contrôles positifs et négatifs. Cependant les résultats obtenus en considérant ces annotations pour l'entraînement ont montré que les modèles associés se contentaient d'utiliser cette annotation, et n'apprenait pas à distinguer les contrôles positifs des négatifs sur d'autres annotations. Or, par l'entraînement sur de multiples sources d'informations, je souhaitais définir des modèles capables d'exploiter des combinaisons complexes d'annotations, plutôt que de proposer une identification triviale basée sur la localisation des variants vis-à-vis de leurs gènes. Cela m'a conduit à écarter cette information particulière ; en revanche, mes modèles de prédictions utilisent tout de même un grand nombre d'annotations correspondant à des prédictions de régions régulatrices (14 parmi 41 annotations au total).

Les étapes d'entraînement des modèles ont également été l'occasion de souligner l'importance de l'approche de validation croisée, étant donnée la nature des variants. En effet les variants ne sont pas indépendants les uns des autres : ils correspondent à des changements à des positions nucléotidiques qui sont plus ou moins proches. Il est donc nécessaire de prendre en compte cette proximité potentielle, surtout lorsque l'on cherche à évaluer la capacité de généralisation de son modèle : si les variants sont pris totalement au hasard, il y a un risque pour que deux variants très proches soient chacun pris pour l'entraînement et pour la validation, facilitant ainsi la tâche du modèle de prédiction, et conduisant à une sur-évaluation de la qualité de ce dernier. J'ai proposé une illustration de cette

sur-évaluation, par la comparaison d'une validation croisée avec séparation aléatoire des variants, et d'une validation croisée basée sur une séparation des bandes cytogénétiques. La première méthode conduisait systématiquement à des sur-évaluations de la qualité des modèles, en comparaison avec la seconde. Il est à noter que cette séparation selon les bandes cytogénétiques est arbitraire : au delà des propriétés biologiques de ces bandes, le plus important est de veiller à ce que les variants utilisés pour l'évaluation soient à une distance importante des variants utilisés pour l'entraînement ; une séparation par chromosome aurait probablement produit des résultats comparables.

Capacité d'interprétation des modèles. Il est intéressant d'avoir un modèle de prédiction capable de fournir pour un variant un score de fonctionnalité, dérivé de l'intégration de données multiples et hétérogènes. Cependant cela ne permet pas de comprendre clairement quelles propriétés biologiques associées à un variant sont effectivement considérées par ce modèle, ni comment ces annotations sont utilisées. La capacité à comprendre les décisions des méthodes d'apprentissage machine représente par ailleurs un impératif pour la confiance que l'on peut accorder à un modèle (Ribeiro et al., 2016, Char et al., 2018), ainsi qu'un impératif légal (Holm, 2019, *Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données)* 2016).

J'ai donc proposé parmi les modules de FINSURF un outil de calcul qui quantifie les contribution des annotations au score prédit pour un variant, et ce de manière spécifique à chaque variant. L'application de cette méthode de calcul sur le jeu d'entraînement des modèles de prédiction a permis de comprendre plus en détails les différentes combinaisons d'annotations que le modèle a été capable d'identifier lors de l'apprentissage. Elle a également permis d'identifier les limites des capacités du modèle à exploiter certaines annotations. Dans le cadre de la démonstration de l'utilisation de FINSURF pour identifier des mutations *de novo* candidates, ces valeurs de contributions ont permis à nouveau d'interpréter comment les annotations des variants sélectionnées ont servis au modèle pour leur assigner un score élevé.

Perspectives

Concernant la mise à disposition à la communauté de la méthode FINSURF : au delà de la disponibilité du code, je proposerai dans un premier temps un tableau de valeurs pour l'ensemble des 558 158 839 positions génomiques identifiées comme régions régulatrices par les jeux d'annotations sélectionnés dans ce projet (PEGASUS, FANTOM, GENEHANCER, FOCS). Ce tableau contiendra pour chacune des positions le score de prédiction maximal associé à une variation nucléotidique de cette position, ainsi qu'une valeur de contribution de cette annotation au score, obtenue depuis le modèle. Puisque ces régions régulatrices sont associées à des prédictions d'associations, l'ensemble permettra d'annoter facilement des positions d'intérêts et d'identifier celles associées à des gènes d'intérêt.

Le modèle qui sera utilisé pour cette première étape sera le modèle HGMD-DM Cytoband-match ; dans mes analyses, il est associé aux meilleurs scores de validation parmi tous les modèles explorés, et semble être le plus adapté pour l'identification de variants candidats dans le cadre de l'étude de variants impliqués dans des maladies génétiques rares.

Dans un second temps, des tableaux similaires pourront être générés pour chacun des autres modèles développés dans le cadre de ce projet, et qui permettent de répondre à des problématiques biologiques différentes. Enfin, il sera peut-être intéressant de fournir un serveur web qui permettra d'identifier les meilleures variants candidats, selon chacun des modèles, et qui permettra également d'explorer pour chaque variant leurs profils fonctionnels (voir chapitre 10).

Dans le cadre d'une application plus concrète de la méthode FINSURF : je travaille actuellement en collaboration avec une équipe de recherche de l'Institut Imagine, pour identifier des variants candidats parmi les mutations *de novo* identifiées chez quatre familles, dont les enfants sont atteints d'une maladie appelée Syndrome d'Ondine. Par ailleurs, j'avais initialement développé le projet FINSURF pour une application potentielle au projet NIHR Bioresources for Rare Diseases, dans le cadre duquel 9 802 patients atteints de maladies rares ont été séquencés (Ouwehand, 2019). Ces patients sont rassemblés en différentes cohortes de phénotypes : par exemple une des cohortes regroupe des patients atteints de maladies de la rétine héritées. Cette cohorte a été décrite dans l'article Carss et al., 2017 : parmi les 722 patients analysés, 440 sont expliqués par une mutation codante ; cela signifie que pour les 40% restant, une hiérarchisation des variants non-codants par les modèles de prédiction FINSURF pourrait aider à identifier des variants candidats

expliquant les phénotypes.

11.2.2 Limites de la méthode FINSURF

Plusieurs modèles développés mais un seul exploité

J'ai proposé plusieurs modèles, basés sur différents jeux de variants fonctionnels, qui permettaient théoriquement d'explorer un ensemble plus large de variants fonctionnels non-codants. Cependant, les différentes étapes d'évaluation et de caractérisation de mes modèles m'ont finalement poussé à n'exploiter qu'un seul d'entre eux dans le chapitre d'application. Le choix du modèle HGMD-DM Cytoband-match s'est fait sur la base de la qualité des prédictions de ce dernier lors des étapes d'entraînement-validation (chapitre 8), ainsi que sur le contexte biologique de l'étape d'application. En effet le modèle a été entraîné à identifier des variants fonctionnels associés à des maladies héréditaires, parmi un ensemble de variants non-fonctionnels aléatoirement échantillonnés dans le génome. L'analyse des mutations *de novo* correspondait donc à un contexte similaire, ce qui a motivé le choix de ce modèle.

Le modèle HGMD-DM Distance-match quant à lui serait plutôt adapté à une évaluation de la fonctionnalité de positions au sein d'une région donnée ; par exemple, pour évaluer quelles positions dans un élément régulateur sont effectivement potentiellement fonctionnelles. Nous avons cependant vu que le taux de faux négatifs est plus élevé que pour le modèle HGMD-DM Cytoband-match. Une évaluation plus complète de ce modèle serait donc nécessaire pour clairement établir sa pertinence, et son application à la détection de sites fonctionnels dans des régions régulatrices.

Les modèles eQTLs-OMIM ont quant à eux été écartés sur la base de leurs qualités de prédiction, et de biais potentiellement présents dans les jeux de variants utilisés. Ces modèles ne sont pas mauvais : l'évaluation des modèles sur les jeux d'entraînement a montré que le modèle eQTLs-OMIM Cytoband-match est meilleur que les autres modèles, et comparable au modèle FIRE (dédié à l'identification d'eQTLs), tandis que le modèle eQTLs-OMIM Distance-match est meilleur que tous les autres modèles. Cependant l'étude de l'importance des annotations dans la qualité des modèles (voir chapitre 9) a montré que ces modèles exploitaient beaucoup le type de mutation considéré (transition, transversion, ou INDEL), indiquant un potentiel biais dans les compositions de ces variants dans les jeux d'entraînement. Il sera donc nécessaire d'étudier plus en détails des modèles de prédictions basés sur ces variants eQTLs ; notamment, leur nature de SNPs fréquent dans

la population limite potentiellement les capacités des modèles à les exploiter pleinement. Nous avons vu que ces variants sont effectivement associés à des régions présentant des propriétés régulatrices, mais il est probable qu'une minorité de ces SNPs soient effectivement fonctionnels. Par ailleurs, l'intérêt de la sélection d'un sous-jeu d'eQTLs associés à des gènes de maladie est à démontrer : les performances comparables avec la méthode FIRE (entraîné à identifier des eQTLs généraux) semble indiquer que les eQTLs associés aux gènes de maladie OMIM n'ont pas des propriétés particulièrement différentes de celles d'autres eQTLs.

