



Intégration d'annotations fonctionnelles dans les tests d'agrégation de variants rares sous vraisemblance rétrospective dans le cadre de devis familiaux

Thèse

Loïc Mangnier

Doctorat en biostatistique
Philosophiæ doctor (Ph. D.)

Québec, Canada

Intégration d'annotations fonctionnelles dans les tests d'agrégation de variants rares sous vraisemblance rétrospective dans le cadre de devis familiaux

Thèse

Loïc Mangnier

Sous la direction de :

Alexandre Bureau, directeur de recherche

Arnaud Droit, codirecteur de recherche

Résumé

Depuis ces vingt dernières années et la démocratisation des technologies de séquençage haut débit de l'ADN, l'accessibilité des données génomiques a permis le développement de nouveaux outils analytiques, notamment dans l'étude des maladies complexes. Au-delà de la simple perspective descriptive, il y a un intérêt croissant des chercheurs à étudier les phénomènes biologiques menant à l'émergence ou au développement de maladies chez l'humain. En effet, la vaste majorité des troubles complexes sont la résultante de la combinaison de variations rares et communes situées dans des régions non codantes du génome, rendant l'interprétation des mécanismes biologiques en jeu difficile. Ce dernier point a été l'occasion pour les chercheurs de caractériser, à travers de nouvelles technologies, certaines régions du génome selon leur profil épigénétique et d'intégrer cette information dans de nouveaux outils statistiques dans une optique de détection de variants impliqués dans les maladies.

Premièrement, à l'heure actuelle les modèles existant pour caractériser les régions non codantes et les lier à leur(s) gènes cibles n'intègrent pas la notion de réseaux d'interaction. Ainsi il devient critique de proposer un modèle computationnel, valide biologiquement, capable de capturer les contacts entre gènes et éléments de régulation à travers des réseaux, afin de mettre en lumière les mécanismes biologiques impliqués dans les maladies complexes.

D'un autre côté, bien que des modèles permettant de tenir compte de la rareté des variants et de leur implication fonctionnelle dans les maladies ont déjà été proposés, ces derniers ne se limitent qu'aux devis cas-témoins d'individus non apparentés, pouvant nécessiter des tailles d'échantillon souvent élevées pour détecter des associations. Il y a alors un réel besoin méthodologique d'étendre ces approches aux études familiales, lorsque ces dernières sont plus performantes pour identifier des variants rares impliqués dans les maladies.

Ainsi, dans cette perspective, nous proposons un nouveau modèle de réseaux intégrant l'information épigénétique au travers des contacts physiques entre gènes et enhancers, appelé pôles de régulation cis. Aussi, parce que cette information peut être exploiter afin de mettre en lumière de nouvelles régions associées avec des maladies, nous proposons un

nouveau test d'association de variants rares, appelé *RetroFun-RVS*, exploitant la structure familiale et permettant l'intégration des annotations fonctionnelles sous forme de réseaux.

Enfin, nous avons démontré que notre modèle de pôles de régulation cis en plus d'être biologiquement valide est capable de mieux cerner les mécanismes épigénétiques impliqués dans l'étiologie de maladies complexes. Par ailleurs, *RetroFun-RVS* en incorporant les pôles de régulation cis, en tant qu'annotations fonctionnelles et l'information familiale, a été démontré une approche puissante pour détecter de nouveaux variants causaux. Dans une perspective mécanistique, proposer des modèles unifiant approches familiales et information fonctionnelle ouvre alors de nouvelles voies dans l'étude des maladies complexes, notamment en permettant de cibler plus précisément les phénomènes de régulation impliqués.

Abstract

Over the past few years, with the development of whole-genome sequencing technologies, improvements have been made in the understanding of complex diseases. Indeed, most phenotypes are characterized by a combination of common and rare variants, mainly located within noncoding regions, making the underlying biological mechanisms difficult to interpret. The increased availability of genomics data has provided opportunities for researchers to characterize noncoding regions and assess their roles in diseases, by incorporating this information in analytical tools to detect new causal variants involved in diseases.

On the one side, to date, methods to detect enhancers and link to their target genes do not integrate complex networks of interactions. Indeed, no network-based model integrating physical contacts between genes and enhancers has been proposed. Thus, it remains crucial to provide a computational model which is biologically relevant and able to capture complex regulatory interactions, in order to gain insights in epigenetics mechanisms involved in the etiology of complex diseases.

On the other side, although rare-variant association tests integrating functional information have already been proposed, models currently available suffer from two major limitations. Firstly, they are restricted to case-control designs of unrelated people, where tens or hundreds thousand of individuals are often required to reach sufficient power, limiting their applicability in practice. Secondly, these approaches have been mainly evaluated using continuous scores. Indeed, no functional annotations corresponding to regions with regulatory impacts have been assessed. To overcome these limitations, there is an important need to provide family-based rare-variant association tests, incorporating functional annotations through active noncoding regions.

Hence, in this thesis, we are firstly proposing Cis-Regulatory Hubs (CRHs), a new 3D-based model integrating contacts between genes and enhancers within networks of complex interactions. Also, we argue that this information can be used to highlight new risk variants involved in complex diseases. Thus, we have developed a new family-based rare variant association test, called *RetroFun-RVS*, integrating 3D-based functional annotations, such as CRHs.

We have demonstrated that CRHs represent biological relevant structures to gain insights into complex disease etiology, such as schizophrenia. Moreover, *RetroFun-RVS*

incorporating CRHs as functional annotations has been shown to be powerful in detecting new risk variants. We argue that these aspects are crucial to highlighting epigenetics mechanisms involved in complex traits. Finally, by proposing an unified framework combining both family-based studies and functional annotations, progress is made in the understanding of the etiology of complex diseases.

Table des matières

Résumé	ii
Abstract	iv
Liste des Tableaux.....	x
Liste des Figures	xi
Remerciements.....	xv
Avant-propos	xvi
Projets principaux.....	xvi
Articles insérés	xvi
Contributions	xvi
Introduction: L'âge d'or de l'(épi)génétique	1
Chapitre 1 : Revue de la littérature.....	4
1.1 Vers un changement de paradigme.....	4
1.2 Épigénétique, génome non codant et enhancers.....	5
1.3 Organisation 3D du génome, méthodes de caractérisation et implication dans la régulation des gènes	7
1.4 Méthodes d'association de variants rares, une solution à la rareté des variants	11
1.5 Retour d'un intérêt pour les études familiales	12
1.6 ESSOR des annotations fonctionnelles dans les approches statistiques	14
Chapitre 2: Problématique et objectifs	15
Chapitre 3: Cis-regulatory hubs: a new 3D model of complex disease genetics with an application to schizophrenia.....	17
3.1 Résumé.....	19
3.2 Abstract.....	20
3.3 Introduction	21
3.4 Results	22

3.4.1 Promoter and distal element interactions create CRHs in neurons.....	22
3.4.2 CRHs are defined by active chromatin and the presence of schizophrenia-relevant genes	28
3.4.3 CRHs containing schizophrenia-associated genes are small and highly expressed.....	31
3.4.4 Multivariate analysis of CRH features with respect to schizophrenia-associated genes.....	31
3.4.5 CRHs are enriched in schizophrenia-associated SNPs and heritability	33
3.4.6 CRHs predict the association between schizophrenia-associated noncoding SNPs and differentially expressed genes.....	34
3.5 Discussion.....	37
3.6 Materials and Methods	39
3.6.1 Hi-C data and pre-processing	39
3.6.2 CRHs.....	40
3.6.3 ABC-Score.....	40
3.6.4 Summary statistics for schizophrenia	40
3.6.5 Schizophrenia-associated genes	41
3.6.6 Partitioning heritability for schizophrenia	41
3.6.7 Linking noncoding SNPs to DEGs in schizophrenia	41
3.6.8 3D features and other analyses	42
3.6.9 Availability of data and materials.....	42
3.7 Supplementary Information.....	42
3.8 Acknowledgements	42
3.9 Conflict of Interest Statement	43
Chapitre 4: RetroFun-RVS: a retrospective family-based framework for rare variant analysis incorporating functional annotations.....	44
4.1 Résumé.....	46

4.2 Abstract.....	47
4.3 Introduction	48
4.4 Material and Methods	50
4.4.1 Notation and Model.....	50
4.5 Numerical Simulations.....	53
4.5.1 Type I Error Simulations.....	54
4.5.2 Empirical Power Simulations.....	54
4.6 Results	57
4.6.1 Simulation of Type I Error Rate	57
4.6.2 Power Comparison Considering Different Strategies to build Functional Annotations.....	59
4.6.3 Power Comparison with Others Affected-Only Methods.....	59
4.7 Discussion.....	61
4.8 Data Availability.....	63
4.9 Funding	63
4.10 Conflict of Interest	63
 Chapitre 5 : RetroFun-RVS, un package R pour l'analyse de variants rares dans les familles, permettant l'intégration d'annotations fonctionnelles.....	64
5.1 Conception et implémentation	64
5.1.1 Type de données	64
5.1.2 Pré-traitement	65
5.1.3 Obtention des valeurs-p	65
5.2 Application aux données de fentes labiales	66
5.2.1 Présentation des données	66
5.2.2 Résultats.....	68
5.2.3 Discussion	70

5.3 Conclusion	71
Chapitre 6 : Discussion	72
6.1 Extension des pôles de régulation cis, perspectives et limites	73
6.2 Les pôles de régulation cis, une utilisation en « cartographie fine »?.....	75
6.3 Vers une plus large applicabilité de RetroFun-RVS	77
6.3.1 Extension de RetroFun-RVS, le traitement de familles consanguines	77
6.3.2 Extension de RetroFun-RVS intégrant des covariables.....	77
6.3.3 Combiner variants rares et communs.....	78
Conclusion.....	79
Bibliographie	80
Annexes.....	94
Annexes du Chapitre 3.....	94
A.1 Supplemental Figures.....	94
A.2 Methods	110
Annexes du Chapitre 4.....	117
B.1 Risk variants dominate protective variants in an affected-only design.....	117
B.2 Score Variance.....	118
B.3 Numerical Simulation.....	119
B.4 Pedigree Structures.....	120
B.5 Adaptation of RVS and RV-NPL including CRHs.....	128
B.6 Results	129
Annexes du Chapitre 5.....	140
Annexes Discussion.....	142
C.1 Extension de RetroFun-RVS permettant l'intégration de covariables.....	142
C.2 Extension de RetroFun-RVS intégrant variants rares et communs	142

Liste des Tableaux

Table 1: Running times (in seconds) for analyzing rare variants in the TAD, in one simulated replicate, using a single 2.10GHz processor	61
Table 2: Répartition des structures familiales dans les données de séquençage de fentes labiales du génome complet	66
Table 3: Number of variants located within each CRH and outside	119

Liste des Figures

Figure 1: Schéma représentant les mécanismes de régulation entre gènes et enhancers .	6
Figure 2: Cis-regulatory hubs (CRHs) are built from activity-by-contact (ABC)-Score methodology	24
Figure 3: Cis-regulatory hub (CRH) are 3D-based networks mainly constituted by distal elements and more local than high order 3D features	27
Figure 4: Cis-regulatory hubs (CRHs) are enriched in transcriptionally active elements and genes associated with schizophrenia.....	30
Figure 5: Features of schizophrenia-associated genes	32
Figure 6: Cis-regulatory hubs (CRHs) are enriched in schizophrenia-associated SNPs, schizophrenia heritability, and capture links between noncoding SNPs and genes differentially expressed in schizophrenia.....	36
Figure 7: Example of pedigree structures considered in the simulation studies.....	54
Figure 8: Overview of functional annotations considered in the simulation studies.....	56
Figure 9: Example of matrix of functional annotations when considering CRHs.	57
Figure 10: Quantile-Quantile plot of ACAT-Combined P-values for RetroFun-RVS_CRHs considering variant dependence.	58
Figure 11: Power evaluation of RetroFun-RVS under different scenarios for 2% risk variants	60
Figure 12: Example des structures familiales présentes dans les données de séquençage du génome complet de fentes labiales.....	67
Figure 13 : Répartition du nombre de paires gène-enhancer par pôle de régulation cis dans les cellules épithéliales humaines	68
Figure 14: Résultats de RetroFun-RVS dans les données de fentes labiales, retirant les familles syriennes de l'analyse.....	70
Figure 15: CRHs exhibit strong overlap between iPSC neurons and post-mortem brain tissues	94
Figure 16: Genome browser view of the CRH encompassing GRIN2A gene	95

Figure 17: Genome browser view of the CRH encompassing GRM3 gene	96
Figure 18: Genome browser view of the CRH encompassing GRIA1 gene.....	97
Figure 19: CRHs in iPSC-derived neurons show “average behavior” compared to post-mortem tissues regarding number of elements	98
Figure 20: In post-mortem tissues, CRHs are mainly composed by distal elements.....	99
Figure 21: Promoters are more connected than distal elements across post-mortem tissues	100
Figure 22: Promoters are more connected than distal elements in DNase-based and Rao methods.....	101
Figure 23: Promoters are inside more complex relationships than distal elements in post-mortem brain tissues.....	102
Figure 24: CRHs mainly overlap active compartments across post-mortem tissues.....	103
Figure 25: In most cases, CRHs overlap one TAD across post-mortem brain tissues....	104
Figure 26: In most cases, CRHs overlap one TAD across our methods.....	105
Figure 27: Enrichment in FIREs for distal elements and promoters in our methods	106
Figure 28: Enrichment in Encode candidate elements for distal elements and promoters in our control methods	107
Figure 29: CRHs in DNase-based and Rao methods show strong enrichments in schizophrenia-associated SNPs	108
Figure 30: Schizophrenia heritability enrichments for DNase-based and Rao methods .	109
Figure 31: Negative association between Intraclass correlation and the number of promoters considered within CRHs for the ABC-Score method	110
Figure 32: (A) Rao-Based method methodology and CRH building. (B) DNase-based method methodology and CRH building.....	114
Figure 33: Expected contribution of a family to the score statistic value for different effect sizes considering either 1 risk variant (RR) or 1 protective variant (1/RR)	118
Figure 34: Pedigree structures for all the 52 families considered in the simulation studies	127

Figure 35: Adaptation of RVS including CRHs	128
Figure 36: Quantile-Quantile plot of ACAT-Combined p-values for RetroFun-RVS_CRHs considering variant independence.....	129
Figure 37: Quantile-Quantile plots of p-values for RetroFun-RVS incorporating no functional annotation.....	130
Figure 38: Quantile-Quantile plots of p-values for RetroFun-RVS_CRHs incorporating the first CRH	131
Figure 39: Quantile-Quantile plots of p-values for RetroFun-RVS_CRHs incorporating the second CRH	131
Figure 40: Quantile-Quantile plots of p-values for RetroFun-RVS_CRHs incorporating the fourth CRH.....	132
Figure 41: Quantile-Quantile plots of ACAT-Combined p-values for RetroFun-RVS_CRHs considering only small pedigrees	133
Figure 42: Quantile-Quantile plots of ACAT-Combined p-values considering variant dependence	134
Figure 43: Quantile-Quantile plots of ACAT-Combined bootstrap-based p-values considering variant dependence	134
Figure 44: Power evaluation of RetroFun-RVS under different scenarios for 2% risk variants considering only small pedigrees.....	135
Figure 45: Power evaluation of RetroFun-RVS under different scenarios for 1% risk variants	136
Figure 46: Power evaluation of ACAT-Combined p-values under different scenarios for 1% risk variants for RetroFun-RVS_CRHs (CRHs), RetroFun-RVS_Pairs (G-E Pairs), RetroFun-RVS_Genes (Genes), and RetroFun-RVS_Sliding–Window (Sliding)	137
Figure 47: Power at 75% risk variants within one CRH between RetroFun-RVS_CRHs and other affected-only competing methods for 1% causal variant	138
Figure 48: Relationship between average genotype values by family and Burden Original statistic at 1% causal	139
Figure 49: Diagramme Quantiles-Quantiles du -log10 des valeurs-p lorsque l'ensemble des familles sont considérées	140

Figure 50: Diagramme Quantiles-Quantiles du -log10 des valeurs-p pour le sous-ensemble des familles syriennes pour lesquelles on observe de la consanguinité. 141

Remerciements

Je tenais tout d'abord à remercier le Pr. Alexandre Bureau, pour m'avoir donné l'opportunité de poursuivre un doctorat et de m'avoir accompagné pendant cette grande étape de ma vie universitaire. Il a su m'initier aux domaines de la statistique génétique et de l'épigénétique. Il a aussi su me transmettre son goût de la rigueur méthodologique et son expertise du domaine. Ce fût riche d'enseignements

Je tiens également à remercier le Pr. Arnaud Droit, pour m'avoir accueilli dans son équipe, donné l'opportunité de collaborer sur différents projets et permis de rencontrer des gens formidables.

Merci au Pr. Steve Bilodeau d'avoir su me prodiguer de précieux conseils et d'avoir fait preuve de patience. Nos discussions ont été agréables et enrichissantes.

Merci au Dr. Charles Joly-Beauparlant, pour m'avoir accompagné dans la réalisation de mon premier article. Il a su m'apporter sa science de l'épigénétique et de précieux conseils, pour le novice que je suis.

Un grand remerciement à Anne-Julie Boucher, pour avoir accepté de m'aider et pris le temps de lire cette thèse. Tes conseils et commentaires ont permis de rendre ce travail meilleur.

Évidemment tout ceci n'aurait pas été possible sans mon père, ma mère ainsi que mon frère. Leur soutien a été sans faille et a pu illuminer des passages parfois difficiles.

Enfin, je remercie ma conjointe Manon pour avoir su trouver les mots justes, avoir toujours cru en moi et permis de me sortir de situations cornéliennes. Tu as toute ma reconnaissance pour cela, je te dédis en partie cette thèse.

Avant-propos

Projets principaux

Ce document est une thèse avec insertion d'articles. Le chapitre 3 est la retranscription d'un article publié dans la revue *Life Science Alliance*. Le chapitre 4 est un article soumis, en cours de révision.

Cette thèse présente mes travaux de doctorat dont l'objectif central était le développement d'approches statistiques permettant une meilleure compréhension des mécanismes biologiques impliqués dans les maladies. Cette thèse se structure essentiellement autour de deux axes : (1) le développement de nouvelles méthodes permettant d'intégrer l'information épigénétique et (2) l'incorporation de l'information précédemment obtenue dans des modèles statistiques, dans une perspective de mieux comprendre l'étiologie des maladies complexes.

Articles insérés

Les articles insérés sont les suivants :

- ***Cis-regulatory hubs: a new 3D model of complex disease genetics with an application to schizophrenia.*** Mangnier et al. publié dans la revue *Life Science Alliance* en janvier 2022.
- ***RetroFun-RVS: a retrospective family-based framework for rare-variant analysis incorporating functional annotations.*** Mangnier & Bureau. L'article a été soumis à la revue *Genetics* mais disponible sur la plateforme *bioRxiv* depuis juin 2022.

Contributions

Contributions à l'article ***Cis-regulatory hubs: a new 3D model of complex disease genetics with an application to schizophrenia.***

LM a effectué les analyses et écrit le manuscrit, CJB a validé les analyses et écrit le manuscrit. AD a supervisé et écrit le manuscrit. SB a validé, participé à la méthodologie et écrit le manuscrit. AB a supervisé, effectué les analyses et écrit le manuscrit. Tous les auteurs ont lu et approuvé la version soumise.

Contributions à l'article ***RetroFun-RVS: a retrospective family-based framework for rare-variant analysis incorporating functional annotations.***

LM et AB ont réalisé les analyses et écrit le manuscrit. Tous les auteurs ont lu et approuvé la version soumise.

Introduction: L'âge d'or de l'(épi)génétique

Le 27 mai 2021, le consortium Telomere-to-Telomere, une collaboration de plus de 30 institutions, a publié la dernière version du génome humain, mettant en lumière 115 nouveaux gènes, pour un total de 19 969 gènes répertoriés à l'heure actuelle (Nurk et al., 2022). Cette initiative n'est que le plus récent exemple en date des efforts fournis durant ces vingt dernières années sur la compréhension du génome humain. En effet, un travail considérable a été effectué aussi bien sur la génération des données que sur le développement d'outils analytiques innovants, permettant d'adresser des problématiques jusqu'ici non résolues. Par exemple, la réduction des coûts de séquençage du génome humain, passés de 1 000 000 \$ en 2001 à 1 000 \$ en 2021 (The Cost of Sequencing a Human Genome, NIH), a mené à une démocratisation des données génétiques et phénotypiques, ouvrant de nouveaux champs applicatifs, notamment dans l'étude des pathologies. Les efforts mis ces dernières années ne sont plus seulement orientés vers le traitement clinique des maladies, mais également vers la compréhension des mécanismes biologiques causaux, visant un impact positif sur la santé des populations. Puisque la quasi-totalité des maladies chez l'humain est influencée sur le plan individuel par des variations génétiques héréditaires (Jackson et al., 2018; Claussnitzer et al., 2020), il devient alors crucial de mieux évaluer le rôle joué par certaines mutations génétiques dans l'apparition ou le développement de maladies chez l'Homme. Par ailleurs, des avancées technologiques et méthodologiques récentes ont permis de révéler le rôle déterminant de l'épigénétique dans l'expression des maladies (Moosavi & Ardekani, 2016). De manière assez générale, l'épigénétique peut être définie comme l'étude de l'impact des phénomènes environnementaux sur l'ADN et des changements dans l'expression de traits qui en découlent. Cependant, cette définition inclut aussi l'étude et la caractérisation du génome non codant. Les enhancers ont alors reçu un intérêt particulier, du fait que ces courts segments d'ADN occupent une place centrale dans le programme de régulation des gènes, amplifiant leur transcription, les impliquant donc possiblement dans l'émergence de maladies chez l'humain. Le développement croissant de la médecine de précision, visant à attribuer le bon traitement, au bon dosage au bon moment au bon individu, pousse alors les chercheurs à identifier les chaînes causales entre mutations génétiques et maladies, tenant

compte de la dimension épigénétique. Des initiatives récentes tentent donc d'intégrer ces notions dans des outils unifiés de traitement (Kolpakov et al., 2011; Genome Enhancer; geneXplain). Il devient alors critique de mieux cerner les mécanismes biologiques impliqués dans la variation phénotypique. Cette thèse vise par conséquent à proposer et à développer des méthodes statistiques et bio-informatiques pour mettre en lumière et interpréter ces mécanismes, notamment dans l'émergence de maladies complexes.

Ainsi, cette thèse sera divisée en six chapitres majeurs.

- Le premier chapitre permettra de définir les concepts fondamentaux nécessaires à la compréhension de cette thèse. Y sera établie une revue de la littérature, pertinente vis-à-vis des notions mobilisées. Nous y présenterons : l'évolution des paradigmes en génétique, passant des études familiales aux études populationnelles; les travaux récents réalisés autour de la caractérisation du génome non codant et de l'organisation 3D, ainsi que son rôle dans les phénomènes de régulation; les défis méthodologiques posés par le rôle joué par les variants rares dans les maladies complexes; l'adaptation des méthodes d'analyses de variants rares aux devis familiaux; et enfin l'utilisation en pratique des annotations fonctionnelles dans les approches à l'échelle populationnelle.
- Les objectifs de la thèse, en lien avec l'état de la connaissance et les questions scientifiques qui en découlent, seront présentés dans le deuxième chapitre.
- Dans le troisième chapitre de cette thèse, nous présenterons les pôles de régulation cis, un nouveau modèle de réseaux permettant l'intégration des contacts 3D entre gènes et enhancers. En effet, il n'existe pas, à l'heure actuelle, de méthode consensuelle sur la détermination des enhancers. Par conséquent, trouver une méthode fiable pour définir des régions non codantes comme enhancers, les lier à leur(s) gène(s) cible(s) et subséquemment créer des réseaux n'assure aucunement que ces derniers sont biologiquement valides et pertinents dans un contexte de maladie. Sur cette base, la méthode proposée sera évaluée selon trois prismes différents, soit la cohérence au regard des structures 3D connues, la validité biologique et la pertinence dans l'étiologie de maladie. Nous avons sélectionné, pour le dernier volet, la schizophrénie pour démontrer l'intérêt de la méthode proposée.

- Le quatrième chapitre sera quant à lui orienté vers le développement d'une nouvelle méthode statistique intégrant l'information familiale tout en permettant l'intégration d'informations fonctionnelles. Bien que la méthode se veuille applicable à tout type de scores (continus ou binaires), l'évaluation de la performance au regard du contrôle de l'erreur de type 1 et de la puissance sera réalisée par études de simulation, en permettant l'intégration des pôles de régulation cis en tant qu'annotation. En tenant compte à la fois de la structure familiale et de l'intégration de scores fonctionnels, l'objectif est alors de mettre en lumière de nouveaux variants causaux impliqués dans des maladies complexes.
- Dans le cinquième chapitre de cette thèse, nous présenterons *RetroFun-RVS*, l'outil permettant la généralisation et la diffusion à un large public de la méthode présentée dans le chapitre quatre, pour analyser des données familiales, intégrant des scores fonctionnels arbitraires. Nous illustrerons la méthode grâce aux données familiales de fentes labiales.
- Pour sa part, le dernier chapitre servira de base à la discussion des résultats mis en lumière dans les sections précédentes. Ce chapitre sera l'occasion de replacer cette thèse au sein du domaine dans lequel elle opère, nous y jaugerons son apport potentiel, ses limites ainsi que les travaux futurs nécessaires à son applicabilité à d'autres questions scientifiques. La conclusion de cette thèse sera l'occasion de mettre en avant les grands résultats ainsi que les leçons tirées du travail réalisé.

Chapitre 1 : Revue de la littérature

1.1 Vers un changement de paradigme

Traditionnellement, les chercheurs se sont surtout intéressés aux maladies mendéliennes : maladies rares, caractérisées par la présence de mutations génétiques, fortement pénétrantes, au niveau d'un seul locus. Les gènes causaux étaient identifiés par analyse de liaison au sein d'études familiales, alors que les allèles causaux étaient validés à postériori dans des études expérimentales. Bien que fastidieuses, ces approches ont permis de mettre en lumière jusqu'en 2000 un septième des gènes impliqués dans les maladies monogéniques (Claussnitzer et al., 2020). À titre d'exemple, ces approches ont permis des avancées majeures dans la compréhension de la maladie d'Huntington (Bates, 2005). Cependant, le développement des méthodes de séquençage haut débit a permis de détecter un large spectre de variations génétiques causales associées avec des expressions phénotypiques diverses dans la population générale. La transition entre des études caractérisées par un nombre restreint d'individus avec des phénotypes homogènes à des études populationnelles définies par une hétérogénéité des phénotypes a permis de révéler des mécanismes causaux jusqu'ici non connus. Notamment avec la possibilité que des mutations génétiques communes ou rares soient impliquées pour une même maladie. Par ailleurs, ce basculement de paradigme technologique et philosophique a été l'occasion d'étudier efficacement les maladies complexes caractérisées par leur structure polygénique, dont l'impact réel des variants sur le trait n'est jamais clairement identifié (Crouch & Bodmer, 2020). En effet, à la différence des troubles mendéliens, les maladies complexes sont souvent caractérisées par la combinaison de variants communs et rares, où l'environnement joue un rôle prépondérant. Dans cette logique, les études d'association pangénomiques ont permis de détecter un nombre important de variants impliqués dans un large éventail de maladies, notamment dans des troubles complexes (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). Cet intérêt croissant pour les études populationnelles a par ailleurs permis le développement des approches de score de risque polygénique ou de score de déséquilibre de liaison, méthodes cherchant respectivement à prédire le risque génétique individuel ou l'héritabilité expliquée par certaines variations génétiques fréquentes (Dudbridge et al., 2013; Bulik-Sullivan et al., 2015). Cependant, du fait de la structure hétérogène de la plupart des maladies, les variants fréquents détectés n'expliquent qu'une très faible part de l'héritabilité des maladies. En effet, il a été démontré

que, dans la majeure partie des troubles, la combinaison de variations génétiques rares situées dans des régions non codantes explique une large proportion de l'héritabilité chez l'humain (Zhang & Lupski, 2015), menant de plus en plus les chercheurs à développer de nouvelles approches, visant à caractériser le génome non codant et tenant compte de la rareté des variants.

1.2 Épigénétique, génome non codant et enhancers

Le développement des méthodes de séquençages du génome complet a permis l'identification de variants situés dans des régions régulatrices impliqués dans des maladies. De manière générale, ces régions sont dites non codantes, c'est-à-dire que bien que parfois transcrrites elles ne sont pas à l'origine de production de protéines. Par exemple, la transcription des enhancers n'encode pas pour la création de protéines associées. Ces séquences jouent cependant un rôle central dans la régulation des gènes (Pennachio et al., 2013). En effet, des variants présents dans ces régions régulatrices peuvent mener à l'apparition de maladie chez l'Homme, notamment la schizophrénie (Roussos et al., 2014). Par conséquent, des efforts considérables ont été mis dans la caractérisation des mécanismes épigénétiques. Par exemple, la présence de variants délétères dans les enhancers a été démontrée comme jouant un rôle central dans l'émergence de maladies chez l'Homme, suscitant un vif intérêt pour ces mêmes éléments de régulation; poussant les chercheurs à caractériser les enhancers selon leur profil biologique ou fonctionnel (8 Enhancer discovery and characterization. Nature, 2019). On est alors en droit de se poser la question : « *par quels mécanismes un enhancer influe-t-il sur la transcription d'un gène ?* » Dans ce contexte, le modèle le plus couramment employé est celui où, par le biais de facteurs de transcriptions spécifiques, le enhancer, du fait du repliement de la chromatine, interagit avec le promoter cible. Ce modèle sous-tend que les enhancers mettent en lumière un profil épigénétique spécifique à leur activité et au tissu dans lequel ils opèrent, incluant marques d'histones (H3K27ac et H3K4me1 pour les enhancers actifs, par exemple), facteurs de transcriptions (p300), accessibilité de la chromatine et contacts 3D (Figure 1). À travers ce type d'informations, des initiatives comme ENCODE ou le Roadmap Epigenomics Consortium ont permis la caractérisation de millions de régions candidates dans plusieurs tissus chez l'Homme ou la souris. Certains auteurs ont par ailleurs proposé de combiner le profil épigénétique de certaines régions du génome avec la présence de contacts 3D avec un gène pour définir des enhancers comme actifs. Ainsi, Fulco et al. (2019) ont proposé le score d'activité par contact, une métrique unifiée intégrant accessibilité de la

chromatine, présence d'H3K27ac et contacts 3D, permettant de discriminer les enhancers actifs des enhancers inactifs en tenant compte de l'activité environnante. Cette méthode validée expérimentalement par des approches de perturbations basées sur la technologie CRISPR-CAS9 a été établie comme pertinente pour lier les variations génétiques non codantes à leurs gènes cibles par exemple (Nasser et al., 2021). En parallèle, des études récentes, par le biais de nouvelles technologies de séquençage, ont rendu possible la détection d'enhancers selon leur profil d'expression (Yao et al., 2022). Ces approches ont été démontrées plus fiables quant à la caractérisation des enhancers. Enfin, la démocratisation du CRISPR-CAS9 a donné lieu à la validation fonctionnelle *in vivo* de certaines régions comme enhancers (Li et al., 2020). L'insertion de nouveaux variants, la suppression ou l'ajout de certains facteurs de transcription au niveau des enhancers ont mis en lumière l'impact de telles modifications sur la régulation des gènes. Bien que représentant une approche étalon, ces méthodes, du fait de limitations techniques majeures, n'ont pu être étendues à l'échelle du génome chez l'Homme (Gasperini et al., 2020). Comme nous l'avons vu jusqu'ici, des avancées considérables ont été réalisées ces dernières années sur la caractérisation des enhancers, cependant l'organisation 3D du génome joue un rôle central dans les phénomènes de régulation, par quels mécanismes et à quelle échelle ?

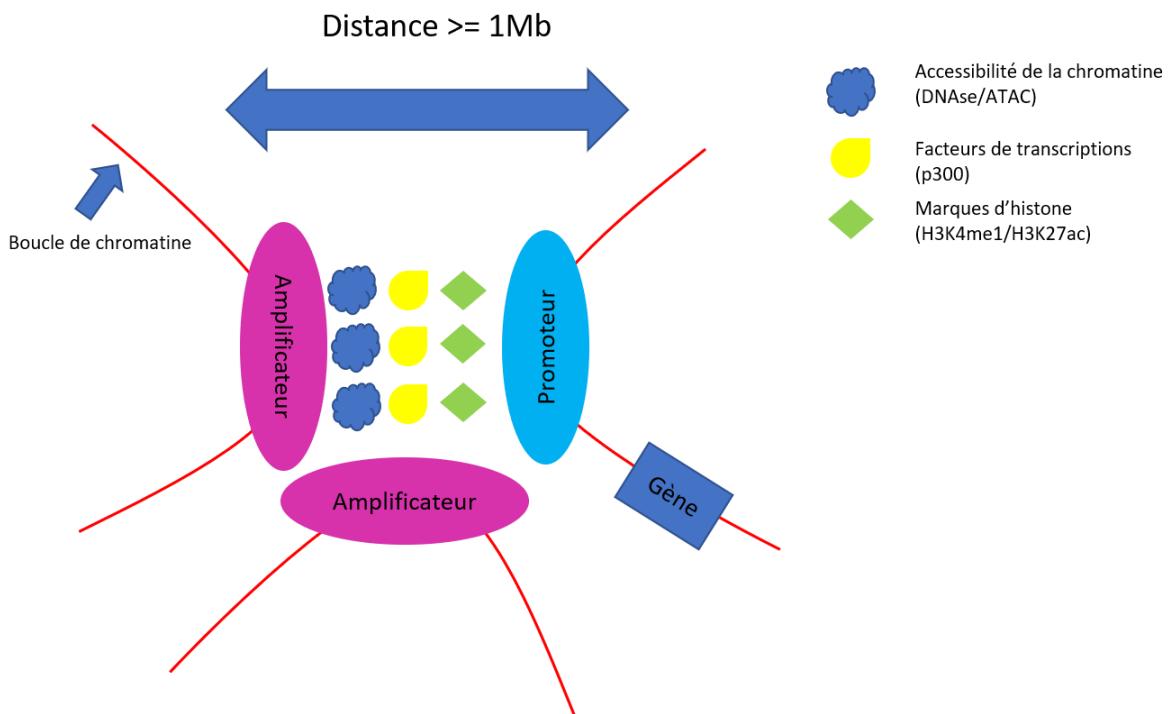


Figure 1: Schéma représentant les mécanismes de régulation entre gènes et enhancers

1.3 Organisation 3D du génome, méthodes de caractérisation et implication dans la régulation des gènes

La démocratisation des technologies de capture de conformation de la chromatine (3C) a permis de mettre en lumière le génome comme une organisation 3D, où, du fait du repliement de la chromatine, gènes et enhancers peuvent être en contacts physiques, même situés à plusieurs mégabases de distance (Kadouke & Blobel, 2009; de Wit & de Laat, 2012). Les liens physiques entre gènes et enhancers dans une structure 3D sont le fondement de la régulation des gènes. Cependant, avant toute chose, il convient de faire un tour d'horizon des technologies de séquençage et des outils permettant la capture et l'analyse des interactions physiques entre régions génomiques.

À ce jour plusieurs grandes familles de méthodes permettent de capturer le repliement de la chromatine, ces approches sont identifiées comme méthodes de capture de conformation de la chromatine. Initialement proposées par Dekker et al., 2002, elles ont leurs équivalents exploitant les technologies de séquençage nouvelle génération (Hi-C) (van Berkum et al., 2010). Dans cette même famille sont à noter, les approches de Capture de Conformation de la Chromatine par Carte (4C) (Simonis et al., 2006), et les méthodes de Capture de Conformation de la Chromatine par Copie de Carbone (5C) (Dostie et al., 2006). Bien que présentant des spécificités techniques propres à chacune, ces méthodes suivent un ensemble d'étapes communes, à savoir :

- Les contacts des brins d'ADN sont capturés en figeant la chromatine dans sa conformation. Ceci est rendu possible par l'emploi dans la plupart des cas de formaldéhyde.
- La présence de sites de restriction va permettre d'isoler par digestion enzymatique les segments en contacts 3D.
- Les fragments d'ADN vont ensuite être ligaturés pour permettre le séquençage.
- Finalement selon l'approche choisie, les fragments d'ADN fragmentés vont être séquencés après amplification PCR (spécifique à un locus) ou sonification (à l'échelle du génome).

Ces étapes permettent ainsi la capture à différentes échelles des contacts physiques entre régions génomiques. Cependant, certaines distinctions techniques entre les différentes stratégies sont à relever.

Historiquement les approches du type 3C ont été les premières à rendre possible la capture des contacts physiques entre régions génomiques. Elles sont qualifiées de stratégies « un contre un », c'est-à-dire que par la nature même de leur protocole elles ne sont capables de capturer efficacement que les contacts entre une seule paire d'éléments à la fois, et ce à des distances modérées (Dekker et al., 2002); supposant une connaissance à priori des régions en interaction. Cependant, lorsque l'on cherche à capturer les interactions physiques à l'échelle du génome, les approches du type 3C, ne sont pas adaptées.

L'amplification par réaction en chaîne par polymérase utilisée dans les approches 3C peuvent mener à certains biais techniques, nécessitant l'emploi de devis expérimentaux précis et rendant l'interprétation des données difficile (Dekker 2006; Simonis et al., 2006). Pour adresser ces limitations, d'autres méthodes ont été proposées. Ainsi, les méthodes du type 4C (et ses dérivées) étendent les approches 3C en exploitant les micropuces (Simonis et al., 2006). Elles permettent ainsi de révéler des régions en interactions 3D avec tous les loci possibles (stratégie « un contre tous »). À la différence des approches 3C, les méthodes 4C ne nécessitent alors aucune connaissance à priori sur les régions entre contact avec le locus d'intérêt. Cependant, de par la nature même du protocole (approche circulaire) (Zhao et al., 2006), la méthode tend à manquer les contacts opérant à des niveaux plus locaux (<50 Kb). Sur ce point et à l'instar du 3C, les technologies 4C-Seq proposent d'étendre le 4C exploitant les méthodes de séquençage nouvelle génération (Splinter et al., 2011).

Dans une logique similaire, van Berkum et al. (2010) ont proposé le Hi-C, une extension du 3C exploitant les technologies de séquençage nouvelle génération, permettant de capturer des contacts entre régions génomiques et ce à l'échelle du génome. D'un point de vue conceptuel, à la différence du 3C, la méthode Hi-C capture l'ensemble des interactions entre toutes les régions génomiques. On parle ainsi de stratégie « tous contre tous ».

Les approches du type 5C combinent méthodes de séquençage haut débit 3C avec des approches de quantification de l'ADN. L'avantage majeur est alors de pouvoir mettre en lumière des réseaux d'interactions physiques entre éléments génomiques (Dostie et al., 2006). À la différence des approches 4C qui se concentrent sur une région génomique

donnée, les approches 5C permettent de reconstruire des structures 3D entre plusieurs paires de sites. Toutefois, lorsque la reconstruction de réseaux d’interactions entre éléments génomiques est d’intérêt, ces méthodes requièrent de l’information à priori notamment sur la localisation des régions régulatrices.

Dans un même ordre d’idée, il peut être pertinent de capturer les contacts physiques entre régions mettant en lumière des profils épigénétiques précis. Ainsi les approches du type ChiA-PET (Fullwood & Ruan, 2009), dont font parties les technologies HiChiP se proposent de détecter l’ensemble des interactions physiques restreintes par la présence de protéines spécifiques. Par exemple, le HiChiP combine méthodes Hi-C et ChiP-Seq pour capturer les interactions entre régions présentant certaines marques d’histones (Munbach et al., 2016).

Enfin, parce que les méthodes présentées jusqu’ici sont des approches du type « en vrac », c’est-à-dire qu’elles agrègent les contacts entre régions opérant dans des cellules différentes. Ainsi, certaines interactions détectées peuvent résulter de contacts « artefact » opérant dans des cellules différentes. Pour pallier cette limite biologique majeure, des approches capturant les interactions entre régions génomiques opérant au niveau d’une même cellule ont été proposées. Ces approches font références aux méthodes du type « une seule cellule » (ou « Single-Cell » en anglais). Le « Single-Cell » Hi-C permet alors de distinguer les contacts entre régions génomiques opérant dans des cellules différentes (Nagano et al., 2013) et ainsi avoir une vision plus claire des mécanismes biologiques au niveau cellulaire.

Parce que le développement méthodologique et les modèles présentés dans cette thèse reposent essentiellement sur les données Hi-C, nous focaliserons dès à présent notre attention sur les outils permettant d’extraire l’information pertinente de ce type de données en vue d’analyses subséquentes. Sur ce point les outils développés tentent de répondre à la question « Quels contacts sont biologiquement valides ? ». Nous présenterons ici deux écosystèmes, Juicer (Durand et al., 2016) et HiC-Pro (Servant et al., 2015) permettant notamment la détection de paires d’interaction valides et des structures 3D de plus grande envergure. Premièrement, Juicer propose d’implémenter l’approche initialement proposée par Rao et al., 2014. Brièvement, la méthode définit les points de contacts enrichis (significatifs en termes statistiques) comme tout contact entre régions génomiques présentant une fréquence d’interaction plus importante que les valeurs attendues des régions environnantes appartenant au même domaine. De son côté, HiC-Pro opte plutôt

pour une approche biologique, basée sur la localisation des sites de restriction et de la distance entre loci pour déterminer les paires d'interaction valides.

À des échelles plus importantes, il a été prouvé que le génome s'organise autour de différentes structures 3D jouant un rôle déterminant dans les phénomènes de régulation et donc dans l'émergence ou le développement de maladies (Anania & Lupianez, 2020). D'une part, les compartiments actifs sont des structures 3D de grande envergure spécifiques au tissu pouvant couvrir plusieurs mégabases. Les approches traditionnellement employées pour la détection des compartiments actifs reposent sur des analyses en composantes principales (ACP) (Pearson, 1901), appliquées sur la matrice de contacts normalisée. L'idée générale est alors de résumer chaque paire possible de régions génomiques à travers ses composantes principales. Le signe de la valeur de la première composante discrimine les compartiments actifs des compartiments inactifs. Il a d'ailleurs été démontré que les compartiments actifs sont corrélés positivement avec des niveaux d'accessibilité de la chromatine, d'expression et de certaines marques d'histones caractérisant les éléments de régulation actifs (Lieberman-Aiden et al., 2009). Des extensions combinant Hi-C et données épigénétiques ont été proposées afin d'affiner la détection des compartiments (Fortin & Hansen, 2015).

