Procédures de tests multiples avec pondérations dans les études d'association pangénomiques

Ludivine Obry

université PARIS-SACLAY

Procédures de tests multiples avec pondérations dans les études d'association pangénomiques

Weighted multiple testing procedures in genome wide association studies

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577,
Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Sciences de la vie et de la santé
Graduate School : Sciences de la vie et santé
Référent : Université d'Évry Val d'Essonne

Thèse préparée dans l'unité de recherche Laboratoire de Mathématiques et Modélisation d'Évry (Université Paris-Saclay, CNRS, Univ Evry) sous la direction de **Cyril Dalmasso**, Maître de Conférence

Thèse soutenue à Évry-Courcouronnes, le 10 octobre 2023, par

Ludivine OBRY

Composition du jury

Elisabeth Petit-Teixeira Professeure des Universités, Université d'Évry Val d'Essonne	Présidente
Mathieu Emily Professeur des Universités, Institut Agro Rennes-Angers	Rapporteur & Examinateur
Pierre Neuvial Directeur de Recherche, CNRS, Institut de Mathématiques de Toulouse	Rapporteur & Examinateur
Philippe Broët Professeur des Universités - Praticien Hospitalier, Université Paris-Saclay	Examinateur
Anne-Louise Leutenegger Chargée de Recherche, INSERM NeuroDiderot, Hôpital Robert Debré	Examinatrice

THESE DE DOCTORAT
NNT : 2023UPASL084

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Procédures de tests multiples avec pondérations dans les études d'association pangénomiques

Mots clés : Études d'association pangénomiques, Taux de fausses découverte, Tests multiples pondérés

Résumé : Avec le développement récent des technologies de séquençage, il est aujourd'hui possible de réaliser des études d'association pangénomiques (GWAS) à très large échelle. Dans ce contexte, l'approche standard consiste à tester chaque marqueur génétique individuellement. Afin de limiter le nombre de faux positifs, des procédures de tests multiples visant à contrôler un risque d'erreur global sont appliquées. Cependant, les approches classiques sont limitées, d'une part, par le fait que la sélection initiale ne tire pas parti des informations a priori et des connaissances d'experts, d'autre part, par la difficulté à identifier des variants rares qui peuvent pourtant avoir des effets importants. L'incorporation de pondérations dans les procédures de tests multiples peut alors être une solution. Dans cette thèse, nous avons évalué différentes procédures de tests multiples avec pondérations dans le contexte spécifique des GWAS. Nous avons également introduit une approche originale permettant d'améliorer la puissance de détection des variants rares tout en maintenant une bonne puissance globale. Nous avons évalué les différentes procédures à travers une étude de simulations dont les résultats montrent les bonnes performances de notre approche par rapport aux procédures existantes. Les différentes méthodes ont été appliquées à un jeu de données réelles.

Title : Weighted multiple testing procedures in genome wide association studies

Keywords : Genome wide association studies, False discovery rate, Weighted MTP

Abstract : With the recent development of sequencing technologies, it is nowadays possible to perform genome-wide association studies (GWAS) on a very large scale. In this context, the standard approach is to test each genetic marker individually. To limit the number of false positives, multiple testing procedures aimed at controlling an overall error risk are applied. However, classical approaches have limitations. Firstly, they do not take advantage of prior information or expert knowledge in the initial selection process. Secondly, identifying rare variants that may have significant effects poses a challenge. Incorporating weights into multiple testing procedures can be a solution. In this thesis, we evaluated some recent weighted multiple testing procedures in the specific context of GWAS. We have also introduced an original approach to improve the detection power of rare variants while maintaining good overall power. We have evaluated the different procedures through a simulation study and the results show the good performance of our approach compared to existing procedures. The different methods were applied to a real dataset.

Remerciements

Pour commencer, je tiens à remercier comme il se doit mon directeur de thèse Cyril Dalmasso, sans qui tout cela n'aurait pas été possible. Je te remercie très sincèrement pour la confiance que tu m'as accordée au cours de ces dernières années et l'incroyable opportunité que tu m'as offerte, je t'en suis infiniment reconnaissante. Merci pour ta bienveillance à mon égard, quelles que soient les circonstances. Je sais pertinemment que ça n'a pas toujours été facile entre mon pessimisme légendaire et mes angoisses. Malgré cela, tu as toujours su trouver les bons mots au bon moment. Je ne te remercierais jamais assez d'avoir cru en moi, parfois plus que moi-même (pour ne pas dire presque tout le temps). Merci d'avoir toujours répondu présent lorsque j'en avais besoin, de m'avoir guidée quand je ne savais plus où aller et ainsi que de m'avoir poussée quand il le fallait. Merci également d'avoir toujours été à l'écoute, d'avoir su me dire de me reposer quand il le fallait. J'ai beaucoup apprécié travailler sous ta direction, tu as été un directeur de thèse exceptionnel (et les mots sont faibles). Ce fut très agréable d'évoluer, d'apprendre et de grandir à tes côtés. Encore une fois merci...

Je tiens à remercier également les rapporteurs de ma thèse Pierre NEUVIAL et Mathieu EMILY pour le temps accordé à la relecture approfondie de mon manuscrit, pour vos commentaires et vos remarques pertinentes qui m'ont permis d'élargir mes réflexions. Je remercie tout autant les membres de mon jury, Anne-Louise LEUTENEGGER, Philippe BROËT et Elisabeth PETIT-TEIXEIRA d'avoir pris de votre temps afin d'évaluer mes travaux effectués durant la thèse. Merci à toutes et tous pour votre présence, votre bienveillance, vos interrogations et l'intérêt que vous avez porté à mes travaux de thèse.

Je tiens également à remercier le Laboratoire de Mathématiques et de Modélisation d'Évry et ses membres de m'avoir accueilli dans leur équipe Statistiques et Génomes. Merci à toutes et à tous, de rendre ce lieu de travail accueillant, chaleureux, sympathique et convivial. Merci aux membres de l'équipe que j'ai eu en tant qu'enseignants (depuis la L1 pour certains). Je pense en particulier à Carène, Cyril et Nathalie, merci à vous pour votre pédagogie, gentillesse et vos conseils tout au long de mon parcours universitaire. Votre implication dans vos enseignements et votre bienveillance ont renforcé mon souhait d'enseigner. Je remercie par ailleurs les doctorants qui animent le laboratoire. Courage à celles et ceux qui soutiendront prochainement, au plaisir de vous revoir.

Parmi les personnes du laboratoire, je tiens à remercier tout particulièrement Margot, Maurice et Franck pour votre trio de choc et votre humour. Merci à vous d'apporter de la joie et de la bonne humeur dans le labo. Mention spéciale pour le rire de Maurice qu'on peut, parfois, entendre de l'autre côté du labo (ou encore lorsque tu râles contre ton PC). Mention spéciale pour notre "célèbre" acteur, Franck et son amour du chocolat et des gâteaux. Merci à toi pour ta bienveillance à mon égard, pour m'avoir fait rire quel que soit le moment de la journée ou l'humeur dans laquelle j'étais au départ. J'espère que tu trouveras une nouvelle personne "à torturer" avec tes "tic-tac" oraux. Merci par également à Nathalie pour ta bienveillance, ta sagesse, ton altruisme et ton côté maternel. Je te remercie de m'avoir toujours laissé ton bureau ouvert pour me plaindre de tout et de rien. Je remercie également Edmond, mon compagnon de bureau. Nous

avons commencé et terminé ensemble cette aventure. Et quelle aventure ! Je suis extrêmement fière de toi et je te remercie pour ton calme, ta patience et toutes les discussions que nous avons pu avoir dans notre bureau (que ce soit le 1er ou le 2nd). Merci à toutes et tous d'avoir contribué à rendre mon expérience de la thèse plus joyeuse. Merci.

Je tiens à remercier les membres de mon comité de thèse, Valérie CHAUDRU et Ismail AHMED, pour avoir accepté de faire partie de mon comité de suivi individuel. Merci pour votre bienveillance, votre écoute, vos conseils, vos avis et questions qui m'ont aidé à conduire mes travaux de thèse. Mention spéciale pour Valérie, je te remercie également pour ta confiance accordée en me donnant l'opportunité de passer de l'autre côté de la barrière en enseignant pour la première fois à l'université.

Je remercie également ma famille (et belle famille) et mes proches pour votre infaillible appui dans cette aventure. Merci maman, papa, Valentin et Marvin d'avoir toujours cru en moi quelle que soit l'aventure entreprise. Mention spéciale pour mon petit numéro vert, merci pour ton admirable soutien et de toujours t'être rendu disponible pour un "petit" coup de téléphone. Je remercie ma famille de cœur : mes amis. Merci à vous tous pour votre soutien, merci de m'avoir changé les idées quand il le fallait, merci d'avoir apporté un sourire sur mon visage dans les moments les plus difficiles. Je tiens à vous exprimer une immense gratitude. Je tiens à remercier tout particulièrement les deux personnes avec qui je vis, merci pour votre patience, votre compréhension et votre appui incommensurable et sans failles. Milles merci à vous deux qui m'avez aidé pour la correction de ce manuscrit.

Enfin, je tiens à remercier tous ceux que je n'ai pas cités et qui pourtant, ont contribué d'une manière ou d'une autre à rendre cette thèse possible.

À toutes et à tous, je vous exprime mes plus sincères remerciements, ainsi que toute ma reconnaissance.