Le choix initial de plusieurs jeux de variants fonctionnels n'est donc pas complètement exploité à l'issue de mon travail de thèse. A une étape de mon projet, j'ai tenté d'établir un méta-modèle de prédiction de variants fonctionnels, par l'entraînement d'un modèle de régression linéaire logistique : les variants étaient résumés aux scores de prédiction obtenus depuis chacun des modèles de forêts aléatoires, et le modèle de régression linéaire logistique proposait des poids optimisés, pour chacun de ces scores. Cependant ce modèle ne permettait pas de classifier correctement les eQTLs, et se focalisait sur l'identification des variants HGMD-DM. Cela peut s'expliquer par les propriétés identifiées lors de l'analyse des valeurs de contributions des annotations (chapitre 9). Nous avons vu que le modèle HGMD-DM Cytoband-match identifiait principalement les variants fonctionnels à partir de leurs scores de conservation élevés, de leur localisation dans des régions promotrices, et dans des sites de fixations de facteurs de transcription. Le modèle eQTLs-OMIM Cytoband-match quant à lui exploite également les annotations de conservation, mais de manière opposée : les variants eQTLs sont identifiés comme moins conservés que les variants contrôles négatifs. Ainsi, il est probablement difficile de réconcilier ces deux modèles, au moins dans le cadre d'une régression linéaire logistique.

Commentaires sur les jeux d'entraînements

La question des jeux d'entraînement est un point important dans la mise en place de modèles d'apprentissage machine. Comme dit précédemment, les variants eQTLs ont été choisis pour leur association mesurable sur l'expression des gènes. Cependant ces variants eQTLs correspondent principalement à des positions polymorphiques dans la population ; la majorité d'entre n'est pas le variant effectivement à l'origine de cette variation d'expression, mais est potentiellement assez proche de celui-ci pour partager des propriétés biologiques associées à la régulation. Ce sont ces propriétés biologiques qui ont été détec-

tées lors des analyses préliminaires (chapitre 7), et exploitées en partie par les modèles de prédiction. Mais comme évoqué dans la section précédentes, il est probable que les modèles soient biaisés par la nature des eQTLs.

Par ailleurs ces eQTLs sont initialement identifiées de manière tissu-spécifiques (The GTEx Consortium et al., 2015). Or certaines des annotations sont initialement disponibles par tissu. Cependant la nature des forêts aléatoires ne permet pas d'exploiter efficacement les données sparses, ce qui m'a poussé à calculer des valeurs médiannes pour les annotations d'états chromatiniens du projet Roadmap Epigenomics par exemple (Kundaje et al., 2015). Il pourrait donc être intéressant de séparer les eQTLs par tissu, afin que le modèle d'apprentissage détecte plus efficacement les annotations de signaux de régions régulatrices spécifiques au tissu considéré.

Concernant les variants HGMD-DM : il est à noter que de nombreuses méthodes utilisent cette base de données pour entraîner leurs modèles de prédiction (Liu et al., 2017, Drubay et al., 2018). Comme je l'ai présenté, cet ensemble de variants est de nombre réduit, et explore une partie restreinte du génome non-codant ; il est donc probable que les modèles FINSURF, ainsi que la plupart des modèles proposés dans la littérature, soient limités dans la capacité à identifier des variants régulateurs dans des régions régulatrices distinctes des régions promoteurs. L'identification expérimentalement validée d'un jeu plus large et divers de variants régulateurs importants serait un atout pour progresser dans ce domaine.

Enfin, j'ai proposé l'utilisation de variants ClinVar identifiés comme non-pathogéniques pour définir les variants contrôles négatifs ; cependant lors de l'assemblage de mes jeux de données d'entraînement, j'ai constaté que les variants ClinVar n'étaient pas complètement disjoints des variants eQTls, ni des variants HGMD-DM (chapitre 7). Etant donnée la nature des ClinVar choisis (non-pathogéniques) et des eQTLs, il n'est pas impossible q'un variant ClinVar soit effectivement identifié comme eQTL dans le cadre du projet GTEX. Concernant les jeux d'entraînement basés sur les variants HGMD-DM, on a vu que plus de 40 variants ClinVar sont détectés comme chevauchant des positions de variants HGMD-DM ; dans ces cas, il est possible que ces variants correspondent à des annotations conflictuelles entre la base de données HGMD et la base de données ClinVar. Ainsi, l'identification de variants non-fonctionnels n'est pas triviale ; dans le cadre d'un apprentissage machine supervisé visant à séparer deux classes, l'introduction de variants ambiguës complexifie la tâche d'apprentissage du modèle. Cependant, la réalité biologique est plus compliquée, et

des degrés de fonctionnalité variables seraient une meilleure façon d'annoter les variants ; ces degrés resteraient cependant à établir, sur une base expérimentale par exemple (par mesure du niveau d'impact d'un variant sur l'expression d'un gène).

Complémentarité des modèles

Toutes les méthodes de prédiction de variants fonctionnels, qu'ils soient codants ou non-codants, cherchent systématiquement à se démarquer les unes des autres en termes de performance. C'est une étape qui permet de justifier la mise à disposition d'une nouvelle méthode à la communauté. J'ai moi-même proposé une comparaison de mes modèles FINSURF à différents modèles identifiés dans la littérature, et ai montré que les modèles FINSURF surpassaient la majorité d'entre eux.

Cependant, il existe très peu de variants fonctionnels non-codants connus. Comme je l'ai souligné dans l'introduction, cela conduit beaucoup des modèles de prédiction à utiliser les mêmes jeux de variants fonctionnels. Par ailleurs la question des variants non-fonctionnels n'est pas triviale non-plus, comme illustré dans la sous-section précédente. Ainsi, la question d'une comparaison rigoureuse et équitables des modèles entre eux se pose. La définition d'un socle commun de variants clairement identifiés comme fonctionnels et non-fonctionnels serait idéal, mais alors ce jeu de variant pourrait également être exploité pour entraîner des modèles supplémentaires.

Plutôt que de chercher à ce qu'une méthode donnée en surpasse une autre, il serait intéressant de clairement caractériser les propriétés identifiées par chacune. Il est en effet probable que chaque méthode contribue à l'identification d'une partie de la fonctionnalité des variants ; les combiner serait alors plus intéressant.

Cette question de la complémentarité n'est pas spécifique aux variants non-codants : la méthode REVEL (Ioannidis, Rothstein et al., 2016) souligne que différentes méthodes dédiées aux variants codants sont plus ou moins corrélées pour différents types de variants fonctionnels codants. Ainsi, certains auteurs (Campa et al., 2017) proposent d'avoir une approche complémentaire et intégrative des méthodes existantes, en combinant ces méthodes pour leur permettre de se compléter.

Cette approche peut effectivement permettre d'augmenter notre capacité à détecter des variants fonctionnels d'intérêt ; la question se pose cependant des taux de faux positifs. Par ailleurs, établir une telle approche nécessiterait de caractériser plus précisément sur quelles propriétés biologiques les différents modèles se complètent. Je souhaite donc souligner

qu'au delà de la comparaison des performances de mes modèles FINSURF, j'ai proposé une approche d'interprétation des modèles, qui permettent de comprendre en détails les prédictions, et d'identifier les propriétés biologiques apprises.

Validations expérimentales et pertinence des modèles de prédiction

La méthode FINSURF a pour objectif d'identifier des variants candidats parmi les variants non-codants identifiables chez un patient. La hiérarchisation par le score de fonctionnalité permet d'identifier les variants les plus fonctionnels, et l'utilisation des annotations de prédictions régulatrices permet de sélectionner ceux associés à des gènes d'intérêts. Les variants ainsi identifiés représentent des candidats pour des validations plus poussées.

Dans le cadre de l'analyse des mutations *de novo* chez les patients atteints d'autisme (voir chapitre 10), j'ai tenté d'aller au delà des prédictions de mon modèle FINSURF, en ajoutant des étapes d'évaluation d'impact des mutations candidates sur des sites de fixation de facteur de transcription, pour sélectionner 14 variants (parmi 622) qui impactent des sites de fixation identifiés dans des signaux de ChIP-seq associés.