D'autre part, les domaines topologiquement associant (topologically associating domains [TAD]) quant à eux sont des régions caractérisées par une fréquence de contacts intra-TAD plus élevée que la fréquence inter-TADs (Dixon et al., 2012). En d'autres termes, étant donné la présence de facteurs de transcription CTCF aux bordures de TADs, les interactions entre éléments d'un même TAD sont favorisées au détriment d'interactions entre éléments de TADs différents (Pombo & Dillon, 2015), ceci privilégiant les phénomènes de régulation entre gènes et enhancers au sein d'un même domaine. Les TADs peuvent être détectés en exploitant plusieurs stratégies. Parmi les plus utilisées en pratique, nous pouvons citer « l'indice de directivité » (Dixon et al., 2012), le « score d'insulation » (Crane et al., 2015) et Arrowhead (Rao et al., 2014; Durand et al., 2016). Ainsi, Dixon et al., 2012 ont initialement proposé une méthode appelée « indice de directivité ». Intuitivement, la méthode propose d'estimer, sur la base d'un modèle de chaînes de Markov cachées, le biais d'interactions « amont » ou « aval » pour définir les bordures de domaines. C'est-à-dire que la présence de biais « aval » définit le début d'un TAD, alors que des biais « amont » la fin du domaine. Le « score d'insulation » quant à lui est une métrique moyennant le nombre de contacts entre une région et l'ensemble des régions l'entourant au

niveau d'une fenêtre génomique donnée (Crane et al., 2015). Il est attendu que les frontières de TADs soient définies par des valeurs de scores d'insulation minimales. Enfin, Arrowhead exploite une stratégie un peu plus complexe que les deux précédentes basée sur la notion de triangles (Rao et al., 2014). Brièvement, si deux points de contacts sont à l'intérieur du même domaine alors la valeur calculée par Arrowhead est approximativement zéro, positive lorsque le point supérieur est en dehors du domaine et négative sinon. Pour une comparaison systématique des méthodes, nous référons le lecteur à Dali & Blanchette, 2017 et Zufferey et al., 2018. De plus, la présence de variations génétiques dans les bordures de TADs peut mener à la fusion de plusieurs domaines (Melo et al., 2020), à l'origine de contacts ectopiques entraînant des phénomènes de détournements d'enhancers, impliqués dans les maladies (Fudenberg & Pollard, 2019). L'organisation 3D du génome ouvre alors de nouvelles voies pour lier variations génétiques situées dans des régions régulatrices à leurs gènes cibles (Nasser et al., 2021). Cependant, comment détecter de tels variants causaux lorsque ces derniers sont rares dans la population générale ?

1.4 Méthodes d'association de variants rares, une solution à la rareté des variants

À la différence des méthodes de liaison où l'on teste la ségrégation d'un variant au sein d'un locus avec une maladie, les méthodes d'association testent la relation statistique entre un génotype et un phénotype d'intérêt. La rareté des variants fait en sorte que les méthodes traditionnelles, testant les variants de manière individuelle, performent peu du fait du faible signal dans la population générale (Madsen & Browning, 2009). Pour adresser cette limitation ont été proposées plusieurs méthodes, appelées test d'association de variants rares, souvent exprimés à travers des modèles de régression pour des traits continus ou dichotomiques. En présence de variants rares, ces méthodes d'association agrègent les variants dans une région pour maximiser le signal et ainsi augmenter la puissance de détection de variants causaux (Li & Leal, 2008; Madsen & Browning, 2009). L'agrégation pouvant se faire à l'échelle du gène, d'un ensemble de gènes ou d'éléments fonctionnellement proches. La plupart des stratégies sont alors basées sur la combinaison des allèles à travers plusieurs variants dans une statistique unifiée; pondérant ou non les variants, dans une optique de priorisation. En permettant un poids plus important pour les mutations rares chez les individus sains et, à contrario, un poids plus faible aux variants

fortement présents, une approche pondérée a été démontrée plus puissante que les démarches non pondérées (Madsen & Browning, 2009). Par exemple, les méthodes, dites de « fardeau », en réduisant le nombre de loci à tester, augmentent la puissance statistique associée. Cependant, ces approches reposent sur une hypothèse fondamentale d'homogénéité de sens des variants. En effet, il est attendu que l'ensemble des variants d'une région est délétère pour le trait en question. La présence de variants à la fois protecteurs et néfastes mène alors à une perte de puissance statistique (Lee et al., 2014). D'autres types de modèles statistiques, appelés méthodes de « noyaux », ont été développés pour adresser cette limitation majeure (Ionita-Laza et al., 2011; Neale et al., 2011). Ces approches peuvent être exprimées à travers des modèles mixtes, supposant un effet fixe pour les variables d'ajustement et des effets aléatoires pour les effets génétiques (Wu et al., 2011). Enfin, des modèles unifiés ont été proposés, le SKAT-O combine par exemple la statistique du « fardeau » et du « noyau » (Lee et al., 2012). Cette approche tient sa justification par la possible présence de régions dominées par des variants seulement néfastes, lorsque d'autres régions combinent variants délétères et protecteurs. Plus récemment, Liu et al. (2019) ont proposé ACAT, une méthode générale permettant de combiner efficacement des statistiques arbitraires, sur la base d'une loi de Cauchy, tout en permettant l'ajustement pour la multiplicité des tests. Initialement proposé dans un contexte de test d'association de variants rares, l'avantage majeur de ACAT, par rapport aux autres méthodes de combinaison (méthode de Fisher ou valeur-p minimale, par exemple), est qu'elle ne nécessite ni méthodes d'échantillonnage, ni indépendance des statistiques ou de modèles explicites de corrélation. Cependant de manière assez générale, les tests d'association de variants rares requièrent un nombre souvent élevé d'individus pour obtenir suffisamment de puissance de détection, lorsqu'appliquées dans la population générale.

1.5 Retour d'un intérêt pour les études familiales

Pour adresser les limitations des tests d'association de variants rares à l'échelle populationnelle, les modèles ont été étendus à des devis familiaux. En plus de réduire l'hétérogénéité génétique induite par les devis cas-témoins d'individus non apparentés, les devis familiaux offrent des puissances plus marquées lorsque l'on s'attend à un enrichissement de variants causaux dans les familles (Laird & Lange, 2006; Li et al., 2006; Ott et al., 2011). Ainsi, certains auteurs ont proposé d'étendre les tests d'association de variants rares aux devis familiaux, exploitant des stratégies différentes : introduisant un nouveau terme aléatoire pour les familles dans des modèles mixtes ou spécifiant un modèle

de travail tenant compte de la structure familiale dans des équations d'estimation généralisées (Chen et al., 2013; Oualkacha et al., 2013; Chen et al., 2016; Wang et al., 2017). Aussi, des extensions intégrant la corrélation intervenant auprès des familles à travers des modèles de copules ont été proposées pour des traits arbitraires (Lakhal Chaieb et al., 2015). En parallèle de cela, l'information extraite des variants ségrégant avec la maladie d'intérêt peut être exploitée, donnant un regain d'intérêt pour les analyses de liaison pour des familles étendues. Des méthodes récentes reposant sur la notion d'identité par descendance (i.e., deux individus ou plus partageant le même locus hérité d'un ancêtre commun sans recombinaison) ou combinant méthodes de liaison et d'agrégation ont été développées (Sul et al., 2016; Bureau et al., 2019; Zhao et al., 2019). Ainsi Rare-variant sharing (RVS) (Bureau et al., 2019) propose de calculer la probabilité de partage d'un variant causal sur la base de l'ensemble des combinaisons possibles des individus atteints d'une même famille. Par ailleurs, RVS a l'avantage majeur de ne pas nécessiter de fréquences d'allèles, lorsque ces dernières sont difficiles à obtenir dans la population générale et peuvent mener à des inflations de faux positifs lorsque mal estimées (Bureau et al., 2019). De son côté RV-NPL (Zhao et al., 2019) repose sur la notion de partage d'haplotypes. En plus d'offrir une robustesse à la présence de génotypages manquant, RV-NPL offre des puissances plus élevées par rapport aux méthodes considérant haplotypes causaux et non causaux, tout en contrôlant efficacement l'erreur de type 1. RareIBD (Sul et al., 2016) est quant à elle une méthode hybride combinant approche basée sur l'identité par descendance et méthodes d'agrégation de type « fardeau », reposant sur l'idée que les variants causaux sont enrichis chez les individus atteints et appauvris chez les individus sains. Une des caractéristiques essentielles de ces approches est alors d'analyser, ou de se restreindre, aux individus atteints d'une même famille. Ne considérer que les individus malades offrent certains avantages pratiques. Premièrement, lorsqu'appliqués aux études familiales, les devis « cas-seulement » requièrent des tailles d'échantillons plus faibles que leurs équivalents populationnels pour atteindre des puissantes équivalentes (Li et al., 2019). Deuxièmement, ne considérer que les atteints d'une même famille permet de mesurer le partage de variants potentiellement causaux, lorsqu'individus malades contribuent plus qu'individus sains (Schaid et al., 2010). Cette notion est centrale dans une optique de mettre en lumière les mécanismes biologiques impliqués. Enfin, cela présente certains avantages pour pallier l'existence potentielle de biais de classification ou d'échantillonnage (Albert et al., 2001), pouvant survenir du fait de phénomènes de pénétrance incomplète. Cependant, il a été démontré que de ne considérer que des individus atteints tend à surestimer les

mesures d'effets, lorsque les variants sont enrichis par rapport à la population générale (Schaid et al., 2010). Par conséquent, des approches de vraisemblance rétrospective, conditionnant sur le statut des individus, ont alors été proposées dans le cas de variants communs (Schaid et al., 2010). Ces approches, en plus de permettre la dérivation de tests d'hypothèse, corrigent efficacement les mesures d'effets. Dans cette logique, Schaid et al. (2013) ont suggéré des approches tenant compte du processus d'échantillonnage dans le cadre de tests d'associations de variants rares. Toutefois, dans une optique de mettre en lumière les mécanismes causaux impliqués dans les maladies, il devient crucial d'intégrer de l'information fonctionnelle externe.

1.6 Essor des annotations fonctionnelles dans les approches statistiques

L'effort des chercheurs ne se concentre maintenant plus uniquement sur la détection des variants causaux, mais aussi sur leurs implications d'un point de vue fonctionnel dans les maladies. Au-delà de la dimension mécanistique, intégrer de l'information externe devient crucial pour mettre en lumière de nouveaux variants causaux, dans une perspective de « cartographie fine » notamment (Schaid et al., 2018). De récents modèles reposant sur des approches de tests d'association de variants rares ont donné lieu à l'intégration d'informations fonctionnelles, sous forme de scores continus ou intégrant la notion de contacts jumelés entre gènes et enhancers (He et al., 2017; Wu & Pan, 2018; Ma & Wei, 2019; Ma et al., 2021). Ces approches adaptatives introduisent alors une double priorisation des variants, sur la base d'une fonction de leur fréquence d'allèle mineur et de scores, souvent combinés, prédisant leur effet délétère (Kircher et al., 2014; Ionita-Laza et al., 2015). De plus, en permettant l'intégration du test original, ces méthodes robustes diminuent la perte de puissance lorsqu'aucune annotation fonctionnelle n'est prédictive, offrant un gain substantiel de puissance quand au moins une annotation est associée avec le trait en question (He et al., 2017). Il faut toutefois noter que ces méthodes, bien que performantes, ne se limitent jusqu'à aujourd'hui qu'aux approches populationnelles.

Chapitre 2: Problématique et objectifs

Comme nous l'avons vu jusqu'ici, l'émergence de nouveaux champs applicatifs, notamment dans la compréhension de l'implication des variations génétiques dans les maladies, a été possible grâce à l'avènement et au développement des méthodes de séquençage haut débit. Dans une optique de découverte de variants, l'utilisation d'information externe paraît essentielle, notamment pour localiser de nouvelles variations situées hors des régions codantes. Cependant, l'utilisation de l'information épigénétique s'est jusqu'à maintenant limitée à définir des éléments de régulation (Roadmap Epigenomics Consortium, 2015) ou à lier variations génétiques non codantes aux gènes cibles (Nasser et al., 2021) par la détection de paires fonctionnelles entre gènes et enhancers (Fulco et al., 2019). Bien que cette information ait été intégrée dans des modèles statistiques (Wu & Pan, 2018; Ma et al., 2021), aucun modèle de réseaux n'a encore été proposé à l'heure actuelle. Cette limite majeure ne tient pas compte de la nature des interactions entre gènes et enhancers, dont l'opération en réseaux a été démontrée chez la drosophile (Espinola et al., 2021). De plus, il a été démontré que le profil d'interaction entre éléments de régulation et gènes était central dans la capacité de ces derniers à faire face aux stimuli environnementaux (Bergman et al., 2022; Tsai et al., 2019). Par conséquent, dans une optique de déterminer les variants génétiques impliqués dans une maladie, proposer un modèle de réseaux, pouvant lier gènes et éléments de régulation de manière indirecte, sur la base des contacts 3D, paraît pertinent. En outre, l'évaluation de la performance des modèles statistiques suggérant d'intégrer de l'information à travers des scores fonctionnels ne s'est restreinte qu'à des scores continus (CADD, Eigen) et à des devis cas-témoins d'individus non apparentés. Aussi, bien souvent, ces méthodes ne presupposent pas la capacité fonctionnelle des régions et testent de manière successive des fenêtres de taille fixe ou variable (Li et al., 2019). Ces stratégies ont deux limites pratiques majeures, soit leur incapacité à cibler directement les mécanismes biologiques impliqués et la nécessité de tester plusieurs milliers d'individus pour obtenir la puissance statistique suffisante. Par ailleurs, les approches familiales, bien que puissantes, souffrent de temps de calcul parfois longs, pouvant limiter leur application à des études à l'échelle du génome. Fort de ce constat, dans l'objectif de mettre en lumière de nouveaux variants impliqués dans les maladies complexes et les mécanismes de régulation associés, proposer un test statistique, computationnellement efficient, unifiant

approches familiales et capacité d'intégrer des annotations fonctionnelles, principalement sous forme de régions, paraît central.

À la lumière des éléments présentés précédemment, nous structurerons cette thèse autour de deux objectifs:

- Développer un nouveau modèle 3D valide de réseaux d'interactions entre gènes et enhancers;
- Proposer un cadre statistique permettant l'intégration de l'information familiale et fonctionnelle, notamment sous forme de régions disjointes correspondant à des réseaux de régulation actifs, afin de mettre en lumière de nouvelles variations génétiques rares impliquées dans les maladies.

Chapitre 3: Cis-regulatory hubs: a new 3D model of complex disease genetics with an application to schizophrenia

Dans ce premier chapitre, nous proposons les pôles de régulation cis, un nouveau modèle 3D mettant en interaction gènes et enhancers actifs. Cette méthode a été validée au regard des autres structures 3D connues ou phénomènes biologiques. Ceci dans l'optique de définir des structures biologiquement valides afin de mettre en lumière les mécanismes de régulation impliqués dans les maladies complexes. Pour illustrer la pertinence des pôles de régulation cis dans l'étiologie de maladies complexes, nous avons sélectionné la schizophrénie. La méthode a montré des enrichissements plus importants dans l'héritabilité de la schizophrénie par rapport à des structures équivalentes. Aussi, nous avons démontré que les pôles de régulation cis en formant des organisations intermédiaires étaient plus efficaces pour lier variations génétiques non codantes aux gènes différentiellement exprimés dans la schizophrénie par rapport aux paires gène-enhancer ou aux domaines topologiquement associant.

Journal

Cet article a été publié dans la revue Life Science Alliance le 5 janvier 2022.

L'article est en accès libre distribué selon les termes de la licence Creative Commons Attribution License (CC BY). L'éditeur autorise son utilisation et sa diffusion.

Accessibilité

Mangnier L, Joly-Beauparlant C, Droit A, Bilodeau S, Bureau A. Cis-regulatory hubs: a new 3D model of complex disease genetics with an application to schizophrenia. *Life Sci Alliance*. 2022 Jan 27;5(5): e202101156. doi: 10.26508/lsa.202101156. PMID: 35086934; PMCID: PMC8807870.

Liste des auteurs

Mangnier Loïc ^{1,2,3,4}, Joly-Beauparlant Charles ^{4,5}, Droit Arnaud ^{3,4,5}, Bilodeau Steve ^{6,7,8}, Bureau Alexandre ^{1,2,3}

- ¹Centre de Recherche CERVO, Quebec City, Canada.
- ²Département de Médecine Sociale et Préventive, Université Laval, Quebec City, Canada.
- ³Centre de Recherche en données Massives de l'Université Laval, Quebec City, Canada.
- ⁴Centre de Recherche du Centre Hospitalier Universitaire de Québec - Université Laval, Quebec City, Canada.
- ⁵Département de Médecine Moléculaire, Université Laval, Quebec City, Canada.
- ⁶Centre de Recherche en données Massives de l'Université Laval, Quebec City, Canada steve.bilodeau@crchudequebec.ulaval.ca.
- ⁷Centre de recherche du Centre Hospitalier Universitaire de Québec - Université Laval, Axe Oncologie, Quebec City, Canada.
- ⁸Département de Biologie Moléculaire, Biochimie Médicale et Pathologie, Faculté de Médecine, Université Laval, Quebec City, Canada.

Contribution

LM a effectué les analyses et écrit le manuscrit, CJB a validé les analyses et écrit le manuscrit. AD a supervisé et écrit le manuscrit. SB a validé, participé à la méthodologie et écrit le manuscrit. AB a supervisé, effectué les analyses et écrit le manuscrit. Tous les auteurs ont lu et approuvé la version soumise.

3.1 Résumé

Le développement récent des technologies de capture de la conformation de la chromatine (3C) a permis de mettre en lumière le rôle joué par l'organisation 3D dans les phénomènes de régulation. Malgré le fait que les régions non codantes aient été démontrées comme jouant un rôle dans l'étiologie de maladies complexes telles que la schizophrénie, aucun modèle computationnel de réseaux entre gènes et éléments de régulation n'a jusqu'ici été proposé. Ainsi, nous proposons les pôles de régulation cis, un modèle 3D mettant en interactions gènes et enhancers, permettant la mise en lumière les mécanismes épigénétiques impliqués dans les maladies complexes. La méthode a été démontrée biologiquement pertinente au regard des phénomènes de régulation et des autres structures 3D connus. Finalement, les pôles de régulation cis représentent une nouvelle structure 3D révélant des interactions complexes entre gènes et enhancers.

3.2 Abstract

The 3D conformation of the chromatin creates complex networks of noncoding regulatory regions (distal elements) and promoters impacting gene regulation. Despite the importance of the role of noncoding regions in complex diseases, little is known about their interplay within regulatory hubs and implication in multigenic diseases such as schizophrenia. Here we show that cis-regulatory hubs (CRHs) in neurons highlight functional interactions between distal elements and promoters, providing a model to explain epigenetic mechanisms involved in complex diseases. CRHs represent a new 3D model, where distal elements interact to create a complex network of active genes. In a disease context, CRHs highlighted strong enrichments in schizophrenia- associated genes, schizophrenia-associated SNPs, and schizophrenia heritability compared with equivalent structures. Finally, CRHs exhibit larger proportions of genes differentially expressed in schizophrenia compared with promoter-distal element pairs or TADs. CRHs thus capture causal regulatory processes improving the understanding of complex disease etiology such as schizophrenia. These multiple lines of genetic and statistical evidence support CRHs as 3D models to study dysregulation of gene expression in complex diseases more generally.

3.3 Introduction

The etiology of complex diseases involves a broad range of causal factors, both genetic and environmental, leading to gene expression changes (Vliet et al, 2007; Do et al, 2017). Models currently used in the etiology of complex diseases suggest that most risk variants are located within noncoding regions explaining a large portion of the heritability (Maurano et al, 2012). Indeed, most risk variants are enriched in distal noncoding regions, disturbing the tissue-specific transcriptional program, and therefore playing a key role in disease etiology (Zhang & Lupski, 2015). The difficulty to assign distal regulatory elements to genes hampered the ability to discover the underlying molecular mechanisms. Consistent with a role of noncoding regions in complex phenotypes, there is also strong evidence on the involvement of 3D chromatin conformation in gene regulation. The 3D genome organization, captured by chromosome conformation assays (van Berkum et al, 2010), revealed the physical proximity between regulatory elements and putative target genes. In addition to chromatin loops connecting promoters to distal noncoding regions (Gorkin et al, 2014; Bouwman & de Laat, 2015; Dekker & Mirny, 2016), the genome is parsed into larger domains including topologically associating domains (TADs) (Dixon et al, 2012) and A/B compartments (Lieberman-Aiden et al, 2009). Interestingly, DNA sequence variations influencing the 3D genome organization are associated with complex disease risks (Gorkin et al, 2019). For example, structural variants disrupting TADs, which are enriched in enhancer-promoter interactions, lead to fused- TADs promoting ectopic promoter-enhancer connections and disruption of the normal transcriptional program (Fudenberg & Pollard, 2019; Melo et al, 2020). However, precisely identifying which genes are affected by a risk variant remains a challenge.

The combination of chromatin interactions and microscopy-based techniques established that groups of genes share the same physical environment (Gizzi et al, 2019). In fact, promoters interact with enhancers inside complex organizations, forming regulatory hub structures (Oudelaar et al, 2019; Campigli Di Giammartino et al, 2020). These hubs exhibit distinct organization from known 3D features, encompassing in most cases few promoters, strongly involved in biological processes (Espinola et al, 2021). In fact, highly interconnected enhancers converge to genes with crucial phenotypic implications, with dynamic enhancer crosstalk at the genome-wide level, occurring more frequently during differentiation (Madsen et al, 2020). Furthermore, super interacting promoters are enriched in lineage-specific genes (Song et al, 2020), known to play a crucial role in diseases,

whereas multiple enhancers connected to a promoter lead to phenotypic robustness in environmental or genetic perturbations (Tsai et al, 2019). At the molecular level, enhancers increase the gene activity through modulation of transcriptional bursting (Fukaya et al, 2016) or indirectly influencing transcription activation (Benabdallah et al, 2019). Interestingly, the organization of genes and noncoding regulatory regions may be pre-established, present in different cells, highly dynamic during differentiation (Rubin et al, 2017; Espinola et al, 2021). However, whether and how promoters and enhancers interacting in hubs are involved in the etiology of complex diseases are still open questions.

Schizophrenia is a complex chronic brain disorder associated with perturbations in the transcriptional programs of neurons (Ruzicka et al, 2020 Preprint). Indeed, schizophrenia is characterized by long-standing delusions and hallucinations strongly reducing life-expectancy (Sullivan et al, 2012). Recent findings suggest that schizophrenia is explained by a polygenic architecture (Smeland et al, 2020), where most of the risk variants are located within non-coding regions. Schizophrenia-risk loci are enriched in active enhancers or promoters in neurons from the adult human frontal lobe (Roussos et al, 2014; Fullard et al, 2018; Girdhar et al, 2018; Hauberg et al, 2020). Also, multiple studies have demonstrated the involvement of 3D organization in the disorder. For example, chromatin loops are enriched in expression quantitative trait loci or schizophrenia-risk variants impacting the proximal gene regulation (Rajarajan et al, 2018). In addition, ultra-rare structural variants in TAD borders lead to gene dysregulations increasing the risk of schizophrenia (Halvorsen et al, 2020). However, the implication of regulatory hubs in the schizophrenia etiology remains to be addressed.

In the present study, we are defining *cis*-regulatory hubs (CRHs) as 3D structures linking one or more gene promoters to networks of distal elements which capture complex patterns of gene regulation. In neurons, CRHs are strongly enriched in schizophrenia-associated genes, SNPs, and heritability compared with equivalent structures.

3.4 Results

3.4.1 Promoter and distal element interactions create CRHs in neurons

To understand regulatory processes involved in complex phenotypes, we built CRHs as bipartite graphs, a natural structure for the 3D contacts between two classes of elements: the promoter of genes and their distal elements. To evaluate their role in schizophrenia etiology, we defined CRHs using chromatin contacts provided by Hi-C data with and without

additional epigenetic features defining classes of distal regulatory elements (See Annexes Chapitre 3 Supplemental Data 1). Because open chromatin regions in the prefrontal cortex of schizophrenia individuals have shown to be enriched in risk variants (Bryois et al, 2018) and that H3K27ac regions are strongly associated with schizophrenia-risk variants (Girdhar et al, 2018), we focused our attention on the activity-by-contact (ABC) model (Fulco et al, 2019) of enhancer–promoter interactions. The approach integrates the frequency of physical contacts between distal elements and promoters (500 bp from an annotated TSS) with the activity defined by the DNase accessibility and the occupancy of H3K27ac (Fig 2, see the Materials and Methods section). The ABC model is a good predictor of differential gene expression (Fulco et al, 2019) and a useful tool to link noncoding variants to their target genes (Nasser et al, 2021). Using available datasets in neurons derived from induced pluripotent stem cells (iPSCs) (Rajarajan et al, 2018), a relevant cell type to study schizophrenia (Sey et al, 2020), we identified 62,658 functional pairs of distal elements and promoters where the ABC score exceeded a threshold of 0.012. The value of the threshold was chosen so that we had, on average, 4.51 distal elements per gene, as recommended by Fulco et al (2019). CRHs were built from these connections between promoters and distal elements (Fig 2A–C). We identified 1,633 CRHs, ranging between 2 and 506 nodes (median of six elements). Postmortem brains are an alternative source of neurons to study 3D contacts. In three samples from postmortem brains: dopaminergic neurons (mentioned as Dopa_1 and Dopa_2 in the Supplemental Data 1) and general neuron population (mentioned as Neu), respectively (Espeso-Gil et al, 2020, See Supplemental Data 1), we observed a strong overlap of the pairs of promoters and distal elements detected by the ABC approach with those found in iPSC-derived neurons. Indeed, we found that most distal elements were shared between iPSC-derived neurons and postmortem brains (Fig 15A). Also, more than 75% of pairs were either strictly found in iPSC-derived neurons (e.g., identical pair) or in an indirect connection within the same CRH (Fig 15B), supporting the reproducibility of the proposed method.

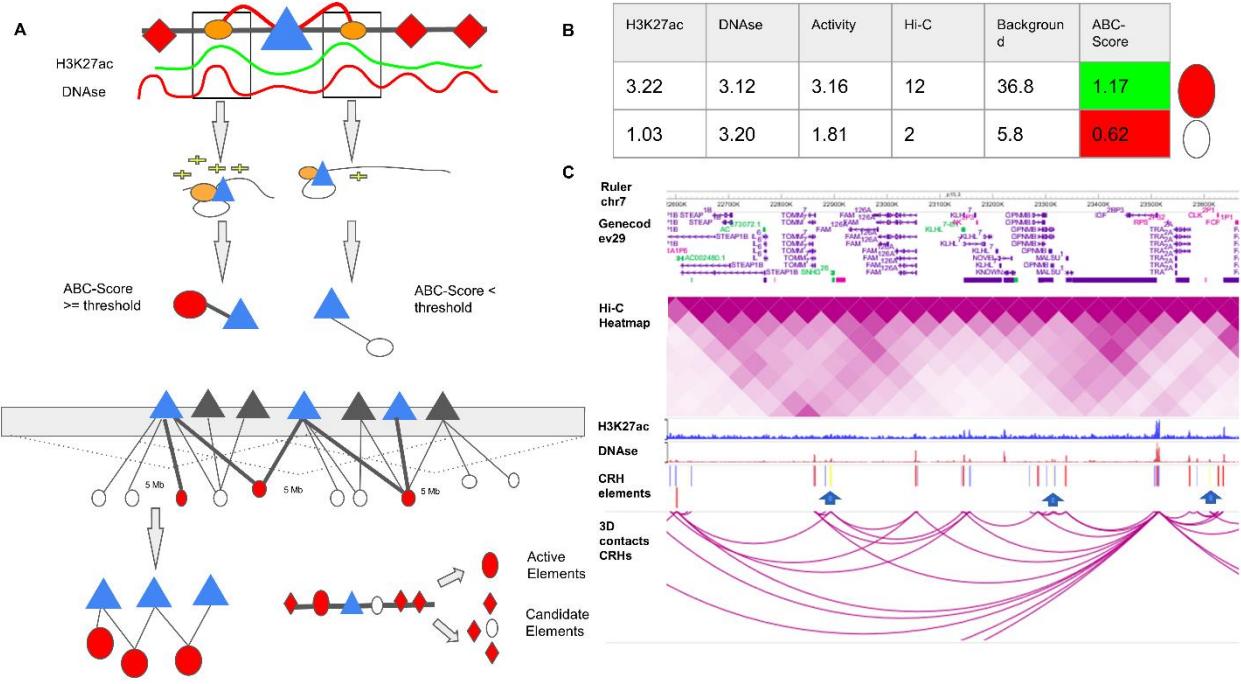


Figure 2: Cis-regulatory hubs (CRHs) are built from activity-by-contact (ABC)-Score methodology

(A) Diagram showing ABC-Score methodology to build functional pairs of promoter (triangles) and distal element (circles). H3K27ac and DNase signals are shown on an arbitrary scale. Among all distal elements (orange circles), active elements (red circles) are discriminated from candidate elements (white circles) based on the value for the ABC-Score. Candidate elements are DNAse accessible regions without H3K27ac signal and non-overlapping active elements. **(B)** Example for the computation of the ABC-Score. Activity is calculated with geometric mean of H3K27ac and DNase signals. Finally, the ABC-Score is the product of Activity by Hi-C signal divided by the background activity, within a 5-Mb window. Here we used a threshold of 1.12 to determine functional connection. All values shown in the table are arbitrary. **(C)** Representation of the physical contacts of a CRH subset on chromosome 7 from the WashU Epigenome Browser (Zhou et al, 2011). Hi-C data in induced pluripotent stem cell-derived neurons are represented. In the CRH element track, we represented distal elements belonging to the CRH (red), promoters (blue) and, elements encompassing noncoding SNPs (yellow bars and blue arrows).

To start investigating the complexity of CRHs in iPSC-derived neurons, we surveyed genes associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014). For example, genes involved in glutamatergic transmission or synaptic plasticity pathways (*GRIA1*, *GRIN2A*, and *GRM3*) exhibited strong differences

regarding CRH complexity (Figs 3A and 16–18). Indeed, *GRIN2A* and *GRM3* were found in relatively simple CRHs of two (one promoter and one enhancer) and three (one promoter and two enhancers) nodes, respectively, whereas *GRIA1* was found in a complex network with 24 genes and 72 regulatory regions. Among the 1,633 CRHs, 15% were pairs of two nodes and therefore constituted monogamous relationships, whereas 85% had 3 elements or more (Fig 3B). We observed comparable results in postmortem brains (Fig 19A and B). Moreover, in iPSC-derived neurons and in post-mortem brain tissues, CRHs contained, on average, a significantly higher number of distal elements than promoters, up to twofold more (median of five distal elements against two promoters, two-sided Wilcoxon signed-rank test, P-value $\leq 2 \times 10^{-16}$) (Figs 3C and 20). Accordingly, promoters were more connected than distal elements as the 80% least connected promoters had at least twice as many connections as the corresponding 80% distal elements (Fig 3D). This result was confirmed in postmortem brain (Fig 21A–C). As expected, the proportion of distal elements was positively correlated with the connections between promoters and distal elements (or complexity) within CRHs (Spearman $\tau = 0.37$, P-value $\leq 2 \times 10^{-16}$), revealing that complex CRHs are significantly associated with a higher proportion of distal elements. The above results suggest that distal regulatory elements and gene promoter regions are organized into complex regulatory structures in neurons. The connectivity between promoters and enhancers is strongly associated with the emergence of tissue-specific phenotypes as they control the transcriptional program (Tsai et al, 2019). Recent studies have shown that highly connected enhancers converge to genes with strong phenotypic impacts (Madsen et al, 2020), whereas promoters enriched in connections are more tissue-specific (Song et al, 2020). Because we expected that connections of genes or distal elements may play a role in disease emergence, we investigated in more detail the organization of genes and distal elements in CRHs of three nodes or more. Thus, we defined two metrics aiming to characterize genes and distal elements involved in these complex relationships (Fig 3A): (1) the proportion of instances where one distal element connects one promoter with at least one other distal element or the reverse: one promoter connects one distal element with at least one other promoter (i.e., 1-1-N with $N > 0$) and (2) the proportion of polygamous elements (i.e., which are not in a monogamous pair or 1-1-N, forming complex shared interactions by promoters or distal elements). Interestingly, most distal elements (63%) were connected to a single promoter, whereas 90% of promoters showed interactions with multiple distal elements. We also observed that 1% of promoters are within 1-1-N relationships versus 5% of distal elements (Fig 3E). This result suggests that distal elements

share a gene more frequently than genes share a distal element, in accordance with previous findings in model organisms (Espinola et al, 2021). Also, comparing CRHs built using the ABC approach with other CRH definitions, we found that promoters were also more connected than distal elements (Fig 22A–C). This result was confirmed across postmortem brain tissues (Fig 23). However, promoters showed fewer connections in our other CRH definitions than the ABC approach. Therefore, the proposed definition of CRHs aligns with previous models suggesting that distal elements interact more specifically (Madsen et al, 2020), whereas promoters are more frequent inside complex relationships.

Next, we wanted to determine the relationship between CRHs and known 3D structures. We focused our analysis on A/B compartments (Lieberman-Aiden et al, 2009), TADs (Dixon et al, 2012), and frequently interacting regions (FIREs) (Schmitt et al, 2016), respectively, segmenting the genome into open and close chromatin, domains of frequent interactions between distal elements and genes, and hotspots for chromatin contacts. In iPSC-derived neurons, the majority of CRHs (76%) shared compartments of the same type, with 46% and 29% for active and inactive compartments, respectively (Fig 3F), whereas only a minor portion (8%) of CRHs overlapped several compartments of different types or were in genomic regions not assigned to a compartment (17%). This result was confirmed in postmortem brains where 47% and 53% of CRHs overlapped active compartments for general neuronal populations and dopaminergic neuronal nuclei (Fig 24). Because it has been shown that A compartments correlated strongly with the presence of genes, accessible chromatin, activating, and repressive histone marks (Lieberman-Aiden et al, 2009), we argue that CRHs are consistent with the open chromatin characteristic associated with functional elements. Moreover, most of CRHs (64%) overlapped a single TAD (Fig 3G). Interestingly, 26% of TADs included two or more CRHs. These observations were confirmed in postmortem brain tissues (Fig 25) and by testing multiple TAD detection algorithms (Fig 26A and B). Last, CRHs were enriched in FIREs compared with candidate CRHs (tissue-specific regions non integrating 3D contacts, see the Materials and Methods section) (two-sided Fisher's exact test, odds ratio = 1.41, P-value $\leq 2.2 \times 10^{-16}$), although only a minor portion of distal elements or promoters overlapped with FIREs (11% and 13%, respectively). The presence of CRHs within compartments and TADs in addition to the enrichment in FIREs was confirmed using the different CRH definitions (Fig 27A and B). Collectively, our results support that CRHs are networks of inter- acting regulatory regions and genes at a finer scale than previously defined chromosome structures. Given the similarity between

CRHs in neurons from iPSC and from postmortem brain tissue, from now on results are restricted to neurons from iPSC.

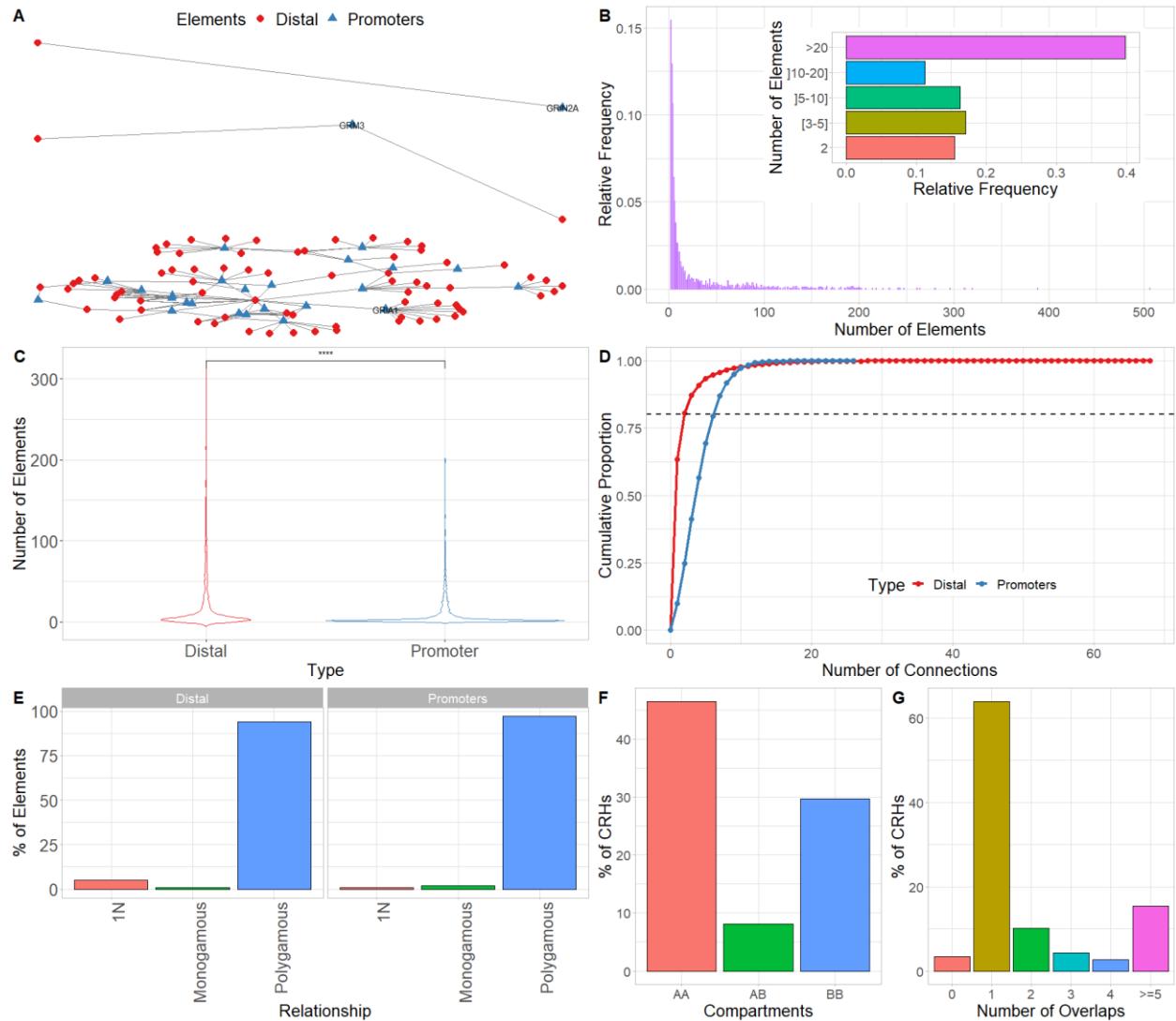


Figure 3: Cis-regulatory hub (CRH) are 3D-based networks mainly constituted by distal elements and more local than high order 3D features

(A) CRHs connecting promoters (blue) to distal elements (red) for *GRIN2A* (top), *GRM3* (middle), and *GRIA1* (bottom) genes. The genes are within monogamous, 1-1-N, and polygamous relationships, respectively. Distal elements are represented by red circles, whereas promoters by blue triangles. **(B)** Distribution of the number of elements (promoters and distal elements) within CRHs. The subpanel shows the number of CRH elements by aggregated categories. **(C)** Distribution of the number of promoters (blue) and distal elements (red) per CRH. We used two-sided Wilcoxon signed-rank test to compare the number of elements. **(D)** Cumulative distribution function of the number of connections for

promoters (blue) and distal elements (red). The dotted line shows the 80th percentile of the number of connections. **(E)** Distribution of the kind of relationship for distal elements (left) and promoters (right). **(F)** Distribution of overlap of CRHs with each compartment type (AA, Active–Active; AB, Active–Inactive; BB, Inactive–Inactive). When CRHs overlap several compartments, we restrict our attention to the farthest elements. The CRHs in genomic regions not assigned to compartments (17%) were omitted from the distribution. **(G)** Distribution of the number of topologically associating domains overlapped by each CRH, when topologically associating domains are detected with the directionality index.

Data information: In **(C)** **** represents P-value ≤ 0.0001 .

3.4.2 CRHs are defined by active chromatin and the presence of schizophrenia-relevant genes

Genes and regulatory elements sharing the same nuclear environment often show coherent transcriptional states and related molecular functions (Campigli et al, 2020). To further characterize the transcriptional activity of CRHs and their involvement in schizophrenia, we overlaid the chromatin states defined by the Roadmap Epigenomics Consortium (Roadmap Epigenomics Consortium et al, 2015). The 18-states model in neurons was sub-divided as follows into three broad categories: (1) Active (1_TssA, 2_TssFlnk, 3_TssFlnkU, 4_TssFlnkD, 5_Tx, 7_EnhG1, 8_EnhG2, 9_EnhA1, 10_EnhA2, and 12_ZNF/Rpts), (2) Weakly Active (6_TxWk, 11_EnhWk, 14_TssBiv, and 15_EnhBiv), and (3) Inactive/Repressor (13_Het, 16_ReprPC, 17_ReprPCWk, and 18_Quies). At the broad category level, we found that most elements (promoters and distal elements) included in CRHs (58%) overlapped Weakly Active regions against 49% for Inactive or Repressor and 53% for Active regions, respectively (Fig 4A). At the individual state level, we observed that 39% of the distal elements included the Quiescent state (Fig 4B) but that CRHs were enriched 2.35-fold (two-sided Fisher's exact test, P-value $\leq 2 \times 10^{-16}$) in active states and depleted in inactive states (two-sided Fisher's exact test, odds ratio = 0.49, P-value $\leq 2 \times 10^{-16}$) compared with candidate CRHs. To confirm the enrichment of CRHs in functional elements, we used ENCODE candidate elements in neurons (The ENCODE Project Consortium et al, 2020). ENCODE candidate elements are regions exhibiting significant signals in H3K4me3, H3K27ac, DNAse, or CCCTC-binding factor (CTCF). CRHs were strongly associated with H3K4me3 (two-sided Fisher's exact test, odds ratio = 1.81, P-value $\leq 2 \times 10^{-16}$), DNAse (two-sided Fisher's exact test, odds ratio = 1.66, P-value $\leq 2 \times 10^{-16}$), and H3K27ac (two-sided Fisher's exact test, odds ratio = 1.44, P-value $\leq 2 \times 10^{-16}$), but not

with CTCF. Our results were supported by other CRH definitions (Fig 28A and B). Then, to extract the global pattern of chromatin states within a CRH, we kept chromatin states representing up to 80% of the total chromatin state signal and observed a striking difference across CRHs. Indeed, 35% of CRHs exhibited a unique combination of chromatin states (e.g., a set of states found only once in CRHs) (Table S1). Also, CRHs characterized by active states were more complex than those strongly defined by quiescent states (18_Quires) (Fig 4C). Considering the above findings, CRHs are enriched in active distal elements and exhibit a variety of chromatin state combinations, suggesting they are important for the control of the transcriptional program of neurons. As CRHs are enriched in active elements in neurons, we postulated that they would be enriched in schizophrenia-relevant genes. First, we identified 8,075 genes associated with schizophrenia (False Discovery Rate ≤ 0.05) using H-Magma (Sey et al, 2020), a statistical approach using 3D noncoding regions with genetic data from genome-wide association study for schizophrenia (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al, 2020 Preprint). We found that 35% of genes significantly associated with schizophrenia are within CRHs compared with 23% for all other genes (1.82-fold enrichment, two-sided Fisher's exact test, P-value $\leq 2.2 \times 10^{-16}$). Moreover, 42% (687/1,633) of CRHs include at least one schizophrenia-associated gene with 23% (376/1,633) harboring several schizophrenia-related genes (mean = 1.77, max = 69) (Fig 4D). Finally, we found that CRHs were enriched in Gene Ontology (GO) biological processes associated with schizophrenia (Fig 4E). Taken together, these results suggest that CRHs are associated with the pathoetiology of schizophrenia, constituting an interesting model for understanding gene regulation and the emergence of complex phenotypes.