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1
INTRODUCTION

Dans ce premier chapitre, les bases de la biologie moléculaire et les variations génétiques sont tout d'abord présentées, suivies des approches méthodologiques permettant d'identifier les variants génétiques impliqués dans les maladies humaines, en particulier les variants de faibles fréquences dans les GWAS. Nous présentons également les enjeux méthodologiques et les objectifs de la thèse.

Sommaire

1.1 . Variations génétiques

1.1.1 . L'ADN

La cellule, unité fondamentale et structurelle de base de tout organisme vivant, renferme en son noyau l'ADN (Acide DesoxyriboNucléique, DNA en anglais) qui est caractérisé par une séquence constituée de 4 nucléotides : l'Adénine (A), la Thymine (T), la Guanine (G) et la Cytosine (C) (Figure 1.1). C'est la succession et l'enchaînement de ces 4 nucléotides qui constituent le message génétique.

Chaque molécule d'ADN est constituée de deux brins complémentaires enroulés en hélice. Les deux brins sont reliés entre eux par des liaisons hydrogènes. La complémentarité des deux brins réside dans l'appariement spécifique des nucléotides : l'adénine se lie toujours à la thymine et la guanine se lie toujours à la cytosine.

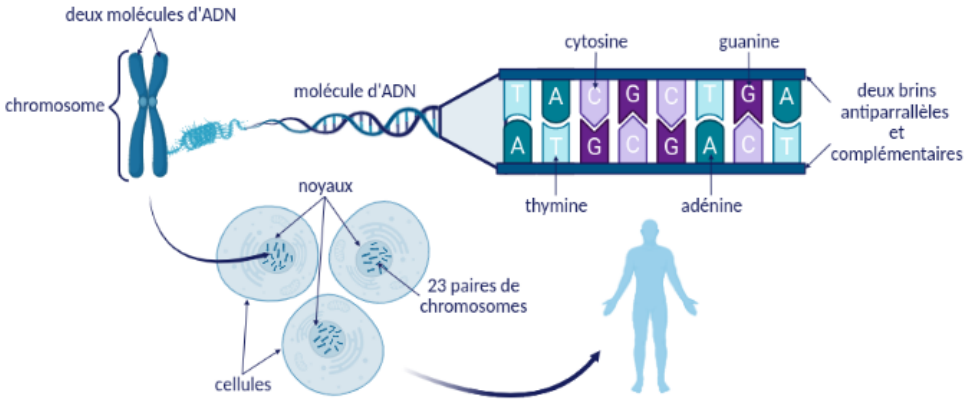

Figure 1.1 – De l'Homme à l'ADN. (Réalisé avec BioRender.com)

Ces molécules d'ADN s'enroulent autour de protéines pour former un chromosome. Alors que les bactéries possèdent un ou deux chromosomes circulaires, les êtres humains en possèdent 46 linéaires groupés par paires. Ils sont hérités des parents et pour chaque paire, un chromosome provient de la mère et l'autre du père. Parmi ces paires de chromosomes, 22 sont dites homologues, c'est-à-dire de tailles et formes égales et possédant la même disposition de gène (à quelques exceptions près). La 23ème paire détermine à elle seule le sexe d'un individu à partir des chromosomes sexuels X et Y (XX : femme, XY : homme).

L'ensemble des chromosomes est contenu dans le noyau de la cellule chez les animaux et les plantes. Si ce matériel génétique, appelé génome, est le même dans chaque cellule d'un individu, on observe tout de même des différences (localisations, morphologiques, fonctions). Prenons l'exemple des cellules sanguines parmi

lesquelles on retrouve les globules rouges, les globules blancs et les plaquettes. Bien qu'elles soient toutes des cellules sanguines, elles diffèrent selon leurs morphologies et fonctions : les globules rouges interviennent dans le transport de l'oxygène dans le sang, les plaquettes jouent un rôle dans la coagulation du sang et les globules blancs qui se décomposent en cinq types cellulaires différents jouent quant à eux un rôle dans l'immunité d'un individu (Figure 1.2). La différence réside dans l'expression des gènes de chacune d'entre elles et varie en fonction du temps et de l'environnement.

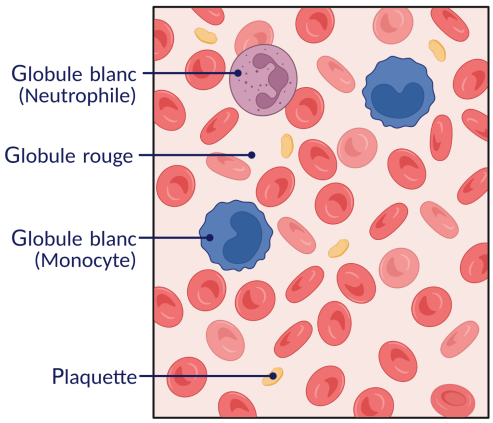

Figure 1.2 – Les cellules sanguines. (Réalisé avec BioRender.com)

1.1.2 . Les gènes

On définit un gène comme une portion de chromosome contenant l'information génétique codant pour une molécule (protéine) ou un caractère observable (phénotype). Chez les eucaryotes, organismes vivants dont le génome est contenu dans un noyau, un gène est structuré en régions codantes et non codantes appelées respectivement exons et introns (Figure 1.3). Les gènes des organismes unicellulaires dépourvus de noyau, autrement appelés procaryotes, ne possèdent pas d'introns.

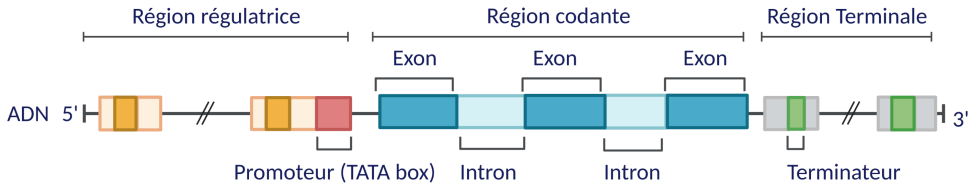

Figure 1.3 – Structure d'un gène eucaryote. (Réalisé avec BioRender.com)

Les exons d'un gène peuvent alors être transcrits ou non en ARN (Acide Ribo-Nucléiques) par le biais de la transcription. Ils sont ensuite traduits en protéines (Figure 1.4). Les introns d'un gène contiennent quant à eux des séquences régulatrices déterminant dans quel type de cellules et à quel moment les exons seront transcrits puis traduits en protéine.

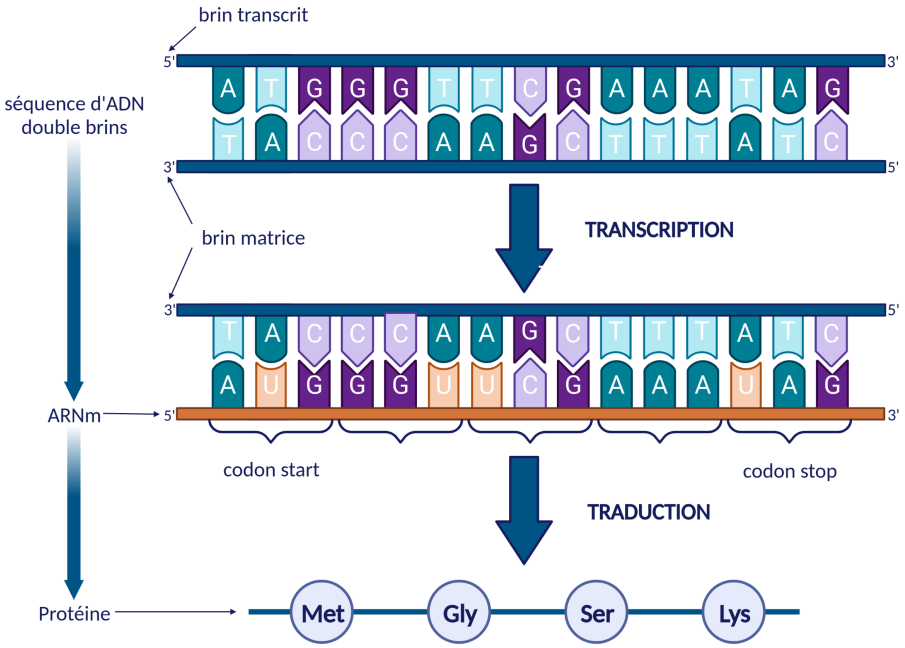

Figure 1.4 – Transcription et traduction d'un gène en protéine. (Réalisé avec BioRender.com)

1.1.3 . Polymorphismes et SNPs

Un gène est dit polymorphe lorsqu'il existe plusieurs variantes de celui-ci au sein d'une population. La plupart des variations génétiques résultent de mutations correspondant à une modification aléatoire spontanée ou induite par un agent mutagène de la séquence d'ADN. Il existe différents types de mutations dont la taille peut varier d'un seul nucléotide à l'ensemble du chromosome.

Les polymorphismes les plus communs sont les SNPs (*Single Nucleotide Polymorphism* en anglais) qui représentent environ 90% de toutes les variations génétiques humaines (Collins et al., 1998; Brookes, 1999). Ils sont caractérisés par une variation ponctuelle d'une seule base dans la séquence d'ADN (Figure 1.5). Ces mutations génétiques ponctuelles résultent le plus couramment d'une substitution d'un nucléotide, mais peuvent aussi découler d'une insertion ou d'une délétion d'une base.