A ma connaissance, il est difficile d'aller plus loin dans les validations *in silico*, et des validations expérimentales s'imposent. Comme évoqué dans l'introduction, différentes méthodes sont disponibles pour tester l'effet d'un variant non-codant sur le potentiel régulateur d'un enhancer (Rodenburg, 2018). En particulier, le développement récent des méthodes d'édition de la séquence par CRISPR-Cas9 représente une avancée majeur dans notre capacité à tester des variants candidats : il est ainsi par exemple possible d'introduire une mutation d'intérêt chez un embryon de souris, pour évaluer les conséquences de cette mutation sur le développement, et sur le comportement de la souris.

Cependant, cette approche impose une conservation en séquence de la région dont on souhaite évaluer la fonctionnalité. Si certains éléments non-codants conservés entre l'homme et la souris sont effectivement identifiés comme des éléments régulateurs (Visel et al., 2007), de récents travaux montrent que les régions régulatrices ne sont pas en majorité conservées en séquence, ni en localisation (Villar et al., 2015, Berthelot et al., 2018). Par ailleurs, la redondance des éléments régulateurs peut conduire à ce qu'une mutation effectivement fonctionnelle soit sauvée par d'autres éléments régulateurs la compensant (Osterwalder et al., 2018).

Ainsi, la question de la pénétrance et de l'expressivité variables d'une mutation limitent notre capacité à identifier formellement des mutations causatives ; la combinaison

de validations expérimentales vient donc s'ajouter aux prédictions *in silico* pour compléter le faisceau d'informations concernant le caractère causal associés aux mutations. Je note par ailleurs que récemment ont été développées des méthodes d'études à haut débit de la fonctionnalité des positions dans des régions régulatrices prédites (comme le STARR-seq), qui permettent donc de comprendre plus en détail les fonctions portées par ces régions (Klein et al., 2018, X. Wang et al., 2018).

Ainsi, le développement des méthodes de validations expérimentales à haut-débit et des projets de séquençage à large échelle nous permettent de mieux comprendre les propriétés des régions du génome non-codant, de comprendre comment les variations nucléotidiques se répartissent dans ces régions, et contribuent à établir des bases de références, à partir desquelles nous pourrons identifier plus clairement les divergences observées pour des patients atteints de maladies génétiques.

Bibliographie compilée

« A haplotype map of the human genome » (2005). In : *Nature* 437.7063. DOI : 10.1038/nature04226.

« A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms » (2001). En. In : *Nature* 409.6822, p. 928. DOI : 10.1038/35057149.

« A second generation human haplotype map of over 3.1 million SNPs » (2007). In : *Nature* 449.7164, p. 851–861. DOI : 10.1038/nature06258.

Acuna-Hidalgo, R., J. A. Veltman et A. Hoischen (2016). « New insights into the generation and role of de novo mutations in health and disease ». In : *Genome Biology* 17.1, p. 241. DOI : 10.1186/s13059-016-1110-1.

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork et al. (2010). « A method and server for predicting damaging missense mutations ». In : *Nature methods* 7.4, p. 248–249. DOI : 10.1038/nmeth0410-248.

Akker, J. van den, G. Mishne, A. D. Zimmer et A. Y. Zhou (2018). « A machine learning model to determine the accuracy of variant calls in capture-based next generation sequencing ». eng. In : *BMC genomics* 19.1, p. 263. DOI : 10.1186/s12864-018-4659-0.

Albert, F. W. et L. Kruglyak (2015). « The role of regulatory variation in complex traits and disease ». In : *Nat Rev Genet* 16.4.

Amberger, J. S., C. A. Bocchini, F. Schiettecatte, A. F. Scott et A. Hamosh (2015). « OMIM.org : Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders ». In : *Nucleic Acids Research* 43.Database issue, p. D789–D798. DOI : 10.1093/nar/gku1205.

An, J.-Y., K. Lin, L. Zhu, D. M. Werling, S. Dong, H. Brand et al. (2018). « Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder ». en. In : *Science* 362.6420, eaat6576. DOI : 10.1126/science.aat6576.

Andersson, R., C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd et al. (2014). « An atlas of active enhancers across human cell types and tissues ». In : *Nature* 507.7493, p. 455–461.

Auer, P. L. et G. Lettre (2015). « Rare variant association studies : considerations, challenges and opportunities ». In : *Genome Medicine* 7.1. doi : 10.1186/s13073-015-0138-2.

Barbosa, M., R. S. Joshi, P. Garg, A. Martin-Trujillo, N. Patel, B. Jadhav et al. (2018). « Identification of rare de novo epigenetic variations in congenital disorders ». En. In : *Nature Communications* 9.1, p. 2064. doi : 10.1038/s41467-018-04540-x.

Belton, J.-M., R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan et J. Dekker (2012). « Hi–C : A comprehensive technique to capture the conformation of genomes ». en. In : *Methods* 58.3, p. 268–276. doi : 10.1016/j.ymeth.2012.05.001.

Bemmel, J. G. v., R. Galupa, C. Gard, N. Servant, C. Picard, J. Davies et al. (2019). « The bipartite TAD organization of the X-inactivation center ensures opposing developmental regulation of Tsix and Xist ». En. In : *Nature Genetics* 51.6, p. 1024. doi : 10.1038/s41588-019-0412-0.

Berthelot, C., D. Villar, J. E. Horvath, D. T. Odom et P. Flicek (2018). « Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression ». en. In : *Nature Ecology & Evolution* 2.1, p. 152–163. doi : 10.1038/s41559-017-0377-2.

Boyle, A. P., E. L. Hong, M. Hariharan, Y. Cheng, M. A. Schaub, M. Kasowski et al. (2012). « Annotation of functional variation in personal genomes using RegulomeDB ». In : *Genome Research* 22.9, p. 1790–1797. doi : 10.1101/gr.137323.112.

El-Brolosy, M. A., Z. Kontarakis, A. Rossi, C. Kuenne, S. Günther, N. Fukuda et al. (2019). « Genetic compensation triggered by mutant mRNA degradation ». En. In : *Nature* 568.7751. doi : 10.1038/s41586-019-1064-z.

El-Brolosy, M. A. et D. Y. R. Stainier (2017). « Genetic compensation : A phenomenon in search of mechanisms ». en. In : *PLOS Genetics* 13.7. doi : 10.1371/journal.pgen.1006780.

Campa, E. Á. de la, N. Padilla et X. de la Cruz (2017). « Development of pathogenicity predictors specific for variants that do not comply with clinical guidelines for the use of computational evidence ». In : *BMC Genomics* 18.Suppl 5. doi : 10.1186/s12864-017-3914-0.

Caron, B., Y. Luo et A. Rausell (2018). « Scoring of pathogenic non-coding variants in Mendelian diseases through supervised learning on ancient, recent and ongoing purifying selection signals in human ». en. In : *bioRxiv*, p. 363903. doi : 10.1101/363903.

Carss, K. J., G. Arno, M. Erwood, J. Stephens, A. Sanchis-Juan, S. Hull et al. (2017). « Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease ». en. In : *The American Journal of Human Genetics* 100.1, p. 75–90. doi : 10.1016/j.ajhg.2016.12.003.

Carvalho-Silva, D., A. Pierleoni, M. Pignatelli, C. Ong, L. Fumis, N. Karamanis et al. (2019). « Open Targets Platform : new developments and updates two years on ». en. In : *Nucleic Acids Research* 47.D1, p. D1056–D1065. doi : 10.1093/nar/gky1133.

Char, D. S., N. H. Shah et D. Magnus (2018). « Implementing Machine Learning in Health Care — Addressing Ethical Challenges ». In : *The New England journal of medicine* 378.11, p. 981–983. doi : 10.1056/NEJMp1714229.

Chawla, N. V., K. W. Bowyer, L. O. Hall et W. P. Kegelmeyer (2002). « SMOTE : Synthetic Minority Over-sampling Technique ». en. In : *Journal of Artificial Intelligence Research* 16, p. 321–357. doi : 10.1613/jair.953.

Chen, T. et C. Guestrin (2016). « Xgboost : A scalable tree boosting system ». In : *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, p. 785–794.

Chèneby, J., M. Gheorghe, M. Artufel, A. Mathelier et B. Ballester (2018). « ReMap 2018 : an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments ». In : *Nucleic Acids Research* 46, p. D267–D275. doi : 10.1093/nar/gkx1092.

Church, D. M., V. A. Schneider, K. M. Steinberg, M. C. Schatz, A. R. Quinlan, C.-S. Chin et al. (2015). « Extending reference assembly models ». In : *Genome Biology* 16.1, p. 13. doi : 10.1186/s13059-015-0587-3.

Clément, Y., P. Torbey, P. Gilardi-Hebenstreit et H. R. Crollius (2018). « Enhancer-gene maps in the human and zebrafish genomes using evolutionary linkage conservation ». en. In : *bioRxiv*, p. 244475. doi : 10.1101/244475.