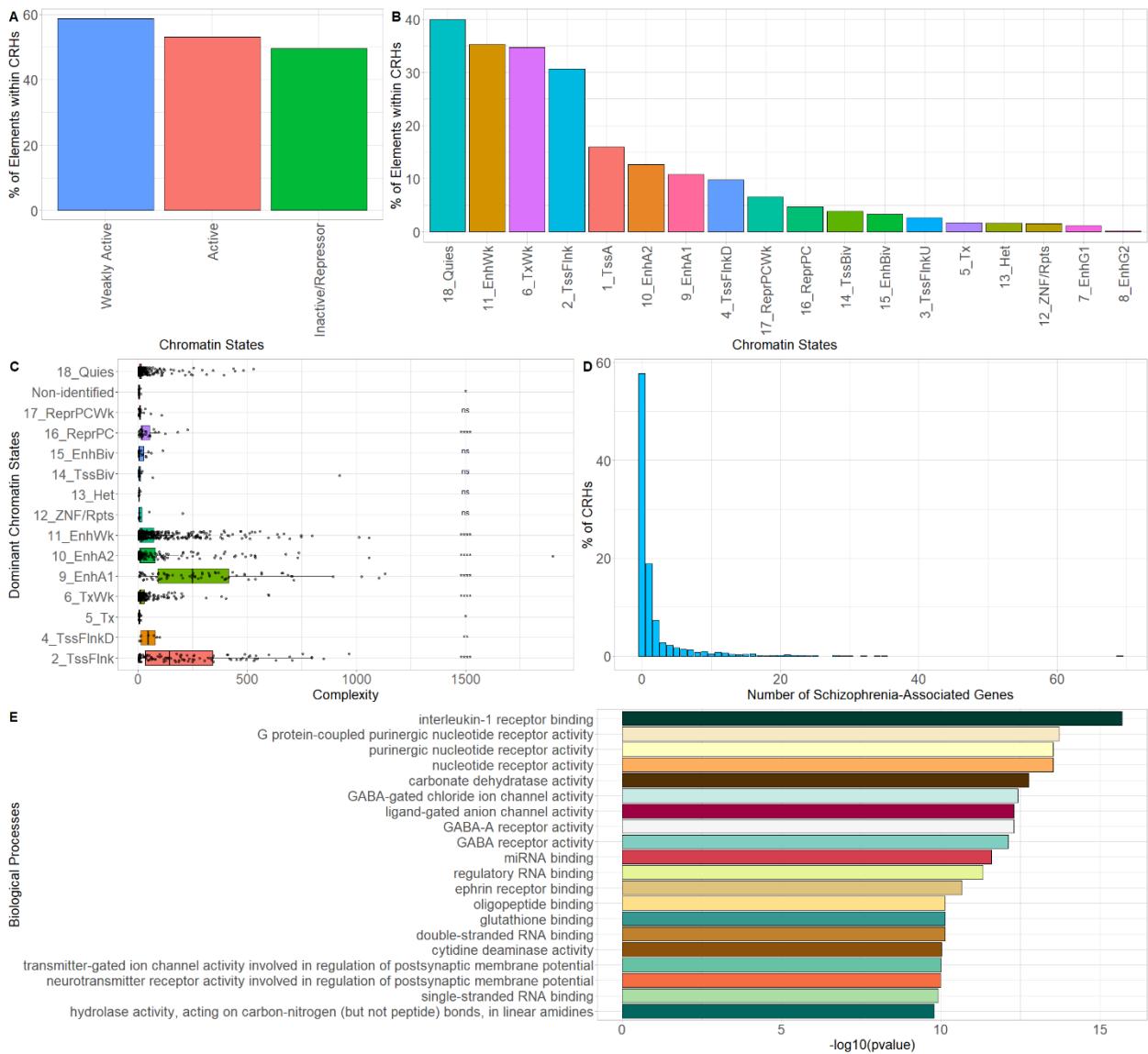


Figure 4: Cis-regulatory hubs (CRHs) are enriched in transcriptionally active elements and genes associated with schizophrenia

(A) Proportion of elements (promoters and distal elements) included within CRHs overlapping chromatin states grouped by activity. **(B)** Proportion of elements (promoters and distal elements) included within CRHs overlapping individual chromatin states. **(C)** Boxplot representing complexity by most present chromatin state within each CRH. Two-sided Wilcoxon rank-sum test was used to compare complexity for each chromatin state with 18_Quies state. **(D)** Distribution of the number of schizophrenia-associated genes per CRH. **(E)** GO enrichment for all genes found within CRHs. The top 20 biological processes are represented.

Data information: In **(D)**, ns, nonsignificant, * represents P-value ≤ 0.05 , ** represents P-value ≤ 0.01 , *** represents P-value ≤ 0.001 , whereas **** represents P-value ≤ 0.0001 .

3.4.3 CRHs containing schizophrenia-associated genes are small and highly expressed

To further characterize CRHs including schizophrenia-associated genes, we examined their characteristics regarding complexity and gene expression levels. Interestingly, CRHs encompassing schizophrenia-associated genes showed larger distances between elements than CRHs not harboring schizophrenia-associated genes (Fig 5A). Also, the number of connections with distal elements was slightly lower for schizophrenia-associated genes than non-associated ones (mean associated genes = 4.39, mean non-associated genes = 4.55, two-tailed t test P-value = 0.005). In addition, schizophrenia-associated genes included within CRHs showed higher expression levels than non-associated genes (median associated = 6.18, median non-associated = 5.87, two-sided Wilcoxon rank-sum test P-value $\leq 2.2 \times 10^{-16}$) (Fig 5B) and were enriched in active distal elements (two-sided Fisher's exact test, odds ratio = 1.79, P-value $\leq 2.2 \times 10^{-16}$). Moreover, schizophrenia-associated genes were more often monogamous genes compared with non-associated ones, showing a 1.34-fold enrichment (two-sided Fisher's exact test, P-value = 0.04). Indeed, 26% of monogamous genes are schizophrenia-associated genes against 20% for non-monogamous ones (Fig 5C). The number of distal elements in CRHs harboring schizophrenia-associated genes was correlated negatively with the proportion of associated genes (Spearman $\tau = -0.47$, P-value $\leq 2.2 \times 10^{-16}$). These results suggest that schizophrenia-associated genes are, in most cases, within small hubs, less connected to distal elements, but expressed at higher levels than non-associated genes.

3.4.4 Multivariate analysis of CRH features with respect to schizophrenia-associated genes

To examine the mutually adjusted influence of the factors examined in the previous sections on schizophrenia-associated genes, we fitted a logistic regression of the status of genes (associated versus non-associated with schizophrenia) on RNA level, the number of connections to distal elements, the 90th percentile of the proportion of active distal elements per gene, and the information regarding monogamy. As expected, the gene status regarding its association with schizophrenia was positively associated with RNA level, the 90th percentile of the proportion of active distal elements, and the monogamy status, whereas it

was negatively associated with the number of connections, confirming our results found with univariate analyses (Fig 5D). Collectively, our results suggest that schizophrenia-associated genes are within small hubs characterized by fewer connections to distal elements and higher transcriptional activity.

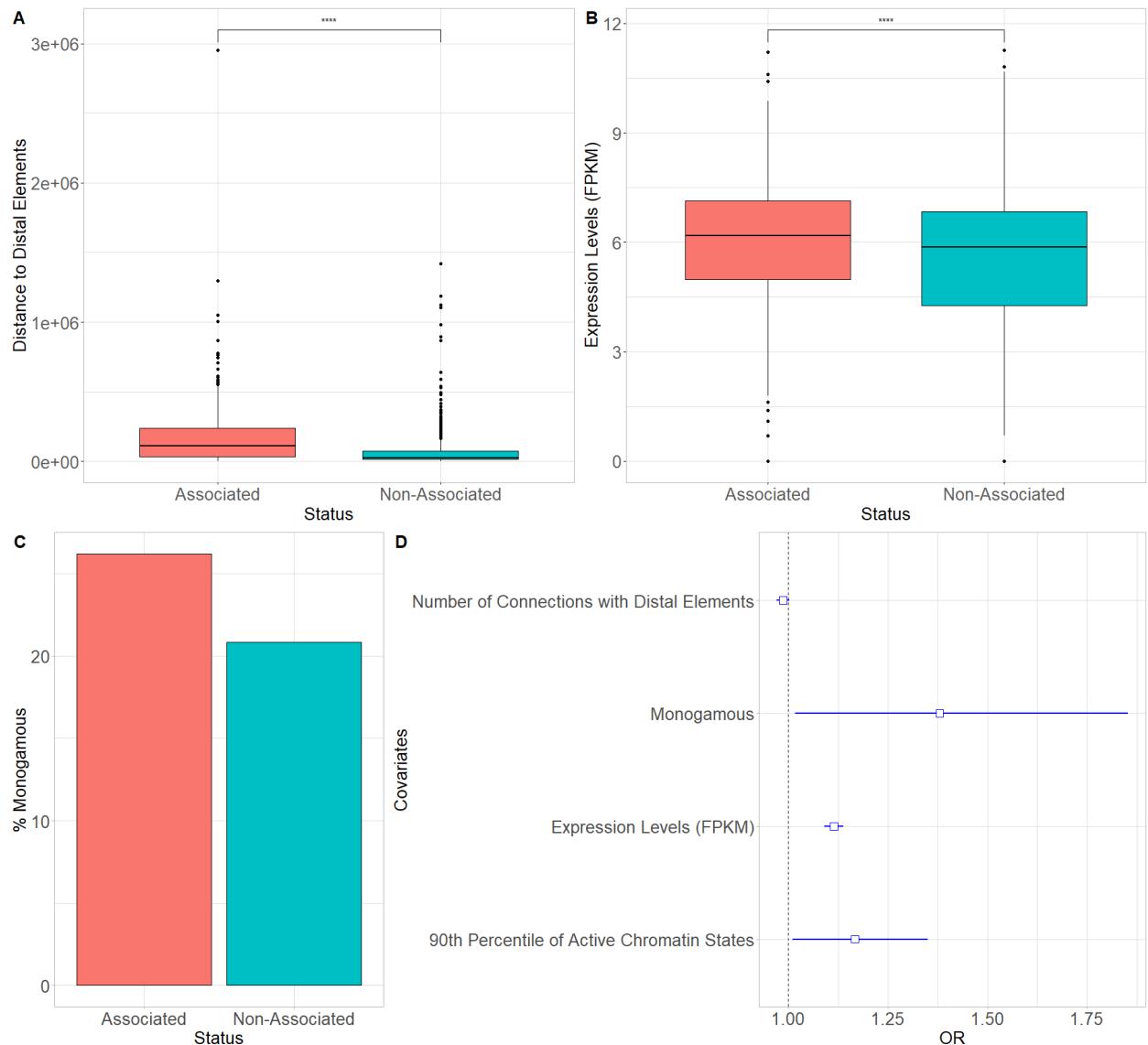


Figure 5: Features of schizophrenia-associated genes

(A) Boxplot of mean distance between elements for cis-regulatory hubs encompassing schizophrenia- associated genes and cis-regulatory hubs not harboring ones. Differences were assessed using two- sided Wilcoxon rank-sum test. **(B)** Boxplot of RNA levels for schizophrenia-associated genes and non- associated ones. Differences were assessed using two- sided Wilcoxon rank-sum test. **(C)** Percentage of monogamous genes which are

associated with schizophrenia or non-associated. **(D)** Odds ratios (OR) and their 95% confidence interval for a logistic regression of the association status of genes with schizophrenia (yes/no). The dotted line represents the null value.

Data information: In **(A)** and **(B)** **** represents P-values $\leq 2 \times 10^{-16}$.

3.4.5 CRHs are enriched in schizophrenia-associated SNPs and heritability

Current models suggest that distal regulatory regions explain a great proportion of the schizophrenia etiology (Roussos et al, 2014). In fact, a wide range of genetic variants affecting the gene expression program are involved in the disorder (Huo et al, 2019). Because we demonstrated the enrichment in schizophrenia-relevant genes within CRHs, we next assessed the presence of schizophrenia-associated SNPs. We collected 99,194 SNPs (after clumping, see the Materials and Methods section) from genome-wide association studies (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al, 2020 Preprint). We mapped them to their corresponding CRH and quantified their enrichments at various association P-value thresholds using the two-sided Fisher's exact test. For instance, there were 2,058 SNPs with a P-value $\leq 1 \times 10^{-4}$. At this significance level, we observed enrichments (odds ratio = 1.29, P-value = 0.04) in CRHs compared with the candidate CRHs (Figs 6A and 29A). Then, we used the same methodology as Nasser et al (2021) to define enrichment in common SNPs overlapping a given functional annotation (proportion of significant SNPs for schizophrenia/proportion of all common SNPs). Consistent with our previous finding, we observed higher fold enrichments (enrichment for elements of interest/enrichment for candidates) for CRHs than for distal elements, becoming stronger with the significance level (Fig 6B). This enrichment was stronger with alternative definitions of CRHs (Fig 29B). Therefore, our results suggest that CRHs are enriched in SNPs for schizophrenia. After demonstrating the relevance of CRHs with schizophrenia associated SNPs, we wondered whether they explained schizophrenia heritability. To this end, we leveraged linkage disequilibrium score regression (LDSC; Finucane et al, 2015) which provides the portion of disease heritability explained by a functional annotation. First, comparing CRHs to equivalent non-tissue-specific noncoding regions, we ensured to maximize the explained heritability by using tissue-specific elements and integrating 3D contacts by conditioning on enhancers, promoters, H3K27ac, and DNase peaks from the LDSC baseline model. In addition, we compared CRHs with equivalent components, defining candidate CRHs as tissue-specific elements equivalent to those found in CRHs but without 3D contacts. Among functional annotations with significant heritability,

the heritability enrichment was higher for CRHs than for non-tissue-specific non-coding regions and candidate CRHs (Fig 6C), with strong enrichment signal for CRHs compared with candidate CRHs (Z-Score CRHs = 2.41, two-sided P-value = 0.01; Z-Score candidate CRHs = -1.84, two-sided P-value = 0.065). CRHs explained 11-fold more heritability than their respective candidate CRHs or up to 44-fold more than non-tissue-specific elements (Table S2). When compared with methods building hubs using only the chromatin contacts and using DNase, CRHs built using the ABC approach performed better regarding schizophrenia heritability, showing enrichments of 3.08 (Fig 6C) against 0.84 (Fig 30A) and 2.98 (Fig 30B), respectively. All heritability enrichment results at the individual level and CRH level for the complete baseline model and his modified version are given for the ABC and control methods in Supplementary Tables (Tables S2 and S3). This result demonstrates a better concordance of the CRH including epigenetic features to explain schizophrenia heritability compared with only using chromatin interactions or combining chromatin interactions with chromatin accessibility. Because we observed that schizophrenia-associated genes are highly expressed, enriched in small hubs, and connected to few distal elements, we defined strata of CRH number of promoters based on the proportion of total variance explained by CRHs (intraclass correlation) through a linear mixed model of the gene expression (Fig 31). Supporting our previous findings, we found that small CRHs (≤ 3 promoters) are more enriched in schizophrenia heritability than medium (>3 and ≤ 25 promoters) or large ones (>25 promoters) (Fig 6D). Overall, these results support that CRHs, especially small ones, are a relevant structure to explain the etiology of schizophrenia.

3.4.6 CRHs predict the association between schizophrenia-associated noncoding SNPs and differentially expressed genes

Because we observed that CRHs are enriched in schizophrenia-associated SNPs and schizophrenia heritability, we wondered whether they represent a useful structure to link noncoding SNPs to genes differentially expressed in schizophrenia. To conduct this investigation, we leveraged information on 8,413 up- and down-regulated genes in all the available cell-types from a large set of schizophrenia patient brain tissues compared with controls (differentially expressed genes or DEGs) from SZBDMulti-Seq (Ruzicka et al, 2020 Preprint) and schizophrenia-associated SNPs from genome-wide association studies (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al, 2020 Preprint). First, CRHs were strongly enriched in DEGs compared with candidate genes, not included in CRHs (two-sided Fisher's exact test, odds ratio = 5.33, P-value $\leq 2 \times 10^{-16}$). CRHs

encompassing at least one DEG exhibited a slightly larger proportion of distal elements compared with CRHs without DEGs (median of 68% compared with 66%, two-sided Wilcoxon rank-sum test, P-value = 8.01×10^{-6}). These results suggest that CRHs may capture the links between SNPs in regulatory regions and DEGs. We tested the hypothesis that links between noncoding SNPs and DEGs are better captured by CRHs than by promoter-distal element pairs and TADs, respectively, the simplest form of CRHs and one of the most studied 3D features in disease etiology (Bryois et al, 2018; Fudenberg & Pollard, 2019), by measuring the proportions of DEGs linked with noncoding SNPs. We first assigned schizophrenia- associated noncoding SNPs to each kind of structure (see the Materials and Methods section) and observed that 30% of CRHs exhibited at least one such assigned SNP in their regulatory regions, compared with 4%, and 95% for pairs and TADs, respectively. In the subset of elements where we observed assigned SNPs, CRHs exhibited a larger proportion of DEGs, exhibiting median proportion of 25% compared with 0% for regulatory regions directly connected with gene promoters and 18% for TADs, respectively (Fig 6E). Therefore, as intermediate structures compared with promoter-distal element pairs and TADs, CRHs better capture links between noncoding SNPs to gene expression variation possibly involved in schizophrenia.

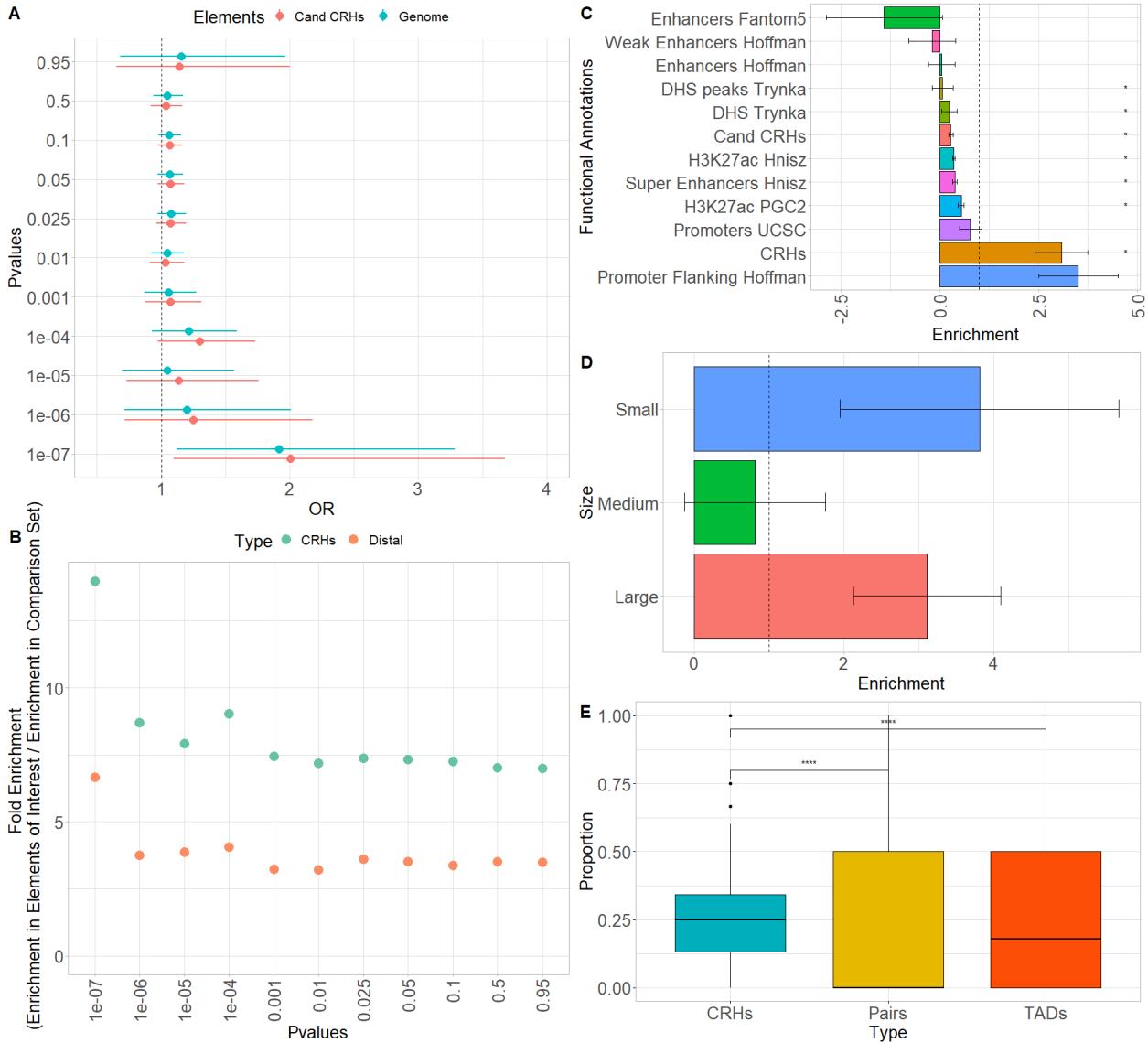


Figure 6: *Cis*-regulatory hubs (CRHs) are enriched in schizophrenia-associated SNPs, schizophrenia heritability, and capture links between noncoding SNPs and genes differentially expressed in schizophrenia

(A) SNP enrichment analysis measured through odds ratios (OR) and their 95% confidence interval for CRHs compared with candidate CRHs and the rest of the genome at different significance levels. The dotted line represents the null value. **(B)** Fold enrichment of distal elements and CRHs compared with their respective candidate sets for different significance levels. **(C)** Schizophrenia heritability enrichment measured with LDSC with error bars for CRHs, candidate CRHs, and non-tissue noncoding elements. The dotted line represents the null value. Errors bars represent the standard errors around the estimates of enrichment. **(D)** Schizophrenia heritability enrichment (measured with LDSC) with error bars for CRHs, considering the number of genes within CRHs. The dotted line represents the null value. **(E)**

Boxplot of DEG proportions within CRHs, promoter distal elements pairs, and topologically associating domains. We considered only elements where we observed noncoding schizophrenia-associated SNPs. Differences were assessed using two-sided Wilcoxon rank-sum test.

Data information: In **(C)** * represents P-value ≤ 0.05 after Bonferroni correction. In **(E)** **** represents P-values $\leq 2 \times 10^{-16}$.

3.5 Discussion

Distal elements play a crucial role in complex diseases, such as schizophrenia. Recent studies have characterized relationships between promoters and distal elements interacting in hubs (Madsen et al, 2020; Song et al, 2020; Espinola et al, 2021). However, their precise contributions to complex disease etiology remain unclear. In this study, we assessed the role of hubs linking promoters to distal elements in a complex disease. Thus, we defined CRHs in neurons as complex networks of gene promoters and distal elements in physical proximity (Fig 2A). CRHs aim to highlight direct and indirect contacts between promoters and distal elements which may not be targeted by other approaches. Our findings confirm the interest in integrating 3D contacts with tissue-specific regions to gain a deeper understanding of regulatory processes involved in complex diseases, where genetic disruptions may impact the transcription program of several genes (Figs 2C and 18). CRHs are enriched in gene promoters and distal elements associated with schizophrenia (Fig 6A) and explain a larger portion of heritability than candidate CRHs (Fig 6C) or other definitions to characterize CRHs (Fig 30A and B). Also, assessing the functional interest of CRHs in schizophrenia etiology, we found that CRHs are more efficient to capture the links between noncoding SNPs to genes differentially expressed in schizophrenia compared with TADs and promoter-distal element pairs. Thus, through CRHs, impacts of polymorphisms on gene expression variation can be better targeted. Therefore, our results establish that CRHs, by integrating interactions between distal elements and gene promoters, constitute a relevant 3D model to study complex diseases such as schizophrenia.

Previous studies suggest that hubs linking genes to enhancers are involved in the emergence of TADs (Espinola et al, 2021) or that highly interconnected enhancers constitute sub-TADs strongly enriched in CTCF (Madsen et al, 2020). Recent studies have either investigated the role of chromatin loops (Rajarajan et al, 2018) or the impact of ultra-rare variants in TAD borders in the emergence of schizophrenia (Halvorsen et al, 2020). CRHs

constitute a more local functional organization than higher order chromatin features (A/B compartments, TADs) (Fig 3F and G) and are enriched in FIREs. In fact, CRHs are strongly enriched in active regions (Fig 4C), defining CRHs as functional hubs with high transcriptional activity. Moreover, CRHs are strongly enriched in schizophrenia-associated genes, which are characterized by higher expression levels (Fig 5B) and active regulatory regions (Fig 5D). These results are in line with those of Sey et al (2020), as they have shown that schizophrenia- associated genes exhibit higher differential expression in schizophrenia. Based on the above lines of evidence, we argue that focusing on CRHs should be prioritized over other levels of 3D organization in a context of complex phenotypes. Thus, through CRHs, impacts of polymorphisms on gene expression variation can be better targeted, aiming to highlight underlying regulatory processes.

Promoters and distal elements involved in CRHs exhibit different connectivity behaviors. Indeed, CRHs harbor more distal elements than genes (Fig 3C), suggesting that within a CRH, genes tend to have more connections compared with distal elements (Fig 3D) (Madsen et al, 2020; Espinola et al, 2021). Espinola et al (2021) have shown that hubs connecting promoters to distal elements encompass a single promoter, whereas Madsen et al (2020) exhibited that enhancers are mostly involved in one-to-one connections. These results suggest that genes have fewer specific relationships, whereas enhancers, strongly connected to promoters, link genes with strong involvement in diseases (Madsen et al, 2020). However, in our data we found that CRH often harbor several genes connected by distal elements, supporting that CRHs can be either promoter hubs, enhancer hubs or multi hubs (Fig 3A) (Campigli et al, 2020). Limitations of CRHs defined from Hi-C data are their dependence on Hi-C resolution and the measure of contacts from multiple cells in bulk, which may lead to spurious merging of CRHs with contacts occurring in distinct cell sub-populations. Future studies using single-cell chromosome conformation will be needed to assess the relevance of CRHs at higher resolution (Nagano et al, 2013).

An important contribution of this study is to establish CRHs as a relevant model to study complex diseases such as schizophrenia. Indeed, we found strong enrichments in schizophrenia-associated SNPs, schizophrenia heritability within CRHs (Fig 6A and C), com- pared with candidate CRHs. Also consistent with this idea, we found that including DNase hypersensitive sites and H3K27ac-enriched regions to the definition of CRH explains a larger portion of schizophrenia heritability than networks built only from chromatin contacts. Moreover, CRHs aim to highlight indirect connections between promoters and

distal elements and our results show they offer an advantage over a pair of enhancer–promoter or larger domains to efficiently link noncoding SNPs to DEGs in schizophrenia. Collectively, these results point to the capability of CRHs to capture complex interplay between regulatory regions, which can help to fine map the functional regions involved in complex diseases, one of the most important challenges in polygenic diseases. Moreover, schizophrenia-associated genes show fewer connections than non-associated ones and are enriched in monogamous relationships (Fig 5C and D). These results suggest that schizophrenia-associated genes are more strongly impacted than other active genes by disruptions of their distal elements because they are regulated by fewer connections to distal elements. Interestingly, we found that hubs encompassing a small number of genes highlight stronger schizophrenia heritability enrichments than medium or larger hubs (Fig 6D). We expect that small hubs or genes weakly connected to distal elements (monogamous, 1-1-N) will be more impacted by disruptions in their distal elements than large hubs or highly connected genes, supporting the model where weakly connected genes are more involved in disease etiology. From this study and others, the emerging model is that a gene with limited connections to distal elements will be more impacted by polymorphisms, whereas highly connected genes will have stronger environmental or genetic resilience to disruptions in their distal elements (Tsai et al, 2019).

Based on these results, we argue that CRHs capture direct and indirect connections between promoters and distal elements, explaining the underlying regulatory processes involved in complex phenotypes. Future studies will demonstrate whether CRHs as a functional 3D model improve detection power of causal genes or pathways to elucidate the underlying causal regulatory processes involved in complex diseases. Indeed, because a substantial portion of schizophrenia heritability remains to be explained, future work will be needed to assess the relevance of CRHs to help detect the rare variants which may be involved. CRHs can be integrated as functional annotation in association tests (He et al, 2017) or proposed as new regions to aggregate variants in pathway-based approaches (Wu & Pan, 2018).

3.6 Materials and Methods

3.6.1 Hi-C data and pre-processing

Hi-C data for neurons from iPSCs at 10 Kb resolution were obtained from PsychENCODE Synapse platform (.hic format, intra-chromosomal). In the present study, we refer to these

data as the neuron Hi-C dataset. Except for Score-FIRE calculation and ABC score, we applied KR- normalization with the Juicer toolbox (Durand et al, 2016) to obtain either a sparse or dense matrix.

3.6.2 CRHs

CRHs were built based on the ABC model (Fulco et al, 2019) to capture active regulatory processes between distal elements and gene promoters. To validate analyses shown in the article, two other methods to build CRHs were also proposed (See Figure 32).

3.6.3 ABC-Score

The ABC model (Fulco et al, 2019) defines active enhancers based on a quantitative score of DNase (ENCSR278FVO), H3K27ac (ENCSR331CCW), and normalized Hi-C contact number. This score is computed relative to a background activity over a 5-Mb window around a candidate element. Here candidate element refers to DNase peaks on which enhancers are defined (Fulco et al, 2019). Then, we set the threshold to 0.012; beyond which a candidate element is considered as a distal element. This value was selected to ensure that the mean number of distal enhancers per promoter is between two and five in the neuron Hi-C dataset (Fulco et al, 2019).

As an extension of the ABC-Score, CRHs were defined as bipartite networks (igraph R package; Csardi & Nepusz, 2006) between promoters and distal elements. Because of the nature of the methodology of the ABC-Score, contacts between distal elements and promoters were restricted. In proposing CRHs based on the ABC-Score, active regulatory phenomena occurring in our tissue were captured.

CRHs are conceived to capture regulatory phenomena based on Hi-C. For the purpose of enrichment analysis for different external validation sources, SNPs or disease heritability, equivalent sets with elements having the same characteristics but in no 3D contacts with promoters were proposed. These elements were referred to as candidate CRHs. Thus, the same approach as Nasser et al (2021) was applied, where candidate distal elements are all DNase peaks which do not overlap ABC distal elements. Also, candidate promoters were all promoters for known hg19 genes not included in CRHs.

3.6.4 Summary statistics for schizophrenia

The original SCZ3 GWAS summary statistics (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al, 2020 Preprint) used in SNP and heritability

enrichments were downloaded from the PGC site <https://www.med.unc.edu/pgc/> results-and-downloads. To assess independent SNPs in enrichment analyses, we used the clumped SNP file keeping the SNPs with the highest association signal with schizophrenia for a given genomic window.

3.6.5 Schizophrenia-associated genes

To assess schizophrenia-associated genes, H-Magma (Sey et al, 2020) was used on iPSC-derived Hi-C neurons (Rajarajan et al, 2018) with schizophrenia SNP summary statistics (The Schizophrenia Working Group of the Psychiatric Genomics Consortium et al, 2020 Preprint) to link noncoding SNPs to their target gene. To determine significant schizophrenia-associated genes, all genes with a P-value lower or equal to 0.05 after Benjamini and Hochberg correction were selected.

3.6.6 Partitioning heritability for schizophrenia

The LDSC regression (Finucane et al, 2015) was used to partition SNP heritability for schizophrenia integrating CRHs. For the main analysis, a modified version of the LDSC baseline model was used, only considering non-neuron-specific annotations corresponding to regulatory regions included within CRHs: promoters, H3K27ac histone marks, DNase I hypersensitive sites, ChromHMM/Segway predictions, super-enhancers, and FANTOM5 enhancers. By proceeding this way, we sought an unbiased comparison of SNP heritability in neuron CRHs compared with candidate CRHs or equivalent sets of genomic features widely used for this purpose. CRHs and candidate CRHs were extended by 500 bp upstream and downstream to consider the background activity and avoid inflating the enrichment signal, as suggested by Finucane et al (2015).

3.6.7 Linking noncoding SNPs to DEGs in schizophrenia

To link noncoding SNPs to differentially expressed genes, we only considered the clumped SNPs because they represent the genetic variations the most associated with schizophrenia within genomic windows, without applying a P-value threshold. Then, the proportion of DEGs was calculated for the subset of CRHs, promoter distal-element pairs and TADs where we observed at least one SNP in one distal element.

3.6.8 3D features and other analyses

All analyses related to 3D features (A/B compartments, TADs, and FIREs), functional enrichments, peak calling, and other technical details are presented in Supplemental Data 1.

3.6.9 Availability of data and materials

Datasets analyzed in this study are publicly available from: PsychENCODE Knowledge Portal (<https://www.synapse.org>, syn13363580, syn20833047) for Hi-C in iPSC, and postmortem brains, respectively. PGC3 (<https://www.med.unc.edu/pgc/> results-and-downloads) for summary statistics, SCREEN (<https://screen.encodeproject.org>) for Encode candidate regulatory elements in neural progenitor cell originated from H9, Roadmap Epigenomics data Portal (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/_jointModel/final) for 18-states model in E007, ENCODE data portal (ENCSR539JGB), and GEO (GSE142670) for reference epigenome and RNA-Seq in neurons. Various analyses shown in this article, as well as additional documentation and CRHs in neurons, are available at: https://github.com/lmangnier/Hi-C_analysis.

3.7 Supplementary Information

Supplementary Information is available at <https://doi.org/10.26508/lsa.202101156>.

3.8 Acknowledgements

We would like to acknowledge Antoine Bodein, Julien Prunier, and Christophe Khun-Narom Tav of the Arnaud Droit Lab for important discussions and feedback during the preparation of this manuscript. We also thank Katie Pollard, Geoff Fudenberg, and Luis Chumpitaz-Diaz from the Pollard Lab for their insightful advice. Data were generated as part of the PsychENCODE Consortium, whose contributors and funding are listed in Supplemental Data 1. Funding: This work was partly funded by the Canadian Statistical Sciences Institute through a Collaborative Research Team grant and by a National Science and Engineering Research Council of Canada Discovery grant to A Bureau. Some data analyses were performed on computing re- sources from Compute Canada.

3.9 Conflict of Interest Statement

The authors declare that they have no conflict of interest.

Chapitre 4: RetroFun-RVS: a retrospective family-based framework for rare variant analysis incorporating functional annotations

Nous avons démontré dans le chapitre précédent que les pôles de régulation cis en formant des réseaux d'interactions 3D entre gènes et enhancers, étaient des structures actives, permettant de mettre en lumière les mécanismes de régulation impliqués dans les maladies complexes. Dans ce chapitre, nous proposons un test d'association de variants rares permettant l'incorporation d'annotations fonctionnelles. Une composante majeure de la méthode proposée est de tirer profit de la structure familiale des données en exploitant seulement les individus atteints. Cette stratégie a été démontrée utile pour mettre en lumière le partage de variants rares potentiellement causaux dans les maladies complexes. Le défi ici est alors de proposer une méthode capable de tenir compte de la nature des données en permettant d'ajuster pour le devis d'étude. Nous proposons alors une approche rétrospective. La méthode sera évaluée à travers plusieurs scénarios dans des études de simulation.

Journal

Cet article a été soumis à la revue *Genetics* le 23 juin 2022. L'article est disponible sur *bioRxiv* depuis cette date.

L'article est en accès libre distribué selon les termes de la licence Creative Commons Attribution License (CC BY). L'éditeur autorise son utilisation et sa diffusion.

Accessibilité

Loïc Mangnier, Alexandre Bureau, *RetroFun-RVS: a retrospective family-based framework for rare variant analysis incorporating functional annotations*, *bioRxiv* 2022.06.21.497085; doi: <https://doi.org/10.1101/2022.06.21.497085>

Liste des auteurs

Loïc Mangnier^{1,2,3}, Alexandre Bureau^{1,2,3}

- ¹Centre de Recherche CERVO, Quebec City, Canada.
- ²Département de Médecine Sociale et Préventive, Université Laval, Quebec City, Canada.
- ³Centre de Recherche en données Massives de l'Université Laval, Quebec City, Canada.

4.1 Résumé

La plupart des variants impliqués dans les maladies complexes sont rares et localisés dans des régions non codantes, rendant l'interprétation des mécanismes biologiques difficiles. Dans cet article nous proposons *RetroFun-RVS*, un test d'association de variants rares, permettant l'intégration d'annotations fonctionnelles, tout en incorporant l'information familiale. Une des subtilités du modèle proposé est de n'exploiter que les individus atteints au sein des familles. En permettant l'intégration du test original, notre méthode est robuste lorsqu'aucune annotation est prédictive pour le trait. De plus, à travers des études de simulation, nous avons exploré différentes stratégies pour incorporer des annotations fonctionnelles. En permettant l'incorporation d'annotations fonctionnelles sous forme de réseaux 3D, la méthode a été démonté plus puissante que les autres stratégies ou méthodes compétitives. Finalement, *RetroFun-RVS*, en exploitant devis familiaux et information fonctionnelle est utile pour mettre en lumière les mécanismes de régulation impliqués dans l'étiologie de maladies complexes.

4.2 Abstract

A large proportion of genetic variations involved in complex diseases are rare and located within non-coding regions, making the interpretation of underlying biological mechanisms difficult. Although technical and methodological progresses have been made to annotate the genome, current disease- rare-variant association tests incorporating such annotations suffer from two major limitations. Firstly, they are restricted to case-control designs of unrelated individuals, which often require tens or hundreds of thousands of individuals to achieve sufficient power. Secondly, they were not evaluated with region-based annotations needed to interpret the causal regulatory mechanisms. In this work we propose RetroFun-RVS, a new retrospective family-based score test, incorporating functional annotations. One of the critical features of the proposed method is to aggregate genotypes while measuring rare variant sharing among affected family members to compute the test statistic. Through extensive simulations, we have demonstrated that RetroFun-RVS integrating networks based on 3D genome contacts as functional annotations reaches greater power over the region-wide test, other strategies to include sub-regions and competing methods. Also, the proposed framework shows robustness to non-informative annotations, keeping a stable power when causal variants are spread across regions. We provide recommendations when dealing with different types of annotations or family structures commonly encountered in practice. In summary, we argue that RetroFun-RVS, by allowing integration of functional annotations corresponding to regions or networks with transcriptional impacts, is a useful framework to highlight regulatory mechanisms involved in complex diseases.

4.3 Introduction

Over the past few years with the democratization of whole-exome or whole genome sequencing data, important progress has been made in the effort to link genetic variations to phenotypes. Indeed, at population scale, Genome-Wide Association Studies (GWAS) have provided useful resources to highlight variants involved in diseases. However, these methods, in addition to requiring tens or hundreds of thousands of individuals, are mainly restricted to common variants, leaving an important part of heritability unexplained (Manolio et al., 2009). In fact, studies have shown that the individual genetic risk is also influenced by rare variants (minor allele frequency (MAF) $\leq 1\%$), (Singh et al., 2022; Sun et al., 2022). In addition to being rare, variants influencing disease risk tend to be located within non-coding regions, making the underlying biological mechanisms difficult to interpret (Zhang and Lupski, 2015). Thus, the tremendous amount of rare variants located within non-coding regions brings new challenges to identify new causal variants involved in diseases, and accounting for their functional impacts remains crucial from a fine-mapping perspective (Schaid et al., 2018).

Methods have been proposed to overcome the challenge of sparsity. Indeed, because variants are rare, methods testing them in an unitary fashion perform badly (Madsen and Browning, 2009). Thus, rare-variant association tests (RVATs) are methods aggregating genotypes across several variant sites within a gene, pathway or regions functionally close. By collapsing variants across over regions, these methods considerably reduce the number of tests throughout the genome, hence increasing statistical power. Among them, burden tests were initially proposed and are powerful when all variants across regions show a homogeneous effect (Li and Leal, 2008; Madsen and Browning, 2009). However, when regions combine both deleterious and protective variants, burden tests comparing cases to controls suffer from a substantial decrease of power. Alternatives to address this limitation have been proposed (Ionita-Laza et al., 2011; Neale et al., 2011). One of the critical features of RVATs is that they can be expressed through regression models, allowing the integration of variant weights, either fixed (based on the MAF), or estimated in a data adaptive manner (Madsen and Browning, 2009; Wu et al., 2011). The multiple ways to define test statistics created a need for combining several p-values within a given region to assess the association with a trait, while adjusting for multiplicity. Liu et al., 2019 have proposed ACAT, a powerful statistical framework combining p-values in an efficient way. One of the major advantages of ACAT over other combination methods (e.g., taking the

minimum (minP), Fisher's method) is that it requires neither resampling procedures, nor independent p-values nor explicit models for correlations. Although these set-based tests have made possible the discovery of new regions involved in complex diseases, they required very large sample sizes of unrelated subjects.