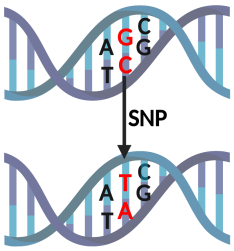

Figure 1.5 – Single Nucleotide Polymorphism. (Réalisé avec BioRender.com)

Les différentes versions d'un gène ou d'un SNP sont appelées des allèles. Un individu diploïde hérite d'une copie de chaque gène de chacun des parents, il possède ainsi deux allèles pour chaque polymorphisme. Si deux allèles sont identiques pour un gène donné, l'individu est dit homozygote pour ce gène. Si ces allèles sont différents, l'individu est dit hétérozygote pour ce gène. Le génotype d'un individu décrit la composition allélique d'un individu pour un gène, un groupe de gènes voire l'ensemble du génome. Par conséquent, le génotype d'un polymorphisme donné correspond à la description des allèles que porte l'individu pour ce polymorphisme.

1.1.4 . Le principe de Hardy-Weinberg

D'après le principe d'Hardy-Weinberg, les fréquences alléliques d'une population sont constantes au fils des générations sous les conditions suivantes :

- Les fécondations se font au hasard dans la population, c'est-à-dire que les couples et gamètes se rencontrent au hasard (panmixie et pangamie).
- Il y a absence de sélection, mutation et migration dans la population.
- Aucun croisement entre générations différentes dans la population.
- La taille de la population est infinie.

Dans le cas d'un SNP bi-allélique, d'allèles A et a notons :

- p, la fréquence de l'allèle A $(0 < p < 1)$
- q, la fréquence de l'allèle a $(0 < q < 1$ et $q = 1 - p)$

On définit l'allèle le plus fréquent comme l'allèle majeur et le moins fréquent comme l'allèle mineur. On note MAF (MAF pour *Minor Allele Frequency* en anglais) la fréquence de l'allèle mineur. L'équilibre de Hardy-Weinberg se traduit par la relation : $p^2 + 2pq + q^2 = 1$ où p^2 est la fréquence du génotype AA, $2pq$ est la fréquence du génotype Aa et q^2 est la fréquence du génotype aa.

Ce principe est exploité dans notre étude de simulation lors de la génération des génotypes (Chapitre 5).

1.1.5 . Déséquilibre de liaison

Au sein du génome, il existe des associations non aléatoires entre allèles situés sur un même chromosome que l'on appelle Déséquilibre de Liaison (DL ou LD pour *Linkage Disequilibrium* en anglais). La présence d'un déséquilibre de liaison entre deux SNPs implique une chance plus élevée que le voudrait le hasard d'être transmis conjointement à la génération suivante. Les déséquilibres de liaison trouvent leurs origines dans les forces évolutives telles que la dérive génétique, la sélection naturelle ou encore la naissance de nouvelles mutations. Ils sont ensuite transmis de génération en génération.

Prenons le cas de deux SNPs bi-alléliques dont le premier possède les allèles A et a de fréquences f_A et f_a respectivement et le second possède les allèles B et b de fréquences f_B et f_b respectivement. Dans le cas où ces SNPs sont indépendants, la fréquence de chaque haplotype (paire d'allèles transmis ensemble) est le produit des fréquences de chacun des allèles, soit :

$$\begin{cases} f_{AB} = f_A \times f_B \\ f_{ab} = f_a \times f_b \\ f_{Ab} = f_A \times f_b \\ f_{aB} = f_a \times f_B \end{cases}$$

Dans le cas contraire, la fréquence de chaque haplotype est différente, c'est ce que l'on appelle le déséquilibre de liaison. En d'autres termes, un déséquilibre de liaison est retrouvé entre deux SNPs lorsque leur transmission simultanée est plus fréquente que celle prédite par le produit de leurs fréquences individuelles.

Différentes mesures du déséquilibre de liaison ont été introduites, la première, notée D, introduite par Robbins (1918) est définie comme la différence entre les fréquences haplotypiques observées et celles attendues dans le cas où les SNPs sont indépendants. Ainsi, le coefficient de déséquilibre de liaison entre les allèles A et B donne :

$$D = f_{AB} - f_A \times f_B$$

Cette mesure, bien que facile à quantifier, est très peu utilisée en pratique car elle très dépendante des fréquences alléliques ce qui rend impossible la comparaison entre différentes valeurs de D. C'est pour cette raison que d'autres mesures normalisées par rapport aux fréquences alléliques ont été introduites.

Lewontin (1964) a introduit une mesure du DL normalisée avec les fréquences alléliques :

$$D' = \frac{D}{D_{max}} \text{ où } \begin{cases} D_{max} = min(f_A \times f_b; f_a \times f_B) \text{ si } D > 0 \\ D_{max} = min(f_a \times f_b; f_A \times f_B) \text{ si } D < 0 \end{cases}$$

Les valeurs de D' varient de -1 à 1. Si D' est nul, cela signifie que les SNPs sont indépendants. Si D' est égal à 1 ou -1, cela indique la présence d'un DL total entre les deux SNPs, c'est-à-dire la présence d'une association préférentielle entre deux allèles et qu'au moins un des haplotypes n'est pas observé dans la population.

La mesure la plus couramment utilisée dans le cadre des études d'association pangénomiques est celle introduite par Hill and Robertson (1968), notée r^2, qui est normalisée avec toutes les fréquences alléliques :

$$r^2 = \frac{D}{(f_A \times f_B \times f_a \times f_b)}$$

Les valeurs de r^2 varient de 0 à 1. Si le r^2 est nul, comme avec la mesure précédente, cela signifie que les SNPs sont indépendants. Si le r^2 est égal à 1, cela indique la présence d'un déséquilibre de liaison total entre les deux SNPs et qu'ils sont parfaitement corrélés.

1.2 . Étude des polymorphismes

L'un des objectifs en génétique est de trouver le ou les gènes responsables du phénotype observé, qui, le plus souvent, correspond à une maladie. L'étude des polymorphismes et leur identification permettent de mieux comprendre les mécanismes impliqués dans les pathologies complexes, d'améliorer les objectifs ou encore de définir des stratégies thérapeutiques. Plusieurs approches existent afin d'identifier ces marqueurs et diffèrent selon le type de maladie étudiée, son mode de transmission et la population d'étude.

1.2.1 . Analyses de liaison

Les analyses de liaison visent principalement à identifier les facteurs génétiques impliqués dans une maladie héréditaire mendélienne. Les maladies mendéliennes sont caractérisées par des mutations rares dans la population, mais élevées dans les familles dans lesquelles il existe un cas. Ce sont des maladies sévères, invalidantes, voire fatales et ayant une pénétrance élevée (probabilité élevée de développer la maladie lorsqu'on est porteur du génotype à risque).

Les analyses de liaison se concentrent sur la co-transmission des allèles et du phénotype au fil des générations au sein d'individus d'une même famille. Si les analyses de liaison ont connu un franc succès jusqu'aux années 2000 et ont permis d'identifier un grand nombre de polymorphismes impliqués dans des maladies mendéliennes, elles sont cependant limitées en ce qui concerne les maladies multifactorielles. En effet, bien que les analyses de liaison puissent être effectuées sur les maladies complexes, elles manquent souvent de puissance.

Contrairement aux maladies mendéliennes, les maladies complexes (ou maladies multifactorielles) ont une très faible agrégation familiale, une pénétrance faible (probabilité faible de développer la maladie lorsqu'on est porteur du génotype à risque) et dépendent de plusieurs facteurs tels que des facteurs environnementaux et génétiques.

1.2.2 . Études d'association pangénomiques

Les études d'association ont été introduites afin de répondre au manque de puissance des analyses de liaison sur les maladies complexes. Elles visent ainsi à identifier les SNPs impliqués dans une maladie au sein d'une large cohorte d'individus généralement non apparentés. Elles se concentrent sur des corrélations potentielles entre les SNPs et la maladie dans toute la population étudiée. Il est cependant possible de réaliser des études d'association à partir de données familiales (Ott et al., 2011; Won et al., 2015).

Historiquement, ces études étaient réalisées sur des portions de génome à partir de gènes candidats, c'est-à-dire sur un ensemble de gènes sélectionnés en fonction d'hypothèses a priori sur leur rôle dans la maladie étudiée. Cette méthode présente l'avantage d'être peu coûteuse et relativement rapide puisque seule une région prédéfinie par des a priori biologiques est étudiée. Elle est cependant limitée puisque dépendante de ces mêmes connaissances a priori qui peuvent parfois être insuffisantes (Tabor et al., 2002; Patnala et al., 2013).

Grâce à l'avancée technologique et scientifique, les analyses sur le génome entier (recherche pangénomique) ont été rendues possibles et moins coûteuses. La recherche pangénomique consiste à réaliser l'étude (analyse de liaison ou étude d'association) sur l'ensemble du génome sans a priori biologique (Risch and Merikangas, 1996). Cette méthode présente l'avantage d'être sans a priori permettant ainsi de découvrir de nouveaux facteurs impliqués ou responsables de la maladie étudiée sans être limités à un ensemble de gènes (ou région) préalablement sélectionnés (Sladek et al., 2007; Barrett et al., 2008).

Les études d'association pangénomiques (GWAS pour *Genome-Wide Association Studies* en anglais) sont largement utilisées actuellement et ont permis de trouver un grand nombre de SNPs impliqués dans de nombreuses maladies, généralement proches de gènes jusqu'alors insoupçonnés. L'avancée des techniques de séquençage a amélioré la connaissance des polymorphismes et a permis de construire un catalogue fourni de SNPs référencés avec de nombreuses informations sur les DL et les haplotypes correspondants.