Crollius, H. R., O. Jaillon, A. Bernot, C. Dasilva, L. Bouneau, C. Fischer et al. (2000). « Estimate of human gene number provided by genome-wide analysis using Tetraodon nigroviridis DNA sequence ». In : *Nature Genetics* 25.2, p. 235. doi : 10.1038/76118.

Cutler, A., D. R. Cutler et J. R. Stevens (2001). « Random Forests ». en. In : *Ensemble Machine Learning : Methods and Applications*. Sous la dir. de C. Zhang et Y. Ma. Boston, MA : Springer US, p. 157–175. doi : 10.1007/978-1-4419-9326-7_5.

Dale, R. K., B. S. Pedersen et A. R. Quinlan (2011). « Pybedtools : a flexible Python library for manipulating genomic datasets and annotations ». en. In : *Bioinformatics* 27.24, p. 3423–3424. doi : 10.1093/bioinformatics/btr539.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson et E. S. Lander (2001). « High-resolution haplotype structure in the human genome ». En. In : *Nature Genetics* 29.2, p. 229. doi : 10.1038/ng1001-229.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo et al. (2011). « The variant call format and VCFtools ». en. In : *Bioinformatics* 27.15, p. 2156–2158. doi : 10.1093/bioinformatics/btr330.

Davydov, E. V., D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow et S. Batzoglou (2010). « Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++ ». en. In : *PLoS Computational Biology* 6.12. Sous la dir. de W. W. Wasserman, e1001025. doi : 10.1371/journal.pcbi.1001025.

Dekker, J., M. A. Marti-Renom et L. A. Mirny (2013). « Exploring the three-dimensional organization of genomes : interpreting chromatin interaction data ». In : *Nature Reviews Genetics* 14.6, p. 390–403. doi : 10.1038/nrg3454.

Denker, A. et W. de Laat (2016). « The second decade of 3C technologies : detailed insights into nuclear organization ». en. In : *Genes & Development* 30.12, p. 1357–1382. doi : 10.1101/gad.281964.116.

Dickinson, M. E., A. M. Flenniken, X. Ji, L. Teboul, M. D. Wong, J. K. White et al. (2016). « High-throughput discovery of novel developmental phenotypes ». In : *Nature* 537.7621, p. 508–514. doi : 10.1038/nature19356.

Dopazo, J., A. Amadoz, M. Bleda, L. Garcia-Alonso, A. Alemán, F. García-García et al. (2016). « 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation ». In : *Molecular Biology and Evolution* 33.5, p. 1205–1218. doi : 10.1093/molbev/msw005.

Dror, I., R. Rohs et Y. Mandel-Gutfreund (2016). « How motif environment influences transcription factor search dynamics : Finding a needle in a haystack ». en. In : *BioEssays* 38.7. doi : 10.1002/bies.201600005.

Drubay, D., D. Gautheret et S. Michiels (2018). « A benchmark study of scoring methods for non-coding mutations ». In : *Bioinformatics (Oxford, England)* 34.10, p. 1635–1641. doi : 10.1093/bioinformatics/bty008.

Dulbecco, R. (1986). « A turning point in cancer research : sequencing the human genome ». en. In : *Science* 231.4742, p. 1055–1056. doi : 10.1126/science.3945817.

Edwards, S. L., J. Beesley, J. D. French et A. M. Dunning (2013). « Beyond GWASs : Illuminating the Dark Road from Association to Function ». In : *American Journal of Human Genetics* 93.5, p. 779–797. doi : 10.1016/j.ajhg.2013.10.012.

Epifanio, I. (2017). « Intervention in prediction measure : a new approach to assessing variable importance for random forests ». In : *BMC Bioinformatics* 18.1, p. 230. doi : 10.1186/s12859-017-1650-8.

Epp, C. D. (1997). « Definition of a gene ». En. In : *Nature* 389.6651, p. 537. doi : 10.1038/39166.

Ernst, J. et M. Kellis (2010). « Discovery and characterization of chromatin states for systematic annotation of the human genome ». en. In : *Nature Biotechnology* 28.8, p. 817–825. doi : 10.1038/nbt.1662.

— (2012). « ChromHMM : automating chromatin-state discovery and characterization ». In : *Nature Methods* 9.3, p. 215–216. doi : 10.1038/nmeth.1906.

Fishilevich, S., R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein et al. (2017). « GeneHancer : genome-wide integration of enhancers and target genes in GeneCards ». In : *Database : The Journal of Biological Databases and Curation* 2017. doi : 10.1093/database/bax028.

Francioli, L. C., P. P. Polak, A. Koren, A. Menelaou, S. Chun, I. Renkens et al. (2015). « Genome-wide patterns and properties of de novo mutations in humans ». In : *Nature genetics* 47.7, p. 822–826. doi : 10.1038/ng.3292.

Frankish, A., M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland et al. (2019). « GENCODE reference annotation for the human and mouse genomes ». In : *Nucleic Acids Research* 47.Database issue, p. D766–D773. doi : 10.1093/nar/gky955.

Frésard, L., C. Smail, N. M. Ferraro, N. A. Teran, X. Li, K. S. Smith et al. (2019). « Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts ». En. In : *Nature Medicine*, p. 1. doi : 10.1038/s41591-019-0457-8.

Gijsberts, C. M., K. A. Groenewegen, I. E. Hoefer, M. J. C. Eijkemans, F. W. Asselbergs, T. J. Anderson et al. (2015). « Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events ». eng. In : *PloS One* 10.7. doi : 10.1371/journal.pone.0132321.

Gilissen, C., A. Hoischen, H. G. Brunner et J. A. Veltman (2011). « Unlocking Mendelian disease using exome sequencing ». In : *Genome Biology* 12.9, p. 228. doi : 10.1186/gb-2011-12-9-228.

Giresi, P. G., J. Kim, R. M. McDaniell, V. R. Iyer et J. D. Lieb (2007). « FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin ». In : *Genome Research* 17.6, p. 877–885. doi : 10.1101/gr.5533506.

Glisovic, T., J. L. Bachorik, J. Yong et G. Dreyfuss (2008). « RNA-binding proteins and post-transcriptional gene regulation ». In : *FEBS letters* 582.14, p. 1977–1986. doi : 10.1016/j.febslet.2008.03.004.

Gregory, S. G., K. F. Barlow, K. E. McLay, R. Kaul, D. Swarbreck, A. Dunham et al. (2006). « The DNA sequence and biological annotation of human chromosome 1 ». En. In : *Nature* 441.7091. doi : 10.1038/nature04727.

Gronau, I., L. Arbiza, J. Mohammed et A. Siepel (2013). « Inference of Natural Selection from Interspersed Genomic Elements Based on Polymorphism and Divergence ». In : *Molecular Biology and Evolution* 30.5, p. 1159–1171. doi : 10.1093/molbev/mst019.

GTEx Consortium, L. Analysts, D. A. C. C. (Laboratory, N. p. Management, B. Collection, Pathology et al. (2017). « Genetic effects on gene expression across human tissues ». en. In : *Nature* 550.7675. doi : 10.1038/nature24277.

Gudbjartsson, D. F., H. Helgason, S. A. Gudjonsson, F. Zink, A. Oddson, A. Gylfason et al. (2015). « Large-scale whole-genome sequencing of the Icelandic population ». en. In : *Nature Genetics* 47.5, p. 435–444. doi : 10.1038/ng.3247.

Guénet, J. L. (2005). « The mouse genome ». en. In : *Genome Research* 15.12, p. 1729–1740. doi : 10.1101/gr.3728305.

Gulko, B., M. J. Hubisz, I. Gronau et A. Siepel (2015). « A method for calculating probabilities of fitness consequences for point mutations across the human genome ». In : *Nat Genet* 47.3.

Guo, M. H., A. Dauber, M. F. Lippincott, Y.-M. Chan, R. M. Salem et J. N. Hirschhorn (2016). « Determinants of Power in Gene-Based Burden Testing for Monogenic Disorders ». In : *The American Journal of Human Genetics* 99.3, p. 527–539. doi : 10.1016/j.ajhg.2016.06.031.

Gurdasani, D., I. Barroso, E. Zeggini et M. S. Sandhu (2019). « Genomics of disease risk in globally diverse populations ». En. In : *Nature Reviews Genetics*. doi : 10.1038/s41576-019-0144-0.

Hait, T. A., D. Amar, R. Shamir et R. Elkon (2018). « FOCS : a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map ». In : *Genome Biology* 19. doi : 10.1186/s13059-018-1432-2.

Hardison, R. C. (2003). « Comparative Genomics ». In : *PLoS Biology* 1. doi : 10.1371/journal.pbio.0000058.