An alternative approach is to exploit family-based studies. In addition to reducing genetic heterogeneity, pedigree-based studies have been shown to have more power than population-based approaches for detecting rare variants, when an enrichment of risk variants among families is expected (Laird & Lange, 2006; Li et al., 2006; Ott et al., 2011). Information provided by variants segregating with the disease, even imperfect, can be exploited to highlight new causal variants, giving a second life to studies in extended pedigrees (e.g., Bureau et al., 2014). Recent methods based on identity-by-descent (IBD) or combining both linkage approaches and RVATs have been developed (Bureau et al., 2019; Sul et al., 2016; Zhao et al., 2019). One important feature of these approaches is that they focus on, or can be restricted to, only affected family members, when these are expected to contribute more information than unaffected subjects (Schaid et al., 2010). Affected-only designs have a long tradition in gene-gene or gene-environment interaction analysis and have been extended to family-based studies, requiring smaller sample sizes to reach equivalent power, compared to considering unrelated case-only individuals, which is an appealing feature in practice (Li et al., 2019). However, selecting individuals retrospectively (based on their phenotype) may lead to highly ascertained sampling schemes resulting in overestimated association measures. Retrospective likelihood, by conditioning on phenotype, have been shown to give accurate estimates for common variants when ascertainment bias is expected (Schaid et al., 2010). Extensions of RVATs have been proposed for retrospective samples of families (Schaid et al., 2013). A limitation of all the above methods is that none of them currently integrates external information on biological mechanisms involved in diseases. How to leverage information on non-coding regulatory elements in the detection of variants influencing disease risk remains an open question. Thus, the increasing interest in using external information for this task, and hence highlighting the biological mechanisms. Recent methods, such as FST (He et al., 2017) or FunSPU (Ma and Wei, 2019) have proposed to adaptively test functional annotations under a general RVAT framework. These methods have shown substantial increases in power when at least one functional score is predictive for the effect of variants on the trait, while they show robustness when no annotations were predictive for variant impact on the trait, revealing new causal variants involved in complex traits. More recently, with the striking

development of methods detecting regulatory elements such as enhancers (Fulco et al., 2019; Yao et al., 2022), progress has been made in associating non-coding SNPs to their target genes (Gazal et al., 2022; Nasser et al., 2021). Subsequently, some authors have proposed to incorporate this information within statistical frameworks. Hence, Ma et al., 2021 have demonstrated that long range 3D interactions between genes and enhancers add information for the integration of non-coding regulatory regions within gene-based frameworks. This model only considers gene-enhancer pairs, consistent with previous studies (Wu and Pan, 2018). Models extending gene-enhancer pairs to Cis-Regulatory Hubs (CRHs), networks encompassing up to several genes and active enhancers have been proposed (Mangnier et al., 2022). CRHs have been shown to be a relevant model in schizophrenia etiology, explaining more heritability than tissue- and non-tissue- specific elements, and being more effective to link noncoding SNPs to differentially expressed genes in schizophrenia compared to Topologically Associated Domains (TADs) or pairs of gene-enhancers. To our knowledge, no study to date has proposed to integrate functional annotations within a RVAT framework for family-based designs, while allowing the incorporation of discontinuous genomic regions involved in 3D-based networks.

In this paper, we propose RetroFun-RVS (Retrospective Functional Rare Variant Sharing), a model, allowing the integration of functional annotations under a family-based design considering only affected individuals. Through extensive simulation studies, we have demonstrated that RetroFun-RVS integrating CRHs as functional annotations is a more powerful approach to detect causal variants over others strategies, while controlling the Type I error rate well. We provided recommendations when dealing with different types of functional scores or pedigree structures. Finally, we have demonstrated that integrating 3D-based functional annotations through networks is a relevant strategy to gain power of detection of causal variants, while highlighting the underlying biological mechanisms involved in diseases.

4.4 Material and Methods

4.4.1 Notation and Model

Suppose that we have N subjects within F families, where n_f is the number of individuals for the f^{th} family. Let's define Y , a binary vector of phenotypes, G , a $N \times p$ matrix of genotypes for rare variants, coded as the number of copies of the minor allele 0, 1, 2. Assuming a log-additive model for the individual SNP effect on disease risk, under

assumption of conditional independence of the phenotypes of different individuals given their genotypes and considering only affected individuals, following Schaid et al., 2010, the retrospective likelihood for one family can be written as:

$$P(G|Y) = \frac{\exp \sum_{i \in D} \sum_{j=1}^p \beta_j x_{ij} P(G)}{\sum_{G^*} \exp \sum_{i \in D} \sum_{j=1}^p \beta_j x_{ij}^* P(G^*)}$$

where D is the subset of affected members in the family, while x_{ij} is a condensed notation for $x(G_{ij})$, the number of minor alleles for variant j in individual i in the multilocus genotype configuration G . We make the assumption that only one copy of the minor allele was introduced once by a family founder, implying x_{ij} can only take the values 0 or 1 in the absence of inbreeding in the family. The software implementation of RetroFun-RVS converts genotypes homozygous for the rare allele to heterozygous genotypes by default (i.e. $x_{ij}=2$ is changed to $x_{ij}=1$). An alternative option is to discard variants with homozygous rare genotypes. In Schaid et al., 2010, $P(G)$ is the unconditional genotype probability and depends on MAF, which needs to be estimated in practice. Instead, we opted for conditioning the probability on the event of observing at least one copy of each RV j present in the family (i.e., $\sum_i x_{ij} \geq 1$) as in Bureau et al., 2019. In addition, we combined this conditional probability with the assumption that the variant frequency tends to 0, hence the probability does not depend on MAF and therefore the computation does not require external variant frequency estimates. In this context, the genotypes can be interpreted as rare variant sharing patterns, hence RVS in the method name. The sum in the denominator is over all genotype configurations respecting the condition within the given pedigree, where G^* denotes one particular configuration. Since we expect that risk variant effects dominate protective variants in the score test statistic when considering only affected individuals (Supplementary methods, annexes du Chapitre 4 and Figure 33), we propose to adapt the retrospective framework for a burden test (Li and Leal, 2008; Madsen and Browning, 2009). So, we can express β_j the effect of the j^{th} variant through $\beta_0 w_j$ where w_j is usually a weighting function to specify variant effects through a function of MAF. From now on and in the following sections we will consider $w_j = \beta(MAF_j, 1, 25)$ to up-weight rare variants.

As suggested by He et al., 2017, the effect for the j^{th} variant can be partitioned with respect to functional annotations $Z_{jk}, k = 1 \dots q$. So, for the Burden test this leads to:

$$\beta_0 = \sum_j w_j \sum_{k=0}^q Z_{jk} \gamma_k$$

with $Z_{j0} = 1$ and γ_0 corresponding to the original burden test parameter. Intuitively, partitioning variant effect allows a modulation of the variant effect based on MAF and functional annotations. Moreover, γ_0 ensures a minimal power when no predictive functional annotations are present for the trait. When at least one annotation is predictive, the partitioned model offers increased power over the original test (He et al., 2017).

Now combining the retrospective likelihood model described by Schaid et al., 2010 and the decomposed variant effect, we obtain:

$$P(G|Y) = \frac{\exp(\sum_{i \in D} \sum_{j=1}^p w_j x_{ij} \sum_{k=0}^q Z_{jk} \gamma_k) P(G)}{\sum_{G^*} \exp(\sum_{i \in D} \sum_{j=1}^p w_j x_{ij}^* \sum_{k=0}^q Z_{jk} \gamma_k) P(G^*)}$$

Thus, for the k^{th} functional annotation the score function summed across the F families is:

$$S_k(\gamma_k) = \sum_{f=1}^F \left(\sum_{j=1}^p w_j Z_{jk} \left(\sum_{i \in D} x_{fij} - \frac{\sum_{G_f^*} \sum_{i \in D} x_{fij}^* \exp(\sum_{j=1}^p w_j Z_{jk} \gamma_k \sum_{i \in D} x_{fij}^*) P(G_f^*)}{\sum_{G_f^*} \exp(\sum_{j=1}^p w_j Z_{jk} \gamma_k \sum_{i \in D} x_{fij}^*) P(G_f^*)} \right) \right)$$

Intuitively, this quantity can be seen as the difference between the observed genotype value and the expected value, weighted by MAF and functional annotations. Setting γ_k to 0, we obtain the score statistic:

$$S_k(0) = \sum_{f=1}^F \left(\sum_{j=1}^p w_j Z_{jk} \left(\sum_{i \in D} x_{fij} - \sum_{G_{fj}^*} \sum_{i \in D} x_{ij}^* P(G_{fj}^*) \right) \right)$$

The genotype probability required, $P(G_{fj})$, is for a single variant configuration in family f and can be computed using RVS (Sherman et al., 2019). $S_k^2(0)$ asymptotically follows a χ^2 , when properly scaled by the variance of $S_k(0)$. This variance can be obtained by combining sharing patterns and observed genotypes within families. Moreover, simplifications may be obtained from assumptions on the linkage disequilibrium structure (See Supplementary methods, annexes du Chapitre 4). However, we observed when only a few variants are expected within a functional annotation, that resampling procedures may be required to control the Type I error rate. We proposed to resample observed variant counts based on family-specific genotype configuration probabilities $P(G_{fj})$. Thus, this bootstrap procedure maintains both the linkage disequilibrium and family structures of the data.

For testing multiple functional scores within a single unified test $H_0: \forall k, \gamma_k = 0$ vs $H_1: \exists k, \gamma_k > 0$, we then propose to combine $q + 1$ single p-values corresponding to the q functional annotations and the original burden with ACAT (Liu et al., 2019). Briefly, ACAT aggregates individual p-values and approximates the test statistic (and the subsequent p-value) based on a Cauchy distribution. So, for $q + 1$ tests in a region of interest, the ACAT statistic can be written as:

$$T_{ACAT} = \sum_{k=0}^q \tan((0.5 - p_k)\pi)$$

4.5 Numerical Simulations

Genotypes were simulated based on observed variant sites and their corresponding MAF for the European population from the 1000 Genome Project database (phase 3). We extracted the 510 rare (MAF $\leq 1\%$) coding non-synonymous and within-enhancer non-coding single nucleotide variants from a region of 800Kb (chr1:24100000-24970000), corresponding to a TAD in iPSC-derived neurons. This TAD has been selected since it encompasses four CRHs showing different complexities (two genes-five enhancers (CRH1); two genes-two enhancers (CRH2); one gene-one enhancer (CRH3); one gene-four enhancers (CRH4); See Table 3). Refer to Mangnier et al., 2022 for more details. Using RarePedSim (Li et al., 2015), we generated sequence data for 270 affected subjects in a primary sample of 52 extended and small pedigrees (Figures 7 and 34) and a secondary sample of 81 small pedigrees (Figure 34). For both Type I error rate and power evaluations, the dichotomous phenotypes were assumed to follow a logistic model without covariates, and with a population prevalence of 1%. Details on pedigree structures and simulations are provided in the Supplementary methods. We focused on evaluating the ACAT-combined p-values.

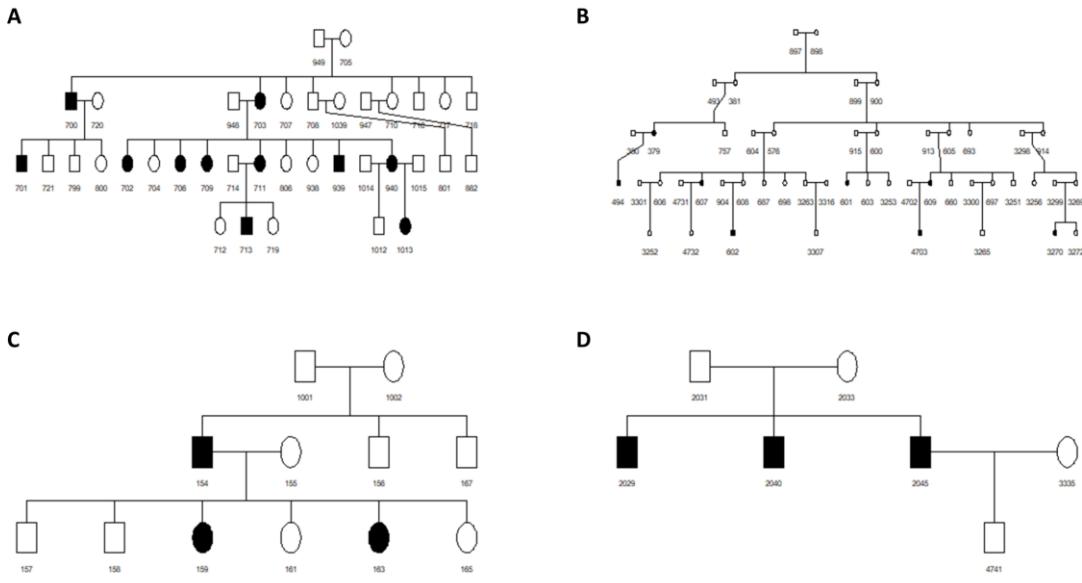


Figure 7: Example of pedigree structures considered in the simulation studies

4.5.1 Type I Error Simulations

To determine whether the proposed framework preserves the desired Type I error rate, genotype data were generated unconditional on the affection status for family members. We specified a null effect for variants observed in families, i.e., odds-ratio (OR) = 1. Generating ten thousand replicates, we first examined the performance of RetroFun-RVS_{CRHs}, which is RetroFun-RVS applied to CRHs and including variants over the entire TAD as global burden, with alternative definitions of regions to be included as functional scores: RetroFun-RVS_{Pairs}, RetroFun-RVS_{Genes}, and RetroFun-RVS_{Sliding-Window}, for the method considering pairs of gene-enhancers, genes and a 10 Kb sliding window, respectively (Figure 8). We also assessed whether the method is well-calibrated in presence of small families. In the alternative scenarios, control of Type I error was assessed generating one thousand replicates. Results for this setting were reported in Supplementary figures.

4.5.2 Empirical Power Simulations

We set 2% of the variants over the entire region to be risk variants as suggested before (Ma et al., 2021), also performing simulations with 1% of risk variants as a sensitivity analysis. Genotypes were generated conditional on the affection status for each pedigree member assuming a multiplicative model with fixed variant effect, i.e., not depending on

the MAF. We considered different scenarios where we varied the proportion of causal variants found in CRHs: 100%, 75% and 50% of causal variants (OR=5) were located within one CRH. The remaining variants being neutral (OR=1). This scenario is expected when variants are concentrated within elements functionally close. These three proportions correspond to the most advantageous scenario where all causal variants are within the same region and two mixed scenarios where signal is spread across the sequence of the region at different degrees. Our first evaluation assessed the gain of power by incorporating CRHs as functional annotations over the test including no scores (referred to as Burden Original). We also compared RetroFun-RVS_{CRHs} with other strategies to incorporate regions as functional annotations: RetroFun-RVS_{Pairs}, RetroFun-RVS_{Genes}, and RetroFunRVS_{Sliding-Window}, for the method considering pairs of gene-enhancers, genes and a 10 Kb sliding window, respectively (Figure 8). An example of functional annotation matrix was given in Figure 9 when considering CRHs. Also, we assessed the performance in terms of power of our method compared to existing approaches namely, RVS (Bureau et al., 2019) and RV-NPL (Zhao et al., 2019) (Figure 35). Power was evaluated as the proportion of p-values less than $\alpha = 8.33e-6$, corresponding to the Bonferroni-adjusted 0.05 significance level when testing six thousand independent regions across the genome, corresponding to three thousand TADs (the average number of TADs found in our previous study across cell-types or tissues (Mangnier et al., 2022)), while permitting the same number of additional domains of interest, i.e., outside TADs, to be tested. Results at lower proportion of risk variants and considering small pedigrees were given in the Supplements.

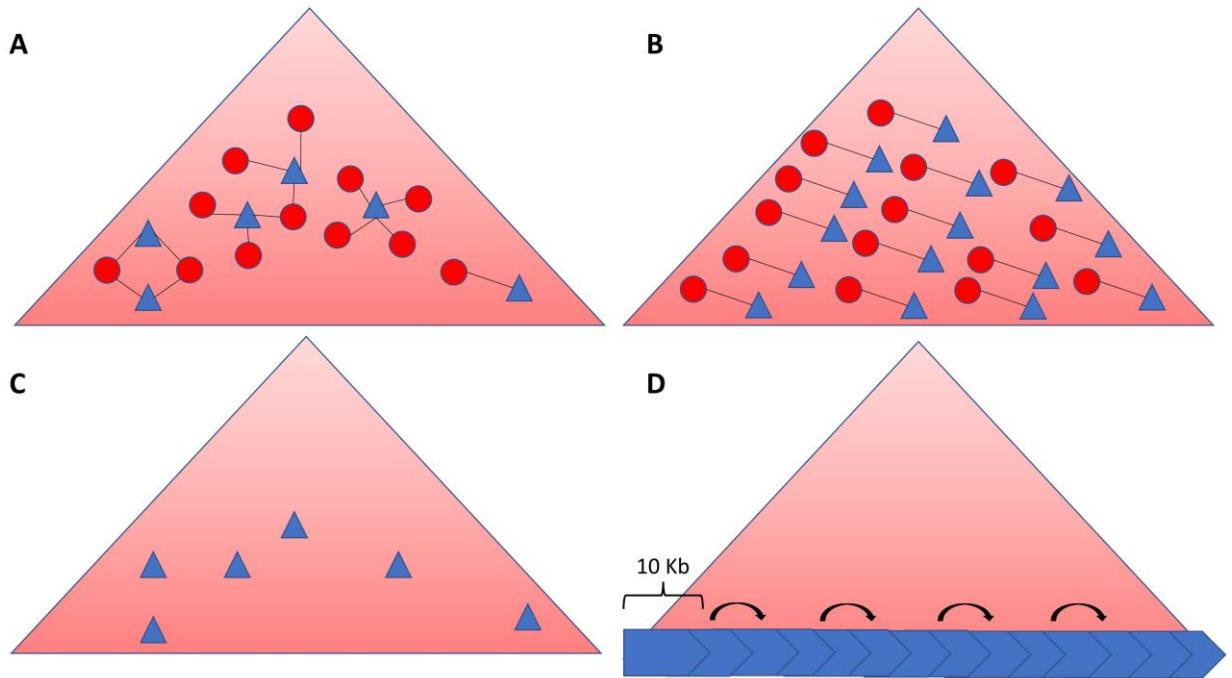


Figure 8: Overview of functional annotations considered in the simulation studies

For all the 4 panels, big red triangles represent the selected TAD for the simulation studies, small blue triangles the genes (exons + promoters), and red circles the enhancers. **(A)** CRHs as functional annotations. **(B)** Pairs as functional annotations. CRHs are split with respect to each gene-enhancer pair. **(C)** Genes as functional annotations. **(D)** 10 Kb sliding windows as functional annotations.

Z

1	1	0	0	0
1	1	0	0	0
1	1	0	0	0
1	1	0	0	0
1	0	1	0	0
1	0	1	0	0
1	0	1	0	0
1	0	0	1	0
.
.
.
1	0	0	0	1

510 x 5

Figure 9: Example of matrix of functional annotations when considering CRHs.

The first column represents the Burden original test, while the 4 other columns CRHs.

4.6 Results

4.6.1 Simulation of Type I Error Rate

The results show that, when we considered CRHs as functional annotations, the Type I error rate was slightly conservative for modest p-values while well controlled for more stringent thresholds, except for the extreme smallest p-value (Figure 10). However, when we considered variants as independent in the variance calculation, Type I error rate was slightly inflated (Figure 36). Moreover, RetroFun-RVS with no functional annotation (i.e., a single test of all RVs in the TAD) was conservative when we considered variant dependence through covariance terms in the variance calculation, while it was well-calibrated assuming variant independence (Figure 37). Results for RetroFun-RVS_{CRHs} for each individual score show that the approach with covariance terms is either well calibrated or slightly conservative (Figures 38-40). In addition, the method shows moderate Type I error rate inflation when applied to small family structures, increasing when assuming variant independence (Figure 41). Furthermore, integrating pairs and genes as functional

annotations, we observed moderate inflation of the Type I error rate in extended pedigrees, even when considering variant dependence, while for 10Kb sliding windows the Type I error rate inflation was severe (Figure 42). We attempted to discard 10 kb windows with few variants, and observed that Type I error control was achieved on windows encompassing 30 variants or more (results not shown). Moreover, the bootstrap procedure applied to RetroFun-RVS_{Pairs}, RetroFun-RVS_{Genes} and, RetroFunRVS_{Sliding-Window} to compute p-values empirically provides good Type I error rate control, while slightly conservative, even for functional annotations encompassing few variants (Figure 43). To summarize, the results show that RetroFun-RVS with asymptotic p-values is a valid approach when CRHs or a large region are considered in extended pedigrees, despite being inflated to various degrees for others strategies or family structures. Bootstrap p-values can be computed in these instances to control the Type I error rate.

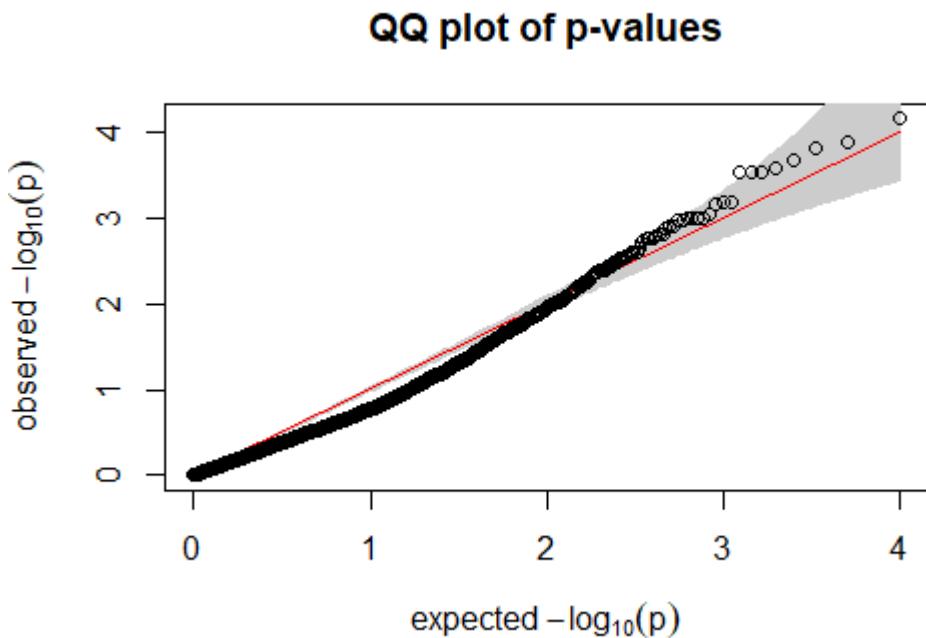


Figure 10: Quantile-Quantile plot of ACAT-Combined P-values for RetroFun-RVS_CRHs considering variant dependence.

Because only a few replicates had p-values for CRH 3, we omitted it in the analysis.

4.6.2 Power Comparison Considering Different Strategies to build Functional Annotations

In the first set of power evaluations, we assessed power under different scenarios of causal variant distributions. Firstly, we compared RetroFun-RVS integrating CRHs with the same method incorporating no functional annotation. Consequently, when 100% and 75% of causal variants were within one CRH, our method RetroFun-RVS_{CRHs} performed better than the original burden test showing gains of 10% and 11%, while at 50% causal the power remains comparable (Figure 11A). Also, considering only small pedigrees, we observed that, even if both RetroFun-RVS_{CRHs} and the original burden test without annotation exhibit lower power, the gain for RetroFun-RVS_{CRHs} becomes higher as the percentage of causal variant within the CRH of interest increases (Figure 44). Congruent results were obtained when a lower proportion of causal variants was considered, showing a minimal power gain of 16% (Figure 45). Therefore, our findings suggest that substantial power gain can be achieved when CRHs are predictive for the effect of variants on the trait, RetroFun-RVS_{CRHs} showing robustness when signal is spread across several CRHs. Then, we compared RetroFun-RVS_{CRHs} to other strategies to integrate regions as functional annotations, namely RetroFun-RVS_{pairs}, RetroFun-RVS_{genes}, and RetroFunRVS_{Sliding-Window}. Our results show that integrating CRHs as functional annotations is a more powerful strategy compared to the other strategies considered (Figure 11B). Globally our results follow the same pattern when decreasing the proportion of causal variants (Figure 46). In summary, RetroFun-RVS_{CRHs} exhibits gain of powers when CRHs show high or modest percentages of causal variants. Also, the method is robust and powerful under the different scenarios that we considered, that are, inclusion of weakly predictive CRHs, small percentages of risk variants, and presence of small families.

4.6.3 Power Comparison with Others Affected-Only Methods

In the second set of power evaluations, we compared RetroFun-RVS_{CRHs} with others affected-only methods, namely RVS (Bureau et al., 2019) and RV-NPL (Zhao et al., 2019). Thus, to proceed to fair comparisons between methods, we adapted RVS and RV-NPL to take CRHs into account (See Supplementary methods, annexes du Chapitre 4). With 2% risk variants, when we considered 75% of causal variants located within one CRHs, we observed that RetroFun-RVS reaches greater power compared to competing methods (Figure 11C), exhibiting significantly shorter computation times (Table 1). At lower

proportions of risk variants, the new method remains more powerful compared to RV-CHP or RVS, and equivalent to RVNPL (Figure 47).

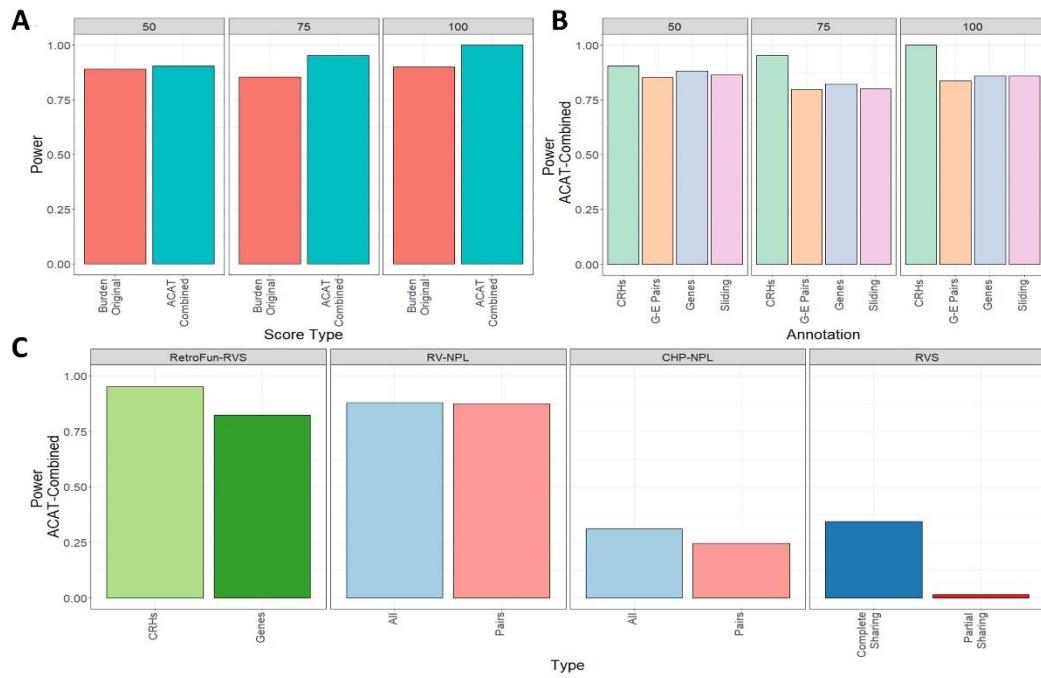


Figure 11: Power evaluation of RetroFun-RVS under different scenarios for 2% risk variants

(A) Power at different proportions of risk variants within the CRH, between RetroFun-RVS_{CRHs} with no functional annotation (Burden Original) and RetroFun-RVS_{CRHs} including the four CRHs (ACAT-Combined). Power was evaluated on the basis of 1,000 replicates.

(B) Power at different proportions of risk variants within the CRH between RetroFun-RVS_{CRHs} (CRHs), RetroFun-RVS_{Pairs} (G-E Pairs), RetroFun-RVS_{Genes} (Genes), and RetroFunRVS_{Sliding-Window} (Sliding). To correct Type I error inflation observed in RetroFun-RVS_{Sliding-Window}, we only considered windows encompassing 30 variants or more. Power was evaluated on the basis of 1,000 replicates. **(C)** Power at 75% risk variants within one CRH between RetroFun-RVS_{CRHs} and other affected-only competing methods. Here we included RetroFun-RV_{genes} to mimic CHP-NPL procedure. Power for RetroFun-RVS_{CRHs} and RetroFunRVS_{Genes} was evaluated on the basis on 1,000 replicates, while for RV-NPL and RVS we generated 200 replicates.

RetroFun-RVS				RV-NPL	CHP-NPL	RVS	
CRHs	G-E Pairs	Genes	Sliding	All + Pairs	All + Pairs	Complete	Partial
1.06	2.02	1.11	11.05	971.4	1823.4	14.26	443.5

Table 1: Running times (in seconds) for analyzing rare variants in the TAD, in one simulated replicate, using a single 2.10GHz processor

For RV-NPL empirical p-values were obtained based on 1 million permutations.

4.7 Discussion

Most of rare genetic variations are located within non-coding regions, making the underlying biological mechanisms through which they impact disease risk difficult to interpret. Over the past few years, efforts were not only made in annotating the genome but also integrating these annotations into statistical frameworks (He et al., 2017; Ma and Wei, 2019). Although such methods have already been developed for unrelated subjects such as case-control samples, to our knowledge, no approach to date has been proposed to integrate functional annotations within family-based designs. In this paper we have presented RetroFun-RVS, a retrospective burden test, integrating functional annotations considering only affected individuals within families. We have shown that binary annotations corresponding to disjoint regions with regulatory impacts, such as CRHs, provide power gains when such regions concentrate causal variants, outperforming other strategies or competing methods (Figure 11), while well controlling the Type I error rate (Figure 10). Since regulatory mechanisms are highly tissue- or context-dependent it can be challenging to have the right tissue for the right trait, and misspecifying the model is likely in practice. Thus, integrating the original burden test in RetroFun-RVS makes it robust, showing stable power when functional annotations poorly predict the trait. Finally, by computing p-values asymptotically, RetroFun-RVS is computationally faster than competing methods, which often require permutation-based approaches or exact probability computations to sharply control the Type I error rate.

The main rationale behind RetroFun-RVS is that risk variants are enriched among affected individuals compared to the expected variant count based on their relationships. Hence, one critical feature of our method is to aggregate genotypes while measuring rare variant sharing among affected family members to compute the test statistic. However, to implement an affected-only analysis, where individuals are selected based on their disease

status, we have adopted a retrospective approach, considering genotypes as random, while conditioning on phenotypes (Schaid et al., 2010). Also, since genotype probabilities do not depend on MAF under the assumption that the variant frequency tends to 0, RetroFun-RVS necessitates only familial information to compute these probabilities, in order to derive the score statistic and its variance (See Material and Methods, annexes du Chapitre 4). This aspect is central, since the variance terms need to be computed only once for the entire set of families, which is computationally efficient even in presence of large pedigrees. Our rare variant assumption however implies that genotypes homozygous for the rare allele are impossible in the absence of inbreeding. Data simulated for Type I error and power assessments did contain the small number of homozygous rare genotypes expected for variants with $MAF = 1\%$. Conversion to heterozygous genotypes did not increase Type I error rate compared to removing the variants with homozygous rare genotypes (results not shown).

Moreover, RetroFun-RVS in its current form is restricted to binary phenotypes and does not allow the integration of individual-level covariates, such as sex, age or genetic principal components. Hence, future work is needed to extend the framework to continuous phenotypes and include covariates. Also, future works are needed to extend RetroFun-RVS when more than one copy of the minor allele is introduced, which can arise in presence of inbreeding.

In addition to being computationally effective, RetroFun-RVS is more powerful than other affected-only competing methods, under certain scenarios (Figure 11C, Figure 47). For example, compared to RVS, on which RetroFun-RVS is built upon, but which can only analyze between one and five rare variants simultaneously in the pedigree sample used in the simulation study, we reached greater power by testing tens of variants together in annotated regions, or even hundreds of variants in the absence of annotations. Although the test is well-calibrated and powerful for extended pedigrees, we have demonstrated that it performs well when applied to small family structures, with modest Type I error rate inflation (Figure 41). It is noteworthy that the simulated variant ORs did not depend on the variant MAF due to limitations of the simulation software. The MAF-dependent variant weighting scheme of RetroFun-RVS was thus misspecified in the power evaluation. Greater power gains of RetroFun-RVS over the competing methods ignoring variant MAF could have been achieved had the variant ORs be inversely related to MAF. Some analyses have shown that Type I error rate or power are highly dependent on the number

of variants present in the region of interest. Indeed, we have observed that when large numbers of variants are considered, RetroFun-RVS might provide conservative results involving some power loss (Figure 35), while a small number of variants tends to offer inflated Type I error rate (Figure 42). Complementary analyses are needed to inspect the empirical relationship between size of region and performance. Therefore, in the meantime we recommend in practice to use the covariance-adjusted model. Finally, bootstrap procedures (Figure 43) might be considered to sharply control type I error rate for small numbers of variants at the expense of longer computing time.

We argue that the performance of the proposed method is strongly dependent to the availability of the relevant tissue for the studied disease. Indeed, regulatory mechanisms operate in a tissue- or cell-type-specific manner. Our framework, by allowing the incorporation of several functional annotations from diverse tissues or cell-types without loss of power, is useful to highlight the underlying biological mechanisms involved in the trait. This aspect is central from a fine-mapping perspective, thus RetroFun-RVS will be an important tool to pinpoint causal variants located within non-coding regions, which could have been missed so far.

4.8 Data Availability

Cis-Regulatory Hubs and Topologically associated domains used in this paper are available on <https://github.com/lmangnier/CRHs>. Variant data were available from the 1000 Genome project: <https://www.internationalgenome.org/dataportal/data-collection/phase-3>. We have implemented RetroFun-RVS in a R package, available on GitHub (<https://github.com/lmangnier/RetroFun-RVS>). The code for the simulations is also available on GitHub (https://github.com/lmangnier/Simulation_RL).

4.9 Funding

Some data analyses were performed on computing resources from Compute Canada. This work was partly funded by the Canadian Statistical Sciences Institute through a Collaborative Research Team grant and by a National Science and Engineering Research Council of Canada Discovery grant to A Bureau.

4.10 Conflict of Interest

The authors declare that they have no conflict of interest.

Chapitre 5 : RetroFun-RVS, un package R pour l'analyse de variants rares dans les familles, permettant l'intégration d'annotations fonctionnelles.

Dans le chapitre précédent, nous avons démontré à travers une étude de simulation que *RetroFun-RVS*, en intégrant la structure familiale en complément des pôles de régulation cis en tant qu'annotations fonctionnelles était une stratégie rapide et puissante pour détecter des variants causaux. En effet, la méthode a permis d'atteindre des puissances plus élevées par rapport aux autres stratégies pour intégrer des annotations fonctionnelles ou méthodes compétitives. Ainsi, dans l'optique de généraliser la méthode et de l'appliquer à des données réelles, nous avons développé et mis à disposition un package R, appelé *RetroFun-RVS*. Pour illustrer les fonctions centrales du package, nous avons donc sélectionné des données familiales de séquençage du génome complet de fentes labiales (Bureau et al., 2019). L'ensemble du code utilisé dans ce chapitre a été rendu disponible sur Github (<https://github.com/lmangnier/RetroFun-RVS>).

5.1 Conception et implémentation

Nous présentons ici les fonctions et étapes centrales du package.

5.1.1 Type de données

RetroFun-RVS accepte en entrée différents types de données : des données génétiques au format .ped, des fichiers d'annotations fonctionnelles et des fichiers de pondération pour chaque variant.

Les données de type pedigree au format ped résument les données génétiques dans un format standard (Plink) (Purcell et al., 2007). Les six premières colonnes sont définies respectivement, par l'identifiant de la famille, l'identifiant de l'individu, l'identifiant du père,

l'identifiant de la mère, le sexe ainsi que le statut (malade ou sain). Le reste des colonnes (2 fois le nombre de variants) représentant les deux allèles pour chaque variant, égale à 1 lorsque l'individu porte l'allèle de référence et 2 lorsqu'il porte l'allèle alternatif.

Les matrices d'annotations fonctionnelles de taille p (variants) * q (annotations fonctionnelles) peuvent être des matrices continues ou binaires. La première colonne doit être égale à 1, afin d'intégrer le test original.

Les poids pour les variants sont des vecteurs de taille p (variants) * 1.

De plus, le package nécessite de précalculer les valeurs attendues des génotypes pour chaque famille, ainsi que les variances et covariances associées, stockées dans un objet du type *data.frame*. Cette information peut être obtenue grâce à *RVS* (Sherman et al., 2019).

5.1.2 Pré-traitement

La fonction *aggregate.geno.by.fam* prend alors en entrée un fichier .ped et effectue les étapes suivantes :

- 1) Combine les allèles par variant pour chaque individu.
- 2) Retire les individus sains pour ne garder que les individus malades dans les familles.
- 3) Filtre les variants présents dans plus de la moitié des individus (variants non rares) et procède à quelques étapes de nettoyage. Le remplacement ou la suppression des variants avec au moins un sujet homozygote peut être défini par l'utilisateur. Un sous-ensemble de familles d'intérêt peut aussi être sélectionné.
- 4) Agrégation des variants par famille.

Retourne une liste avec les variants agrégés par famille et l'indice des variants.

5.1.3 Obtention des valeurs-p

La fonction *RetroFun-RVS* est la fonction centrale du package et permet d'obtenir les valeurs-p combinées et individuelles pour chaque annotation fonctionnelle, au niveau d'une région génomique donnée. La fonction prend en entrée un *data.frame* avec les valeurs attendues, variances et covariances pour chaque famille, la liste retournée par *aggregate.geno.by.fam*, ainsi que la matrice d'annotations fonctionnelles et la matrice de pondération des variants.

La statistique du score, ainsi que la variance associée sont obtenues grâce aux fonctions *compute.Burden.by.Annot* et *compute.Var.by.Annot*. L'indépendance des variants peut être spécifiée par l'utilisateur dans le calcul de la variance du score.

L'objet retourné est alors une liste avec les valeurs-p pour chaque annotation, ainsi que les valeurs-p combinées par ACAT (Liu et al., 2019) ou par la méthode de Fisher.

5.2 Application aux données de fentes labiales

5.2.1 Présentation des données

Les données de séquençage du génome complet ont été générées chez 160 individus dans 54 familles étendues provenant des Philippines, États-Unis, Guatemala et Syrie. Les données comportent essentiellement des paires ou triplés d'individus apparentés proches (Voir Table 2). Au total 153 individus atteints ont été séquencés pour 7 non-atteints. Cependant, parce que pour 8 individus d'une famille syrienne, aucun ancêtre commun n'était séquencé, un sous-ensemble de 6 individus atteints apparentés a été inclus à la place, portant le total à 151 atteints. Le séquençage, l'alignement et l'appel de variants ont été réalisés suivant la méthodologie présentées dans Holzinger et al. (2017). Ensuite, les variants d'épissage, non-synonymes et non-sens présents dans les exons des gènes, ainsi que l'ensemble des variants non-codants ont été extraits. Les données ont été filtrées suivant la même procédure que Bureau, Younkin et al. (2014), permettant une fréquence maximale d'allèle mineur de 1% dans les échantillons de populations provenant du projet des 1000 Genomes (<https://www.internationalgenome.org/data>), portant le nombre total de variants considérés dans l'analyse à 3 797 938. De plus, en lien avec la méthodologie présentée dans le chapitre 4, les variants présents dans plus de la moitié des individus ont été retirés, aussi les variants enregistrant au moins un individu homozygote ont été supprimés.

Nombre d'atteints	2	3	4	5	6
Fréquence	33	5	11	4	1

Table 2: Répartition des structures familiales dans les données de séquençage de fentes labiales du génome complet

Différentes familles illustrant différentes structures familiales sont données en Figure 12.

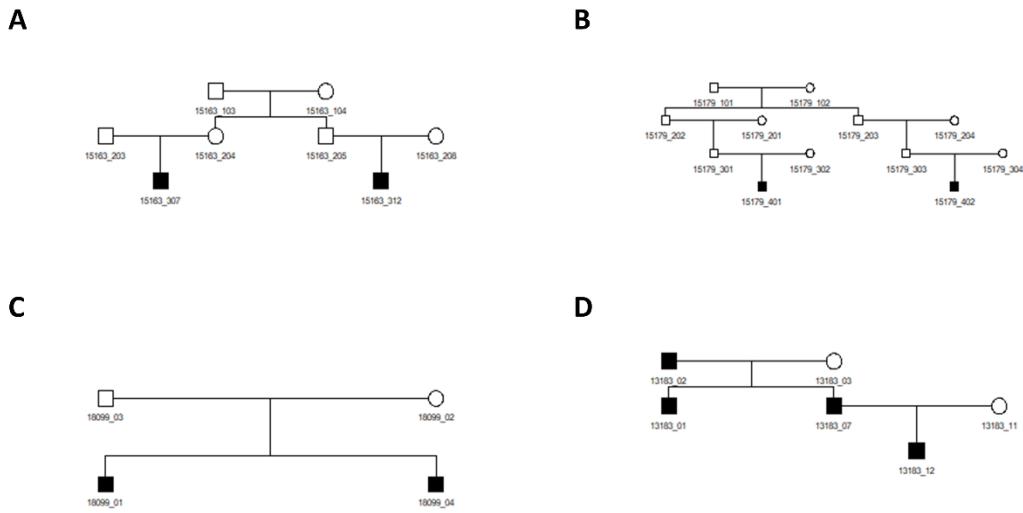


Figure 12: Example des structures familiales présentes dans les données de séquençage du génome complet de fentes labiales

En complément, les pôles de régulation cis utilisés dans ce chapitre ont été générés dans les cellules épithéliales humaines à une résolution de 10Kb. À noter que les domaines topologiquement associant avaient déjà été générés par Rao et al. (2014). En effet, il a été démontré que certains variants rares localisés dans des régions régulatrices actives dans ce tissu sont associés à l'émergence de fentes labiales chez l'humain (Schaffer et al., 2019). Les données Hi-C (Rao et al., 2014) (GSE63525), ainsi que les marques d'histones H3K27ac et données d'accessibilité de la chromatine ont été utilisées (ENCSR460EGF) pour générer les pôles de régulation cis. Ainsi, 329 domaines topologiquement associant ont été retenus dans l'analyse. Par ailleurs, dans un souci de simplification de l'analyse 361 pôles de régulation cis ne chevauchant qu'un seul domaine ont été considérés, représentant le tiers du nombre total des réseaux (361/1125). Ces derniers mettant en lumière des organisations variées (Figure 13).