Deux projets ont largement contribué à la caractérisation d'un grand nombre de SNPs le long du génome ainsi que leur structure de corrélation pour différentes populations : le projet Hapmap et le projet 1000 Genomes.

Projet Hapmap : Lancé en 2002, le projet Hapmap a pour objectif d'identifier l'ensemble des SNPs humain ainsi que d'établir une carte des blocs de DL et d'haplotypes la plus complète possible pour différentes populations. Ce projet ambitieux s'est déroulé en trois grandes phases (HapMap et al., 2005, 2007, 2010).

Projet 1000 Genomes : Lancé en 2008, le projet 1000 Genomes, partage le même objectif que le projet HapMap, c'est-à-dire caractériser l'ensemble des SNPs humain en établissant une carte des blocs de DL et d'haplotypes la plus complète possible pour différentes populations (Project et al., 2010). Contrairement au projet HapMap, les SNPs ayant de faibles MAF(1%<MAF<5%) ont été inclus dans le projet 1000 Genomes. Ainsi, au cours de ce projet, 2500 individus issus de 28 populations différentes ont été étudiés.

Ces deux projets ont permis l'émergence de banques de données gratuites contenant un large catalogue de SNPs référencés et de SNP représentatifs de régions en DL (tagSNP). L'identification d'un tagSNP permet de réduire la dimension des données. La Figure 1.6 illustre la construction de blocs haplotypiques à partir desquels il est possible de définir des tagSNPs. Dans cet exemple, quatre SNPs sont nécessaires et suffisants pour étudier la région concernée. Par conséquent, l'utilisation de tagSNPs permet de réduire le nombre de SNPs à génotyper afin d'obtenir de l'information sur l'ensemble du génome.

L'approche la plus courante dans les GWAS est celle des études cas-témoins (*Case Control Study* en anglais) dans lesquelles deux cohortes sont comparées (Figure 1.7). La première est composée d'individus présentant le phénotype d'intérêt (les cas). La seconde est composée d'individus ne présentant pas le phénotype d'intérêt (les témoins). Le phénotype observé est binaire : soit l'individu est malade, soit il ne l'est pas. Tous les individus sont alors génotypés pour un grand nombre de variants localisés le long du génome. Puis les fréquences des variants sont comparées entre les deux groupes. Un allèle est considéré comme associé au phénotype s'il est significativement plus fréquent chez les Cas que chez les Témoins.

Les associations identifiées peuvent être de deux types différents : directes ou indirectes (Figure 1.8). Une association est directe lorsque le SNP considéré comme associé est directement causal (ou fonctionnel), c'est-à-dire qu'il est associé et a une influence sur le phénotype. Une association est indirecte lorsque le SNP considéré comme associé est en réalité en fort DL avec un SNP causal et qu'il n'a

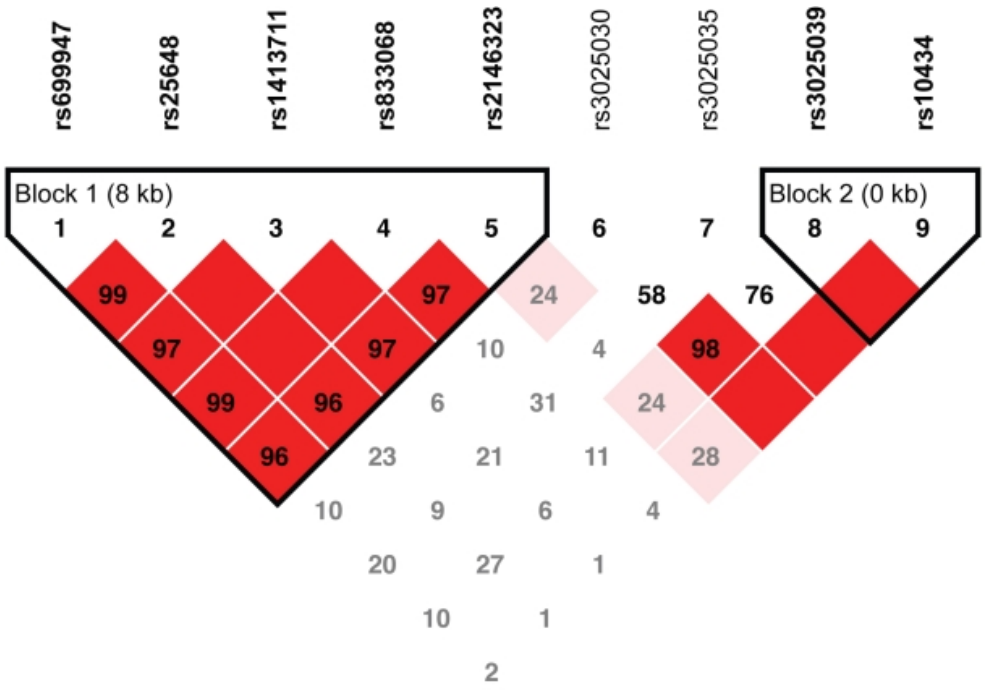

Figure 1.6 – Cartographie du déséquilibre de liaison produit à l'aide de Haploview issue du travail de Fang et al. (2009). Les valeurs représentent les mesures de DL par paire de SNP. Les carrés rouge foncé indiquent des valeurs de D' supérieurs ou égale à 0.80 ($D' \geq 80$) tandis que les carrés vides (sans présence de valeurs) indiquent des valeurs de D' égale à 1 ($D' = 1$).

pas d'influence directe sur le phénotype. Des analyses fonctionnelles sont par la suite nécessaires afin de valider les résultats obtenus.

Dans les GWAS, le phénotype observé n'est pas toujours binaire, il peut être continu comme la taille des individus, le taux biologique (hémoglobine, cholestérol, etc), un rendement, un poids, etc.

1.3 . Analyses statistiques dans les GWAS

1.3.1 . Génotypage des SNPs

Le génotypage d'individus dans le cadre d'une étude d'association pangénomique peut être obtenu par différentes méthodes telles que les puces à ADN ou encore par des techniques de séquençage à haut débit (NGS pour *Next-Generation Sequencing* en anglais). Ces méthodes permettent de génotyper plusieurs individus pour des centaines de milliers de SNPs simultanément en un temps limité.

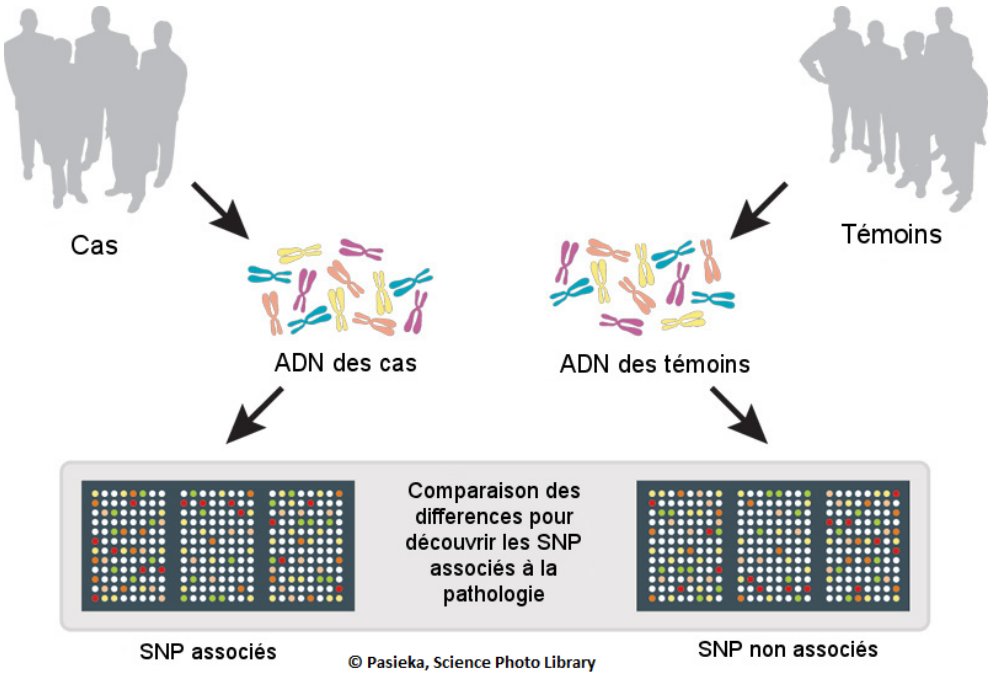

Figure 1.7 – Études d'association Cas-Témoins

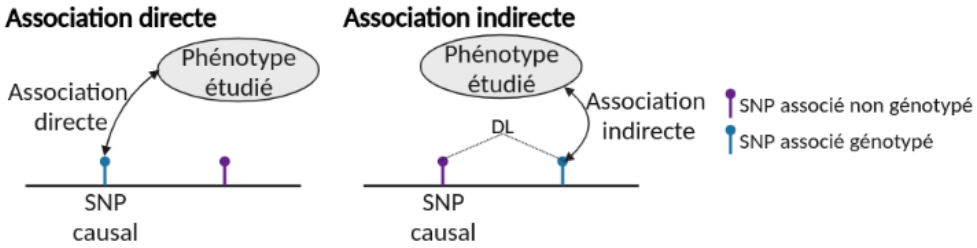

Figure 1.8 – Associations directes et indirectes dans les GWAS. (Réalisé avec BioRender.com)

Les puces à ADN : Le génotypage par puce à ADN est la méthode la plus couramment utilisée dans les GWAS. Si différents types de puces existent, leur principe reste le même ; l'ADN des individus précédemment récolté, fragmenté et amplifié est disposé sur un support solide contenant des milliers de sondes choisies, connues et marquées par un nucléotide radioactif ou un fluorochrome. Ces fragments d'ADN s'apparient ensuite avec leur sonde complémentaire si cette dernière est présente dans l'échantillon biologique (étape appelée hybridation). La puce est ensuite lavée afin d'éliminer les fragments ne s'étant pas hybridés avec une sonde puis scannée pour quantifier l'intensité du signal émis à chaque hybridation. Cette dernière étape permet de déterminer les génotypes des individus pour les SNPs présents sur la puce.