Hinrichs, A. S., D. Karolchik, R. Baertsch, G. P. Barber, G. Bejerano, H. Clawson et al. (2006). « The UCSC Genome Browser Database : update 2006 ». In : *Nucleic Acids Research* 34.Database issue, p. D590–D598. doi : 10.1093/nar/gkj144.

Holm, E. A. (2019). « In defense of the black box ». en. In : *Science* 364.6435, p. 26–27. doi : 10.1126/science.aax0162.

Hout, C. V. V., I. Tachmazidou, J. D. Backman, J. X. Hoffman, B. Ye, A. K. Pandey et al. (2019). « Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank ». In : *bioRxiv*, p. 572347. doi : 10.1101/572347.

Huang, Y.-F., B. Gulko et A. Siepel (2017). « Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data ». eng. In : *Nature Genetics* 49.4. doi : 10.1038/ng.3810.

Hwang, K.-B., I.-H. Lee, H. Li, D.-G. Won, C. Hernandez-Ferrer, J. A. Negron et al. (2019). « Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings ». En. In : *Scientific Reports* 9.1, p. 3219. doi : 10.1038/s41598-019-39108-2.

« Initial sequencing and analysis of the human genome » (2001). En. In : *Nature* 409.6822, p. 860. doi : 10.1038/35057062.

« Initial sequencing and comparative analysis of the mouse genome » (2002). En. In : *Nature* 420.6915, p. 520. doi : 10.1038/nature01262.

« Integrating common and rare genetic variation in diverse human populations » (2010). In : *Nature* 467.7311, p. 52–58. doi : 10.1038/nature09298.

International Human Genome Sequencing Consortium (2004). « Finishing the euchromatic sequence of the human genome ». eng. In : *Nature* 431.7011, p. 931–945. doi : 10.1038/nature03001.

Ioannidis, N. M., J. R. Davis, M. K. DeGorter, N. B. Larson, S. K. McDonnell, A. J. French et al. (2017). « FIRE : functional inference of genetic variants that regulate gene expression ». en. In : *Bioinformatics*. doi : 10.1093/bioinformatics/btx534.

Ioannidis, N. M., J. H. Rothstein, V. Pejaver, S. Middha, S. K. McDonnell, S. Baheti et al. (2016). « REVEL : An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants ». English. In : *The American Journal of Human Genetics* 99.4, p. 877–885. doi : 10.1016/j.ajhg.2016.08.016.

Ionita-Laza, I., K. McCallum, B. Xu et J. D. Buxbaum (2016). « A spectral approach integrating functional genomic annotations for coding and noncoding variants ». In : *Nature Genetics* 48.2, p. 214–220. doi : 10.1038/ng.3477.

Iossifov, I., M. Ronemus, D. Levy, Z. Wang, I. Hakker, J. Rosenbaum et al. (2012). « De Novo Gene Disruptions in Children on the Autistic Spectrum ». In : *Neuron* 74.2, p. 285–299. doi : 10.1016/j.neuron.2012.04.009.

Iulio, J. d., I. Bartha, E. H. M. Wong, H.-C. Yu, V. Lavrenko, D. Yang et al. (2018). « The human noncoding genome defined by genetic diversity ». en. In : *Nature Genetics*, p. 1. doi : 10.1038/s41588-018-0062-7.

Javierre, B. M., O. S. Burren, S. P. Wilder, R. Kreuzhuber, S. M. Hill, S. Sewitz et al. (2016). « Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters ». en. In : *Cell* 167.5, 1369–1384.e19. doi : 10.1016/j.cell.2016.09.037.

Jolma, A., T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge et al. (2010). « Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities ». en. In : *Genome Research* 20.6. doi : 10.1101/gr.100552.109.

Jolma, A., Y. Yin, K. R. Nitta, K. Dave, A. Popov, M. Taipale et al. (2015). « DNA-dependent formation of transcription factor pairs alters their binding specificity ». en. In : *Nature* 527.7578, p. 384–388. doi : 10.1038/nature15518.

Jónsson, H., P. Sulem, B. Kehr, S. Kristmundsdottir, F. Zink, E. Hjartarson et al. (2017). « Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland ». En. In : *Nature* 549.7673, p. 519. doi : 10.1038/nature24018.

Karczewski, K. J., L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang et al. (2019). « Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes ». In : *bioRxiv*, p. 531210. doi : 10.1101/531210.

Karolchik, D., A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler et al. (2004). « The UCSC Table Browser data retrieval tool ». eng. In : *Nucleic Acids Research* 32.Database issue, p. D493–496. doi : 10.1093/nar/gkh103.

Kent, W. J., A. S. Zweig, G. Barber, A. S. Hinrichs et D. Karolchik (2010). « BigWig and BigBed : enabling browsing of large distributed datasets ». en. In : *Bioinformatics* 26.17, p. 2204–2207. doi : 10.1093/bioinformatics/btq351.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler et al. (2002). « The Human Genome Browser at UCSC ». en. In : *Genome Research* 12.6, p. 996–1006. doi : 10.1101/gr.229102.

Khan, A., O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee et al. (2018). « JASPAR 2018 : update of the open-access database of transcription factor binding profiles and its web framework ». In : *Nucleic Acids Research* 46, p. D260–D266. doi : 10.1093/nar/gkx1126.

Kikuta, H., M. Laplante, P. Navratilova, A. Z. Komisarczuk, P. G. Engström, D. Fredman et al. (2007). « Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates ». In : *Genome Research* 17.5, p. 545–555. doi : 10.1101/gr.6086307.

Kinsella, R. J., A. Kähäri, S. Haider, J. Zamora, G. Proctor, G. Spudich et al. (2011). « Ensembl BioMarts : a hub for data retrieval across taxonomic space ». In : *Database : The Journal of Biological Databases and Curation* 2011. doi : 10.1093/database/bar030.

Kircher, M., D. M. Witten, P. Jain, B. J. O'Roak, G. M. Cooper et J. Shendure (2014). « A general framework for estimating the relative pathogenicity of human genetic variants ». en. In : *Nature Genetics* 46.3, p. 310–315. doi : 10.1038/ng.2892.

Klein, J. C., A. Keith, V. Agarwal, T. Durham et J. Shendure (2018). « Functional characterization of enhancer evolution in the primate lineage ». In : *Genome Biology* 19.1, p. 99. doi : 10.1186/s13059-018-1473-6.

Koscielny, G., P. An, D. Carvalho-Silva, J. A. Cham, L. Fumis, R. Gasparyan et al. (2017). « Open Targets : a platform for therapeutic target identification and

validation ». In : *Nucleic Acids Research* 45.D1, p. D985–D994. DOI : 10.1093/nar/gkw1055.

Kuhn, R. M., D. Haussler et W. J. Kent (2013). « The UCSC genome browser and associated tools ». In : *Briefings in Bioinformatics* 14.2, p. 144–161. DOI : 10.1093/bib/bbs038.

Kumar, P., S. Henikoff et P. C. Ng (2009). « Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm ». eng. In : *Nature Protocols* 4.7, p. 1073–1081. DOI : 10.1038/nprot.2009.86.

Kundaje, A., W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi et al. (2015). « Integrative analysis of 111 reference human epigenomes ». In : *Nature* 518.7539, p. 317–330. DOI : 10.1038/nature14248.

Kuz'min, V. E., P. G. Polishchuk, A. G. Artemenko et S. A. Andronati (2011). « Interpretation of QSAR Models Based on Random Forest Methods ». en. In : *Molecular Informatics* 30.6-7. DOI : 10.1002/minf.201000173.

Landrum, M. J., J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church et al. (2014). « ClinVar : public archive of relationships among sequence variation and human phenotype ». In : *Nucleic Acids Research* 42.Database issue, p. D980–D985. DOI : 10.1093/nar/gkt1113.

Leblond, C. S., F. Cliquet, C. Carton, G. Huguet, A. Mathieu, T. Kergrohen et al. (2019). « Both rare and common genetic variants contribute to autism in the Faroe Islands ». En. In : *npj Genomic Medicine* 4.1, p. 1. DOI : 10.1038/s41525-018-0075-2.

Lee, J. J., R. Wedow, A. Okbay, E. Kong, O. Maghzian, M. Zacher et al. (2018). « Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment ». In : *Nature genetics* 50. DOI : 10.1038/s41588-018-0147-3.

Lek, M., K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell et al. (2016). « Analysis of protein-coding genetic variation in 60,706 humans ». In : *Nature* 536.7616, p. 285–291. DOI : 10.1038/nature19057.

Li, H. (2011). « Tabix : fast retrieval of sequence features from generic TAB-delimited files ». en. In : *Bioinformatics* 27.5, p. 718–719. DOI : 10.1093/bioinformatics/btq671.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer et al. (2009). « The Sequence Alignment/Map format and SAMtools ». eng. In : *Bioinformatics (Oxford, England)* 25.16. DOI : 10.1093/bioinformatics/btp352.