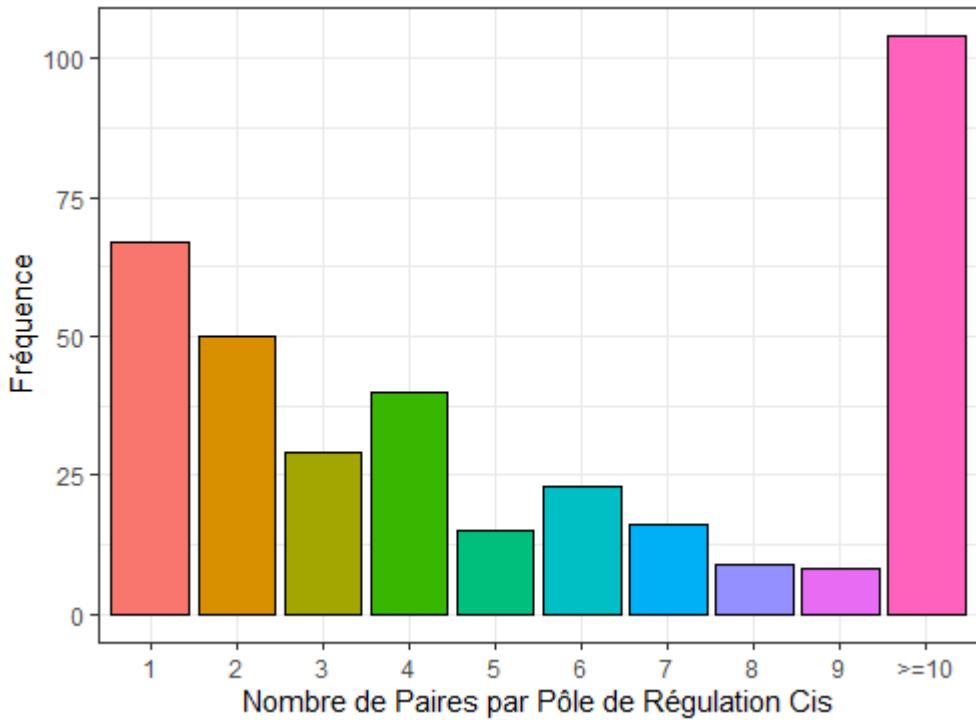


Figure 13 : Répartition du nombre de paires gène-enhancer par pôle de régulation cis dans les cellules épithéliales humaines

Les matrices d'annotations ont été générées de telle sorte que $\forall j : I(v_j \in Z_k)$. En d'autres termes, la matrice d'annotations fonctionnelles est une matrice binaire égale à 1 lorsque le variant j appartient au pôle de régulation k et 0 sinon (Voir Figure 9 du Chapitre 4). Par ailleurs, l'hétérogénéité ethnique des familles retenues dans l'analyse implique des structures génétiques différentes, rendant l'estimation des fréquences d'allèles mineurs difficile. Par conséquent, l'approche non-pondérée sera ici considérée. Les valeurs-p combinées par ACAT seront calculées au niveau de chaque domaine, tel que présenté dans le chapitre 4. Nous corrigerais pour la multiplicité, ajustant par la méthode de Bonferroni au seuil $\alpha=0.0001$ ($0.05/361$).

5.2.2 Résultats

Tel que discuté dans la section précédente, nous avons prétraité et formaté les données grâce à la fonction `agg.genos.by.fam` (Voir ligne de code 1). Nous retrouvons en moyenne dans les TADs 140.2 variants (écart-type = 105.07), pour un nombre moyen de variants par familles de 5.23 (écart-type = 5.09). Par ailleurs, les pôles de régulation cis mettent en lumière un nombre moyen de variants de 3 (écart-type = 4.27) où 44% sont situés dans les enhancers.

```
Res = agg.genos.by.fam(pedfile,correction = "remove")
```

Ligne de code 1

Ensuite, tel que présenté dans le chapitre précédent, *RetroFun-RVS* repose en partie sur le package R *RVS* (Sherman et al., 2019) notamment dans le calcul des valeurs attendues par famille et les variances/covariances associées à l'aide de la fonction *compute.null* (Voir ligne de code 2) appliquée sur l'ensemble des configurations possibles et les probabilités de partage renvoyées par la fonction *RVsharing* de *RVS*.

```
Null = compute.null(pedigree.configurations, pedigree.probas)
```

Ligne de code 2

Parce que de la consanguinité est attendue dans les familles syriennes, 14 familles ont été supprimées ne gardant que les 40 familles restantes dans l'analyse principale. Cependant, les résultats considérant l'ensemble des familles seront donnés en annexes (Annexes du Chapitre 5). En effet, la présence de consanguinité ou de relations cryptiques entraîne la violation de certaines hypothèses derrière *RVS* (introduction d'une seule copie de l'allèle mineur). Ces points seront discutés plus tard dans la discussion.

Après l'obtention des valeurs attendues et de la variance dans l'ensemble des familles, *RetroFun-RVS* a été appliquée dans les 329 domaines topologiquement associant et en considérant les 361 pôles de régulation cis en tant qu'annotations fonctionnelles, considérant la dépendance entre variants (Voir ligne de code 3).

```
RetroFun_RVS(Null, Res, Z, W, independence=F)
```

Ligne de code 3

Bien que la méthode offre un bon contrôle de l'erreur de type I (Figure 14A), nous n'avons pu identifier aucun domaine pour lequel des valeurs-p inférieures à 0.0001 étaient enregistrées. Cependant au sein du domaine mettant en lumière la plus forte association avec le phénotype (valeur-p combinée= 0.0003; Figure 14B), le signal est généré par un pôle de régulation cis contenant 11 gènes et 2 enhancers pour un total de 52 variants rares analysés (valeur-p au niveau de l'ensemble du domaine = 0.36; valeur-p au niveau du pôle de régulation cis = 0.0001).

5.2.3 Discussion

L'incapacité de la méthode à détecter du signal au seuil de significativité considéré peut provenir par exemple (1) d'un nombre d'atteints par famille trop faible, (2) d'un nombre de régions trop faible à tester permettant de détecter du signal ou (3) du choix d'un tissu ou type cellulaire peu pertinent à la maladie. Fort de ce constat d'autres stratégies d'analyse doivent être considérées dans des travaux futurs dans l'optique d'avoir une compréhension plus fine de l'étiologie du trouble. De plus, en présence de consanguinité ou de relation cryptique, comme c'est le cas ici dans les familles syriennes, une extension du modèle permettant notamment la distinction entre la présence d'une ou deux copies de l'allèle mineur doit être considérée pour adresser cet aspect.

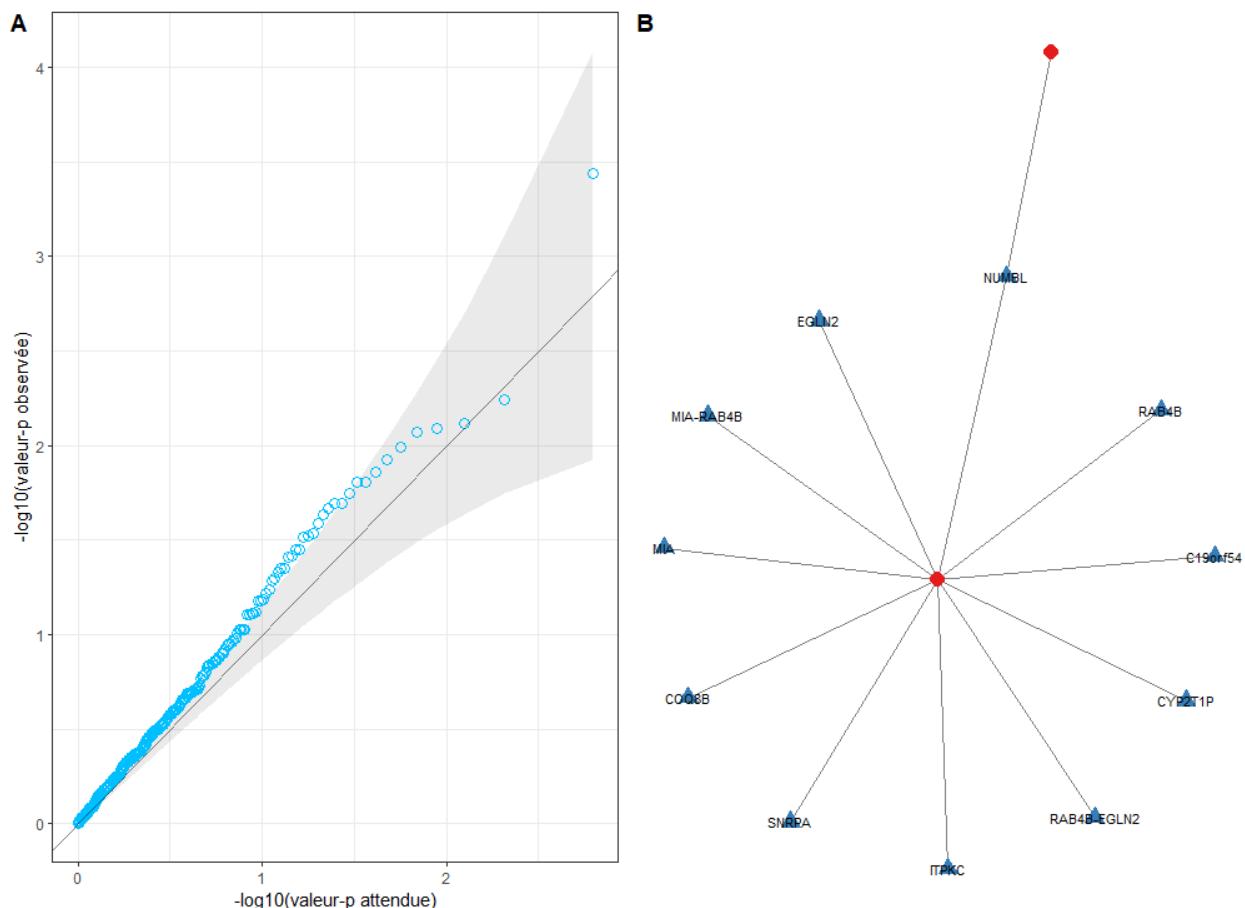


Figure 14: Résultats de RetroFun-RVS dans les données de fentes labiales, retirant les familles syriennes de l'analyse

(A) Diagramme quantile-quantile des valeurs-p combinées avec ACAT considérant les 40 familles non-syriennes. Les résultats sont rapportés en $-\log_{10}$. **(B)** Pôle de régulation cis sur

le chromosome 19 le plus associé avec les fentes labiales, impliquant 11 gènes (triangles bleus) et deux enhancers (cercles rouges).

5.3 Conclusion

RetroFun-RVS est un package R simple d'utilisation permettant d'intégrer des annotations fonctionnelles dans un test d'association de variants rares appliqué aux familles. L'obtention des valeurs-p est rapide et les résultats obtenus intuitifs. En effet, nous avons montré que dans le cas d'annotations fonctionnelles correspondant à des régions fonctionnellement valides, il est facile de cibler la ou les annotations fonctionnelles enrichies en variants causaux et ainsi gagner en compréhension dans les mécanismes biologiques impliqués dans les maladies.

Chapitre 6 : Discussion

La démocratisation des données de séquençage de l'ADN de haut débit à l'échelle du génome a permis de mettre en lumière le profil épigénétique de nouvelles régions. En effet, grâce aux technologies de DNase-Seq, d'ATAC-Seq ou de ChIP-Seq nous sommes désormais capables de cibler les régions associées avec des phénomènes de régulation. En parallèle, l'avènement et la diffusion des technologies de Capture de Conformation de la Chromatine (3C) ont ouvert de nouvelles voies de compréhension sur le rôle joué par l'organisation 3D du génome dans la régulation des gènes et plus largement dans l'émergence de maladies. La mise à disposition à grande échelle des données d'épigénétique a permis le développement d'outils permettant de caractériser plus finement les régions non codantes (Fulco et al., 2019). Dans un contexte de maladies, cette information est centrale pour élucider les mécanismes de régulation impliqués dans l'émergence ou le développement de maladies chez l'humain. Bien que des approches statistiques intégrant de l'information fonctionnelle ont déjà été proposées (He et al., 2017; Ma & Wei, 2019; Ma et al., 2021), aucune méthode à ce jour n'est applicable aux devis familiaux, lorsque ces derniers sont plus puissants que les approches populationnelles pour mettre en lumière le partage de variants rares. Fort de ce constat, dans la perspective d'élucider les mécanismes épigénétiques par lesquels les maladies peuvent émerger, nous avons tout d'abord proposé un nouveau modèle 3D de réseaux complexes entre gènes et enhancers, appelé pôles de régulation cis. Cette nouvelle structure 3D a été démontrée informative dans l'étiologie de maladies complexes, telles que la schizophrénie. Ainsi, les pôles de régulations cis représentent une ressource pertinente pour détecter de nouveaux variants causaux impliquées dans les traits complexes. Sur cette base, nous avons proposé *RetroFun-RVS*, une méthode statistique rétrospective permettant l'intégration de l'information fonctionnelle sous forme de réseaux 3D et incorporant la structure familiale. Une caractéristique importante de la méthode est qu'elle exploite seulement les individus atteints dans les familles, dans une optique de maximiser le partage de variants potentiellement causaux. En comparaison des autres stratégies pour inclure des annotations fonctionnelles ou des méthodes compétitives, *RetroFun-RVS* en intégrant les pôles de régulation cis a été démontrée plus puissante pour détecter des variants de risque. Aussi nous avons démontré la robustesse de la méthode aux annotations non informatives, caractéristique intéressante en pratique. Finalement, en combinant approches familiales et

information fonctionnelle sous forme de réseaux 3D d'éléments de régulation actifs, nous sommes désormais capables non seulement de déterminer quel variant ou quelle région est associé à une maladie, mais aussi d'avoir une compréhension plus fine des mécanismes biologiques impliqués. À la lumière des limites relevées pour les méthodes présentées dans cette thèse ou à travers leur application à des données réelles, nous proposons ici quelques extensions des modèles dans une logique d'une plus large applicabilité en pratique.

6.1 Extension des pôles de régulation cis, perspectives et limites

En tant que réseaux combinant contacts 3D, accessibilité de la chromatine et facteurs de transcriptions, les pôles de régulation cis peuvent être considérés comme des réseaux multi-couches mettant en lumière les interactions entre enhancers actifs et gènes. Cependant, lorsque la dynamique des maladies peut résulter de contacts entre promoters, enhancers ou impliquant d'autres types de régions régulatrices (silencers par exemple), intégrer d'autres types d'interactions paraît essentiel dans une perspective mécanistique, notamment sur la compréhension des phénomènes de régulation impliqués. En effet, certains auteurs ont démontré que le profil d'interaction entre promoters ou enhancers pouvait être impliqué dans l'émergence de maladies complexes (Madsen et al., 2020; Song et al., 2020). Aussi, pour un même type d'élément de régulation, il peut exister différentes sous-typologies d'éléments pouvant résulter à des implications variées en termes de régulation (enhancers actifs et enhancers « positionnés » par exemple). Ainsi, avoir des réseaux ne se limitant pas qu'aux enhancers actifs, mais intégrant contacts entre promoters, enhancers, intégrant d'autres types d'éléments de régulation ou différents sous-types d'éléments permettrait d'avoir une compréhension plus large des phénomènes épigénétiques impliqués dans l'émergence des maladies. D'un point de vue pratique, l'emploi de marques d'histones spécifiques aux éléments de régulation considérés (H3K4me3 pour les promoters, H3K27me3 pour les silencers ou H3K4me1 pour les enhancers « positionnés ») permettrait d'étendre les pôles de régulation cis respectivement, aux contacts promoters-promoters, promoters-silencers et promoters-enhancers « positionnés ». Cependant, intégrer les contacts entre enhancers requiert alors la modification du modèle actuel, centré autour du gène au profit d'une approche orientée autour du enhancer (modèle « enhancer-centric »). Sur ce point, Hecker et al. (2022) ont proposé une extension du score d'activité par contacts (Fulco et al., 2019) permettant

d'intégrer la dimension « enhancer-centric » des phénomènes de régulation. Bien que prometteur ce modèle demande cependant à être validé expérimentalement.

Les phénomènes de régulation sont connus pour être des événements dynamiques mettant en jeu différents éléments à différents temps données (Voss & Hager, 2014). Par exemple, de nouvelles méthodes de microscopie ont démontré que les interactions gènes-enhancers sont des processus dynamiques précédant notamment la transcription des gènes (Espinola et al., 2021). Aussi, il y a un attrait croissant des chercheurs pour l'intégration d'information biologique tenant compte des différents processus opérant à différentes échelles (appelés omiques, ex : protéomique, transcriptomique, métabolomique, etc...). Des modèles « multi-omiques » intégrant la nature longitudinale des phénomènes biologiques ont déjà été proposés (Bodein et al., 2019). Cependant bien que ces modèles aient mené à des interprétations plus fines des mécanismes en jeu (Bodein et al., 2022), ces approches souffrent de deux limites majeures : (1) ils n'incluent par exemple pas la dimension épigénétique à travers les contacts 3D et (2) n'ont pas été évalués dans des contextes de maladies. Fort de ce constat, dans une logique d'avoir une compréhension plus fine des événements biologiques intervenant à différents niveaux « omiques », une extension naturelle des pôles de régulation cis pourrait impliquer l'ajout de couches supplémentaires d'information notamment longitudinales dans l'étude de maladies chez l'humain.

Par ailleurs, une des limites des pôles de régulation cis tel que proposé dans cette thèse est de n'intégrer que des données « en vrac », c'est-à-dire là où les contacts 3D observés sont la résultante de l'amalgame de différents contacts au niveau de plusieurs cellules. Bien que permettant d'atteindre des résolutions plus fines, certaines interactions gènes-enhancers peuvent résulter de contacts opérant dans des cellules différentes, formant des réseaux « artefacts ». Une extension naturelle des pôles de régulation cis aux données du type « single-cell » (Nagano et al., 2013), permettrait d'adresser cette limitation technique. Brièvement, les technologies « single-cell » sont une nouvelle famille de méthodes de séquençage permettant de capturer les phénomènes biologiques (expression, contacts 3D, marques d'histone ou accessibilité de la chromatine) spécifiques à la cellule dans lesquels ils opèrent. Par exemple pour le 3D, cette dimension cellulaire implique un signal souvent « creux » nécessitant la mise au point de méthodes permettant d'agglomérer les contacts survenant dans des cellules différentes. Des méthodes de « clustering » ont été proposées pour adresser cette limitation (Zhou et al., 2019; Zhang et al., 2022). En

intégrant ce type de données dans la détection des enhancers actifs et de manière subséquente dans la formation des pôles de régulation cis, on permettrait la détection des interactions entre enhancers et promoters de manière plus précise, tout en limitant la formation de réseaux non représentatifs d'évènements épigénétiques. Sur cet aspect, Hecker et al. (2022) ont étendu le score d'activité par contact aux données « single-cell » d'accessibilité de la chromatine.

Enfin, comme nous venons de le voir, l'extension des pôles de régulation cis nécessite l'emploi de nouveaux types de données dans une optique d'affiner notre compréhension des mécanismes biologiques impliqués dans les maladies. Cette question de l'accessibilité de la donnée est fondamentale en épigénétique mais plus largement en biologie. De par leur nature spécifique au tissu ou au contexte, caractériser plus finement les phénomènes épigénétiques en ajoutant des couches supplémentaires d'information nécessite d'avoir les bonnes données dans le bon tissu, entraînant des problématiques de coûts lorsque le tissu d'intérêt est difficile à obtenir ou que le séquençage se fait de manière longitudinale. Même si à travers des initiatives comme le Roadmap Epigenomics Consortium ou ENCODE, un large éventail de données est disponible dans un spectre étendu de tissus, ces données ne se limitent pour l'instant qu'à des données en « vrac » et la mise à disposition de données « single-cell » n'est pas tout à fait généralisée.

6.2 Les pôles de régulation cis, une utilisation en « cartographie fine » ?

En combinant approches familiales et annotations fonctionnelles, *RetroFun-RVS* permet de mettre en lumière les régions fonctionnellement actives associées avec la maladie d'intérêt. Au-delà de la simple perspective descriptive, la méthode pourrait servir de base pour localiser le ou les variants de risque situés dans les régions non codantes et ainsi résoudre les mécanismes biologiques impliqués. En effet, une des limites majeures des approches actuelles, e.g., score de déséquilibre de liaison (Bulik-Sullivan et al., 2015) est de ne reposer que sur les statistiques récapitulatives des études d'association pangénomiques, avec le désavantage majeur d'être en incapacité de révéler le ou les variants causaux. En effet, il a été démontré que du fait de limitations techniques et méthodologiques, la plupart des variants significativement associés avec un trait d'intérêt ne sont pas ceux menant à la maladie (Broekema et al., 2020). Dans cette logique, l'intérêt de plus en plus croissant pour les approches de « cartographie fine » ont permis d'améliorer notre compréhension de

l'implication de certains variants potentiellement causaux pour un large spectre de maladies (Schaid et al., 2018). Ainsi, la méthode proposée pourrait servir de base pour affiner notre compréhension des phénomènes de régulation résultant à l'émergence ou au développement de maladies chez l'humain. L'intuition derrière une approche de « cartographie fine » est alors de restreindre l'ensemble possible des variants incluant le ou les variants causaux. Bien que les approches traditionnelles reposent sur une hypothèse irréaliste, qu'un seul variant est causal dans la région associée, des méthodes alternatives permettent de tenir compte de la potentialité que plusieurs variants soient causaux (Hutchinson et al., 2020). Pour se faire, les modèles intègrent, sur la base de statistiques d'études d'association pangénomique, un large éventail d'informations externes incluant, déséquilibre de liaison, ascendance et annotations fonctionnelles, le plus souvent dans des approches bayésiennes (Schaid et al., 2018). Dans cette logique Paintor, en permettant l'intégration de régions fonctionnellement actives, a permis une réduction substantielle des intervalles de crédibilité tout en révélant de nouveaux variants impliqués dans des traits métaboliques (Kichaev et al., 2014).

Fort de ce constat, les pôles de régulation cis, en liant gènes et enhancers actifs à l'intérieur de réseaux 3D, pourrait permettre d'identifier le ou les gènes impactés par la présence de variants non-codant localisés dans des régions fonctionnellement actives. Intégrer les pôles de régulation cis dans des approches de « cartographie fine » serait une approche pertinente pour mettre en lumière de nouveaux variants impliqués dans les maladies. À ce jour, aucune annotation fonctionnelle reposant sur les réseaux 3D n'a été incluse dans de telles approches. Du point de vue du gène, cet aspect est central, lorsque des variants situés dans des enhancers éloignés peuvent par des phénomènes de « réaction en chaîne », impacter le gène d'intérêt. De plus, une récente étude de Bergman et al. (2022) a démontré que le profil d'interactions des gènes avec les enhancers pouvait être prédictif de leur réponse aux phénomènes de régulation. Cette notion rejoue celle initialement proposée par Tsai et al. (2019) où les auteurs affirment que la fréquence d'interaction d'un gène avec les enhancers est associée avec une certaine robustesse aux stimuli externes. Cette composante pourrait permettre d'affiner les pôles de régulation cis pour permettre de cibler les gènes les plus susceptibles d'être impactés par des dysfonctionnements dans leurs éléments de régulation et ainsi élucider le ou les variants impliqués dans les maladies.

6.3 Vers une plus large applicabilité de RetroFun-RVS

6.3.1 Extension de RetroFun-RVS, le traitement de familles consanguines

Comme nous l'avons remarqué dans le chapitre cinq, l'application de *RetroFun-RVS* aux données de séquençage du génome complet de fentes labiales (Bureau et al., 2019) a permis de mettre en lumière des pôles de régulation cis associé avec le trait. Cependant, le signal détecté peut provenir de la présence de faux positifs pouvant résulter de différents phénomènes : la présence de familles avec peu d'atteints, un tissu ou type cellulaire peut pertinent pour le trait en question, la présence de consanguinité au sein des familles ou de relations cryptiques. Sur ce dernier point, le modèle de partage tel que présenté par Bureau et al. (2019) peut être modifié en permettant l'introduction dans les familles de plus d'une copie de l'allèle mineur. Une extension distinguant le partage d'une ou de deux copies de l'allèle permettrait de passer outre les hypothèses actuelles du modèle en vue de sa généralisation en présence de consanguinité ou de relations cryptiques.

6.3.2 Extension de RetroFun-RVS intégrant des covariables

D'un autre côté, il y a un intérêt croissant pour les chercheurs à considérer et intégrer la notion d'ascendance dans les études d'association pangénomiques ouvrant de nouveaux champs de possibilité dans la compréhension des maladies (the EArly Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium, 2015; Peterson et al., 2019). En effet, il a été démontré que l'ascendance joue un rôle critique dans la différence observée entre les structures génétiques de populations diverses. Par conséquent, ne pas tenir compte de cette information peut tendre à des inflations de faux positifs, des interprétations erronées ou une perte de puissance statistique (Peterson et al., 2019). Fort de ce constat, dans une optique de représentativité, proposer des modèles statistiques applicables dans des populations d'origines ethniques différentes paraît crucial. Pour se faire, une approche traditionnellement utilisée en pratique, consiste à intégrer l'ascendance en intégrant les composantes principales des génotypes comme covariables. L'avantage de cette approche est de fournir un outil intuitif et rapide pour visualiser et intégrer les structures génétiques opérant au niveau populationnel. Cependant, de manière plus générale, permettre l'intégration de covariables relatives au sexe, à l'âge ou à tout autre type d'informations pouvant influer l'apparition ou le développement de la maladie paraît pertinent. Ainsi, le

modèle présenté en chapitre quatre pourrait s'étendre en permettant l'intégration de covariables en addition de la matrice de génotypes (Voir Annexes C.1).

6.3.3 Combiner variants rares et communs

Comme nous l'avons dans les chapitres quatre et cinq, *RetroFun-RVS* en permettant l'intégration de l'information familiale et d'annotations fonctionnelles est une approche rapide et puissante pour détecter de nouvelles régions associées avec des maladies complexes. Cependant, la plupart des traits sont la résultante de la présence de variants communs et rares. Ainsi étendre *RetroFun-RVS* en permettant de considérer dans le même cadre statistique variants fréquents et non-fréquents dans la population générale permettrait de mettre en lumière de nouvelles régions impliquées dans l'étiologie de maladies complexes; avec l'avantage majeur de pouvoir affiner notre connaissance sur la part jouée respectivement par les variants communs et rares. Combiner l'effet des variants communs et rares pourrait être obtenu en exploitant différentes stratégies (Voir Annexes C.2).

Conclusion

À l'heure actuelle, une des questions centrales en sciences est l'élucidation des mécanismes biologiques par lesquels les maladies complexes émergent ou se développent chez l'Homme. La démocratisation des méthodes de séquençage de l'ADN haut débit à l'échelle du génome a rendu possible le développement d'outils statistiques et bio-informatiques, dans une perspective de mieux comprendre les chaînes causales impliquées dans l'étiologie de maladies complexes. Cependant, comme nous l'avons vu tout au long de cette thèse, les maladies complexes sont caractérisées par la présence de variants rares, pour la plupart localisés dans des régions non codantes, rendant l'interprétation des mécanismes biologiques sous-jacents difficile.

À ce titre, pour adresser cette double problématique, de la rareté et du manque d'interprétabilité des variants découverts, l'objectif central de cette thèse est alors de fournir de nouvelles méthodes analytiques dans une perspective de mieux cibler les variants causaux localisés dans les régions non codantes.

Parce qu'il n'y a à ce jour pas de modèle computationnel de réseaux de régulation 3D, nous avons proposé au sein de cette thèse de nouvelles structures 3D mettant en interaction gènes et enhancers, appelés pôles de régulation cis. En plus de représenter des organisations biologiquement valides, nous avons démontré que les pôles de régulation cis étaient un modèle pertinent dans l'étiologie de maladies complexes, notamment dans la schizophrénie.

Ensuite, dans une logique de détecter de nouveaux variants rares impliqués dans les troubles complexes, nous avons décidé d'incorporer les pôles de régulation cis en tant qu'annotations fonctionnelles dans des tests d'association de variants rares appliqués aux familles. Une des subtilités de la méthode proposée est qu'elle ne se restreint qu'aux individus atteints d'une même famille. Nous avons donc proposé *RetroFun-RVS*, un cadre statistique rétrospectif unifiant méthodes d'association de variants rares et incorporation d'annotations fonctionnelles. Nous avons démontré que la méthode incorporant les pôles de régulation cis était rapide et puissante pour détecter de nouveaux variants causaux. La méthode a été rendue disponible et illustrer dans un contexte de fentes labiales.

Le prolongement des méthodes proposées dans cette thèse a été discuté au prisme des défis actuels ou futurs du domaine de la génétique.

Bibliographie

- Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. Al, Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A. A., Barnes, I. H. A., Barozzi, I., ... Zimmerman, J. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583(7818), 699–710. <https://doi.org/10.1038/s41586-020-2493-4>
- Anania, C., & Lupiáñez, D. G. (2020). Order and disorder: Abnormal 3D chromatin organization in human disease. *Briefings in Functional Genomics*, 19(2), 128–138. <https://doi.org/10.1093/bfgp/elz028>
- Bates, G. P. (2005). History of genetic disease: The molecular genetics of Huntington disease - a history | Learn Science at Scitable. *Nature Reviews Genetics*, 6(6), 755–773. <http://www.nature.com/scitable/content/History-of-genetic-disease-The-molecular-genetics-15297>
- Benabdallah, N. S., Williamson, I., Illingworth, R. S., Kane, L., Boyle, S., Sengupta, D., Grimes, G. R., Therizols, P., & Bickmore, W. A. (2019). Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Molecular Cell*, 76(3), 473-484.e7. <https://doi.org/10.1016/j.molcel.2019.07.038>
- Bergman, D. T., Jones, T. R., Liu, V., Ray, J., Jagoda, E., Siraj, L., Kang, H. Y., Nasser, J., Kane, M., Rios, A., Nguyen, T. H., Grossman, S. R., Fulco, C. P., Lander, E. S., & Engreitz, J. M. (2022). Compatibility rules of human enhancer and promoter sequences. In *Nature* (Vol. 607, Issue July). Springer US. <https://doi.org/10.1038/s41586-022-04877-w>
- Bodein, A., Chapleur, O., Droit, A., & Lê Cao, K. A. (2019). A Generic Multivariate Framework for the Integration of Microbiome Longitudinal Studies With Other Data Types. *Frontiers in Genetics*, 10(November), 1–18. <https://doi.org/10.3389/fgene.2019.00963>
- Bodein, A., Scott-Boyer, M. P., Perin, O., Lê Cao, K. A., & Droit, A. (2022). Interpretation of network-based integration from multi-omics longitudinal data. *Nucleic Acids Research*, 50(5), E27. <https://doi.org/10.1093/nar/gkab1200>
- Bouwman, B. A. M., & de Laat, W. (2015). Getting the genome in shape: The formation of loops, domains and compartments. *Genome Biology*, 16(1), 1–9. <https://doi.org/10.1186/s13059-015-0730-1>
- Broekema, R. V., Bakker, O. B., & Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biology*, 10(1). <https://doi.org/10.1098/rsob.190221>

Bryois, J., Garrett, M. E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G. D., Shieh, A. W., Buil, A., Fullard, J. F., Roussos, P., Sklar, P., Akbarian, S., Haroutunian, V., Stockmeier, C. A., Wray, G. A., White, K. P., Liu, C., Reddy, T. E., Ashley-Koch, A., ... Crawford, G. E. (2018). Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. *Nature Communications*, 9(1), 3121. <https://doi.org/10.1038/s41467-018-05379-y>

Bulik-Sullivan, B., Loh, P. R., Finucane, H. K., Ripke, S., Yang, J., Patterson, N., Daly, M. J., Price, A. L., Neale, B. M., Corvin, A., Walters, J. T. R., Farh, K. H., Holmans, P. A., Lee, P., Collier, D. A., Huang, H., Pers, T. H., Agartz, I., Agerbo, E., ... O'Donovan, M. C. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3), 291–295. <https://doi.org/10.1038/ng.3211>

Bureau A, Parker MM, Ruczinski I, Taub MA, Marazita ML, Murray JC, Mangold E, Noethen MM, Ludwig KU, Hetmanski JB, Bailey-Wilson JE, Cropp CD, Li Q, Szymczak S, Albacha-Hejazi H, Alqosayer K, Field LL, Wu-Chou YH, Doheny KF, Ling H, Scott AF, Beaty TH. Whole exome sequencing of distant relatives in multiplex families implicates rare variants in candidate genes for oral clefts. *Genetics*. 2014 Jul;197(3):1039-44. doi: 10.1534/genetics.114.165225. Epub 2014 May 2. PMID: 24793288; PMCID: PMC4096358.

Bureau, A., Begum, F., Taub, M. A., Hetmanski, J. B., Parker, M. M., Albacha-Hejazi, H., Scott, A. F., Murray, J. C., Marazita, M. L., Bailey-Wilson, J. E., Beaty, T. H., & Ruczinski, I. (2019). Inferring disease risk genes from sequencing data in multiplex pedigrees through sharing of rare variants. *Genetic Epidemiology*, 43(1), 37–49. <https://doi.org/10.1002/gepi.22155>

Cardozo-Gizzi, A. M., Cattoni, D. I., Fiche, J. B., Espinola, S. M., Gurgo, J., Messina, O., Houbron, C., Ogiyama, Y., Papadopoulos, G. L., Cavalli, G., Lagha, M., & Nollmann, M. (2019). Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Molecular Cell*, 74(1), 212-222.e5. <https://doi.org/10.1016/j.molcel.2019.01.011>

Chen, H., Meigs, J. B., & Dupuis, J. (2013). Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genetic Epidemiology*, 37(2), 196–204. <https://doi.org/10.1002/gepi.21703>

Chen, M.-H., & Yang, Q. (2016). RVFam: an R package for rare variant association analysis with family data. *Bioinformatics*, 32(4), 624–626. <https://doi.org/10.1093/bioinformatics/btv609>

Campigli Di Giammartino D, Polyzos A, Apostolou E (2020) Transcription factors: Building hubs in the 3D space. *Cell Cycle* 19: 2395–2410. 10.1080/15384101.2020.1805238

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., Kathiresan, S., Kenny, E. E., Lindgren, C. M., MacArthur, D. G., North, K. N., Plon, S. E., Rehm, H. L., Risch, N., Rotimi, C. N., Shendure, J., Soranzo, N., & McCarthy, M. I.

(2020). A brief history of human disease genetics. *Nature*, 577(7789), 179–189. <https://doi.org/10.1038/s41586-019-1879-7>

Consortium, T. S. W. G. of the P. G., Ripke, S., Walters, J. T., & O'Donovan, M. C. (2020). Mapping genomic loci prioritises genes and implicates synaptic biology in schizophrenia. *MedRxiv*, 2020.09.12.20192922. <https://www.medrxiv.org/content/10.1101/2020.09.12.20192922v1%0Ahttps://www.medrxiv.org/content/10.1101/2020.09.12.20192922v1.abstract>

Crouch, D. J. M., & Bodmer, W. F. (2020). Polygenic inheritance, GWAS, polygenic risk scores, and the search for functional variants. *Proceedings of the National Academy of Sciences of the United States of America*, 117(32), 18924–18933. <https://doi.org/10.1073/pnas.2005634117>

Csardi, G., & Nepusz, T. (n.d.). *The igraph software package for complex network research*. 9.

Dali, R., & Blanchette, M. (2017). A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Research*, 45(6), 2994–3005. <https://doi.org/10.1093/nar/gkx145>

de Wit, E., & de Laat, W. (2012). A decade of 3C technologies: Insights into nuclear organization. *Genes and Development*, 26(1), 11–24. <https://doi.org/10.1101/gad.179804.111>

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing Chromosome Conformation. *Science*, 295(5558), 1306–1311.

Dekker, J. (2006). The three “C” s of chromosome conformation capture: Controls, controls, controls. *Nature Methods*, 3(1), 17–21. <https://doi.org/10.1038/nmeth823>

Dekker, J., & Mirny, L. (2016). The 3D Genome as Moderator of Chromosomal Communication. *Cell*, 164(6), 1110–1121. <https://doi.org/10.1016/j.cell.2016.02.007>

Di Giammartino, D. C., Polyzos, A., & Apostolou, E. (2020). Transcription factors: building hubs in the 3D space. *Cell Cycle*, 19(19), 2395–2410. <https://doi.org/10.1080/15384101.2020.1805238>

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376–380. <https://doi.org/10.1038/nature11082>

Do, C., Shearer, A., Suzuki, M., Terry, M. B., Gelernter, J., Greally, J. M., & Tycko, B. (2017). *Genetic – epigenetic interactions in cis: a major focus in the post-GWAS era*. 1–22. <https://doi.org/10.1186/s13059-017-1250-y>

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., & Dekker, J. (2006).

Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10), 1299–1309. <https://doi.org/10.1101/gr.5571506>

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genetics*, 9(3). <https://doi.org/10.1371/journal.pgen.1003348>

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems*, 3(1), 95–98. <https://doi.org/10.1016/j.cels.2016.07.002>

Espeso-Gil, S., Halene, T., Bendl, J., Kassim, B., Ben Hutta, G., Iskhakova, M., Shokrian, N., Auluck, P., Javidfar, B., Rajarajan, P., Chandrasekaran, S., Peter, C. J., Cote, A., Birnbaum, R., Liao, W., Borrman, T., Wiseman, J., Bell, A., Bannon, M. J., ... Akbarian, S. (2020). A chromosomal connectome for psychiatric and metabolic risk variants in adult dopaminergic neurons. *Genome Medicine*, 12(1), 1–19. <https://doi.org/10.1186/s13073-020-0715-x>

Espinola, S. M., Götz, M., Bellec, M., Messina, O., Fiche, J. B., Houbron, C., Dejean, M., Reim, I., Cardozo Gizzi, A. M., Lagha, M., & Nollmann, M. (2021). Cis-regulatory chromatin loops arise before TADs and gene activation, and are independent of cell fate during early Drosophila development. *Nature Genetics*, 53(4), 477–486. <https://doi.org/10.1038/s41588-021-00816-z>

Fortin, J. P., & Hansen, K. D. (2015). Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biology*, 16(1), 1–23. <https://doi.org/10.1186/s13059-015-0741-y>

Fudenberg, G., & Pollard, K. S. (2019). Chromatin features constrain structural variation across evolutionary timescales. *Proceedings of the National Academy of Sciences*, 116(6), 2175–2180. <https://doi.org/10.1073/pnas.1808631116>

Fukaya, T., Lim, B., & Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell*, 166(2), 358–368. <https://doi.org/10.1016/j.cell.2016.05.025>

Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., & Engreitz, J. M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12), 1664–1669. <https://doi.org/10.1038/s41588-019-0538-0>

Fullard, J. F., Hauberg, M. E., Bendl, J., Egervari, G., Cirnar, M. D., Reach, S. M., Motl, J., Ehrlich, M. E., Hurd, Y. L., & Roussos, P. (2018). An atlas of chromatin accessibility in the adult human brain. *Genome Research*, 28(8), 1243–1252. <https://doi.org/10.1101/gr.232488.117>

- Fullwood, M. J., & Ruan, Y. (2009). ChIP-based methods for the identification of long-range chromatin interactions. *Journal of Cellular Biochemistry*, 107(1), 30–39. <https://doi.org/10.1002/jcb.22116>
- Gasperini, M., Tome, J. M., & Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, 21(5), 292–310. <https://doi.org/10.1038/s41576-019-0209-0>
- Gazal, S., Weissbrod, O., Hormozdiari, F. et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nat Genet* 54, 827–836 (2022). <https://doi.org/10.1038/s41588-022-01087-y>
- Girdhar, K., Hoffman, G. E., Jiang, Y., Brown, L., Kundakovic, M., Hauberg, M. E., Francoeur, N. J., Wang, Y. chih, Shah, H., Kavanagh, D. H., Zharovsky, E., Jacobov, R., Wiseman, J. R., Park, R., Johnson, J. S., Kassim, B. S., Sloofman, L., Mattei, E., Weng, Z., ... Akbarian, S. (2018). Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome. *Nature Neuroscience*, 21(8), 1126–1136. <https://doi.org/10.1038/s41593-018-0187-0>
- Gorkin, D. U., Qiu, Y., Hu, M., Fletez-Brant, K., Liu, T., Schmitt, A. D., Noor, A., Chiou, J., Gaulton, K. J., Sebat, J., Li, Y., Hansen, K. D., & Ren, B. (2019). Common DNA sequence variation influences 3-dimensional conformation of the human genome. In *bioRxiv*. <https://doi.org/10.1101/592741>
- Gorkin, D. U., Leung, D., & Ren, B. (2014). The 3D Genome in Transcriptional Regulation and Pluripotency. *Cell Stem Cell*, 14(6), 762–775. <https://doi.org/10.1016/j.stem.2014.05.017>
- Halvorsen, M., Huh, R., Oskolkov, N., Wen, J., Netotea, S., Giusti-Rodriguez, P., Karlsson, R., Bryois, J., Nystedt, B., Ameur, A., Kähler, A. K., Ancalade, N., Farrell, M., Crowley, J. J., Li, Y., Magnusson, P. K. E., Gyllensten, U., Hultman, C. M., Sullivan, P. F., & Szatkiewicz, J. P. (2020). Increased burden of ultra-rare structural variants localizing to boundaries of topologically associated domains in schizophrenia. *Nature Communications*, 11(1), 1842. <https://doi.org/10.1038/s41467-020-15707-w>
- Hauberg, M. E., Creus-muncunill, J., Bendl, J., Kozlenkov, A., Zeng, B., Corwin, C., Chowdhury, S., Kranz, H., Hurd, Y. L., Wegner, M., Børglum, A. D., Dracheva, S., Ehrlich, M. E., & Fullard, J. F. (2020). Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nature Communications*. <https://doi.org/10.1038/s41467-020-19319-2>
- He, Z., Xu, B., Lee, S., & Ionita-Laza, I. (2017). Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. *American Journal of Human Genetics*, 101(3), 340–352. <https://doi.org/10.1016/j.ajhg.2017.07.011>
- He, Z., Xu, B., Lee, S., & Ionita-Laza, I. (2017). Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. *American Journal of Human Genetics*, 101(3), 340–352. <https://doi.org/10.1016/j.ajhg.2017.07.011>

- Hecker, D., Ardkani, F. B., & Schulz, M. H. (2022). The adapted Activity-By-Contact model for enhancer-gene assignment and its application to single-cell data. *BioRxiv*, 2022.01.28.478202.
<http://biorxiv.org/content/early/2022/01/31/2022.01.28.478202.abstract>
- Huo, Y., Li, S., Liu, J., Li, X., & Luo, X. J. (2019). Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nature Communications*, 10(1). <https://doi.org/10.1038/s41467-019-10866-4>
- Holzinger, E. R., Li, Q., Parker, M. M., Hetmanski, J. B., Marazita, M. L., Mangold, E., & Bailey-Wilson, J. E. (2017). Analysis of sequence data to identify potential risk variants for oral clefts in multiplex families. *Molecular Genetics and Genomic Medicine*, 5, 570–579.
- Hutchinson, A., Asimit, J., & Wallace, C. (2020). Fine-mapping genetic associations. *Human Molecular Genetics*, 29(R1), R81–R88. <https://doi.org/10.1093/hmg/ddaa148>
- Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., & Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics*, 7(2). <https://doi.org/10.1371/journal.pgen.1001289>
- Ionita-Laza, I., McCallum, K., Xu, B., & Buxbaum, J. D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, 48(2), 214–220. <https://doi.org/10.1038/ng.3477>
- Jackson, M., Marks, L., May, G. H. W., & Wilson, J. B. (2018). The genetic basis of disease. *Essays in Biochemistry*, 62(5), 643–723.
<https://doi.org/10.1042/EBC20170053>
- Kadouke, S., & Blobel, G. A. (2009). Chromatin loops in gene regulation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1789(1), 17–25. <https://doi.org/10.1016/j.bbagr.2008.07.002>
- Kichaev, G., Yang, W. Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., & Pasaniuc, B. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genetics*, 10(10). <https://doi.org/10.1371/journal.pgen.1004722>
- Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), 310–315. <https://doi.org/10.1038/ng.2892>
- Kolpakov F, Poroikov V, Selivanova G, K. A. (2011). GeneXplain — Identification of Causal Biomarkers and Drug Targets in Personalized Cancer Pathways. *J Biomol Tech*.
- Laird, N. M., & Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, 7(5), 385–394. <https://doi.org/10.1038/nrg1839>

Lakhal-Chaieb, L., Oualkacha, K., Richards, B. J., & Greenwood, C. M. T. (2016). A rare variant association test in family-based designs and non-normal quantitative traits. *Statistics in Medicine*, 35(6), 905–921. <https://doi.org/10.1002/sim.6750>

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., & Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, 91(2), 224–237. <https://doi.org/10.1016/j.ajhg.2012.06.007>

Lee, S., Abecasis, G. R., Boehnke, M., & Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics*, 95(1), 5–23. <https://doi.org/10.1016/j.ajhg.2014.06.009>

Li B, Wang GT, Leal SM. Generation of sequence-based data for pedigree-segregating Mendelian or Complex traits. *Bioinformatics*. 2015 Nov 15;31(22):3706-8. doi: 10.1093/bioinformatics/btv412. Epub 2015 Jul 14. PMID: 26177964; PMCID: PMC4757949.