Les NGS : Le séquençage englobe diverses méthodes, dont la technologie de séquençage Illumina qui est celle principalement utilisée. Si différentes technologies existent, elles sont cependant composées de trois grandes étapes communes : La préparation de librairies (fragmentation de l'ADN et ajout de séquences spécifiques connues auxquelles les fragments d'ADN vont se lier), suivie de l'amplification de ces librairies et enfin le séquençage par synthèse de ces librairies amplifiées.

Bien que les NGS permettent le séquençage d'un grand très grand nombre de SNPs en peu de temps, elles sont plus difficiles à analyser que les puces à ADN et peuvent par conséquent être plus couteuses en termes de temps.

Une fois les génotypes des individus déterminés, une étape de contrôle de qualité strict des données est nécessaire afin de limiter la découverte de résultats faussés (McCarthy et al., 2008). Cette étape est primordiale compte tenu du nombre important de SNPs étudiés dans les GWAS où l'erreur de génotypage peut avoir un impact significatif sur les résultats.

1.3.2 . Contrôle qualité des données

Nous présentons ci-dessous les principales étapes non exhaustives pour le contrôle qualité des données couramment utilisées (Balding, 2006; Turner et al., 2011).

Données manquantes (Call Rate) : Les SNPs et individus présentant une fréquence de données manquante élevée sont supprimés de l'analyse. En effet, un pourcentage trop élevé de données manquantes pour les SNPs et individus peut traduire des erreurs expérimentales et/ou de génotypage mais aussi une mauvaise qualité de l'échantillon et/ou des puces. Un seuil de 5%, parfois 10%, est généralement appliqué mais un seuil plus strict tel que 1% peut être appliqué dans le cas d'échantillons petits.

Modèle de Hardy-Weinberg : Les SNPs ne vérifiant pas le modèle de Hardy-Weinberg dans la population sont supprimés de l'analyse. Pour cela, un test de conformité au modèle de Hardy-Weinberg (test de chi-deux) est réalisé afin de mesurer les écarts au modèle et d'identifier les SNPs dont les mesures s'écartent ainsi du modèle. Un écart important par rapport à l'équilibre peut indiquer des erreurs potentielles de génotypage ou encore une stratification de la population. Généralement, les SNPs ayant une p-valeur inférieure à 10^{-5} (parfois 10^{-3}) sont supprimés, selon les données un seuil plus strict peut être appliqué.

Stratification : Il y a une stratification de la population dans les données lorsque plusieurs sous-groupes d'individus ayant des fréquences alléliques différentes sont présents. L'Analyse en Composante Principale (ACP, PCA pour *Principal Component Analysis* en anglais) est l'approche la plus utilisée pour visualiser la structure de la population et détecter si la population est stratifiée ou non. Lorsqu'il y a la présence d'une stratification dans nos données, il est nécessaire de la prendre en compte dans l'analyse. Des méthodes ont été développées pour cela (Price et al., 2010b).

Fréquences alléliques : Les SNPs ayant une faible MAF ont une faible puissance de détection, c'est pourquoi ils sont supprimés de l'analyse. Le seuil choisi varie selon l'étude, un seuil de 5% ou 1% est couramment appliqué. Cependant, une perte de l'information est obtenue lors de cette étape de filtrage. C'est dans ce contexte que nous avons développé notre procédure qui permet de favoriser la détection de variants peu fréquents et donc de réduire la perte de l'information.

1.3.3 . Modèle génétique

Les tests statistiques reposant sur un modèle génétique présupposé, il est nécessaire de le définir en amont. Dans le cas d'un SNP bi-allélique, d'allèles A et a, il existe trois génotypes possibles : AA, Aa et aa. Pour chacun d'eux, le phénotype obtenu dépendra des effets des allèles sur celui-ci. Il existe différents modèles génétiques définissant les relations entre génotype et phénotype (Thomas, 2004) :

- **Modèle récessif :** l'allèle A est récessif par rapport à a lorsque deux copies de cet allèle sont nécessaires afin d'augmenter le risque de présenter le phénotype d'intérêt. On a alors : $\mathbb{P}(Y|aA) = \mathbb{P}(Y|aa)$ où Y représente le phénotype.

- **Modèle dominant :** l'allèle A est dominant par rapport à a lorsqu'une seule copie de cet allèle suffit à augmenter le risque de présenter le phénotype d'intérêt. On a alors : $\mathbb{P}(Y|aA) = \mathbb{P}(Y|AA)$ où Y représente le phénotype.

  - On définit la **dominance complète** lorsque le génotype est composé d'un allèle dominant et d'un allèle récessif, et que le phénotype obtenu dépend uniquement de l'allèle dominant.

  - On définit la **dominance incomplète** lorsque le génotype est composé de deux allèles dominants et différents et le phénotype obtenu est alors un phénotype intermédiaire. Par exemple, une fleur aura une couleur rose si elle présente deux allèles dominants dont l'un est un allèle "pétale rouge" et l'autre un allèle "pétale blanc".

— **Modèle additif :** Dans ce modèle, le risque de présenter le phénotype d'intérêt dépend du nombre de copies de l'allèle à risque (allèle alternatif). Si l'allèle A est un allèle à risque (c'est-à-dire associé au phénotype), la présence d'une seule copie dans le génotype à un effet moitié moindre que la présence de deux de ses copies. Le risque de développer la maladie est par conséquent proportionnel au nombre de copies de l'allèle à risque.

— **Modèle co-dominant :** Dans ce modèle, chaque allèle a un effet propre sur le phénotype d'intérêt. Les modèles additifs sont un cas particulier des modèles co-dominants. On a alors $\mathbb{P}(Y|aA) \neq \mathbb{P}(Y|aa) \neq \mathbb{P}(Y|AA)$ où Y représente le génotype.

1.3.4 . Tests d'hypothèses

Une fois les génotypes et les différentes étapes de filtrage réalisées, l'analyse des SNPs peut être effectuée. L'objectif est d'identifier les SNPs significativement associés au phénotype étudié. La population d'étude (individus apparentés ou non) ainsi que le type de phénotype (quantitatif ou catégoriel) conditionnent le test statistique appliqué pour chaque marqueur. Pour la suite de ce manuscrit, nous nous concentrerons sur les études d'association d'une population d'individus non apparentés. Nous décrivons ici brièvement le principe des tests d'hypothèses utilisées dans les GWAS, une description plus détaillée des tests statistiques et leurs applications dans ce contexte est disponible dans le Chapitre suivant (Chapitre 2 page 35).

Dans les GWAS, l'approche la plus courante consiste à tester chaque SNP séparément. Ainsi, pour chaque SNP présent dans les données, un test statistique est réalisé. Pour éviter une augmentation des résultats faussement significatifs, des procédures de tests multiples sont alors appliquées dans le but de contrôler un risque d'erreur global.

Deux principaux critères sont utilisés dans le cadre des études d'association pangénomiques dont une description précise est disponible dans la Section 2.2.3. Nos travaux se sont concentrés sur les procédures de tests multiples contrôlant le FDR (*False Discovery Rate* en anglais), critère d'erreur défini comme la proportion de SNPs considérés à tort comme associés parmi l'ensemble des SNPs identifiés. Une fois le critère d'erreur sélectionné, une procédure de tests multiples contrôlant celui-ci est appliquée.

1.3.5 . Visualisation des résultats

Le Manhattan plot est un graphique couramment utilisé afin de visualiser l'ensemble des résultats issus d'une GWAS. Ces diagrammes, dont un exemple est présenté Figure 1.9, permettent d'identifier facilement les résultats significatifs. Les

associations qui sont exprimées en $-log_{10}$ des p-valeurs issues des tests statistiques sont présentées sur l'axe des ordonnées en fonction des coordonnées génomiques (axe des abscisses). Chaque point du graphique représente ainsi un SNP. En général, les associations les plus fortes forment des pics nets où les SNPs corrélés proches présentent tous le même signal.