Liu, X., C. Li et E. Boerwinkle (2017). « The performance of deleteriousness prediction scores for rare non-protein-changing single nucleotide variants in human genes ». In : *Journal of medical genetics* 54.2, p. 134–144. doi : 10.1136/jmedgenet-2016-104369.

Lonsdale, J., J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad et al. (2013). « The Genotype-Tissue Expression (GTEx) project ». In : *Nature Genetics* 45.6, p. 580–585. doi : 10.1038/ng.2653.

Louhelainen, J. (2016). « SNP Arrays ». In : *Microarrays* 5.4. doi : 10.3390/microarrays5040027.

Maldonado, S., J. López et C. Vairetti (2019). « An alternative SMOTE oversampling strategy for high-dimensional datasets ». In : *Applied Soft Computing* 76, p. 380–389. doi : 10.1016/j.asoc.2018.12.024.

McCarthy, D. J., P. Humburg, A. Kanapin, M. A. Rivas, K. Gaulton, J.-B. Cazier et al. (2014). « Choice of transcripts and software has a large effect on variant annotation ». In : *Genome Medicine* 6.3. doi : 10.1186/gm543.

McLaren, W., L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann et al. (2016). « The Ensembl Variant Effect Predictor ». en. In : *Genome Biology* 17.1. doi : 10.1186/s13059-016-0974-4.

McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek et F. Cunningham (2010). « Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor ». en. In : *Bioinformatics* 26.16, p. 2069–2070. doi : 10.1093/bioinformatics/btq330.

McRae, J. F., S. Clayton, T. W. Fitzgerald, J. Kaplanis, E. Prigmore, D. Rajan et al. (2017). « Prevalence and architecture of de novo mutations in developmental disorders ». In : *Nature* 542.7642. doi : 10.1038/nature21062.

McVicker, G., D. Gordon, C. Davis et P. Green (2009). « Widespread Genomic Signatures of Natural Selection in Hominid Evolution ». en. In : *PLOS Genetics* 5.5, e1000471. doi : 10.1371/journal.pgen.1000471.

Michaelson, J. J., Y. Shi, M. Gujral, H. Zheng, D. Malhotra, X. Jin et al. (2012). « Whole-genome sequencing in autism identifies hot spots for de novo germline mutation ». In : *Cell* 151.7, p. 1431–1442. doi : 10.1016/j.cell.2012.11.019.

Mifsud, B., F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira et al. (2015). « Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C ». In : *Nat Genet* 47.6, p. 598–606.

Mishina, Y., R. Murata, Y. Yamauchi, T. Yamashita et H. Fujiyoshi (2015). « Boosted random forest ». In : *IEICE Transactions on Information and systems* 98.9, p. 1630–1636.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer et B. Wold (2008). « Mapping and quantifying mammalian transcriptomes by RNA-Seq ». en. In : *Nature Methods* 5.7, p. 621–628. doi : 10.1038/nmeth.1226.

Naville, M., M. Ishibashi, M. Ferg, H. Bengani, S. Rinkwitz, M. Krecsmarik et al. (2015). « Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome ». In : *Nat Commun* 6.

Neph, S., M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson et al. (2012). « BEDOPS : high-performance genomic feature operations ». en. In : *Bioinformatics* 28.14, p. 1919–1920. doi : 10.1093/bioinformatics/bts277.

Neveling, K., I. Feenstra, C. Gilissen, L. H. Hoefsloot, E.-J. Kamsteeg, A. R. Mensenkamp et al. (2013). « A Post-Hoc Comparison of the Utility of Sanger Sequencing and Exome Sequencing for the Diagnosis of Heterogeneous Diseases ». In : *Human Mutation* 34.12, p. 1721–1726. doi : 10.1002/humu.22450.

Nguyen, N. T. T., B. Contreras-Moreira, J. A. Castro-Mondragon, W. Santana-Garcia, R. Ossio, C. D. Robles-Espinoza et al. (2018). « RSAT 2018 : regulatory sequence analysis tools 20th anniversary ». In : *Nucleic Acids Research* 46.Web Server issue, W209–W214. doi : 10.1093/nar/gky317.

O'Roak, B. J., P. Deriziotis, C. Lee, L. Vives, J. J. Schwartz, S. Girirajan et al. (2011). « Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations ». In : *Nature genetics* 43.6, p. 585–589. doi : 10.1038/ng.835.

Osterwalder, M., I. Barozzi, V. Tissières, Y. Fukuda-Yuzawa, B. J. Mannion, S. Y. Afzal et al. (2018). « Enhancer redundancy provides phenotypic robustness in mammalian development ». In : *Nature* 554.7691, p. 239–243. doi : 10.1038/nature25461.

Ouwehand, W. H. (2019). « Whole-genome sequencing of rare disease patients in a national healthcare system ». en. In : *bioRxiv*. doi : 10.1101/507244.

Palczewska, A. M., J. Palczewski, R. Marchese Robinson et D. Neagu (2014). « Interpreting random forest classification models using a feature contribution method ». en. In : *Integration of Reusable Systems*. Sous la dir. de T. Bouabana-Tebibel et S. H. Rubin. T. 263. Springer, 193–218 (26).

Paul, D. S., N. Soranzo et S. Beck (2014). « Functional interpretation of non-coding sequence variation : Concepts and challenges ». In : *Bioessays* 36.2. doi : 10.1002/bies.201300126.

Pearson, H. (2006). *What is a gene ?* En. News. doi : 10.1038/441398a.

Pertea, M. et S. L. Salzberg (2010). « Between a chicken and a grape : estimating the number of human genes ». In : *Genome Biology* 11.5, p. 206. doi : 10.1186/gb-2010-11-5-206.

Petrovski, S., Q. Wang, E. L. Heinzen, A. S. Allen et D. B. Goldstein (2013). « Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes ». en. In : *PLOS Genetics* 9.8. doi : 10.1371/journal.pgen.1003709.

Pinto, D., A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan et al. (2010). « Functional impact of global rare copy number variation in autism spectrum disorders ». eng. In : *Nature* 466.7304, p. 368–372. doi : 10.1038/nature09146.

Pohl, A. et M. Beato (2014). « bwtool : a tool for bigWig files ». en. In : *Bioinformatics* 30.11. doi : 10.1093/bioinformatics/btu056.

Pollard, K. S., M. J. Hubisz, K. R. Rosenbloom et A. Siepel (2010). « Detection of nonneutral substitution rates on mammalian phylogenies ». In : *Genome Research* 20.1, p. 110–121. doi : 10.1101/gr.097857.109.

Pruitt, K. D., T. Tatusova et D. R. Maglott (2007). « NCBI reference sequences (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins ». In : *Nucleic Acids Research* 35.Database issue, p. D61–D65. doi : 10.1093/nar/gkl842.

Quinlan, A. R. et I. M. Hall (2010). « BEDTools : a flexible suite of utilities for comparing genomic features ». en. In : *Bioinformatics* 26.6, p. 841–842. doi : 10.1093/bioinformatics/btq033.

Ramírez, F., F. Dündar, S. Diehl, B. A. Grüning et T. Manke (2014). « deepTools : a flexible platform for exploring deep-sequencing data ». In : *Nucleic Acids Research* 42.Web Server issue, W187–W191. doi : 10.1093/nar/gku365.

Raney, B. J., T. R. Dreszer, G. P. Barber, H. Clawson, P. A. Fujita, T. Wang et al. (2014). « Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser ». en. In : *Bioinformatics* 30.7, p. 1003–1005. doi : 10.1093/bioinformatics/btt637.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson et al. (2014). « A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping ». en. In : *Cell* 159.7, p. 1665–1680. doi : 10.1016/j.cell.2014.11.021.

Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (2016). fr.

Rentoft, M., D. Svensson, A. Sjödin, P. I. Olason, O. Sjöström, C. Nylander et al. (2019). « A geographically matched control population efficiently limits the number of candidate disease-causing variants in an unbiased whole-genome analysis ». In : *PLOS ONE* 14. doi : 10.1371/journal.pone.0213350.

Rentzsch, P., D. Witten, G. M. Cooper, J. Shendure et M. Kircher (2019). « CADD : predicting the deleteriousness of variants throughout the human genome ». en. In : *Nucleic Acids Research* 47.D1, p. D886–D894. doi : 10.1093/nar/gky1016.

RepeatMasker Open-4.0 (2013 - 2015). http://www.repeatmasker.org. Accessed : 2010-09-30.

Ribeiro, M. T., S. Singh et C. Guestrin (2016). « "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier ». In : *arXiv :1602.04938 [cs, stat]*. arXiv : 1602.04938.