Li, B., & Leal, S. M. (2008). Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *American Journal of Human Genetics*, 83(3), 311–321. <https://doi.org/10.1016/j.ajhg.2008.06.024>

Li, K., Liu, Y., Cao, H., Zhang, Y., Gu, Z., Liu, X., Yu, A., Kaphle, P., Dickerson, K. E., Ni, M., & Xu, J. (2020). Interrogation of enhancer function by enhancer-targeting CRISPR epigenetic editing. *Nature Communications*, 11(1), 1–16. <https://doi.org/10.1038/s41467-020-14362-5>

Li, M., Boehnke, M., & Abecasis, G. R. (2006). Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *American Journal of Human Genetics*, 78(5), 778–792. <https://doi.org/10.1086/503711>

Li, W., Baumbach, J., Mohammadnejad, A., Brasch-Andersen, C., Vandin, F., Korbel, J. O., & Tan, Q. (2019). Enriched power of disease-concordant twin-case-only design in detecting interactions in genome-wide association studies. *European Journal of Human Genetics*, 27(4), 631–636. <https://doi.org/10.1038/s41431-018-0320-2>

Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A. C., Boerwinkle, E., & Lin, X. (2019). Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. *The American Journal of Human Genetics*, 104(5), 802–814. <https://doi.org/https://doi.org/10.1016/j.ajhg.2019.03.002>

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Grinter, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289–293. <https://doi.org/10.1126/science.1181369>

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., & Lin, X. (2019). ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *American Journal of Human Genetics*, 104(3), 410–421. <https://doi.org/10.1016/j.ajhg.2019.01.002>

Ma, S., Dagleish, J., Lee, J., Wang, C., Liu, L., Gill, R., Buxbaum, J. D., Chung, W. K., Aschard, H., Silverman, E. K., Cho, M. H., He, Z., & Ionita-Laza, I. (2021). Powerful gene-based testing by integrating long-range chromatin interactions and knockoff genotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 118(47), 1–12. <https://doi.org/10.1073/pnas.2105191118>

Ma, Y., & Wei, P. (2019). FunspU: A versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. *PLoS Genetics*, 15(4), 1–21. <https://doi.org/10.1371/journal.pgen.1008081>

Madsen, B. E., & Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2). <https://doi.org/10.1371/journal.pgen.1000384>

Madsen, J. G. S., Madsen, M. S., Rauch, A., Traynor, S., Van Hauwaert, E. L., Haakonsson, A. K., Javierre, B. M., Hyldahl, M., Fraser, P., & Mandrup, S. (2020). Highly interconnected enhancer communities control lineage-determining genes in human mesenchymal stem cells. *Nature Genetics*, 52(11), 1227–1238. <https://doi.org/10.1038/s41588-020-0709-z>

Mangnier Loic, Charles Joly-Beauparlant, Arnaud Droit, Steve Bilodeau, Alexandre Bureau. (2022). Cis-regulatory hubs: a new 3D model of complex disease genetics with an application to schizophrenia. *Life Science Alliance*, 5(5), 1–12. [https://doi.org/10.26508/lса.202101156](https://doi.org/10.26508/lsa.202101156)

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753. <https://doi.org/10.1038/nature08494>

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., ... Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190–1195. <https://doi.org/10.1126/science.1222794>

Melo, U. S., Schöpflin, R., Acuna-Hidalgo, R., Mensah, M. A., Fischer-Zirnsak, B., Holtgrewe, M., Klever, M. K., Türkmen, S., Heinrich, V., Pluym, I. D., Matoso, E., Bernardo de Sousa, S., Louro, P., Hülsemann, W., Cohen, M., Dufke, A., Latos-Bieleńska, A., Vingron, M., Kalscheuer, V., ... Mundlos, S. (2020). Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. *American Journal of Human Genetics*, 106(6), 872–884. <https://doi.org/10.1016/j.ajhg.2020.04.016>

- Moosavi, A., & Ardekani, A. M. (2016). Role of epigenetics in biology and human diseases. *Iranian Biomedical Journal*, 20(5), 246–258. <https://doi.org/10.22045/ibj.2016.01>
- Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, 13(11), 919–922. <https://doi.org/10.1038/nmeth.3999>
- Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., & Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469), 59–64. <https://doi.org/10.1038/nature12593>
- Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., ... Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593(7858), 238–243. <https://doi.org/10.1038/s41586-021-03446-x>
- Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., & Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3). <https://doi.org/10.1371/journal.pgen.1001322>
- Ott, J., Kamatani, Y., & Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, 12(7), 465–474. <https://doi.org/10.1038/nrg2989>
- Oualkacha, K., Dastani, Z., Li, R., Cingolani, P. E., Spector, T. D., Hammond, C. J., Richards, J. B., Ciampi, A., & Greenwood, C. M. T. (2013). Adjusted Sequence Kernel Association Test for Rare Variants Controlling for Cryptic and Family Relatedness. *Genetic Epidemiology*, 37(4), 366–376. <https://doi.org/10.1002/gepi.21725>
- Oudelaar, A. M., Harrold, C. L., Hanssen, L. L. P., Telenius, J. M., Higgs, D. R., & Hughes, J. R. (2019). A revised model for promoter competition based on multi-way chromatin interactions at the α -globin locus. *Nature Communications*, 10(1). <https://doi.org/10.1038/s41467-019-13404-x>
- Paternoster, L., Standl, M., Waage, J., Baurecht, H., Hotze, M., Strachan, D. P., Curtin, J. A., Bønnelykke, K., Tian, C., Takahashi, A., Esparza-Gordillo, J., Alves, A. C., Thyssen, J. P., Den Dekker, H. T., Ferreira, M. A., Altmaier, E., Sleiman, P. M. A., Xiao, F. L., Gonzalez, J. R., ... Weidinger, S. (2015). Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nature Genetics*, 47(12), 1449–1456. <https://doi.org/10.1038/ng.3424>
- Paul S. Albert, Duminda Ratnasinghe, Joseph Tangrea, and S. W. (2001). Limitations of the Case-only Design for Identifying Gene-Environment Interactions Paul. *American Journal of Epidemiology*, 154(8), 1–10. <https://www.imedpub.com/ethnomedicine/>

- Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559–572. <https://doi.org/10.1080/14786440109462720>
- Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: Five essential questions. *Nature Reviews Genetics*, 14(4), 288–295. <https://doi.org/10.1038/nrg3458>
- Peterson, R. E., Kuchenbaecker, K., Walters, R. K., Chen, C. Y., Popejoy, A. B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R. J., Brick, L., Carey, C. E., Martin, A. R., Meyers, J. L., Su, J., Chen, J., Edwards, A. C., Kalungi, A., Koen, N., Majara, L., ... Duncan, L. E. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. *Cell*, 179(3), 589–603. <https://doi.org/10.1016/j.cell.2019.08.051>
- Pombo, A., & Dillon, N. (2015). Three-dimensional genome architecture: Players and mechanisms. *Nature Reviews Molecular Cell Biology*, 16(4), 245–257. <https://doi.org/10.1038/nrm3965>
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007 Sep;81(3):559-75. doi: 10.1086/519795. Epub 2007 Jul 25. PMID: 17701901; PMCID: PMC1950838.
- Rajarajan, P., Borrman, T., Liao, W., Schrode, N., Flaherty, E., Casiño, C., Powell, S., Yashaswini, C., LaMarca, E. A., Kassim, B., Javidfar, B., Espeso-Gil, S., Li, A., Won, H., Geschwind, D. H., Ho, S. M., MacDonald, M., Hoffman, G. E., Roussos, P., ... Akbarian, S. (2018). Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. *Science*, 362(6420). <https://doi.org/10.1126/science.aat4311>
- Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), 1665–1680. <https://doi.org/10.1016/j.cell.2014.11.021>
- ReproGen Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R. B., ... Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11), 1228–1235. <https://doi.org/10.1038/ng.3404>
- Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., ... Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), 317–330. <https://doi.org/10.1038/nature14248>

Rodriguez-Esteban, R., & Jiang, X. (2017). Differential gene expression in disease: A comparison between high-throughput studies and the literature. *BMC Medical Genomics*, 10(1), 1–10. <https://doi.org/10.1186/s12920-017-0293-y>

Roussos, P., Mitchell, A. C., Voloudakis, G., Fullard, J. F., Pothula, V. M., Tsang, J., Stahl, E. A., Georgakopoulos, A., Ruderfer, D. M., Charney, A., Okada, Y., Siminovitch, K. A., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Plenge, R. M., Raychaudhuri, S., Fromer, M., ... Sklar, P. (2014). A Role for Noncoding Variation in Schizophrenia. *Cell Reports*, 9(4), 1417–1429. <https://doi.org/10.1016/j.celrep.2014.10.015>

Roussos, P., Mitchell, A. C., Voloudakis, G., Fullard, J. F., Pothula, V. M., Tsang, J., Stahl, E. A., Georgakopoulos, A., Ruderfer, D. M., Charney, A., Okada, Y., Siminovitch, K. A., Worthington, J., Padyukov, L., Klareskog, L., Gregersen, P. K., Plenge, R. M., Raychaudhuri, S., Fromer, M., ... Sklar, P. (2014). A Role for Noncoding Variation in Schizophrenia. *Cell Reports*, 9(4), 1417–1429. <https://doi.org/10.1016/j.celrep.2014.10.015>

Rubin, A. J., Barajas, B. C., Furlan-magaril, M., Lopez-pajares, V., Mumbach, M. R., Howard, I., Kim, D. S., Boxer, L. D., Cairns, J., Spivakov, M., Wingett, S. W., Shi, M., Zhao, Z., Greenleaf, W. J., Kundaje, A., Snyder, M., Chang, H. Y., Fraser, P., & Khavari, P. A. (2017). Lineage-specific dynamic and pre-established enhancer – promoter contacts cooperate in terminal differentiation. *Nature Publishing Group*, 49(10). <https://doi.org/10.1038/ng.3935>

Ruzicka, W. B., Mohammadi, S., Davila-Velderrain, J., Subburaju, S., Tso, D. R., Hourihan, M., & Kellis, M. (2020). Single-cell dissection of schizophrenia reveals neurodevelopmental-synaptic axis and transcriptional resilience. *MedRxiv*, 2020.11.06.20225342. <https://doi.org/10.1101/2020.11.06.20225342>

Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8), 491–504. <https://doi.org/10.1038/s41576-018-0016-z>

Schaid, D. J., McDonnell, S. K., Riska, S. M., Carlson, E. E., & Thibodeau, S. N. (2010). Estimation of genotype relative risks from pedigree data by retrospective likelihoods. *Genetic Epidemiology*, 34(4), 287–298. <https://doi.org/10.1002/gepi.20460>

Schaid, D. J., Mcdonnell, S. K., Sinnwell, J. P., & Thibodeau, S. N. (2013). Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. *Genetic Epidemiology*, 37(5), 409–418. <https://doi.org/10.1002/gepi.21727>

Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511, 421–427 (2014). <https://doi.org/10.1038/nature13595>

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., Li, Y., Lin, S., Lin, Y., Barr, C. L., & Ren, B. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active

Regions in the Human Genome. *Cell Reports*, 17(8), 2042–2059.
<https://doi.org/10.1016/j.celrep.2016.10.061>

Sergey, Nurk and Sergey, Koren and Arang, Rhie and Mikko, Rautiainen and Andrey, V. Bzikadze and Alla, Mikheenko and Mitchell, R. Vollger and Nicolas, Altemose and Lev, Uralsky and Ariel, Gershman and Sergey, Aganezov and Savannah, J. Hoyt and Mark, D. and. (2022). The complete sequence of a human genome. *Science*, 376(6588), 44–53.

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., Heard, E., Dekker, J., & Barillot, E. (2015). HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, 16(1), 1–11. <https://doi.org/10.1186/s13059-015-0831-x>

Sey, N. Y. A., Hu, B., Mah, W., Fauni, H., McAfee, J. C., Rajarajan, P., Brennand, K. J., Akbarian, S., & Won, H. (2020). A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nature Neuroscience*, 23(4), 583–593. <https://doi.org/10.1038/s41593-020-0603-0>

Shaffer JR, LeClair J, Carlson JC, Feingold E, Buxó CJ, Christensen K, Deleyiannis FWB, Field LL, Hecht JT, Moreno L, Orioli IM, Padilla C, Vieira AR, Wehby GL, Murray JC, Weinberg SM, Marazita ML, Leslie EJ. Association of low-frequency genetic variants in regulatory regions with nonsyndromic orofacial clefts. *Am J Med Genet A*. 2019 Mar;179(3):467-474. doi: 10.1002/ajmg.a.61002. Epub 2018 Dec 24. PMID: 30582786; PMCID: PMC6374160.

Sherman, T, Fu, J, Scharpf, R.B, Bureau, A, Ruczinski, I, Detection of rare disease variants in extended pedigrees using RVS, Bioinformatics, Volume 35, Issue 14, July 2019, Pages 2509–2511, <https://doi.org/10.1093/bioinformatics/bty976>

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., De Wit, E., Van Steensel, B., & De Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, 38(11), 1348–1354. <https://doi.org/10.1038/ng1896>

Singh, T., Poterba, T., Curtis, D. et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature* 604, 509–516 (2022). <https://doi.org/10.1038/s41586-022-04556-w>

Smeland, O. B., Frei, O., Dale, A. M., & Andreassen, O. A. (2020). The polygenic architecture of schizophrenia — rethinking pathogenesis and nosology. *Nature Reviews Neurology*, 16(7), 366–379. <https://doi.org/10.1038/s41582-020-0364-0>

Song, M., Pebworth, M. P., Yang, X., Abnousi, A., Fan, C., Wen, J., Rosen, J. D., Choudhary, M. N. K., Cui, X., Jones, I. R., Bergenholz, S., Eze, U. C., Juric, I., Li, B., Maliskova, L., Lee, J., Liu, W., Pollen, A. A., Li, Y., ... Shen, Y. (2020). Cell-type-specific 3D epigenomes in the developing human cortex. *Nature*, 587(7835), 644–649. <https://doi.org/10.1038/s41586-020-2825-4>

- Splinter, E., de Wit, E., Nora, E. P., Klous, P., van de Werken, H. J. G., Zhu, Y., Kaaij, L. J. T., van IJcken, W., Gribnau, J., Heard, E., & de Laat, W. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes and Development*, 25(13), 1371–1383.
<https://doi.org/10.1101/gad.633311>
- Sul, J. H., Cade, B. E., Cho, M. H., Qiao, D., Silverman, E. K., Redline, S., & Sunyaev, S. (2016). Increasing Generality and Power of Rare-Variant Tests by Utilizing Extended Pedigrees. *American Journal of Human Genetics*, 99(4), 846–859.
<https://doi.org/10.1016/j.ajhg.2016.08.015>
- Sullivan, P. F., Daly, M. J., & O'Donovan, M. (2012). Genetic architectures of psychiatric disorders: The emerging picture and its implications. *Nature Reviews Genetics*, 13(8), 537–551. <https://doi.org/10.1038/nrg3240>
- Sun, B.B., Kurki, M.I., Foley, C.N. et al. Genetic associations of protein-coding variants in human disease. *Nature* 603, 95–102 (2022). <https://doi.org/10.1038/s41586-022-04394-w>
- van Berkum, N. L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L. A., Dekker, J., & Lander, E. S. (2010). Hi-C: A method to study the three-dimensional architecture of genomes. *Journal of Visualized Experiments*, 39, 1–7.
<https://doi.org/10.3791/1869>
- Vliet, J. Van, Oates, N. A., & Whitelaw, E. (2007). Review *Epigenetic mechanisms in the context of complex diseases*. 64, 1531–1538. <https://doi.org/10.1007/s00018-007-6526-z>
- Voss, T. C., & Hager, G. L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics*, 15(2), 69–81.
<https://doi.org/10.1038/nrg3623>
- Wang, X., Zhang, Z., Morris, N., Cai, T., Lee, S., Wang, C., Yu, T. W., Walsh, C. A., & Lin, X. (2017). Rare variant association test in family-based sequencing studies. *Briefings in Bioinformatics*, 18(6), 954–961. <https://doi.org/10.1093/bib/bbw083>
- Working, S., & Consortium, G. (2014). *Biological insights from 108 schizophrenia-associated genetic loci*. <https://doi.org/10.1038/nature13595>
- Wu, C., & Pan, W. (2018). Integration of enhancer-promoter interactions with GWAS summary results identifies novel schizophrenia-associated genes and pathways. *Genetics*, 209(3), 699–709. <https://doi.org/10.1534/genetics.118.300805>
- Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1), 82–93. <https://doi.org/10.1016/j.ajhg.2011.05.029>
- Yao, L., Liang, J., Ozer, A., Leung, A. K.-Y., Lis, J. T., & Yu, H. (2022). A comparison of experimental assays and analytical methods for genome-wide identification of active enhancers. *Nature Biotechnology*. <https://doi.org/10.1038/s41587-022-01211-7>

- Zhang, F., & Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human Molecular Genetics*, 24(R1), R102–R110. <https://doi.org/10.1093/hmg/ddv259>
- Zhang, R., Zhou, T., & Ma, J. (2022). Multiscale and integrative single-cell Hi-C analysis with Higashi. *Nature Biotechnology*, 40(2), 254–261. <https://doi.org/10.1038/s41587-021-01034-y>
- Zhao, Z., Tavoosidana, G., Sjölander, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., & Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38(11), 1341–1347. <https://doi.org/10.1038/ng1891>
- Zhao, L., He, Z., Zhang, D., Wang, G. T., Renton, A. E., Vardarajan, B. N., Nothnagel, M., Goate, A. M., Mayeux, R., & Leal, S. M. (2019). A Rare Variant Nonparametric Linkage Method for Nuclear and Extended Pedigrees with Application to Late-Onset Alzheimer Disease via WGS Data. *American Journal of Human Genetics*, 105(4), 822–835. <https://doi.org/10.1016/j.ajhg.2019.09.006>
- Zhao, L., Zhang, Z., Rodriguez, S. M. B., Vardarajan, B. N., Renton, A. E., Goate, A. M., Mayeux, R., Wang, G. T., & Leal, S. M. (2020). A quantitative trait rare variant nonparametric linkage method with application to age-at-onset of Alzheimer's disease. *European Journal of Human Genetics*, 28(12), 1734–1742. <https://doi.org/10.1038/s41431-020-0703-z>
- Zhou, J., Ma, J., Chen, Y., Cheng, C., Bao, B., Peng, J., Sejnowski, T. J., Dixon, J. R., & Ecker, J. R. (2019). Robust single-cell Hi-C clustering by convolution- And random-walk-based imputation. *Proceedings of the National Academy of Sciences of the United States of America*, 116(28), 14011–14018. <https://doi.org/10.1073/pnas.1901423116>
- Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E. A., Koebbe, B. C., Nielsen, C., Hirst, M., Farnham, P., Kuhn, R. M., Zhu, J., Smirnov, I., Kent, W. J., Haussler, D., Madden, P. A. F., Costello, J. F., & Wang, T. (2011). The human epigenome browser at Washington University. *Nature Methods*, 8(12), 989–990. <https://doi.org/10.1038/nmeth.1772>
- Zufferey, M., Tavernari, D., Oricchio, E., & Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, 19(1), 1–18. <https://doi.org/10.1186/s13059-018-1596-9>

Annexes

Annexes du Chapitre 3

A.1 Supplemental Figures

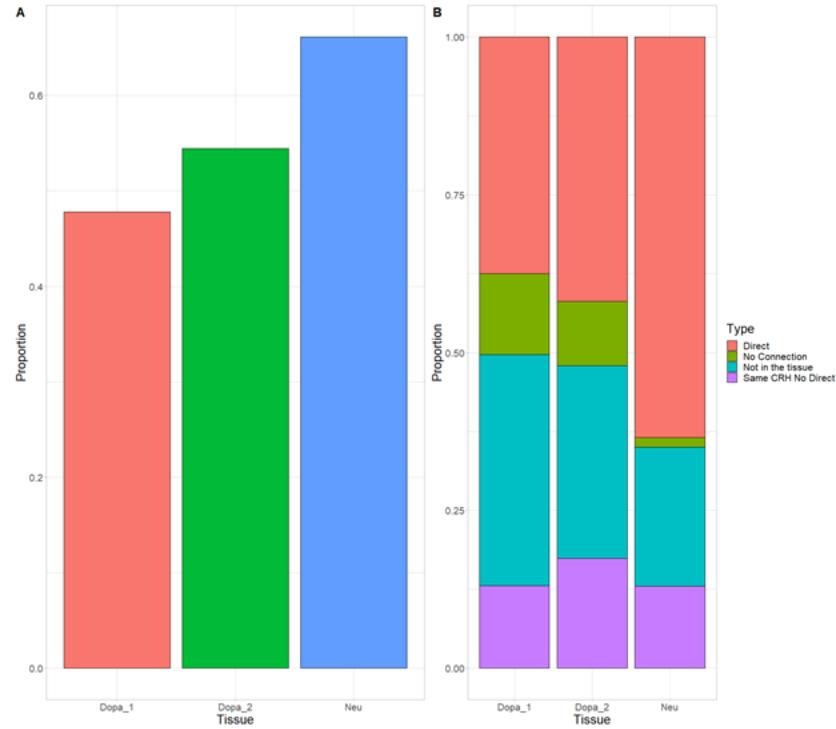


Figure 15: CRHs exhibit strong overlap between iPSC neurons and post-mortem brain tissues

(A) Proportion of distal elements observed in iPSC neurons found in the post-mortem brain tissues. **(B)** Proportion of pairs of promoter-distal element observed in iPSC-derived neurons strictly found in one of other post-mortem tissue (Direct), found in the tissue but not either as direct or indirect connection (No Connection), found in the tissue within the same CRH (Same CRH No Direct) or not found in the tissue. Results are reported relatively to the number of pairs observed in iPSC-derived neurons.

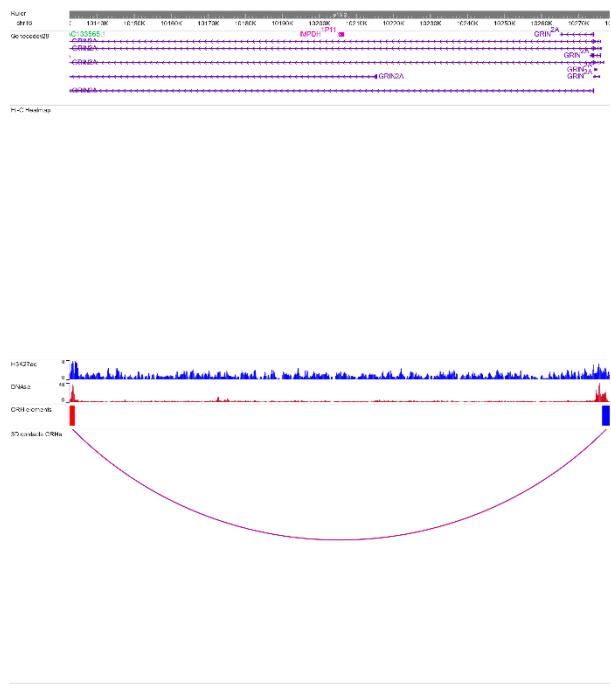


Figure 16: Genome browser view of the CRH encompassing GRIN2A gene

See caption of Figure 2.1C for explanations. Due to the resolution and short distance between CRH elements, no 3D contacts are represented in the Hi-C Heatmap track.

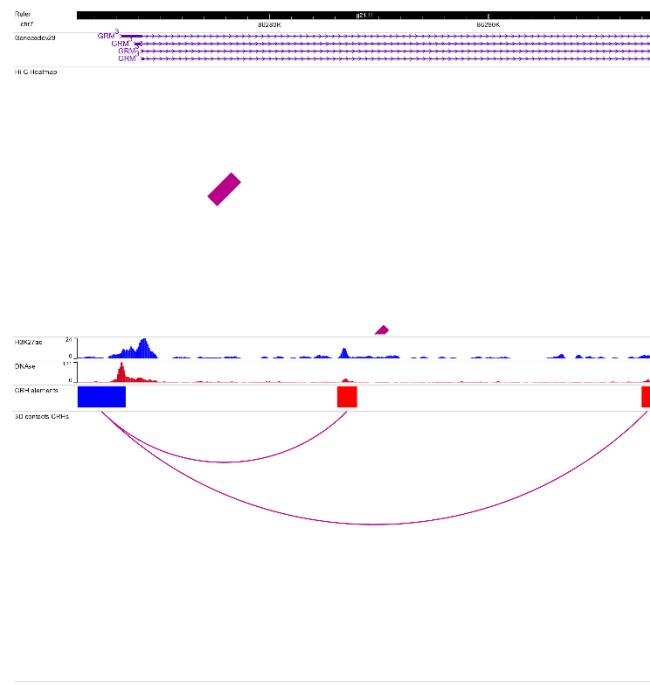


Figure 17: Genome browser view of the CRH encompassing GRM3 gene

See caption of Figure 2.1C for explanations.

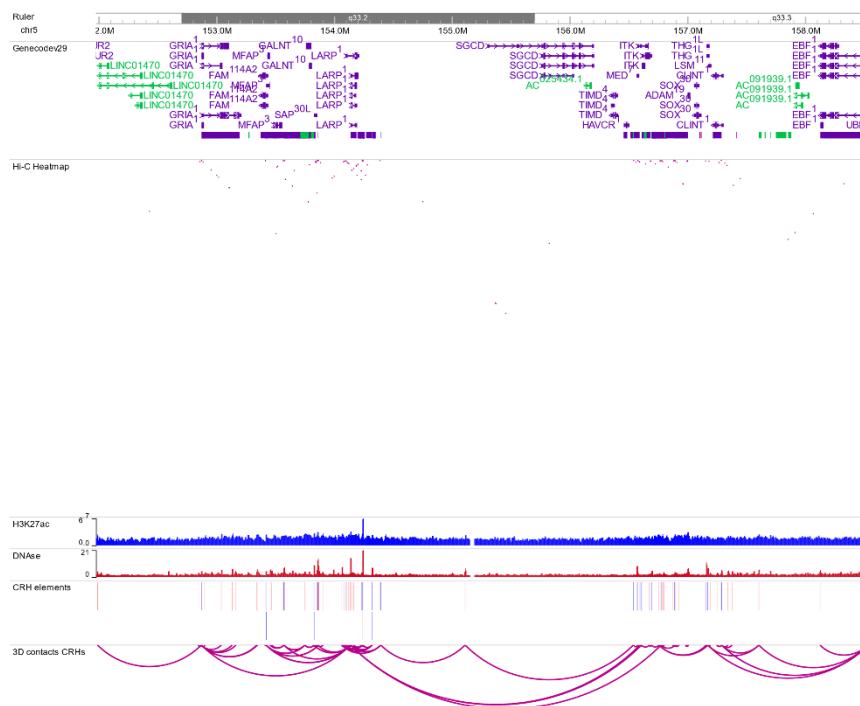


Figure 18: Genome browser view of the CRH encompassing *GRIA1* gene

See caption of Figure 2.1C for explanations.

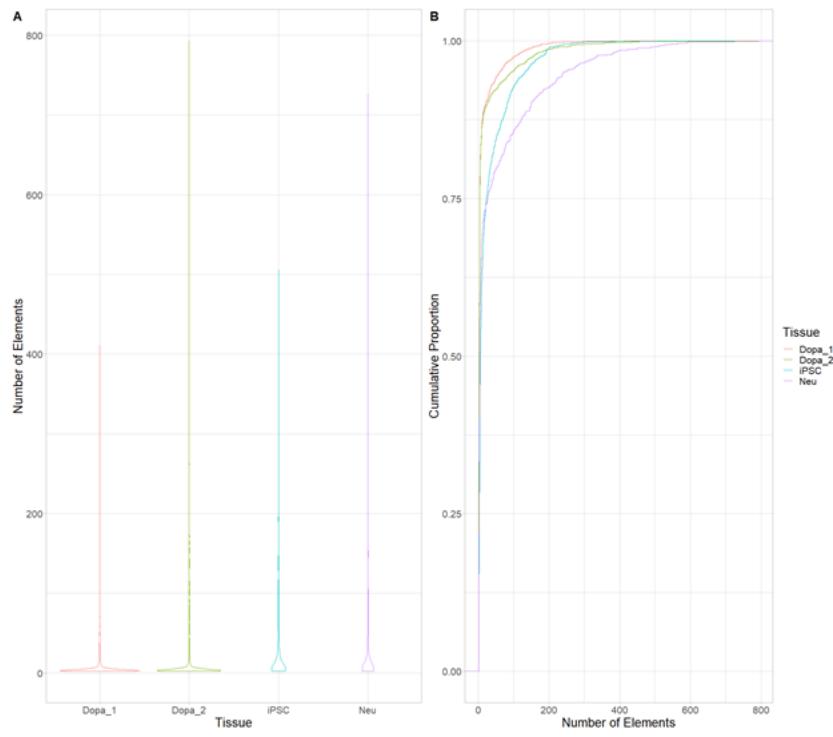


Figure 19: CRHs in iPSC-derived neurons show “average behavior” compared to post-mortem tissues regarding number of elements

(A) Violin plots of the number of elements included in CRHs by brain tissue. **(B)** Cumulative distribution of the number of elements within CRHs by tissue.

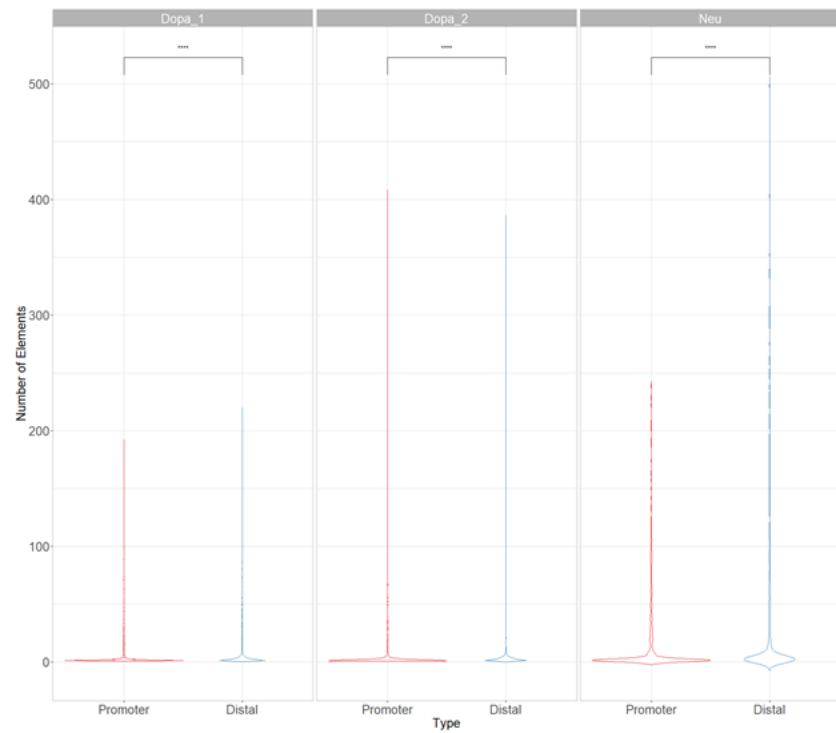


Figure 20: In post-mortem tissues, CRHs are mainly composed by distal elements

Violin plots of the number of promoters or distal elements across post-mortem brain tissues.

Difference between number of elements were assessed with Wilcoxon signed-rank test.

Data Information: **** represent p-value ≤ 0.0001 .

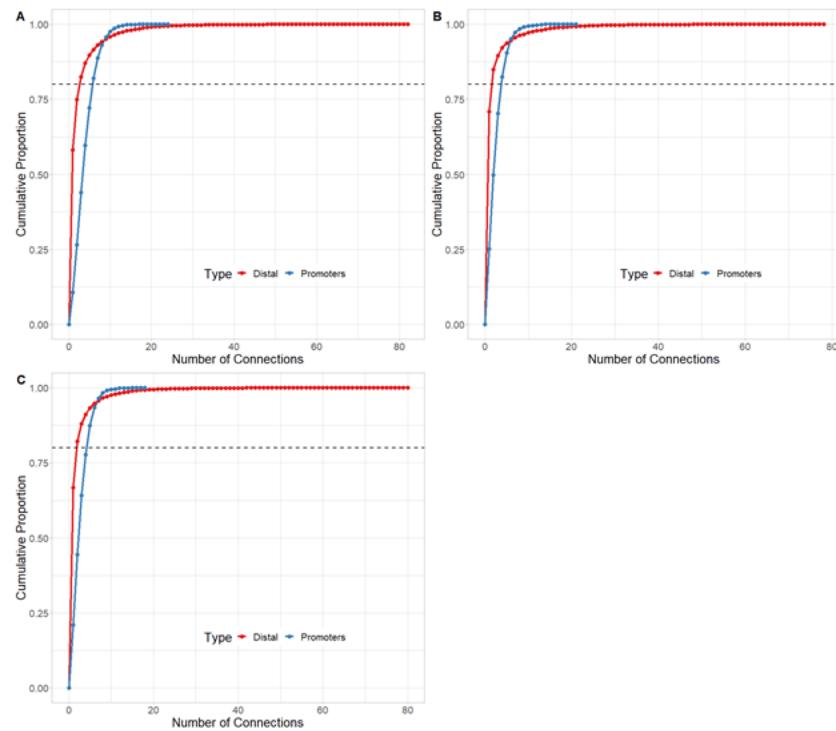


Figure 21: Promoters are more connected than distal elements across post-mortem tissues

Cumulative distribution function of the number of connections by promoter and distal element for **(A)** Neu, **(B)** Dopa_1, and **(C)** Dopa_2. Dotted line shows the number of connections where 80% are less or equal to this value.

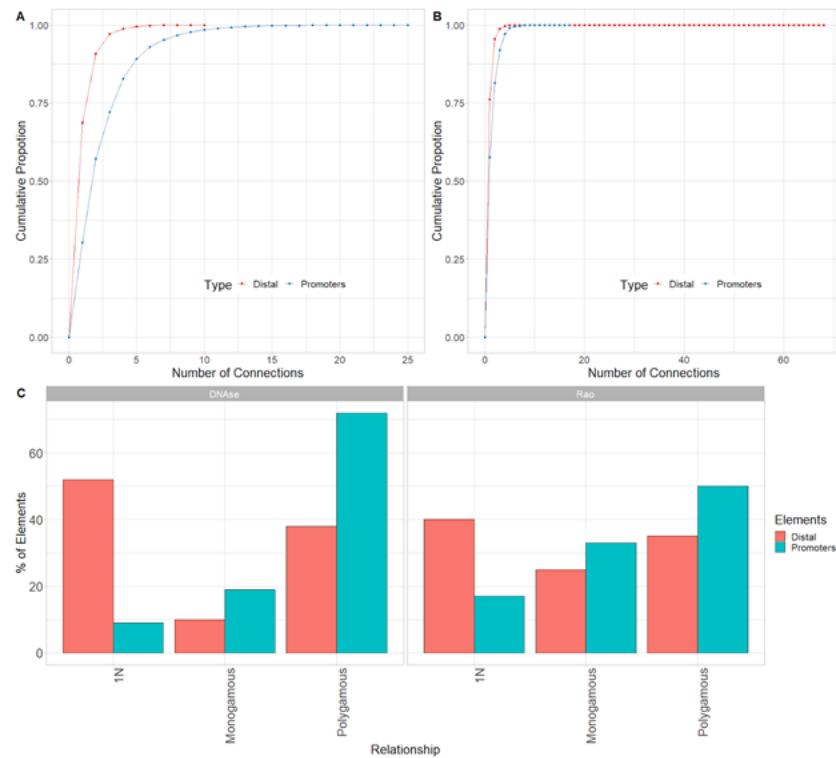


Figure 22: Promoters are more connected than distal elements in DNAse-based and Rao methods

(A) Cumulative proportion of the number of connections for promoters or distal elements for the DNAse method. **(B)** Cumulative proportion of the number of connections for promoters or distal elements for the Rao method. **(C)** Distribution of kind of relationship for the DNAse method (**Left**) and Rao method (**Right**).

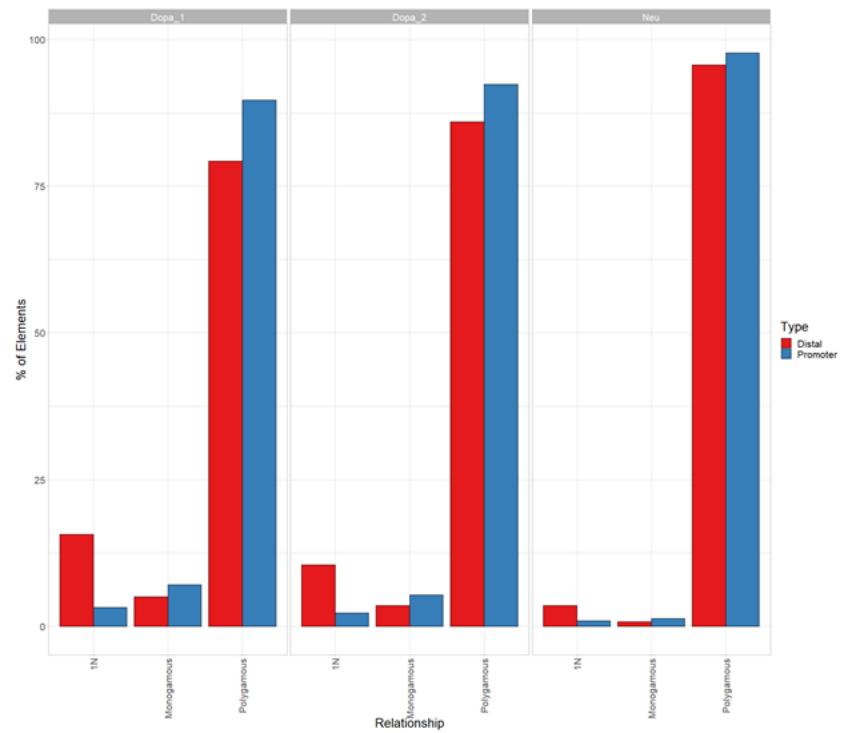


Figure 23: Promoters are inside more complex relationships than distal elements in post-mortem brain tissues
Complexity analysis for promoters and distal elements across post-mortem brain tissues.

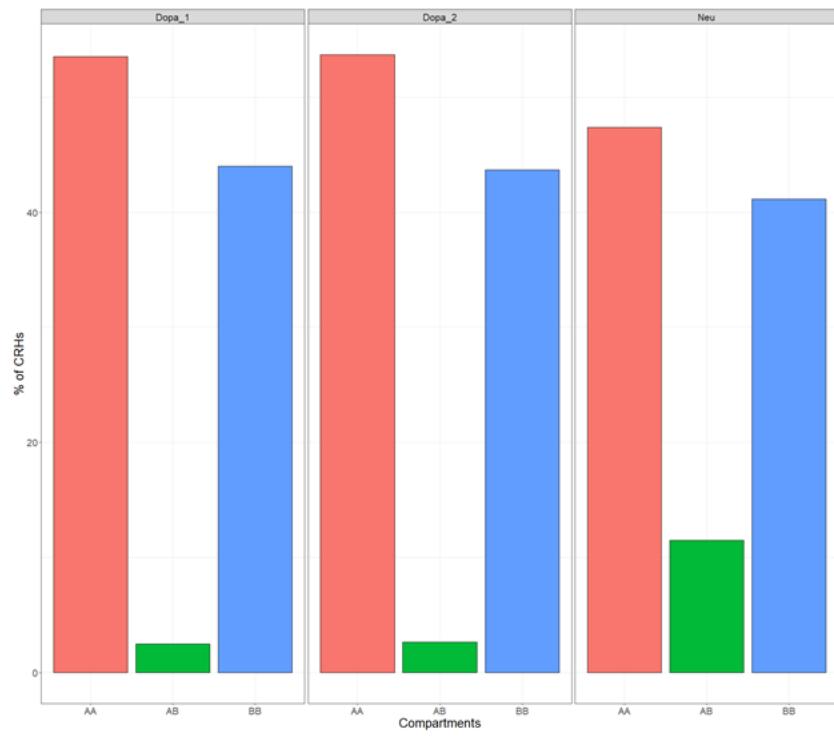


Figure 24: CRHs mainly overlap active compartments across post-mortem tissues

Distribution of the kind of compartment overlapped by CRHs across post-mortem brain tissues. Interestingly, dopaminergic neuron samples showed identical overlapping proportions.