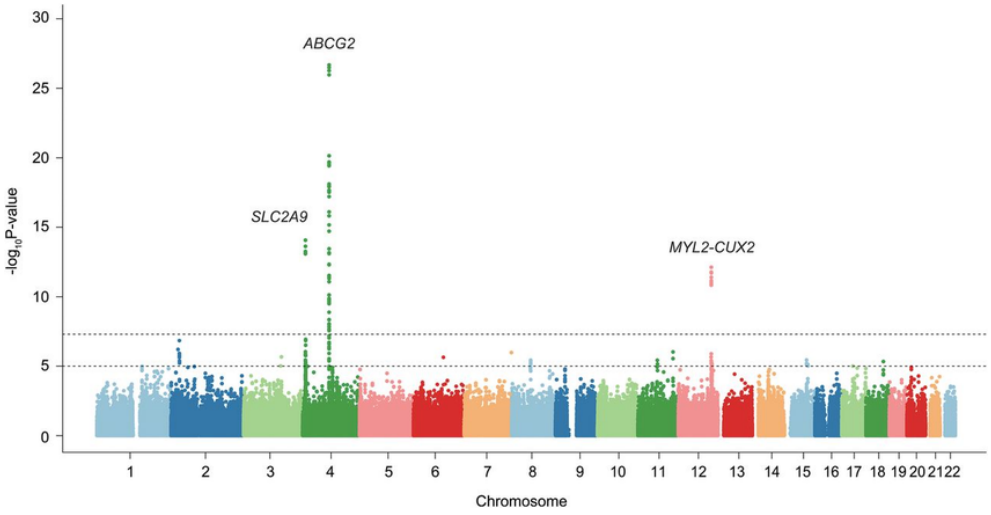

Figure 1.9 – Exemple de Manhattan plot issu des travaux de Matsuo et al. (2016).

1.4 . Analyse des variants rares

Les procédures de tests multiples dans les GWAS ont permis d'identifier des centaines de variants génétiques impliqués dans de nombreuses maladies. Cependant, ces analyses ne permettent pas d'expliquer toute la variabilité observée et seule une petite fraction des variations phénotypiques a été expliquée, reflétant une grande partie de l'héritabilité manquante (Maher, 2008; Manolio et al., 2009; Tam et al., 2019). De nombreuses raisons expliquant cette héritabilité manquante ont été proposées, telles que les variants communs ayant de faibles effets génétiques qui restent à découvrir, l'identification difficile de la variation génétique de dominance (c'est-à-dire l'effet de dominance d'un allèle sur un autre), l'épistasie ou encore les variants rares ayant des effets génétiques forts qui sont mal détectés par les puces de génotypage (Eichler et al., 2010; Zuk et al., 2014).

Les variants rares sont mal couverts par les puces de génotypage. En effet, l'idée générale était que les variants génétiques fréquents expliquent une grande partie de l'héritabilité dans les maladies communes et sont plus faciles à détecter dans les populations, raison pour laquelle les variants rares sont généralement filtrés (Panagiotou et al., 2010; Riancho, 2012; Korte and Farlow, 2013). De plus, ces variants

étant présents en faible nombre par définition, la différence entre une erreur de génotypage, une erreur statistique ou la présence d'un SNP associé ayant une faible MAF n'est pas évidente à identifier. Pour s'assurer que l'identification d'un SNP peu fréquent n'est pas une erreur éventuelle, une taille d'échantillon relativement importante est préconisée. Cependant, cela a un coût et selon la fréquence de la maladie dans la population, cela n'est pas toujours possible. C'est pourquoi, les marqueurs dont la MAF est inférieure à un seuil spécifique (généralement 1% ou 5%) sont retirés des analyses afin de limiter les problèmes dus à des tailles d'échantillon trop petites. Cependant, il a été montré que des variants rares peuvent être fortement impliqués dans certaines maladies. Ainsi, tout le potentiel de ces études n'a pas été exploité puisqu'une partie de l'héritabilité manquante peut être partiellement expliquée par des variants rares plus difficiles à détecter (Manolio et al., 2009; Auer and Lettre, 2015; Bandyopadhyay et al., 2017). Pourtant, ces variants sont susceptibles d'avoir des effets génétiques plus importants que les variants communs (Janssens et al., 2007; Bodmer and Bonilla, 2008; Marouli et al., 2017).

Ces variants manquant de puissance, ils sont très difficiles à identifier. Il existe une littérature abondante et actuelle au sujet de la recherche et l'analyse de variants rares et de nombreuses méthodes ont été proposées pour améliorer leur détection. La stratégie communément employée consiste à utiliser des approches d'agrégations dans lesquelles les variants rares d'une même région génétique, susceptibles d'avoir une fonction similaire, sont regroupés en une seule et même statistique. De cette manière, au lieu de tester chaque SNP individuellement, ces approches testent l'effet cumulé des variants contenus dans un groupe, permettant ainsi l'augmentation de la puissance lorsque plusieurs variants d'un groupe sont associés au phénotype étudié.

Les approches d'agrégation peuvent être classées en plusieurs catégories présentées ci-dessous (Lee et al., 2014) :

Les tests avec charges (*burden tests* en anglais) supposent que tous les variants d'une région génétique sont causaux et affectent par conséquent le phénotype dans la même direction d'effet (c'est-à-dire que ce sont des SNPs protecteurs ou délétères). Parmi ces méthodes, on retrouve différentes procédures telles que CAST (pour *Cohort Allelic Sums Test* en anglais) (Morgenthaler and Thilly, 2007), CMC (pour *Combined Multivariate and Collapsing* en anglais) (Li and Leal, 2008) ou encore WST (pour *Weighted Sum Test* en anglais) (Madsen and Browning, 2009). Ces méthodes permettent d'améliorer la détection des variants rares associés lorsque les SNPs agrégés sont effectivement causaux, cependant, si cela n'est pas le cas, ces procédures perdent en puissance.

Les burden tests adaptatifs quant à eux sont relativement robustes en présence d'une mauvaise classification préalable des variants lors de la construction des groupes. On retrouve parmi ces méthodes les procédures telles que KBAC (pour *Kernel Based Adaptive Cluster* en anglais) (Liu and Leal, 2010), VT (pour *Variable Threshold* en anglais) (Price et al., 2010a). Cependant, si un grand nombre de variants rares sont présents et/ou que les SNPs causaux ont différentes directions d'effets (effet délétère ou protecteur), ces procédures perdent en puissance.

Les non-burden tests ou **variance component tests** quant à eux ne supposent pas que tous les SNPs regroupés soient causaux, et sont par conséquent relativement robustes à la présence de variants causaux et non causaux dans les clusters construits et à la présence à la fois de SNPs protecteurs et délétères. Ces méthodes sont également robustes lorsque peu de SNPs causaux sont présents dans les groupes construits. Cependant, en présence d'une grande proportion de SNPs causaux partageant la même direction d'effet (effet délétère ou protecteur), elles se retrouvent moins puissantes que les burden tests classiques. On retrouve parmi ces méthodes la procédure populaire SKAT (pour *Sequence Kernel Association Test* en anglais) (Wu et al., 2011) ou encore C-alpha (Neale et al., 2011).

Les tests combinés quant à eux permettent la présence de variants non causaux et ayant des tailles d'effets différentes et donc des directions d'effets différents (effet délétère ou protecteur) dans les groupes construits. Ces méthodes nécessitent cependant un temps de calcul plus long qu'avec les autres catégories de méthodes. On retrouve parmi ces méthodes les procédure MiST (pour *Mixed-effects Score Test* en anglais) (Sun et al., 2013) ou encore SKAT-O (Lee et al., 2012).

Toutes les méthodes issues des catégories précédemment présentées, utilisent une approche de regroupement d'hypothèses et nécessitent par conséquent une spécification préalable des régions génétiques. De plus, l'utilisation de stratégies de regroupements implique la perte de l'information individuelle des marqueurs. Autrement dit, si un groupe est identifié comme étant associé avec la pathologie étudiée, il est impossible de savoir lequel ou lesquels sont réellement responsables puisqu'on considère le groupe comme une seule et même unité. Pourtant, l'information individuelle des marqueurs est ce qui nous intéresse tout particulièrement. En effet, l'identification de ces derniers pourrait être utilisée pour classer les individus présentant un risque élevé, prédisant ainsi la susceptibilité d'une maladie. Cela pourrait également permettre de trouver de nouvelles cibles thérapeutiques afin de limiter la progression de certaines maladies et améliorer les traitements et diagnostiques existants. Pour conserver l'information sur les marqueurs individuels, il a été démontré que la prise en compte des stratégies de pondération dans les procédures de tests multiples est un moyen efficace d'augmenter la puissance de détection des variants rares à effet génétique fort (Dalmasso et al., 2008b).

1.5 . Objectifs de la thèse

Plusieurs approches de tests multiples pondérés ont été introduites ces dernières années dans le but d'augmenter la puissance de détection globale. Il existe différentes manières d'incorporer des pondérations dans les procédures de tests multiples. L'introduction d'informations conditionnées par les connaissances a priori sur la maladie étudiée, ou l'utilisation des données elles-mêmes sont des exemples. Pour ce faire, on utilise des covariables qui vont modéliser l'information ajoutée dans les tests statistiques classiques. L'utilisation de ces poids va alors permettre de mettre en avant préférentiellement certaines hypothèses. Par exemple, dans les GWAS, il est courant d'utiliser les MAF comme covariable afin de mettre en avant les variants de faibles fréquences. De cette manière, la puissance de détection de ces variants sera augmentée et celle des variants fréquents diminuée.

Dalmasso et al. (2008b) ont par exemple étendu une stratégie de tests multiples classique dans le cadre du contrôle du critère d'erreur FWER en ajoutant des pondérations dépendantes de la MAF. De ce fait, la puissance globale de détection se retrouve alors augmentée. Il en va de même concernant la détection des variants de faibles fréquences alléliques. Cette procédure intègre ainsi les fréquences alléliques dans la construction des poids, et de cette façon l'information individuelle des marqueurs est gardée. C'est tout naturellement que nous nous sommes intéressés au cours de cette thèse à l'extension de cette approche dans le cadre du contrôle du FDR, critère moins strict que le FWER.