Rice, A. M. et A. McLysaght (2017). « Dosage sensitivity is a major determinant of human copy number variant pathogenicity ». en. In : *Nature Communications* 8, p. 14366. doi : 10.1038/ncomms14366.

Riera, C., N. Padilla et X. de la Cruz (2016). « The Complementarity Between Protein-Specific and General Pathogenicity Predictors for Amino Acid Substitutions ». eng. In : *Human Mutation* 37.10. doi : 10.1002/humu.23048.

Rodenburg, R. J. (2018). « The functional genomics laboratory : functional validation of genetic variants ». en. In : *Journal of Inherited Metabolic Disease* 41.3, p. 297–307. doi : 10.1007/s10545-018-0146-7.

Rojano, E., P. Seoane, J. A. G. Ranea et J. R. Perkins (2018). « Regulatory variants : from detection to predicting impact ». en. In : *Briefings in Bioinformatics*. doi : 10.1093/bib/bby039.

Ronemus, M., I. Iossifov, D. Levy et M. Wigler (2014). « The role of *de novo* mutations in the genetics of autism spectrum disorders ». en. In : *Nature Reviews Genetics* 15.2, p. 133–141. doi : 10.1038/nrg3585.

Rousseeuw, P. J. (1987). « Silhouettes : A graphical aid to the interpretation and validation of cluster analysis ». In : *Journal of Computational and Applied Mathematics* 20. doi : 10.1016/0377-0427(87)90125-7.

Roy, S., C. Coldren, A. Karunamurthy, N. S. Kip, E. W. Klee, S. E. Lincoln et al. (2018). « Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines : A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists ». In : *The Journal of Molecular Diagnostics* 20.1, p. 4–27. doi : 10.1016/j.jmoldx.2017.11.003.

Saabas, A. (2015). https://github.com/andosa/treeinterpreter.

Sabo, P. J., M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson, H. Cao et al. (2006). « Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays ». En. In : *Nature Methods* 3.7, p. 511. doi : 10.1038/nmeth890.

Sanders, S. J., M. T. Murtha, A. R. Gupta, J. D. Murdoch, M. J. Raubeson, A. J. Willsey et al. (2012). « De novo mutations revealed by whole exome sequencing are strongly associated with autism ». In : *Nature* 485.7397, p. 237–241. doi : 10.1038/nature10945.

Santiago-Algarra, D., L. T. Dao, L. Pradel, A. España et S. Spicuglia (2017). « Recent advances in high-throughput approaches to dissect enhancer function ». In : *F1000Research* 6. doi : 10.12688/f1000research.11581.1.

Sebat, J., B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh et al. (2007). « Strong Association of De Novo Copy Number Mutations with Autism ». In : *Science (New York, N.Y.)* 316.5823, p. 445–449. doi : 10.1126/science.1138659.

Shendure, J., G. M. Findlay et M. W. Snyder (2019). « Genomic Medicine–Progress, Pitfalls, and Promise ». English. In : *Cell* 177.1, p. 45–57. doi : 10.1016/j.cell.2019.02.003.

Shendure, J. et H. Ji (2008). « Next-generation DNA sequencing ». en. In : *Nature Biotechnology* 26.10, p. 1135–1145. doi : 10.1038/nbt1486.

Shihab, H. A., M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day et al. (2015). « An integrative approach to predicting the functional effects of non-

coding and coding sequence variation ». In : *Bioinformatics* 31.10, p. 1536–1543. DOI : 10.1093/bioinformatics/btv009.

Shlyueva, D., G. Stampfel et A. Stark (2014). « Transcriptional enhancers : from properties to genome-wide predictions ». In : *Nat Rev Genet* 15.

Siepel, A., G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom et al. (2005). « Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes ». In : *Genome Research* 15.8, p. 1034–1050. DOI : 10.1101/gr.3715005.

Sinsheimer, R. L. (1989). « The Santa Cruz Workshop—May 1985 ». In : *Genomics* 5.4, p. 954–956. DOI : 10.1016/0888-7543(89)90142-0.

Sladek, R., G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre et al. (2007). « A genome-wide association study identifies novel risk loci for type 2 diabetes ». en. In : *Nature* 445.7130. DOI : 10.1038/nature05616.

Smedley, D., M. Schubach, J. O. Jacobsen, S. Köhler, T. Zemojtel, M. Spielmann et al. (2016). « A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease ». en. In : *The American Journal of Human Genetics* 99.3, p. 595–606. DOI : 10.1016/j.ajhg.2016.07.005.

Song, L. et G. E. Crawford (2010). « DNase-seq : a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells ». In : *Cold Spring Harbor protocols* 2010.2, pdb.prot5384. DOI : 10.1101/pdb.prot5384.

Stenson, P. D., M. Mort, E. V. Ball, K. Evans, M. Hayden, S. Heywood et al. (2017). « The Human Gene Mutation Database : towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies ». en. In : *Human Genetics* 136.6, p. 665–677. DOI : 10.1007/s00439-017-1779-6.

Stitziel, N. O., A. Kiezun et S. Sunyaev (2011). « Computational and statistical approaches to analyzing variants identified by exome sequencing ». In : *Genome Biology* 12. DOI : 10.1186/gb-2011-12-9-227.

Taylor, J. C., H. C. Martin, S. Lise, J. Broxholme, J.-B. Cazier, A. Rimmer et al. (2015). « Factors influencing success of clinical genome sequencing across a broad spectrum of disorders ». In : *Nature genetics* 47.7, p. 717–726. DOI : 10.1038/ng.3304.

Team, H. (p.d.). *Hail*. https://github.com/hail-is/hail/releases/tag/0.2.13.

The 1000 Genomes Project Consortium (2015). « A global reference for human genetic variation ». en. In : *Nature* 526.7571, p. 68–74. DOI : 10.1038/nature15393.

The ENCODE Project Consortium (2012). « An integrated encyclopedia of DNA elements in the human genome ». en. In : *Nature* 489.7414, p. 57–74. DOI : 10.1038/nature11247.

The GTEx Consortium, K. G. Ardlie, D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young et al. (2015). « The Genotype-Tissue Expression (GTEx) pilot analysis : Multitissue gene regulation in humans ». en. In : *Science* 348.6235, p. 648–660. DOI : 10.1126/science.1262110.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen et al. (2012). « The accessible chromatin landscape of the human genome ». In : *Nature* 489.7414, p. 75–82. DOI : 10.1038/nature11232.

« TRRUST v2 » (p.d.). « TRRUST v2 : an expanded reference database of human and mouse transcriptional regulatory interactions ». In : 46.Database issue. DOI : 10.1093/nar/gkx1013.

Turner, T. N., B. P. Coe, D. E. Dickel, K. Hoekzema, B. J. Nelson, M. C. Zody et al. (2017). « Genomic Patterns of De Novo Mutation in Simplex Autism ». English. In : *Cell* 171.3, 710–722.e12. DOI : 10.1016/j.cell.2017.08.047.

Turner, T. N. et E. E. Eichler (2018). « The Role of De Novo Noncoding Regulatory Mutations in Neurodevelopmental Disorders ». English. In : *Trends in Neurosciences* 0.0. DOI : 10.1016/j.tins.2018.11.002.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine et al. (2013). « From FastQ data to high confidence variant calls : the Genome Analysis Toolkit best practices pipeline ». In : *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* 11.1110, p. 11.10.1–11.10.33. DOI : 10.1002/0471250953.bi1110s43.

Vance, K. W. et C. P. Ponting (2014). « Transcriptional regulatory functions of nuclear long noncoding RNAs ». en. In : *Trends in Genetics* 30.8. DOI : 10.1016/j.tig.2014.06.001.

Velde, K. J. van der, E. N. de Boer, C. C. van Diemen, B. Sikkema-Raddatz, K. M. Abbott, A. Knopperts et al. (2017). « GAVIN : Gene-Aware Variant INterpretation for medical sequencing ». In : *Genome Biology* 18. DOI : 10.1186/s13059-016-1141-7.

Villar, D., C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli et al. (2015). « Enhancer Evolution across 20 Mammalian Species ». In : *Cell* 160.3, p. 554–566. DOI : 10.1016/j.cell.2015.01.006.

Visel, A., S. Minovitsky, I. Dubchak et L. A. Pennacchio (2007). « VISTA Enhancer Browser–a database of tissue-specific human enhancers ». en. In : *Nucleic Acids Research* 35.Database, p. D88–D92. doi : 10.1093/nar/gkl822.

Wang, X., L. He, S. M. Goggin, A. Saadat, L. Wang, N. Sinnott-Armstrong et al. (2018). « High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human ». En. In : *Nature Communications* 9.1, p. 5380. doi : 10.1038/s41467-018-07746-1.