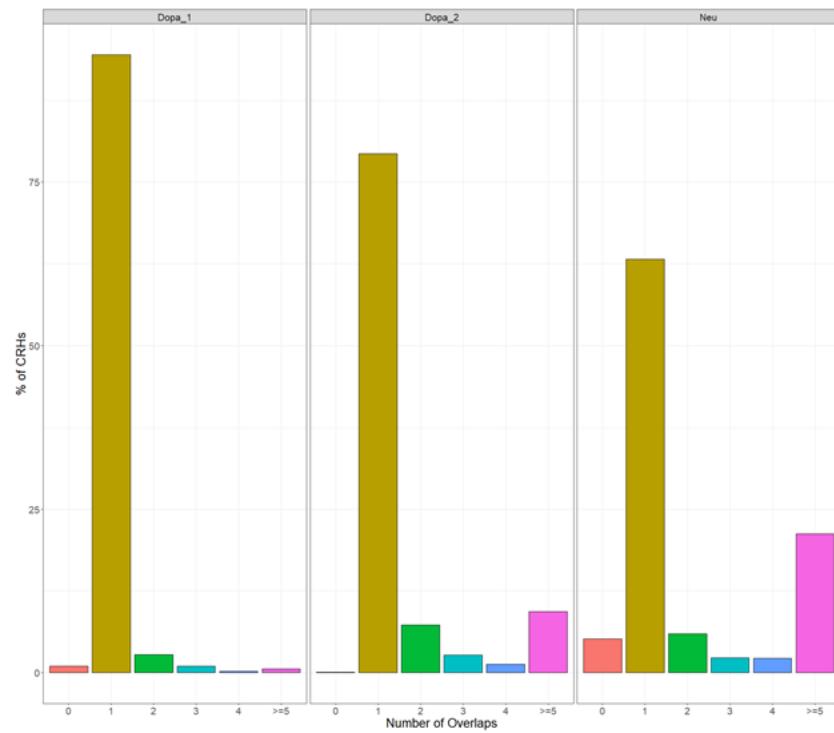


Figure 25: In most cases, CRHs overlap one TAD across post-mortem brain tissues

Distribution of the number of TADs overlapped by CRHs across post-mortem brain tissues, when TADs are detected with the directionality index.

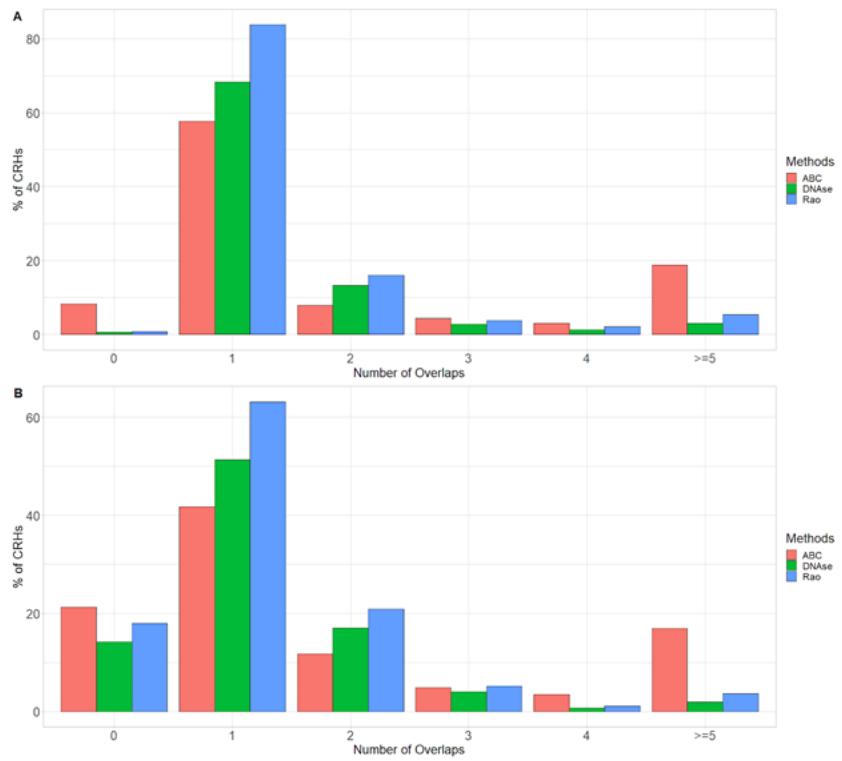


Figure 26: In most cases, CRHs overlap one TAD across our methods

(A) Distribution of the number of TADs overlapped by CRHs built with our different methods, when TADs are detected with the insulation score (Crane et al. 2015). **(B)** Distribution of the number of TADs overlapped by CRHs built with our different methods, when TADs are detected with the Arrowhead algorithm (Rao et al. 2014).

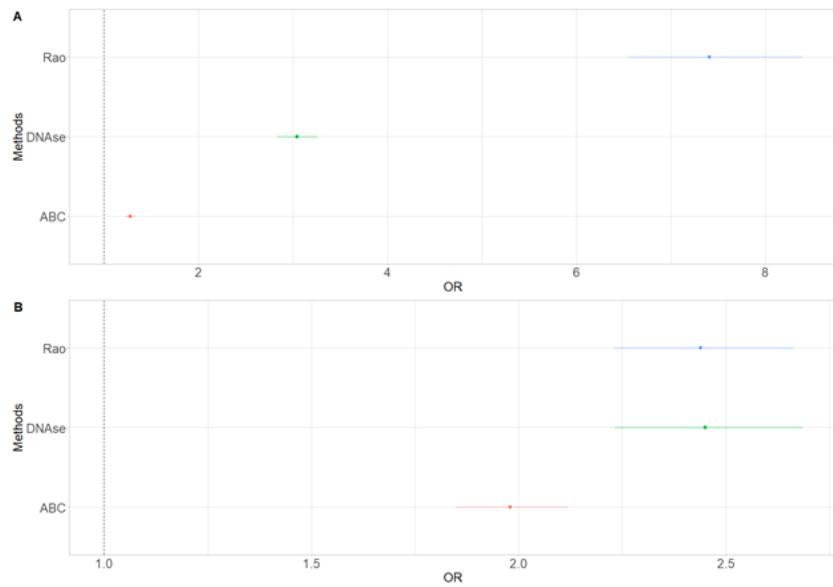


Figure 27: Enrichment in FIREs for distal elements and promoters in our methods

(A) Enrichment in FIREs for distal elements as measured with odds ratio (OR) and their 95% confidence interval. The dotted line represents the null value. **(B)** Enrichment in FIREs for promoters as measured with odds ratio (OR) and their 95% confidence interval. The dotted line represents the null value.

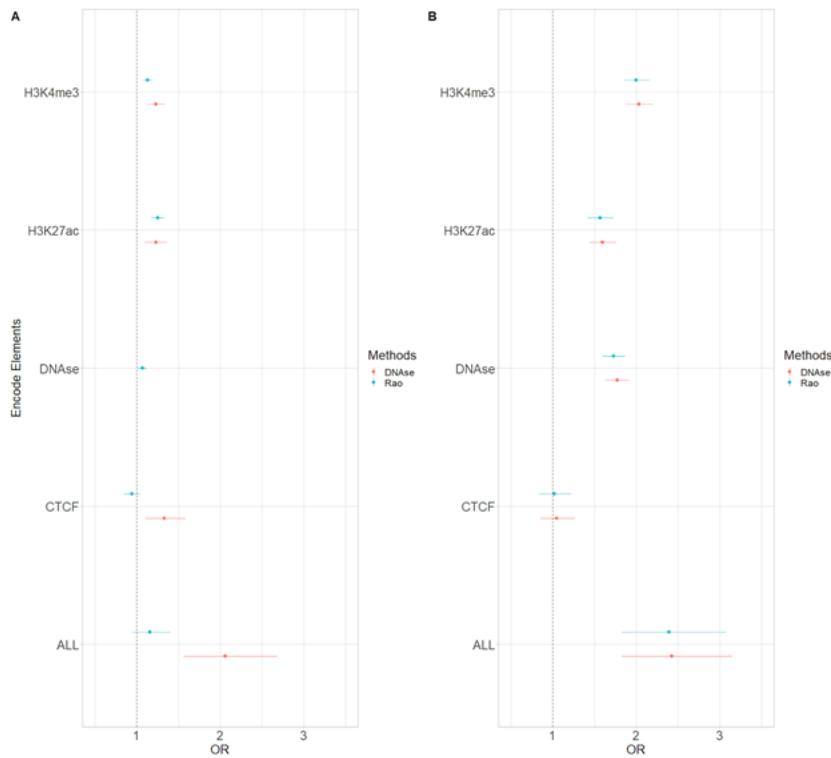


Figure 28: Enrichment in Encode candidate elements for distal elements and promoters in our control methods

(A) Enrichments in Encode Elements for distal elements as measured with odds ratio (OR) and their 95% confidence intervals for our control methods to build CRHs. The dotted line represents the null value. The ALL category encompasses all significant regions for all Encode elements. Since the DNAse method is built entirely on DNAse we removed DNAse candidate regions. **(B)** Enrichments for promoters in Encode Elements as measured with odds ratio (OR) and their confidence intervals for our control methods to build CRHs. The dotted line represents the null value. The ALL category encompasses all significant regions for all Encode elements.

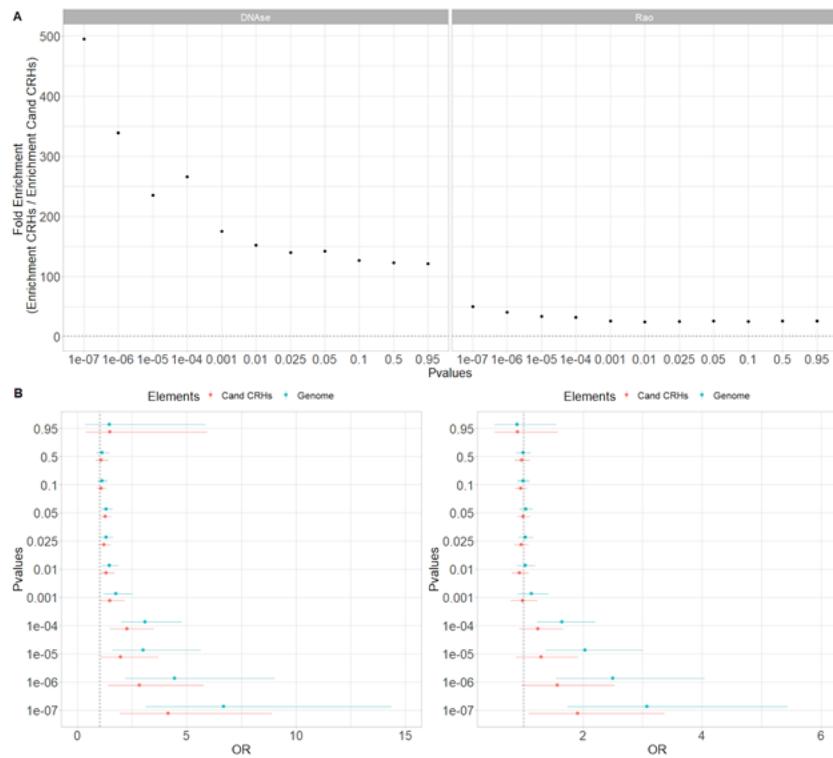


Figure 29: CRHs in DNAse-based and Rao methods show strong enrichments in schizophrenia-associated SNPs

(A) SNP enrichments as measured with odds ratio (OR) with their 95% confidence intervals for our control methods to build CRHs. **(Left)** DNAse method **(Right)** Rao method. The dotted line represents the null value. **(B)** Fold enrichment for our control methods to build CRHs. The dotted line represents the null value.

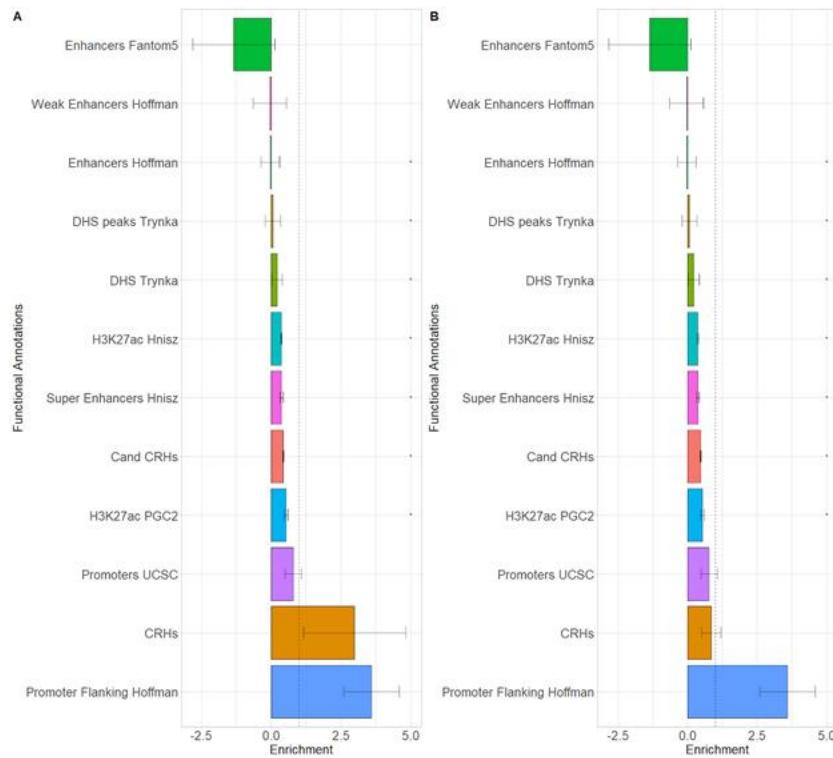


Figure 30: Schizophrenia heritability enrichments for DNAse-based and Rao methods

Schizophrenia heritability for our control methods **(A)** DNAse method **(B)** Rao method with their error bars. The dotted line represents the null.

Data Information: * represent p-value ≤ 0.05 after Bonferroni correction.

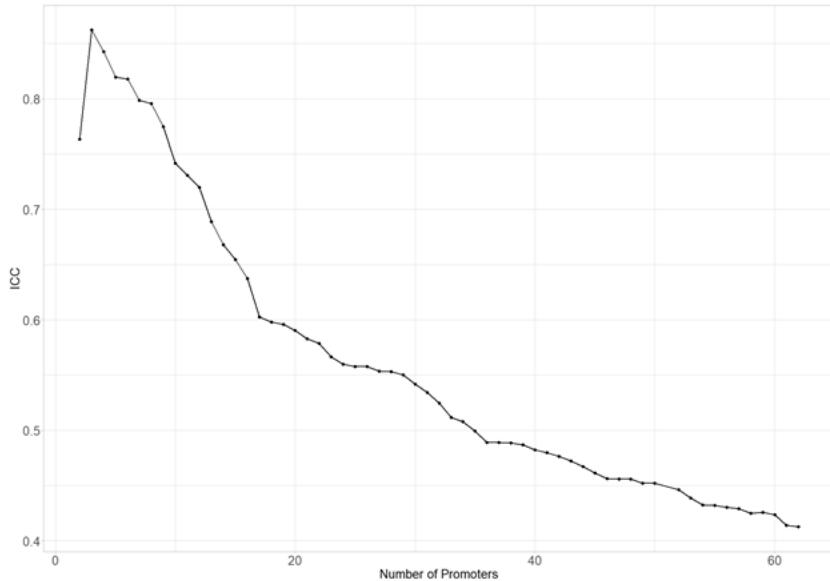


Figure 31: Negative association between Intraclass correlation and the number of promoters considered within CRHs for the ABC-Score method

Intraclass correlation (ICC) evolution of gene expression respectively of the number of promoters within the CRH.

A.2 Methods

3D Features

A/B Compartments: A/B compartments were defined at a 500 kb resolution of the contact matrix (using a 100 kb resolution had little impact on the results). The first principal component (PC) of a suitably normalized Hi-C contact matrix over a chromosome arm captures the plaid pattern of A/B compartments (Lieberman-Aiden et al., 2009). As GC content is higher in A compartments than in B compartments, the correlation of the first PC with GC content was used to orient the first PC so that positive values correspond to the A compartments and negative values to B compartments (Imakaev et al., 2012). The transformation applied to the ratio observed/expected (O/E) contact matrices was selected that 1) maximized the number of autosomal chromosome arms where the first PC had the highest correlation with GC content over the first three PCs and 2) had the highest correlation

of the first PC with GC content in these chromosome arms. The transformation was selected among the following three: O/E - 1 with clipping of values below percentile 1 and above percentile 99, log (O/E) and log (O/E) with clipping of values below percentile 1 and above percentile 99. The last transformation was selected based on the above criteria, with 30/40 autosomal chromosome arms where the first PC correlated the most strongly with GC content

and correlations between 0.38 and 0.87 within these chromosome arms. The A/B compartment for the remaining 10 chromosome arms (6q, 8q, 9p, 10q, 12p, 18p, 18q, 19q, 20p

and 21p) were set to missing.

TADs calling: TADs were called using the directionality index (Dixon et al., 2012), insulation score (Crane et al., 2015) or with Arrowhead algorithm from Juicer software (Rao et al., 2014;

Durand et al., 2016)

Directionality Index (DI) was computed as presented by Dixon et al., 2012 at 10 Kb resolution.

Briefly, for each 10 Kb bins the number of upstream and downstream contacts were calculated. A bias toward upstream regions at the end of a TAD was expected and conversely,

a bias toward downstream regions, at the beginning of a TAD was expected. As mentioned by

Gorkin et al., 2019, the original approach to computing the DI using a 200 Kb window size was applied to capture more local features. DI values for each 10Kb bins were used to build a

Hidden Markov Model and predict upstream bias, downstream bias, and no bias states, respectively. Regions switching from upstream bias to downstream bias were called topological boundaries.

Insulation Score (INS) was computed as presented by Crane et al., 2015. Simply, for each
10

Kb bin, the average number of contacts in 400Kb windows upstream and downstream on
O/E

matrices was computed. A local minimum for INS at TADs borders was expected. INS was
normalized at the chromosome level to take account of differences between chromosomes.
Then INS was scaled between 0 and 1, where 0 is complete insulation and 1 is no insulation
respectively.

Arrowhead TADs were annotated using Arrowhead (Rao et al., 2014; Durand et al., 2016)
at 10

Kb resolution.

Frequently Interacting Regions (FIREs) (Schmitt et al., 2016) was computed with
FIREcaller R package (Crowley et al., 2021) at 10Kb resolution, with minor adjustments to
fit
our data format.

Functional Enrichment analysis

Genes inside CRHs were used for GO enrichment analysis at the CRH or gene level. To do
so,

at the CRH level the clusterProfiler R package (Wang et al., 2012) was used with the
compareCluster function to perform over-representation tests.

Peak calling

Peak calling for ChIP-Seq data was performed with MACS2 (Zhang et al., 2008) software
through

the following command:

```
macs2 callpeak \
-t bamfile \
-n alias \
```

```
-f BAM \
-g hs \
-p .1 \
--call-summits \
--outdir outputdirectory
```

Genome Build

All coordinates in the human genome are reported using build hg19.

Control Methods

As mentioned in the main analysis, results were controlled using two other methods to determine distal elements and CRHs. In doing so, a large spectrum of regulatory processes was

captured. Thus, alternative CRHs were defined through the Rao and DNase methods, described as follows.

Rao: Since it has been shown that significant 3D peaks are enriched in enhancers, all distal elements were defined by 3D significant peaks linking promoters (Rao et al., 2014).

Briefly, as shown by Rao et al., 2014, a significant 3D peak is defined by comparing the number of contacts in this 3D peak relative to four neighborhood regions: horizontal, vertical, lower-left and donut, respectively.

DNase-based: Since it has been shown that non-coding regions are open chromatin regions, DNase peaks were added as an additional biological layer to the 3D peaks from the Rao method. In this method, distal elements are all DNase peaks on 10Kb 3D peaks in 3D contact with a promoter. Due to the methodology, several DNase peaks can be observed on the same 10Kb peak. Thus, each peak was considered as an individual distal element.

Also mentioned in the main method section, we defined candidate CRHs in order to assess comparison for enrichment analysis. For the Rao method, we considered all significant 3D peaks in no contact with a promoter as candidate elements. Based on the same rationale, DNase candidate elements are all DNase peaks in no 3D contact with a promoter.

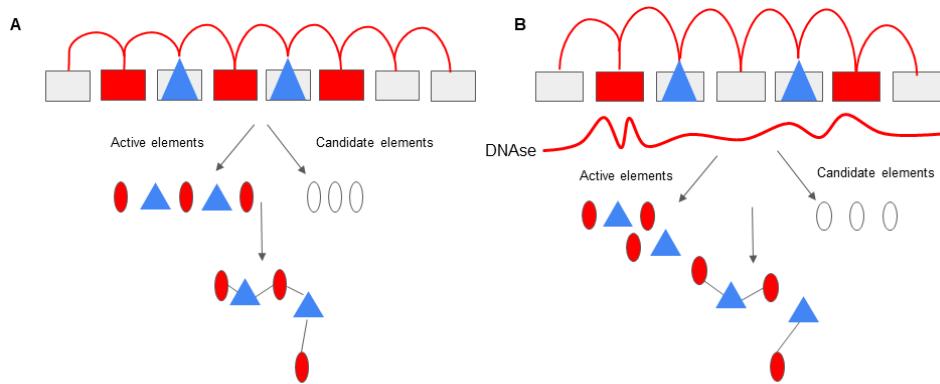


Figure 32: (A) Rao-Based method methodology and CRH building. (B) DNase-based method methodology and CRH building

Data Information: Promoters are represented by blue triangles, distal regulatory elements by red circles and, candidate elements by white circles.

HiC Mapping, Filtering, and Normalization for post-mortem brains

Raw Hi-C sequence fastq files for postmortem dopaminergic neuronal nuclei (NeuN+/Nurr1+) and the general neuronal (NeuN+) populations were obtained from the PsychENCODE Synapse platform. We referred in supplementary figures to Dopa_1 and Dopa_2 for postmortem dopaminergic neuronal nuclei samples and Neu for the general neuronal sample, respectively. They were mapped to the human genome sequence (hg19) in 10 kb bins using distiller (<https://github.com/mirnylab/distiller-nf>). Genome-wide iterative correction (i.e., KR normalization) was performed using cooler (<https://github.com/mirnylab/cooler>). We used the hicConvertFormat tool from the HiCExplorer package (<https://hicexplorer.readthedocs.io>) to convert .cool files into

ginteractions files, which after preprocessing were read by the Juicer toolbox and converted to .hic files.

Acknowledgments

The PsychEncode project is supported by: U01MH103392, U01MH103365, U01MH103346, U01MH103340, U01MH103339, R21MH109956, R21MH105881, R21MH105853, R21MH103877, R21MH102791, R01MH111721, R01MH110928, R01MH110927, R01MH110926, R01MH110921, R01MH110920, R01MH110905, R01MH109715, R01MH109677, R01MH105898, R01MH105898, R01MH094714, P50MH106934, U01MH116488, U01MH116487, U01MH116492, U01MH116489, U01MH116438, U01MH116441, U01MH116442, R01MH114911, R01MH114899, R01MH114901, R01MH117293, R01MH117291, R01MH117292 awarded to: Schahram Akbarian (Icahn School of Medicine at Mount Sinai), Gregory Crawford (Duke University), Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (University of Southern California), Mark Gerstein (Yale University), Daniel Geschwind (University of California, Los Angeles), Fernando Goes (Johns Hopkins University), Thomas M. Hyde (Lieber Institute for Brain Development), Andrew Jaffe (Lieber Institute for Brain Development), James A. Knowles (University of Southern California), Chunyu Liu (SUNY Upstate Medical University), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Panos Roussos (Icahn School of Medicine at Mount Sinai), Stephan Sanders (University of California, San Francisco), Nenad Sestan (Yale University), Pamela Sklar (Icahn School of Medicine at Mount Sinai), Matthew State (University of California, San Francisco), Patrick Sullivan (University of North Carolina), Flora Vaccarino (Yale University), Daniel Weinberger (Lieber Institute for Brain Development), Sherman Weissman (Yale University), Kevin White (University of Chicago), Jeremy Willsey (University of California, San Francisco), and Peter Zandi (Johns Hopkins University).

Références additionnelles :

Crane, E., Bian, Q., McCord, R. P., Lajoie, B. R., Wheeler, B. S., Ralston, E. J., . . . Meyer, B. J.

(2015, 7). Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*, 523, 240–244. doi:10.1038/nature14450

Crowley, C., Yang, Y., Qiu, Y., Hu, B., Abnousi, A., Lipiński, J., . . . Li, Y. (2021). FIREcaller: Detecting frequently interacting regions from Hi-C data. *Computational and Structural Biotechnology Journal*, 19, 355–362. doi:10.1016/j.csbj.2020.12.026

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., . . .

. . .

Mirny, L. A. (2012, 10). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9, 999–1003. doi:10.1038/nmeth.2148

Schmitt, A. D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C. L., . . . Ren, B. (2016). A Compendium of

Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Reports*, 17, 2042–2059. doi:10.1016/j.celrep.2016.10.061

Yu G, Wang L, Han Y, He Q (2012). “clusterProfiler: an R package for comparing biological themes among gene clusters.” *OMICS: A Journal of Integrative Biology*, 16(5), 284-287. doi:

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9. doi:10.1186/gb-2008-9-9-r137

Annexes du Chapitre 4

B.1 Risk variants dominate protective variants in an affected-only design

To assess whether in an affected-only design, the contribution to the score statistic of a risk variant dominates the contribution of a protective variant with equal opposite effect, we considered two different scenarios corresponding to two different family structures where (1) one pair of first cousins are affected and (2) one pair of second cousins are affected. Here we perform the calculations for one variant and for a pair of first cousins. Let's define the sharing probability under the null $p_0 = 1/15$. Hence, the expected value under the null for this family configuration is given by: $\lambda_0 = 2*p_0 + 1*(1-p_0) = 1.06$. As an example, assuming a relative risk (RR) of 50 the expected statistic value under the alternative for a risk variant is $(1-\lambda_0)*P(G_j = 1) + (2-\lambda_0)*P(G_j = 2)$, where $P(G_j = 2) = 1/(1-p_0)/(50*p_0+1)$, while for a protective variant $P(G_j = 2) = 1/(1 - p_0)/(1/50 * p_0 + 1)$, which gives 0.71 and -0.06 respectively. Thus, we observed that for a risk variant the statistic increases with effect size while for protective variant, the statistic remains stable. We can conclude that risk variants contribute more to the score statistic than protective variants (See Figure 31 below).

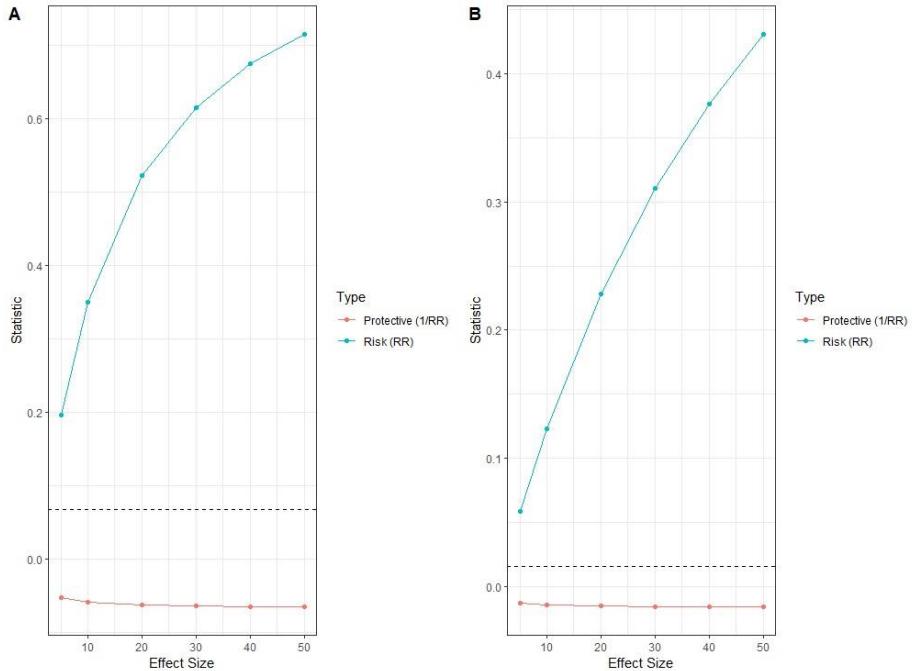


Figure 33: Expected contribution of a family to the score statistic value for different effect sizes considering either 1 risk variant (RR) or 1 protective variant (1/RR)

(A) Pair of affected first cousins, **(B)** Pair of affected second cousins. The sharing probabilities under the null, assuming variant frequency tending to 0, are $1/15$ and $1/63$, respectively, and indicated by the dashed horizontal lines

B.2 Score Variance

The score variance is given by:

$$\begin{aligned} Var[S_k(0)] &= Var\left[\sum_j w_j Z_{jk} \sum_i X_{fij}\right] \\ &= \sum_f \sum_j w_j^2 Z_{jk}^2 Var[X_{f.j}] + \sum_{j \neq j'} w_j w_{j'}' Z_{jk} Z_{j'k} Cov[X_{f.j}, X_{f.j'}] \end{aligned}$$

where X is the random variable corresponding to the minor allele count x and $X_{fj} = P_i X_{fij}$.

Developing variance and covariance components respectively we obtained:

$$Var[X_{f.1}] = \sum_{X_{f.1}} X_{f.1}^2 P(X_{f.1}) - (\sum_{X_{f.1}} X_{f.1} P(X_{f.1}))^2$$

And

$$Cov[X_{f.1}, X_{f.2}] = \sum_{X_{f.1}} \sum_{X_{f.2}} X_{f.1} X_{f.2} P(X_{f.1}; X_{f.2}) - (\sum_{X_{f.1}} X_{f.1} P(X_{f.1}))^2$$

Since these components are identical for all variants and variant pairs, they need to be computed only once for all variants.

Three possible cases can arise for the covariance term:

1. Variants are in perfect linkage disequilibrium and we have $Cov[X_{f.1}, X_{f.2}] = Var[X_{f.1}]$. This is treated in practice by removing one of the two variants since their genotypes are identical.
2. Variant independence can be assumed and $Cov[X_{f.1}, X_{f.2}] = 0$.
3. General case of imperfect linkage disequilibrium. The term $P(X_{f.1}; X_{f.2})$ would be difficult to compute since we condition on $X_j \geq 1, j = 1, \dots, p$. Instead, we set $Cov[X_{f.1}, X_{f.2}]$ to its upper bound

$$Cov[X_{f.1}, X_{f.2}] = \min(Var[X_{f.1}], Var[X_{f.2}])$$

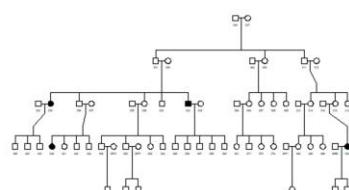
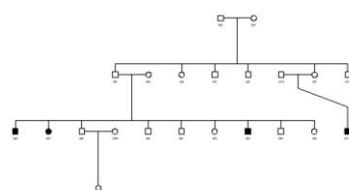
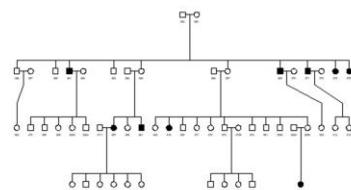
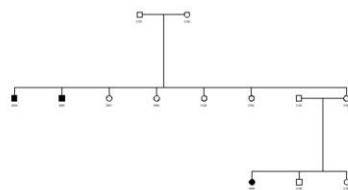
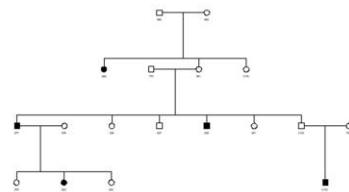
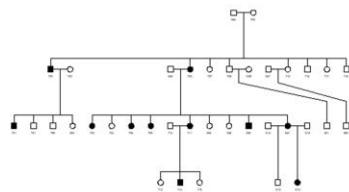
to obtain a conservative approximation of the variance of the score $S_k(0)$.

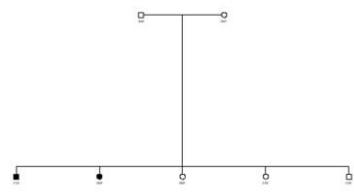
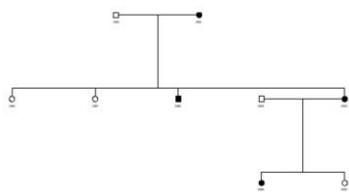
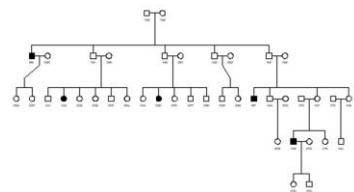
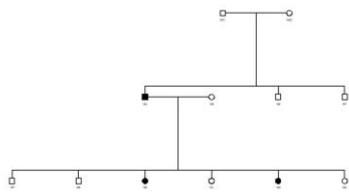
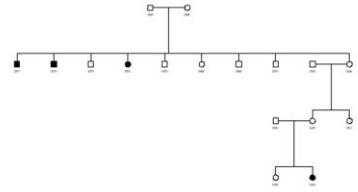
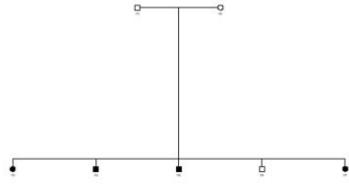
B.3 Numerical Simulation

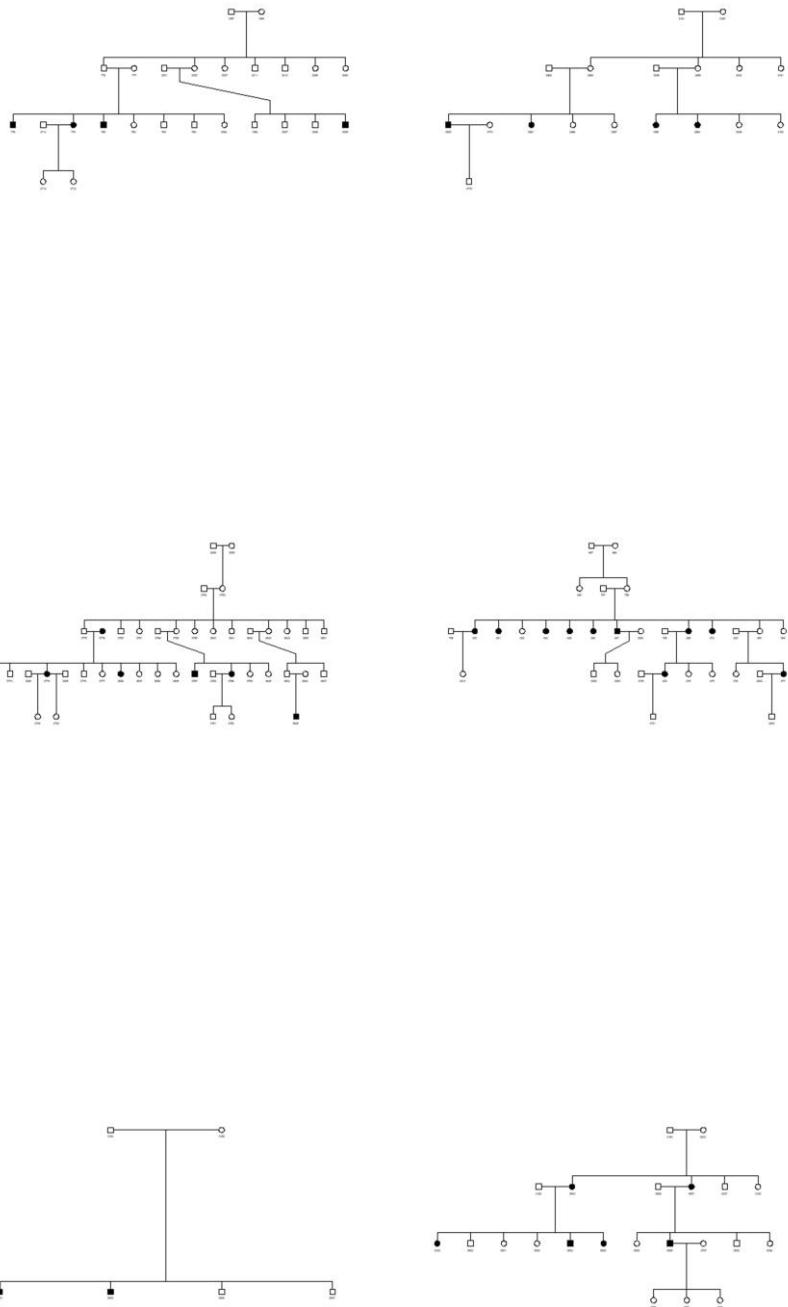
CRH1	CRH2	CRH3	CRH4	Out
158	65	2	41	244

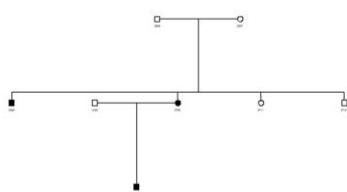
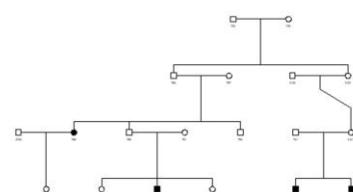
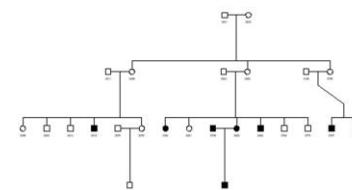
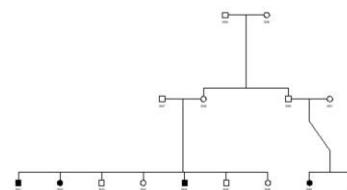
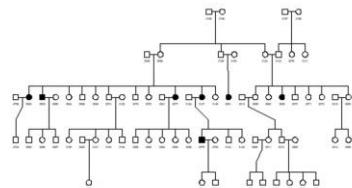
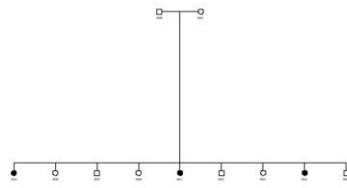
Table 3: Number of variants located within each CRH and outside

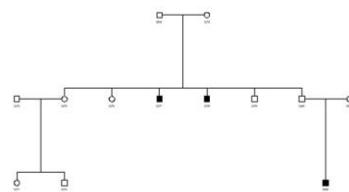
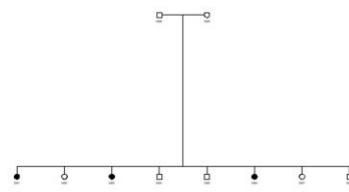
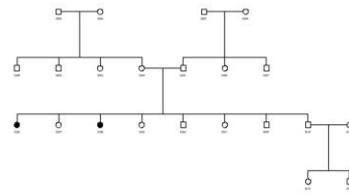
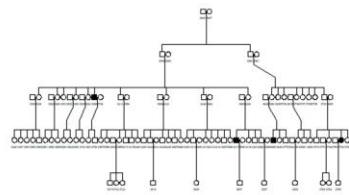
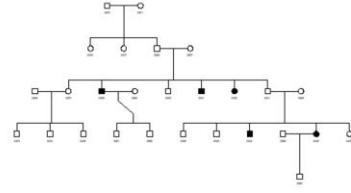
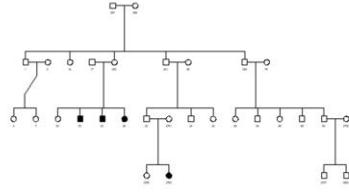
B.4 Pedigree Structures

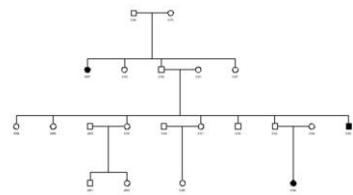
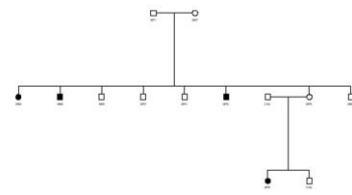
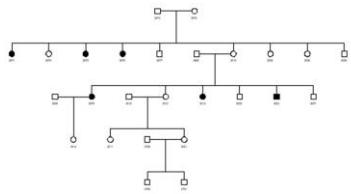
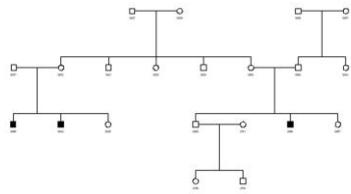
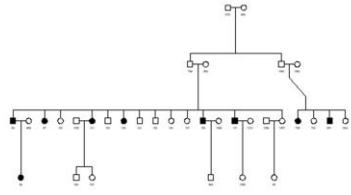
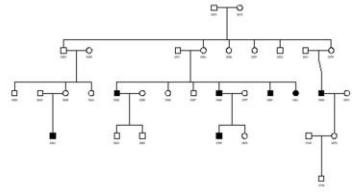


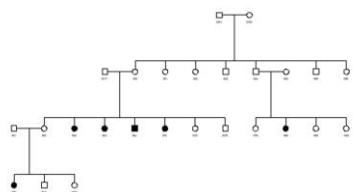
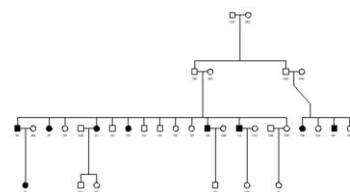
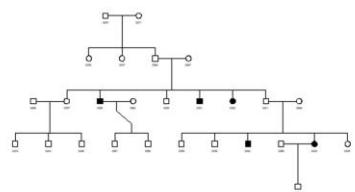
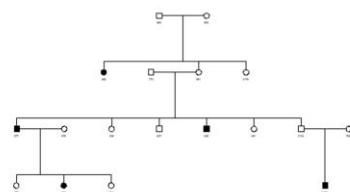
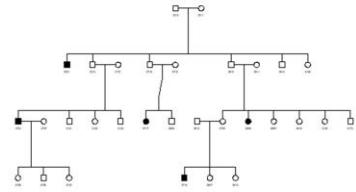
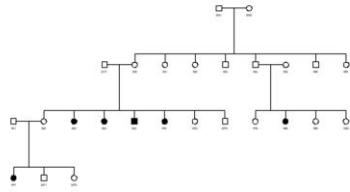












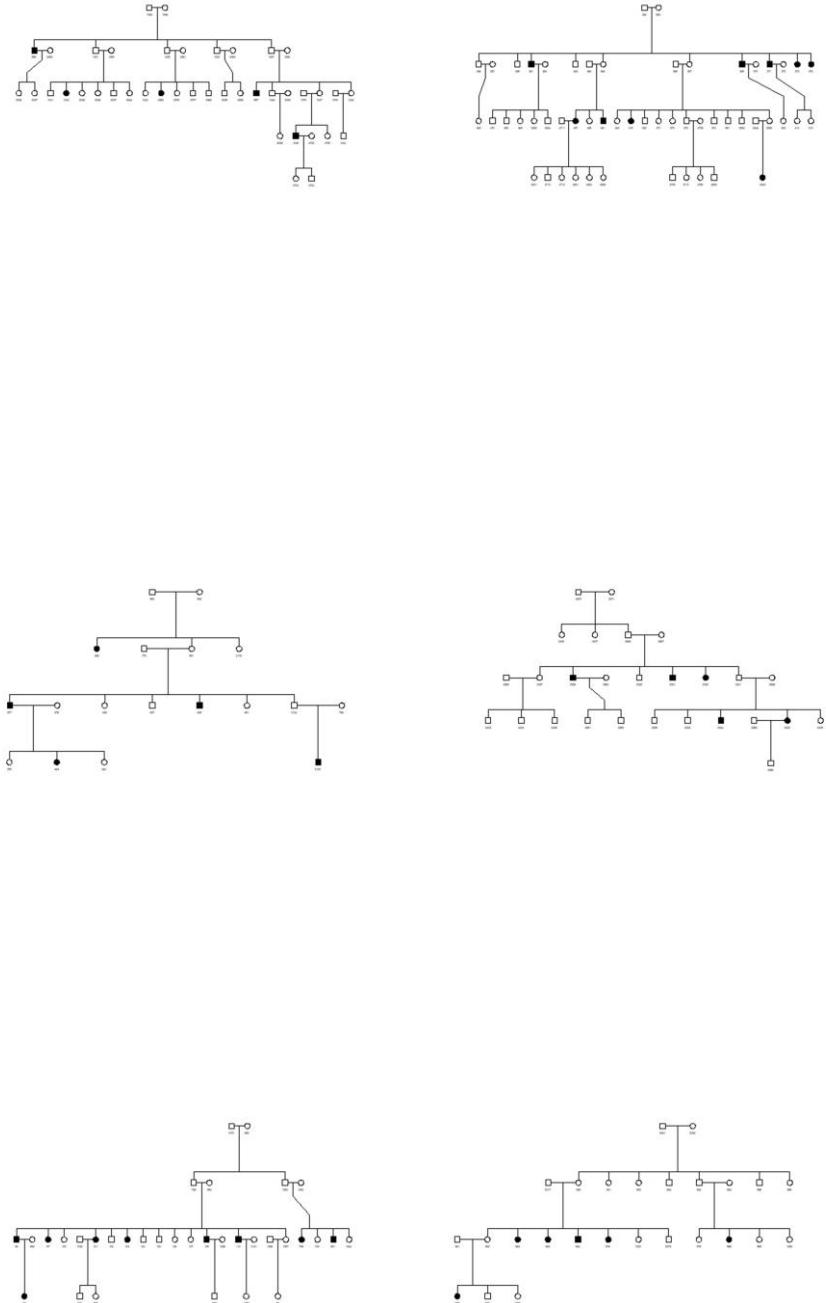


Figure 34: Pedigree structures for all the 52 families considered in the simulation studies

Affected subjects are indicated by filled squares or circles. 12 pedigree structures are repeated. It is worth noting that when we considered only small pedigrees, we restricted to the 17 families with fewer than 10 individuals. However, we repeated these families several times to ensure to have the same total number of affected subjects as in the primary sample.