Récemment, de nouvelles approches exploitant les covariables afin de maximiser la puissance de détection globale ont vues le jour. Ces procédures adaptatives utilisent des méthodes d'optimisations afin d'obtenir la meilleure puissance de détection. Dans ce contexte spécifique des GWAS, nous avons défini deux objectifs : **l'évaluation de procédures pondérées actuelles contrôlant le FDR** ainsi que le **développement d'une procédure adaptative permettant de favoriser les variants rares qui sont mal détectés par les approches classiques**. Deux versions d'une méthode favorisant la détection des variants rares en combinant l'utilisation de covariable telle que la MAF et l'optimisation de la puissance sont ainsi proposées dans ce manuscrit.

Pour évaluer les procédures dans ce contexte spécifique des GWAS, une étude de simulation extensive et approfondie a été réalisée au cours de la thèse. Parmi les procédures utilisant des covariables informatives afin de définir les poids, nous avons considéré wBH (Genovese et al., 2006), FDRreg (Scott et al., 2015), IHW (Ignatiadis et al., 2016), swfdr (Boca and Leek, 2018), AdaPT (Lei and Fithian, 2018), SABHA (Li and Barber, 2018), AdaFDR (Zhang et al., 2019) et CAMT (Zhang and Chen, 2020). Pour une évaluation complète, nous avons également

inclus deux procédures non pondérées dans l'étude de simulation : BH (Benjamini and Hochberg, 1995) et qvalue (Storey and Tibshirani, 2003).

Dans le chapitre suivant, nous détaillons le cadre statistique des tests multiples dans les études d'association pangénomiques (Chapitre 2). Nous décrivons le contexte statistique des tests multiples avec pondérations dans le chapitre suivant (Chapitre 3). Dans ces deux chapitres, nous présentons différentes procédures récentes et performantes ainsi que celles étudiées au cours de la thèse. Dans le chapitre suivant, nous présentons l'approche développée durant cette thèse (Chapitre 4). Nous détaillons dans le chapitre qui suit, l'étude de simulation mise en place pour évaluer les différentes procédures (Chapitre 5). Le Chapitre 6 présente les résultats obtenus à l'issue de l'étude de simulation et de l'application sur un jeu de données (public) réel. Une discussion sur les résultats obtenus suivie d'une conclusion et de perspectives éventuelles sont présentées dans le dernier chapitre (Chapitre 7).

CHAPITRE 2
CADRE STATISTIQUE

Les sections précédentes nous ont permis de traiter les connaissances de bases de la génétique ainsi que le cadre des études d'association pangénomiques. À présent, nous nous concentrons sur le contexte statistique et plus précisément les tests statistiques appliqués dans ces études. Dans un premier temps, nous présentons le principe du test statistique simple. Puis, nous discutons des limites de son application dans le cadre des GWAS afin d'introduire ensuite le contexte des tests multiples dans ce cadre.

Sommaire

2.1 . Tests d'hypothèses simples

L'un des principaux objectifs des GWAS est d'identifier des variants génétiques associés à la maladie étudiée. L'approche la plus courante dans les GWAS, consiste à tester chaque SNP séparément (Bush and Moore, 2012). Ainsi, pour m SNPs, m tests seront effectués.

2.1.1 . Hypothèses

Le principe d'un test d'hypothèse consiste à choisir entre deux hypothèses statistiques en contrôlant un risque d'erreur. Une hypothèse statistique est un énoncé portant sur les caractéristiques d'une loi de probabilité. Parmi les deux hypothèses : l'hypothèse nulle (notée H_0) est supposée vraie et l'hypothèse alternative (notée H_1) est celle que l'on cherche à démontrer. À l'issue du test, une décision est réalisée : rejeter ou non H_0. Dans le cas où l'hypothèse nulle n'est pas admise (c'est-à-dire rejetée), alors l'hypothèse alternative, sera retenue.

Pour effectuer un test statistique dans le cadre des GWAS, on définit les deux hypothèses suivantes : l'hypothèse nulle ($H_{0,i}$) : "Aucune association entre le SNP i et la maladie" qui est testée contre l'hypothèse alternative ($H_{1,i}$) : "Association entre le SNP i et la maladie" ($i = 1, \ldots, m$).

2.1.2 . Risques d'erreurs

À l'issue de ce test, une décision est prise : celle de rejeter ou non H_0. Lors de la prise de décision, quatre résultats sont possibles (Tableau 2.1).

Décision \ Réalité	H_0 vraie	H_0 fausse (H_1 vraie)
H_0 acceptée	$1 - \alpha$	Risque d'erreur de type II (β)
H_0 rejetée (conclut H_1)	Risque d'erreur de type I (α)	$1 - \beta$ (puissance)

Table 2.1 – Toutes les issues possibles et les risques associés lors de la prise de décision dans un test d'hypothèse.

Parmi ces quatre issues, deux types d'erreur peuvent se produire :

- **L'erreur de type I** consiste à rejeter à tort H_0. Cette erreur correspond à une fausse découverte, appelée également Faux Positif (FP). La probabilité α de commettre cette erreur est le risque de première espèce.
- **L'erreur de type II** consiste à ne pas rejeter H0 alors qu'elle est fausse. Cette erreur correspond à un Faux Négatif (FN) c'est-à-dire à une découverte manquante. La probabilité β de commettre cette erreur est le risque de seconde espèce.

2.1.3 . Niveau de test et puissance

L'objectif d'un test d'hypothèse simple est de contrôler le risque de première espèce α, c'est-à-dire de garantir que α soit plus petit qu'un seuil fixé à l'avance. Ce seuil, que l'on appelle également seuil de signification ou niveau du test, est fixé de manière arbitraire (souvent à 5%). Par exemple, un niveau de 5% pour α indique qu'il est acceptable d'avoir 5% de probabilité de rejeter à tort H_0.

La puissance, définie par $1-\beta = \mathbb{P}(\text{rejeter } H_0 | H_0 \text{ fausse })$, est la probabilité de prendre la bonne décision lorsque l'on rejette l'hypothèse nulle. En d'autre termes, la puissance est la capacité d'un test statistique à correctement rejeter une hypothèse nulle (c'est-à-dire conclure à raison une association). Plus β est petit, plus le test sera puissant. On cherche par conséquent à effectuer des tests avec un maximum de puissance tout en contrôlant le risque d'erreur α. Cette puissance dans les GWAS dépend de nombreux facteurs tels que :

- **les MAF :** plus le SNP a une fréquence faible, moins le test est puissant.
- **les tailles d'échantillons :** plus la taille de l'échantillon est grande, plus la puissance est élevée.
- **la force de l'association :** plus l'association est forte, plus la puissance est élevée.

La puissance dépend également du seuil α fixé pour le test. En effet, ces deux éléments sont liés. Ainsi, lorsqu'un test statistique est réalisé, ces deux paramètres sont à prendre en considération.

2.1.4 . Statistique de test et prise de décision

Dès lors que les hypothèses à tester ont été choisies et le niveau de test fixé, la statistique de test T_i, qui est un résumé des données, peut être calculée. À partir de cette statistique de test, on calcule une p-valeur p_i correspondant au degré de signification du test.

La p-valeur représente la probabilité d'obtenir une valeur de la statistique de test au moins aussi extrême que celle observée lorsque l'hypothèse nulle est vraie. La prise de décision se fait par le biais de la comparaison de la p-valeur au seuil de signification du test fixé. Par conséquent, une fois que p_i a été calculé, la règle de décision suivante est appliquée :

- Si $p_i < \alpha$, alors $H_{0,i}$ est rejetée. On conclut alors à une association entre le SNP i et le phénotype étudié.

- Si $p_i \geq \alpha$, alors $H_{0,i}$ n'est pas rejetée. On conclut alors que le SNP i n'est pas associé au phénotype étudié.

Notons que la p-valeur peut également être interprétée comme le plus petit niveau de signification pour lequel $H_{0,i}$ est rejetée.

2.2 . Tests multiples

2.2.1 . Notations

Dans le cadre des GWAS, un grand nombre de tests statistiques sont réalisés. Notons m le nombre total d'hypothèses nulles testées. Parmi elles, m_0 hypothèses nulles sont vraies, tandis que m_1 hypothèses nulles sont fausses, c'est-à-dire que m_1 hypothèses alternatives sont vraies. Lorsqu'une procédure de tests multiples est appliquée, R hypothèses nulles sont rejetées (c'est-à-dire R variants sont considérés comme associés) et $W = m - R$ hypothèses nulles ne sont pas rejetées. Les différents résultats de m hypothèses testées durant la prise de décision sont résumés dans le tableau 2.2. Parmi eux, seule la variable aléatoire R (et donc W) peut être observée, tandis que le nombre de Vrais Positifs (VP), de faux positifs (FP), de Vrais Négatifs (VN) et de faux négatifs (FN) sont des variables aléatoires inobservables.

Réalité / Décision	H_0 vraie	H_0 fausse (H_1 vraie)	Total
H_0 acceptée	VN	FN	W=m-R
H_0 rejetée (conclut H_1)	FP	VP	R
Total	m_0	m_1	m

Table 2.2 – Toutes les issues possibles lors de la prise de décision suite aux tests de m hypothèses.