Wang, Z., X. Liu, B.-Z. Yang et J. Gelernter (2013). « The Role and Challenges of Exome Sequencing in Studies of Human Diseases ». In : *Frontiers in Genetics* 4. doi : 10.3389/fgene.2013.00160.

Ward, L. D. et M. Kellis (2012). « Interpreting noncoding genetic variation in complex traits and human disease ». In : *Nat Biotech* 30.11.

Wasserman, W. W. et A. Sandelin (2004). « Applied bioinformatics for the identification of regulatory elements ». En. In : *Nature Reviews Genetics* 5.4, p. 276. doi : 10.1038/nrg1315.

Werling, D. M., H. Brand, J.-Y. An, M. R. Stone, L. Zhu, J. T. Glessner et al. (2018). « An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder ». En. In : *Nature Genetics* 50.5, p. 727. doi : 10.1038/s41588-018-0107-y.

Whalen, S., R. M. Truty et K. S. Pollard (2016). « Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin ». en. In : *Nature Genetics* 48.5, p. 488–496. doi : 10.1038/ng.3539.

Wikipedia contributors (2019). *Receiver operating characteristic — Wikipedia, The Free Encyclopedia*. [Online ; accessed 2-July-2019].

Wilkinson, M. F. (2019). « Genetic paradox explained by nonsense ». EN. In : *Nature* 568.7751. doi : 10.1038/d41586-019-00823-5.

Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson et al. (2014). « Defining the role of common variation in the genomic and biological architecture of adult human height ». en. In : *Nature Genetics* 46.11, p. 1173–1186. doi : 10.1038/ng.3097.

Worthey, E. A., A. N. Mayer, G. D. Syverson, D. Helbling, B. B. Bonacci, B. Decker et al. (2011). « Making a definitive diagnosis : successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease ».

In : *Genetics in Medicine : Official Journal of the American College of Medical Genetics* 13.3. DOI : 10.1097/GIM.0b013e3182088158.

Wright, C. F., D. R. FitzPatrick et H. V. Firth (2018). « Paediatric genomics : diagnosing rare disease in children ». en. In : *Nature Reviews Genetics* 19.5. DOI : 10.1038/nrg.2017.116.

Xi, W. et M. A. Beer (2018). « Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy ». en. In : *PLOS Computational Biology* 14.12. DOI : 10.1371/journal.pcbi.1006625.

Xu, D., O. Gokcumen et E. Khurana (2019). « Loss-of-function tolerance of enhancers in the human genome ». en. In : *bioRxiv*, p. 608257. DOI : 10.1101/608257.

Xue, A., Y. Wu, Z. Zhu, F. Zhang, K. E. Kemper, Z. Zheng et al. (2018). « Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes ». In : *Nature Communications* 9.1, p. 2941. DOI : 10.1038/s41467-018-04951-w.

Yates, B., B. Braschi, K. A. Gray, R. L. Seal, S. Tweedie et E. A. Bruford (2017). « Genenames.org : the HGNC and VGNC resources in 2017 ». In : *Nucleic Acids Research* 45.Database issue, p. D619–D625. DOI : 10.1093/nar/gkw1033.

Yuen, R. K., D. Merico, H. Cao, G. Pellecchia, B. Alipanahi, B. Thiruvahindrapuram et al. (2016). « Genome-wide characteristics of de novo mutations in autism ». In : *NPJ Genomic Medicine* 1, p. 16027. DOI : 10.1038/npjgenmed.2016.27.

Zerbino, D. R., S. P. Wilder, N. Johnson, T. Juettemann et P. R. Flicek (2015). « The ensembl regulatory build. » eng. In : *Genome biology* 16.1, p. 56–56. DOI : 10.1186/s13059-015-0621-5.

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell, J. Bhai et al. (2018). « Ensembl 2018 ». en. In : *Nucleic Acids Research* 46.D1, p. D754–D761. DOI : 10.1093/nar/gkx1098.

Zerbino, D. R., N. Johnson, T. Juettemann, S. P. Wilder et P. Flicek (2014). « WiggleTools : parallel processing of large collections of genome-wide datasets for visualization and statistical analysis ». en. In : *Bioinformatics* 30.7, p. 1008–1009. DOI : 10.1093/bioinformatics/btt737.

Zhou, J., C. Y. Park, C. L. Theesfeld, A. K. Wong, Y. Yuan, C. Scheckel et al. (2019). « Whole-genome deep-learning analysis identifies contribution of noncoding

mutations to autism risk ». En. In : *Nature Genetics* 51.6, p. 973. DOI : 10.1038/s41588-019-0420-0.

RÉSUMÉ

Le séquençage de génome complet est utilisé de façon croissante chez les patients atteints de maladies génétiques pour diagnostiquer les mutations responsables. Cependant, pour une grande proportion de génomes de patients séquencés, aucun gène associé au phénotype ne présente de mutation codante. Dans ces cas, il est possible qu'une mutation non-codante, localisée dans une région cis-régulatrice, modifie l'expression d'un gène impliqué dans la maladie. Malgré l'existence de méthodes pour annoter et prédire des séquences régulatrices sur la base de propriétés biochimiques et épigénétiques, il reste difficile de définir des critères objectifs pour sélectionner efficacement des mutations candidates parmi les millions de variants non-codants présents chez chaque patient. De plus, les gènes cibles de ces séquences de régulation ne sont généralement pas connus, si bien qu'il est difficile de croiser une mutation non-codante avec le phénotype du patient.

Je propose ici une stratégie d'apprentissage supervisé par forêts aléatoires, adaptée aux jeux de données complexes et hétérogènes, pour classer et sélectionner des mutations non-codantes dérégulant des gènes responsables de maladies. Une innovation notable de mon approche est de prendre en compte des données d'associations entre régions non-codantes et gènes cibles. Par ailleurs, je propose une méthode d'extraction des règles biologiques identifiées par le modèle pour chaque mutation évaluée, ce qui permet une sélection éclairée des mutations hiérarchisées.

Je discute les propriétés fonctionnelles identifiées par le modèle d'apprentissage, à partir d'exemples de variations non-codantes associées à des maladies mendéliennes. J'illustre également le potentiel de cette méthode notamment par une analyse de 255 106 variants de novo identifiés par séquençage complet chez 1902 enfants souffrant de troubles du spectre autistique, et chez lesquels aucune mutation codante pathogénique n'a été identifiée.

Cette méthode permet ainsi de hiérarchiser des mutations, dont les plus prometteuses deviennent des hypothèses testables expérimentalement pour confirmer leur implication dans le développement des maladies considérées. Ainsi pour les projets de séquençage génome-complets de cohortes de patients, une application systématique de notre méthode contribuerait à une meilleure compréhension des mécanismes de régulation de l'expression des gènes, et à une amélioration du diagnostic des patients.

Mots clés : bioinformatique ; génomique fonctionnelle ; génomique humaine ; régulation de l'expression génique ; apprentissage machine

ABSTRACT

Whole genome sequencing is increasingly used in patients with genetic diseases to diagnose the mutations responsible for the phenotype. However, for a large proportion of sequenced genomes, none of the genes associated with the phenotype have a coding mutation. In these cases, it is possible that a non-coding mutation, located in a cis-regulatory region, modifies the expression of a gene involved in the disease. Despite the existence of methods for annotating and predicting regulatory sequences on the basis of biochemical and epigenetic properties, it remains difficult to define objective criteria for effectively selecting candidate mutations from the millions of non-coding variants present in each patient. In addition, the target genes of these regulatory sequences are generally not known, making it difficult to associate a non-coding mutation with the patient's phenotype.

I propose here a supervised learning strategy using random forests, adapted to complex and heterogeneous data sets, to classify and select non-coding mutations deregulating genes responsible for diseases. A notable innovation of my approach is to take into account data of associations between non-coding regions and target genes. In addition, I propose a method for extracting the biological rules identified by the model for each mutation evaluated, allowing an informed selection of candidate mutations.

I discuss the functional properties identified by the learning model, based on examples of non-coding variations associated with Mendelian diseases. I also illustrate the potential of this method by analyzing 255,106 de novo variants identified by complete sequencing in 1,902 children with autism spectrum disorders, in whom no pathogenic coding mutations have been identified.

This method thus makes it possible to prioritize mutations, the most promising of which become experimentally testable hypotheses to confirm their involvement in the development of the diseases in question. Thus, for whole genome sequencing projects of patient cohorts, a systematic application of our method would contribute to a better understanding of the mechanisms regulating gene expression, and to an improvement in patient diagnosis.

Keywords : bioinformatics ; functional genomics ; human genomics ; gene expression regulation ; machine learning