B.5 Adaptation of RVS and RV-NPL including CRHs

Given the computational complexity of RVS is exponential in the number of variants tested together, we had to restrict the RVS tests to short windows. In order to adapt RVS to take CRHs into account, variants were sorted such that all variants belonging to the first CRHs encountered in the region are in consecutive order, followed by the variants in the second CRH and so on until the last CRH (Figure 33). Then, at the end, variants outside CRHs are listed. Windows were defined as sets of consecutive variants in that order, such that windows can span multiple elements of the same CRH (e.g. elements 1 and 2 of the blue CRH on (Figure 33). Computation of RVS tests over windows of 5 variants were attempted. When memory was insufficient, the rightmost variant was removed and the computation reattempted. Variants were removed until computation was successful. The window reduction was only needed for the partial sharing RVS test; the complete sharing RVS test computation was always successful on windows of 5 variants.

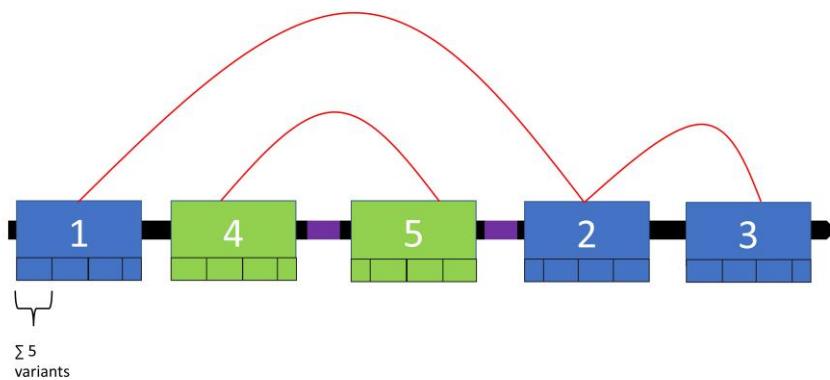


Figure 35: Adaptation of RVS including CRHs

The blue and green regions are two distinct CRHs and the black regions are outside any CRH. Purple rectangles correspond to windows outside CRHs.

We also adapted RV-NPL to integrate CRHs. Variants belonging to each CRH were specified through the option *include-vars* using **rvnpl collapse** when generating results from both RV-NPL and CHP-NPL. Moreover, to obtain an unified p-value, we applied ACAT on p-values from each CRH for RV-NPL and CHP-NPL and p-values from each window for the RVS tests.

B.6 Results

Type I Error Rate

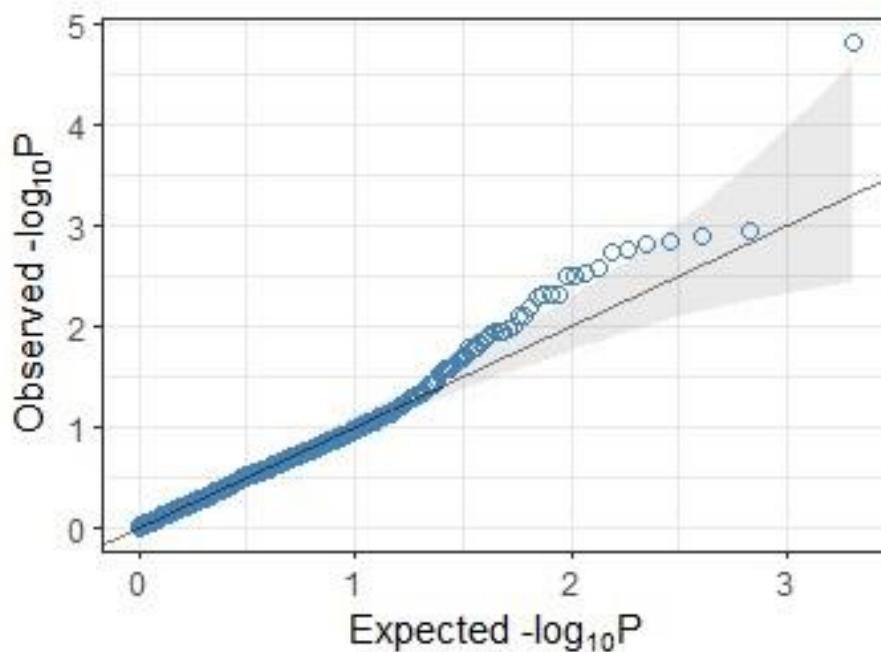


Figure 36: Quantile-Quantile plot of ACAT-Combined p-values for RetroFun-RVS_CRHs considering variant independence

Because only a small proportion of replicates had p-values for CRH 3, we omitted it in the analysis.

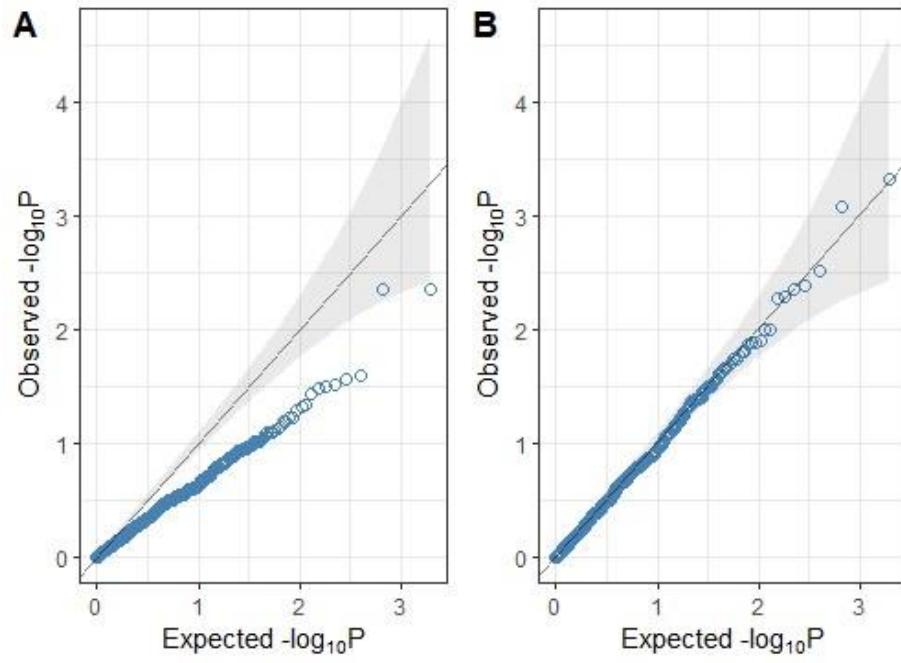


Figure 37: Quantile-Quantile plots of p-values for RetroFun-RVS incorporating no functional annotation

considering: **(A)** variant dependence and **(B)** variant independence. Because only a small proportion of replicates had p-values for CRH 3, we omitted it in the analysis.

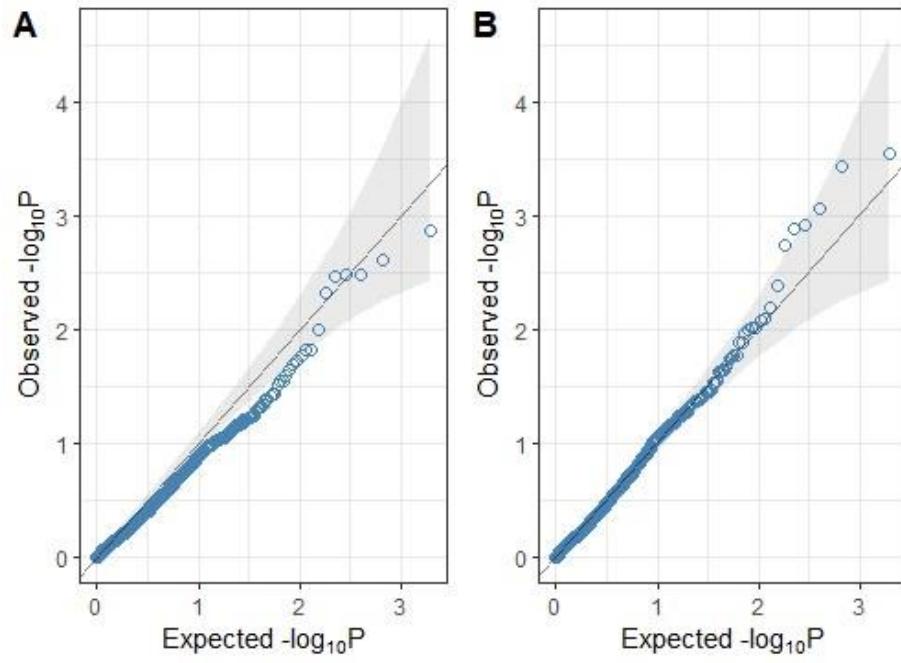


Figure 38: Quantile-Quantile plots of p -values for RetroFun-RVS_CRHs incorporating the first CRH considering: **(A)** variant dependence and **(B)** variant independence.

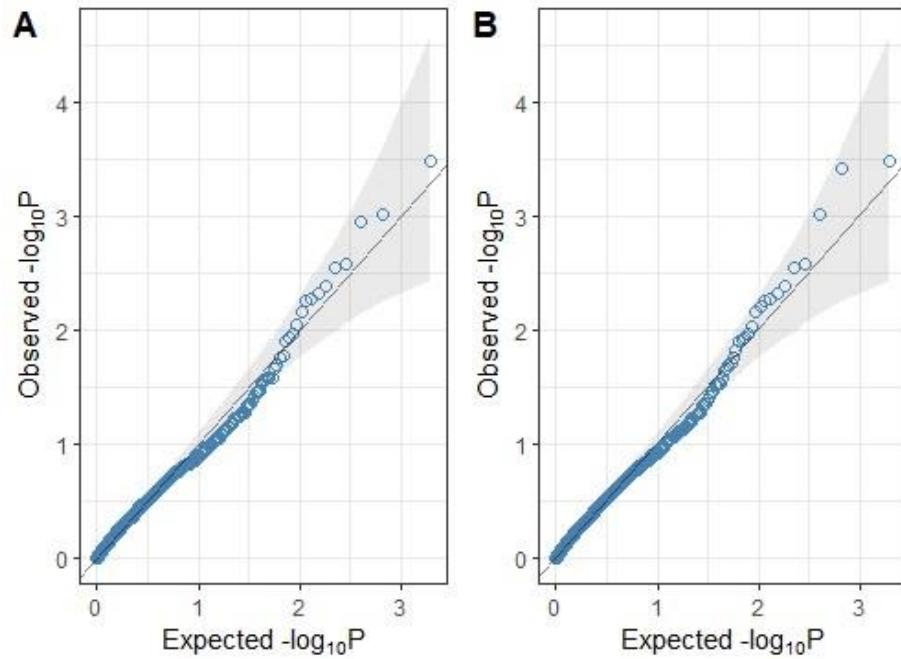


Figure 39: Quantile-Quantile plots of p -values for RetroFun-RVS_CRHs incorporating the second CRH considering: **(A)** variant dependence and **(B)** variant independence.

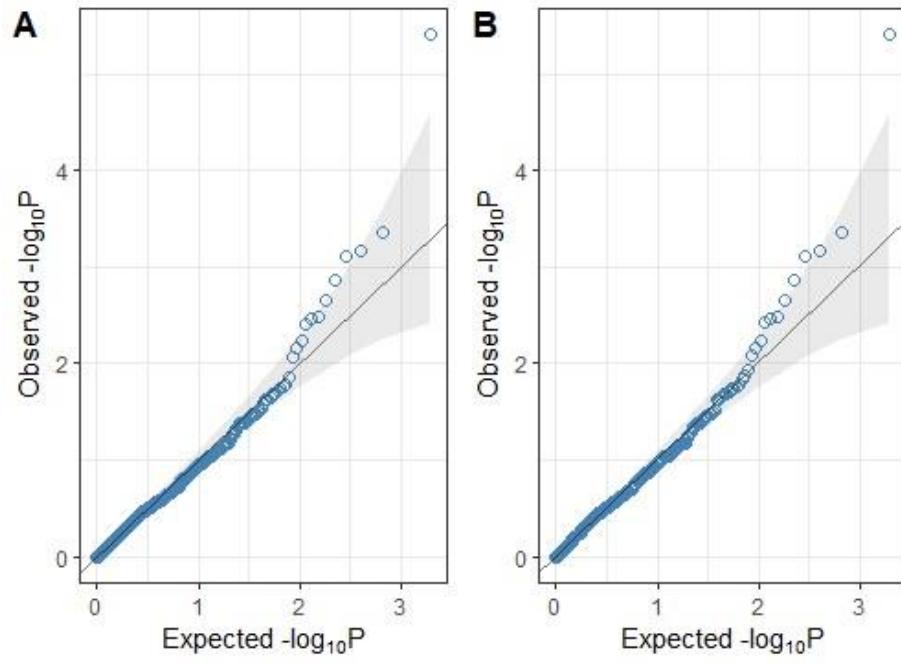


Figure 40: Quantile-Quantile plots of p -values for RetroFun-RVS_CRHs incorporating the fourth CRH considering: **(A)** variant dependence and **(B)** variant independence.

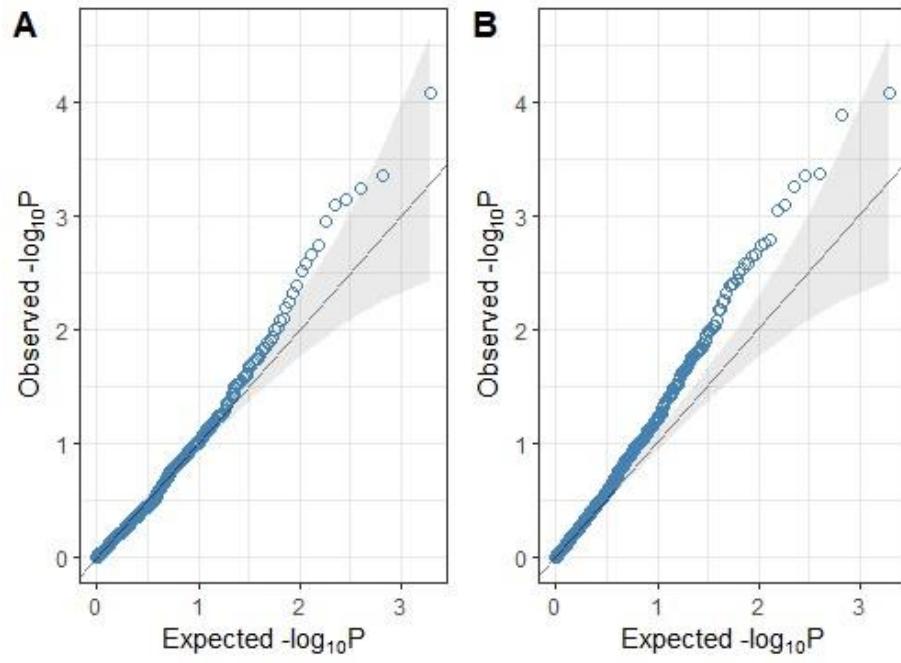


Figure 41: Quantile-Quantile plots of ACAT-Combined p-values for RetroFun-RVS_CRHs considering only small pedigrees

with: **(A)** variant dependence and **(B)** variant independence.

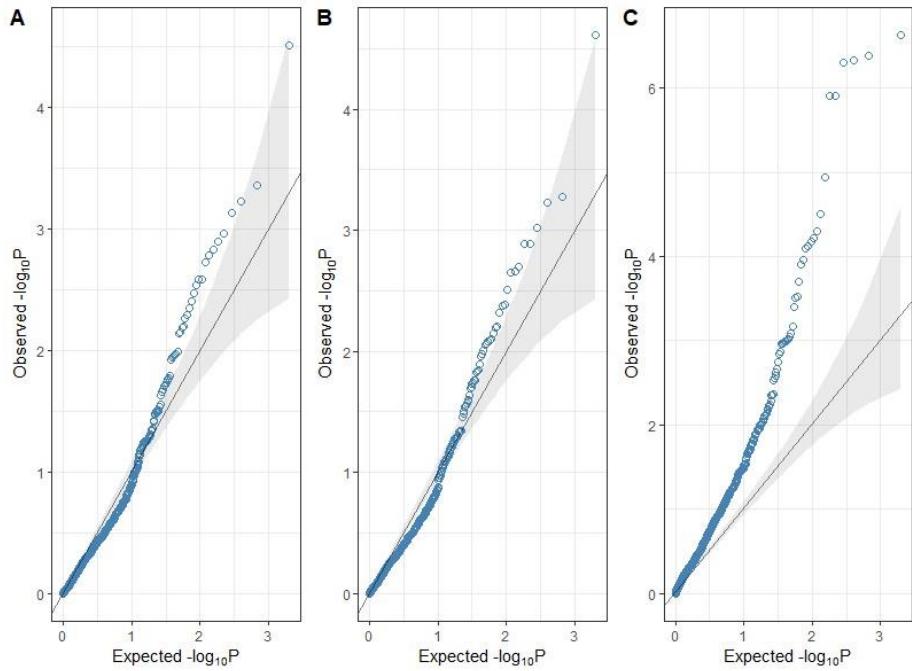


Figure 42: Quantile-Quantile plots of ACAT-Combined p -values considering variant dependence

for: **(A)** RetroFun-RVS_{Pairs}, **(B)** RetroFun-RVS_{Genes}, and **(C)** RetroFun-RVS_{Sliding-Windows}.

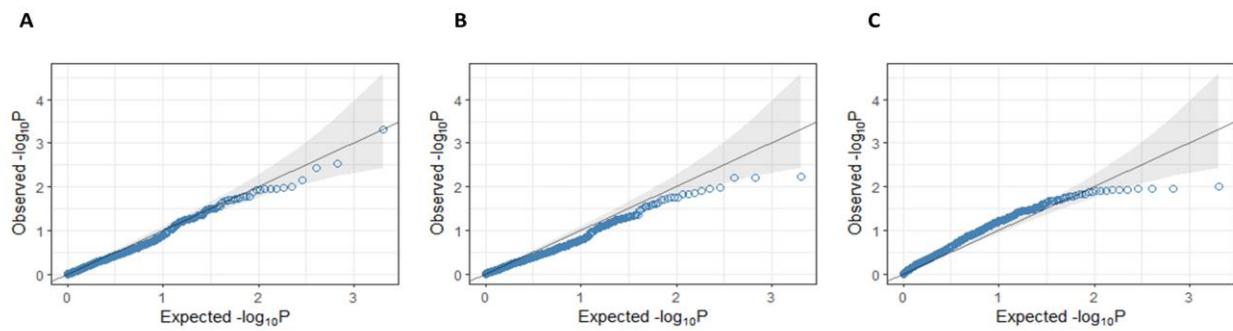


Figure 43: Quantile-Quantile plots of ACAT-Combined bootstrap-based p -values considering variant dependence

for: **(A)** RetroFun-RVS_{Pairs}, **(B)** RetroFun-RVS_{Genes}, and **(C)** RetroFun-RVS_{Sliding-Windows}. P-values were computed empirically based on 1,000 bootstraps.

Power

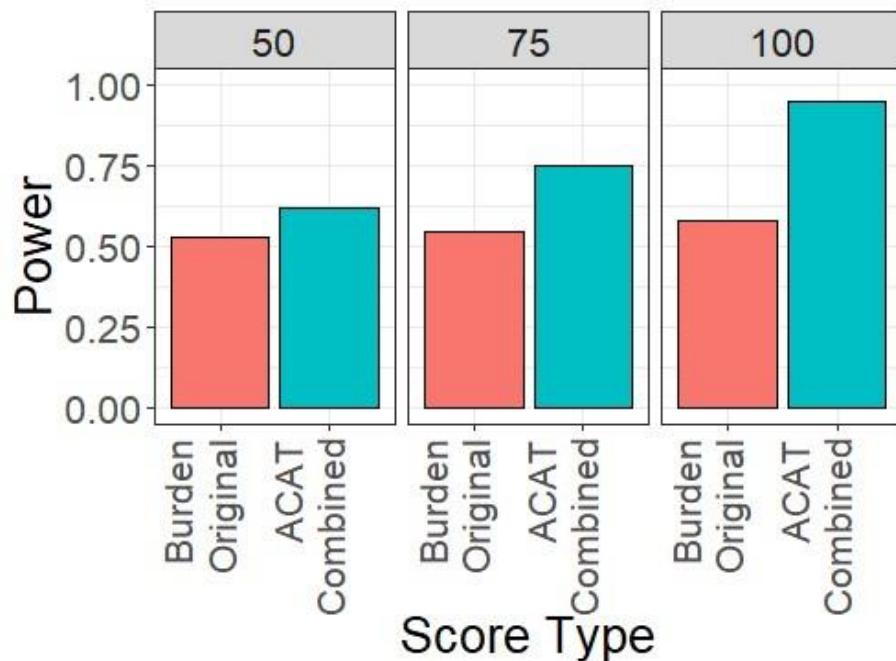


Figure 44: Power evaluation of RetroFun-RVS under different scenarios for 2% risk variants considering only small pedigrees

Power at different proportions of risk variants within the CRH, between RetroFun-RVS_{CRHs} with no functional annotation (Burden Original) and RetroFun-RVS_{CRHs} including the four CRHs (ACAT-Combined). Power was evaluated on the basis of 1,000 replicates.

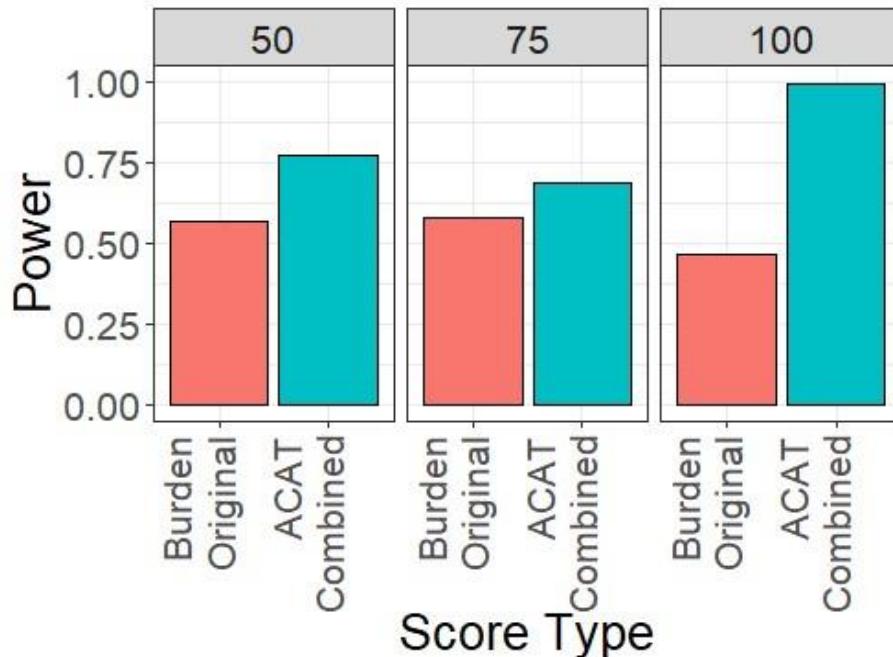


Figure 45: Power evaluation of RetroFun-RVS under different scenarios for 1% risk variants

Power at different proportions of risk variants within the CRH, between RetroFun-RVS_{CRHs} with no functional annotation (Burden Original) and RetroFun-RVS_{CRHs} including the four CRHs (ACAT-Combined). Power was evaluated on the basis of 1,000 replicates. We noticed that the Burden Original statistic varies across 75% or 50% causal scenarios and 100% causal. Investigating this aspect, we found that this may due to simulation specificity, where in 100% causal scenario, statistic values are smaller compared to the two others scenarios (average statistic values of 2,748,881; 3,080,295 and 3,110,373, for 100%, 75% and 50% causal, respectively), while variances remain stable (average statistic values of 116,071; 113,284; 116,541). We hypothesize that could be due to preferential choice of small families in 100% causal compared to 75% causal or 50% causal (average Pearson correlation: 0.83). See Figure 48.

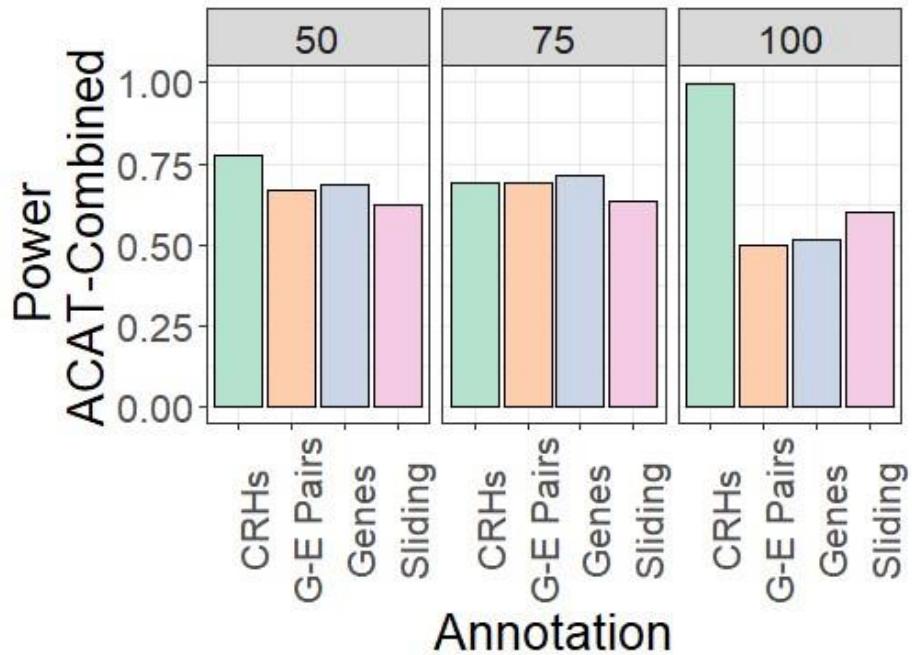


Figure 46: Power evaluation of ACAT-Combined p -values under different scenarios for 1% risk variants for RetroFun-RVS_CRHs (CRHs), RetroFun-RVS_Pairs (G-E Pairs), RetroFun-RVS_Genes (Genes), and RetroFun-RVS_Sliding-Window (Sliding)

To correct Type I error inflation observed in RetroFun-RVS_{Sliding-Window}, we only considered windows encompassing 30 variants or more. Power was evaluated on the basis of 1,000 replicates.

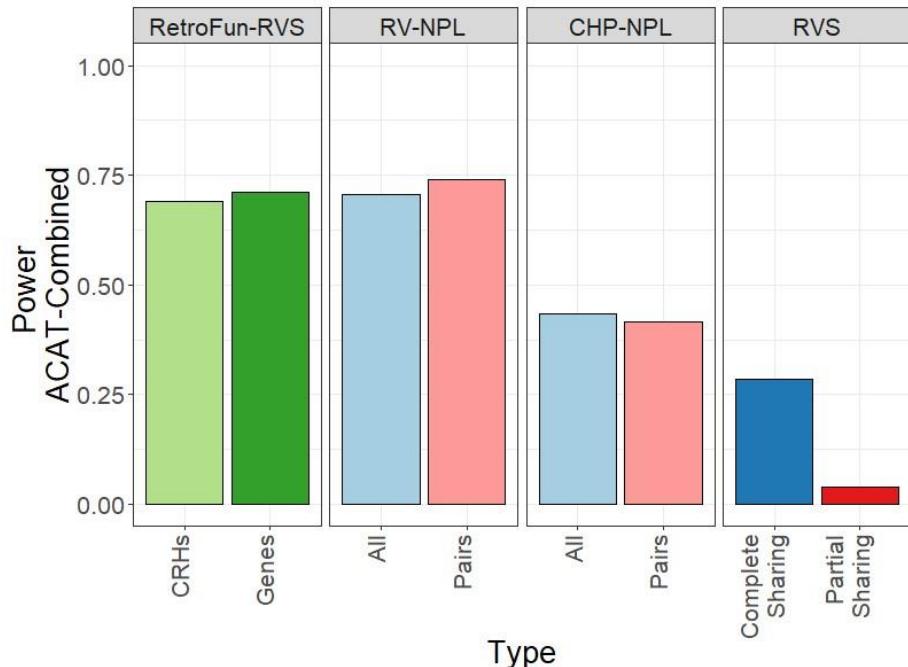


Figure 47: Power at 75% risk variants within one CRH between RetroFun-RVS_CRHs and other affected-only competing methods for 1% causal variant

Here we included RetroFun-RV_{genes} to mimic CHP-NPL procedure. Power for RetroFun-RVS_{CRHs} and RetroFun-RVS_{Genes} was evaluated on the basis on 1,000 replicates while for RV-NPL and RVS we generated 200 replicates.

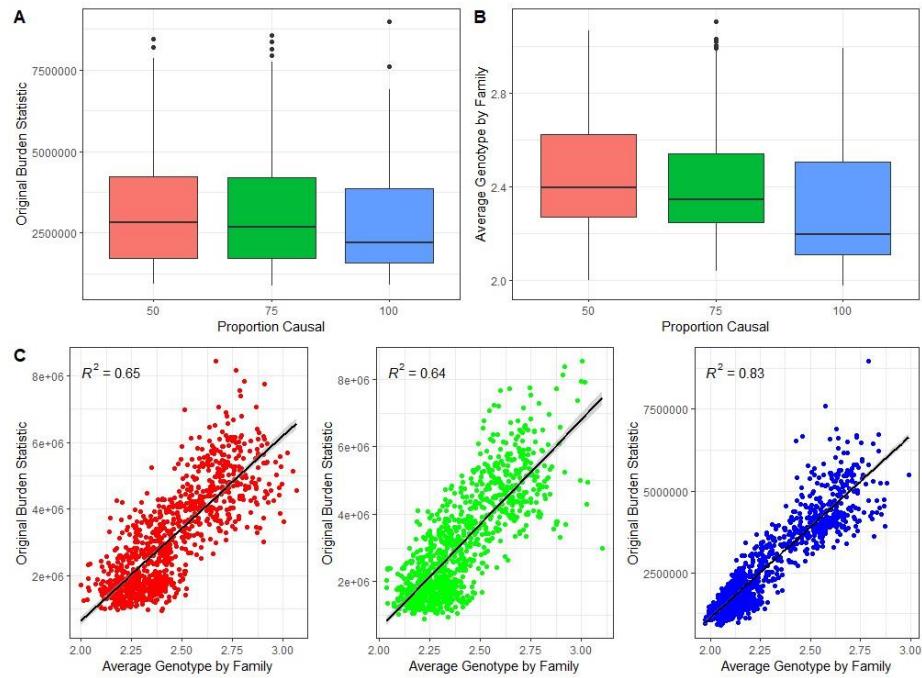


Figure 48: Relationship between average genotype values by family and Burden Original statistic at 1% causal

(A) Boxplots across the 1,000 replicates of the Burden Original statistic for each scenario.

(B) Boxplots across the 1,000 replicates of the average genotype value by family. **(C)**

Regression line across the 1,000 replicates between the average genotype value by family and the Burden Original statistic.

Annexes du Chapitre 5

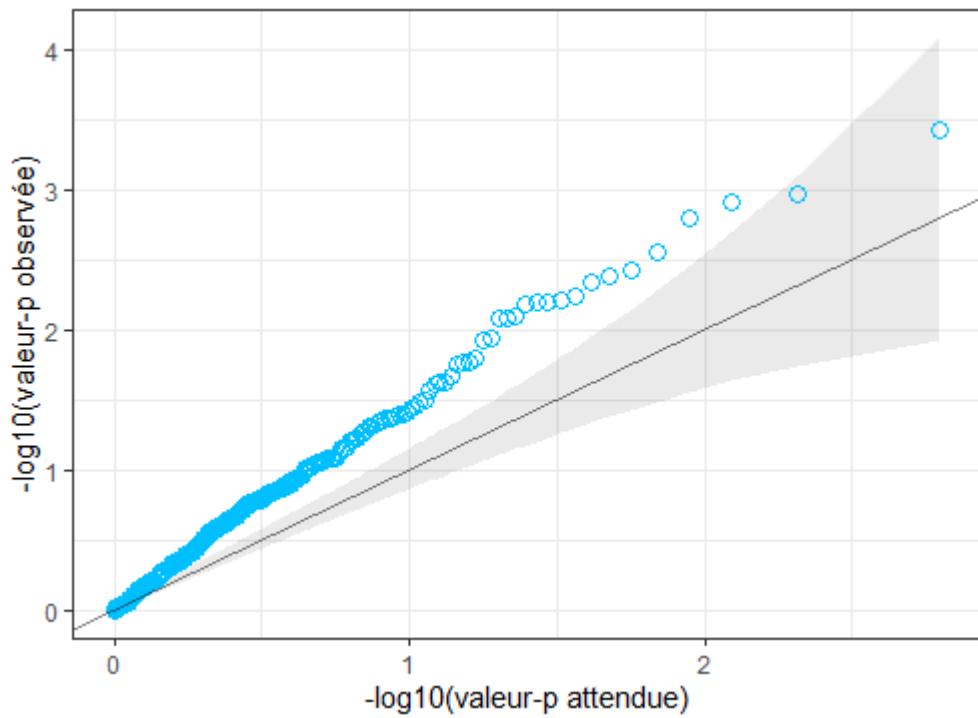


Figure 49: Diagramme Quantiles-Quantiles du $-\log_{10}$ des valeurs-p lorsque l'ensemble des familles sont considérées

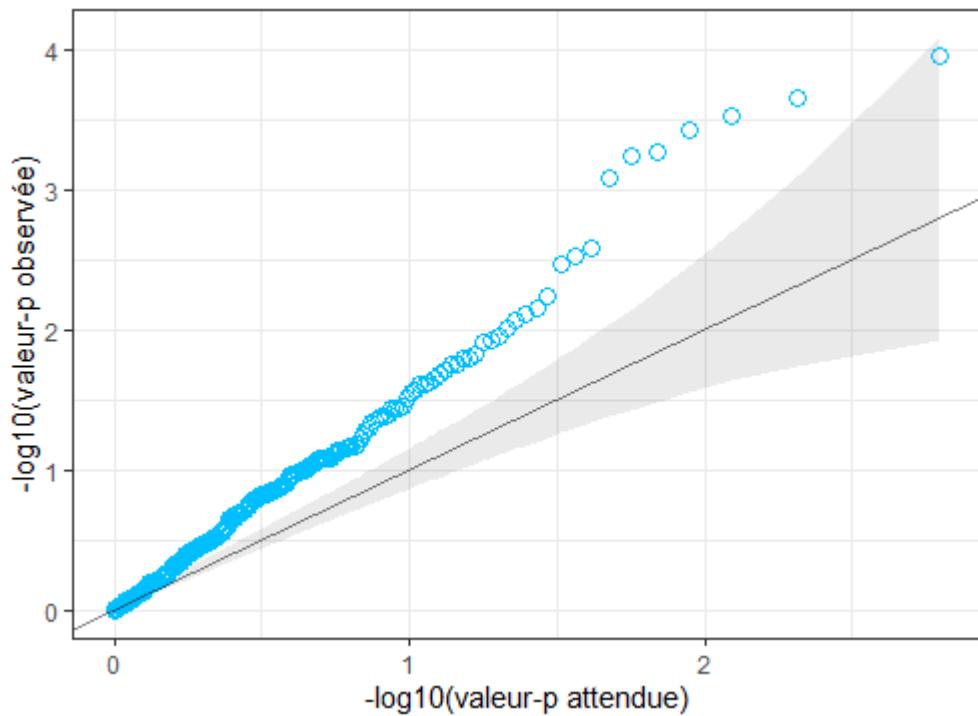


Figure 50: Diagramme Quantiles-Quantiles du $-\log_{10}$ des valeurs-p pour le sous-ensemble des familles syriennes pour lesquelles on observe de la consanguinité.

Annexes Discussion

C.1 Extension de RetroFun-RVS permettant l'intégration de covariables

De manière plus formelle, en supposant un vecteur de paramètres θ de taille $n \times m$ pour les c covariables, le modèle de vraisemblance rétrospective peut être étendu à :

$$P(G|Y, C) = \frac{\exp(\sum_{i \in D} \sum_{j=1}^p \beta_j x_{ij} + \sum_{l=1}^m \theta_l c_{il}) P(G|C)}{\sum_{G^*} \exp(\sum_{i \in D} \sum_{j=1}^p \beta_j x_{ij}^* + \sum_{l=1}^m \theta_l c_{il}) P(G^*|C)}$$

Le vecteur de paramètres θ de taille m est le vecteur de log-risques relatifs et c_{il} la $i^{ème}$ covariable pour le $i^{ème}$ individu. À noter qu'en supposant l'indépendance entre covariables et composantes génétiques, le modèle de vraisemblance rétrospective présenté ci-dessus peut être simplifié en :

$$P(G|Y, C) = \frac{\exp(\sum_{i \in D} \sum_{j=1}^p \beta_j x_{ij}) P(G)}{\sum_{G^*} \exp(\sum_{i \in D} \sum_{j=1}^p \beta_j x_{ij}^*) P(G^*)}$$

Cependant, la présence de corrélation entre certaines covariables et composantes génétiques pourrait jouer un rôle dans l'obtention de la valeur attendue sous l'hypothèse nulle, notamment dans l'estimation de $P(G|C)$.

C.2 Extension de RetroFun-RVS intégrant variants rares et communs

En supposant qu'un variant est défini comme rare lorsque sa fréquence d'allèle mineur est inférieure à un certain seuil et commun lorsque située au-dessus, le vecteur de paramètres de log-risques relatifs β pour l'effet des variants peut être décomposé en β_1 et β_2 , les vecteurs d'effets pour les variants rares et communs respectivement. En supposant l'absence de déséquilibre de liaison et d'interaction entre les effets des variants sur l'échelle multiplicative, en réutilisant la notation du chapitre quatre, le modèle de vraisemblance rétrospective pour une famille peut être réécrit en :

$$P(G|Y) = \frac{\exp(\sum_{i \in D} \sum_{j=1}^p \beta_{1j} x_{1ij} + \sum_{j'=1}^{p'} \beta_{2j'} x_{2ij'}) P(G_1) P(G_2)}{\sum_{G^*} \exp(\sum_{i \in D} \sum_{j=1}^p \beta_{1j} x_{1ij}^* + \sum_{j'=1}^{p'} \beta_{2j'} x_{2ij'}^*) P(G_1^*) P(G_2^*)}$$

Où x_{1ij} est le nombre d'allèles mineurs pour le variant rare j chez l'individu i et $x_{2ij'}$ le nombre d'allèles mineurs pour le variant commun j' chez l'individu i . En lien avec la

méthode présentée dans le chapitre quatre la probabilité de partage $P(G_1)$ peut être obtenue grâce à RVS (Sherman et al., 2019), alors que $P(G_2)$ peut être considérée connue ou estimée dans les familles (Schaid et al., 2010). Le modèle est alors équivalent au modèle de variants multiples proposé par Schaid et al. (2010). De plus, des fonctions de pondération différentes pour variants rares et communs peuvent être considérées en pratique (Ionita-Laza et al., 2013).

Une seconde approche consisterait à combiner les valeurs-p pour chaque annotation fonctionnelle considérant variants rares et communs. Soit en réutilisant ACAT (Liu et al., 2019), la valeur-p combinée pour l'annotation k peut être obtenue grâce à :

$$T_{ACATk} = \sum_{r=1}^2 \tan((0.5 - p_{kr})\pi)$$

Avec r=1 pour la valeur-p obtenue considérant les variants rares et r=2 les variants communs.

Références additionnelles:

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013 Jun 6;92(6):841-53. doi: 10.1016/j.ajhg.2013.04.015. Epub 2013 May 16. PMID: 23684009; PMCID: PMC3675243.