2.2.2 . Limites des tests d'hypothèses simples

Les GWAS étant des études à large échelle où de nombreux SNPs sont étudiés, de nombreux tests statistiques sont effectués simultanément. Ainsi, l'application d'un test d'hypothèse unique au niveau α pour chacune des hypothèses testées conduirait à rejeter à tort un grand nombre d'hypothèses nulles (Györffy et al., 2005; Walters, 2016). Le tableau 2.3 montre le nombre attendu de FP en fonction de différents nombres d'hypothèses nulles testées (m) qui sont supposées toutes vraies et testées au niveau $\alpha = 5\%$. À travers celui-ci on observe que le nombre attendu de FP augmente avec le nombre de tests réalisés.

m	100	1000	10000	100000	1000000
$\mathbb{E}(FP) = m \times \alpha$	5	50	500	5000	50000

Table 2.3 – Nombre attendu de faux positifs en fonction de m en supposant que toutes les hypothèses nulles soient vraies et testées au niveau $\alpha = 5\%$.

En pratique, dans les GWAS des centaines de milliers de SNPs sont testés, un nombre aussi élevé de FP n'est raisonnablement pas acceptable puisque cela impliquerait un grand nombre de résultats à essayer de confirmer dans le cadre d'études complémentaires de confirmation. Ainsi, afin de limiter l'augmentation des FP, il est nécessaire d'appliquer une autre stratégie consistant à contrôler d'autres critères d'erreur globaux prenant en compte l'ensemble des m hypothèses testées (Balding, 2006).

2.2.3 . Critères d'erreurs

Il existe une multitude de critères d'erreurs utilisés dans les tests multiples que l'on peut classer en deux catégories (Dudoit and Van der Laan, 2008; Dickhaus, 2014) :

— **Les critères reposant sur la distribution du nombre de FP :** parmi lesquels on retrouve par exemple le FWER (pour *Family-Wise Error Rate*), le gFWER (pour *Generalized Family-Wise Error Rate*), le PCER (pour *Per-Comparison Error Rate*), le PFER (pour *Per-Family-Wise Error Rate*), le mPFER (pour *Median-based Per-Family Error Rate*), le QNFP (pour *Quantile Number of False Positives*), etc.

— **Les critères reposant sur la distribution de la proportion de FP :** parmi lesquels on retrouve par exemple le TPPFP (pour *Tail Probability for the Proportion of False Positives*), le FDR (pour *False-Discovery Rate*), le PEFR (pour *Proportion of Excepted False Positives*), le QPFP (pour *Quantile Proportion of False Proportion*), le pFDR (pour *positive False Discovery Rate*), etc.

Parmi cette liste non exhaustive de critères d'erreurs globaux, deux d'entre eux sont principalement utilisés dans les GWAS et sont décrits ci-dessous : le FWER et le FDR. De nombreuses procédures basées sur le contrôle de ces critères ont été développées (Pounds, 2006).

Family-Wise Error Rate : Historiquement, le premier critère d'erreur introduit a été le FWER. Ce critère qui repose sur la distribution du nombre d'erreurs de type I, est défini comme la probabilité de rejeter à tort au moins une hypothèse nulle. En d'autres termes, le FWER est la probabilité d'obtenir au moins un FP :

$$FWER = P(FP > 0) \tag{2.1}$$

Le FWER a été considérablement utilisé dans les GWAS pour contrôler la multiplicité à l'aide de méthodes telles que la procédure classique de Bonferroni. Pour tenir compte de la structure de corrélation induite par les DL entre les SNP, différentes approches ont été proposées pour dériver des seuils significatifs basés sur une estimation du nombre effectif de SNP indépendants (Pe'er et al., 2008; Dudbridge and Gusnanto, 2008; Gao et al., 2008; Duggal et al., 2008; Galwey, 2009; Li et al., 2012; Xu et al., 2014). Cependant, les stratégies de tests multiples basées sur le contrôle du FWER sont connues pour être trop conservatrices lorsque le nombre de tests est élevé. Le FDR apparaît alors comme une stratégie alternative intéressante dans le contexte de grande dimension, comme les GWAS, et est devenu de plus en plus populaire.

False Discovery Rate : Le FDR a été introduit par Benjamini et Hochberg comme un critère d'erreur moins strict que le FWER. Les auteurs ont également introduit une procédure contrôlant ce critère en supposant que les hypothèses nulles soient indépendantes et uniformément distribuées. Elle est présentée dans la suite de ce manuscrit (Section 2.2.5 page 44). Le FDR est défini comme l'espérance de la proportion d'hypothèses nulles rejetées à tort parmi toutes les hypothèses rejetées. En d'autres termes, le FDR est l'espérance de la proportion de FP (FDP pour *False Discovery Proportion* en anglais) :

$$FDR = \mathbb{E}(FDP) = \mathbb{E}\left(\frac{FP}{\max(R, 1)}\right) \tag{2.2}$$

Il existe une relation entre le FWER et le FDR, en particulier lorsque toutes les hypothèses nulles sont vraies (Foulkes, 2009). Lorsque cela est le cas, nous avons $V = R$ et

$$V/R = \begin{cases} 0 & \text{si } V = 0 \\ 1 & \text{si } V \geq 1 \end{cases}$$

Ainsi,

$$\begin{aligned} \mathbb{E}(V/R) &= 0 \times \mathbb{P}(V = 0) + 1 \times \mathbb{P}(V \geq 1) \\ &= \mathbb{P}(V \geq 1) \\ &= \text{FWER} \end{aligned}$$

Dans ce cas-là le FDR est égal au FWER et le fait de contrôler le FDR revient alors à contrôler le FWER. En d'autres termes, toute procédure contrôlant le FWER contrôle nécessairement le FDR, notons néanmoins que la réciproque n'est pas vraie. Le FWER est plutôt conseillé dans un cadre décisionnel où l'objectif est de capturer des variants avec un contrôle strict des résultats obtenus. Quant au FDR, il est plutôt recommandé dans un cadre exploratoire où l'objectif est de capturer le maximum de variants impliqués dans le phénotype étudié. C'est pourquoi les procédures contrôlant le FDR sont devenues de plus en plus populaires dans un contexte exploratoire tel que les GWAS, où l'obtention de quelques faux positifs peut être considérée comme acceptable (Brzyski et al., 2017; Brinster et al., 2018).

Bien que les corrélations entre SNP puissent détériorer considérablement les performances de nombreuses procédures FDR (Owen, 2005; Sarkar, 2006; Qiu et al., 2005; Efron, 2007; Neuvial, 2008), les procédures FDR classiques restent valables sous différentes hypothèses de dépendance (Benjamini and Yekutieli, 2001; Farcomeni, 2007; Wu et al., 2009). En particulier, Sabatti et al. (2003) ont observé que la validité tient pour la procédure classique de Benjamini et Hochberg dans les études cas-témoins. Ainsi, en dépit de son hypothèse d'indépendance dans les données, il a été démontré que le FDR est robuste sous des formes spécifiques de dépendance. Dans un contexte d'études d'association pangénomiques avec des tests corrélés, les procédures basées sur le FDR ont permis d'obtenir une puissance plus élevée que la stratégie basée sur le FWER, même à un niveau de FDR strict (Otani et al., 2018).

2.2.4 . Classes de procédures

Procédures en une étape : Dans ces procédures, toutes les hypothèses nulles sont testées en utilisant un même seuil α préalablement fixé. Ainsi, l'ordre dans lequel les hypothèses sont testées dans ces procédures n'a pas d'importance et le seuil utilisé lors de la prise de décision pour le rejet d'une hypothèse est indépendant du rejet des autres hypothèses. La procédure initialement proposée et largement utilisée est la procédure de Bonferroni (1936). Cette méthode, contrôlant le FWER, consiste à rejeter une hypothèse i lorsque p-valeur$_i < \frac{\alpha}{m}$ où m est le nombre d'hypothèses testées.

Procédures séquentielles : Si dans les procédures en une étape toutes les hypothèses sont testées de la même manière, ce n'est pas le cas dans les procédures séquentielles où chaque hypothèse est testée à un niveau spécifique (Tamhane and Dunnett, 1999). En effet, dans les procédures séquentielles, le seuil d'un test d'hypothèse dépend des résultats obtenus précédemment. Elles peuvent être divisées en deux catégories selon l'ordre dans lequel les hypothèses sont triées et testées :

Procédures séquentielles descendantes : Dans ces procédures, les p-valeurs calculées pour chacune des hypothèses (les hypothèses nulles) sont ordonnées de manière croissante. Une fois les hypothèses triées, l'hypothèse ayant la plus petite valeur est testée. À l'issue de ce test, si l'hypothèse n'est pas rejetée, aucune hypothèse ne l'est et la procédure s'arrête. Si l'hypothèse est rejetée, la deuxième hypothèse ayant la plus petite p-valeur est testée et ainsi de suite jusqu'à ce qu'une hypothèse ne soit pas rejetée. Par exemple, Holm en 1979 a proposé une procédure séquentielle descendante contrôlant le FWER, nommée procédure de Holm ou procédure de Holm-Bonferroni (Holm, 1979). Cette procédure consiste à tester chaque hypothèse nulle i à un niveau spécifique $\alpha_i^* = \frac{\alpha}{m+1-i}$ $(i = 1, \ldots, m)$ après qu'elles soient ordonnées dans l'ordre croissant. Cette procédure présente l'avantage d'être plus puissante et moins conservatrice que la procédure de Bonferroni.

Procédures séquentielles ascendantes : Dans ces procédures, les p-valeurs calculées pour chacune des hypothèses (les hypothèses nulles) sont ordonnées de manière décroissante. Une fois les hypothèses triées, l'hypothèse ayant la plus grande valeur est testée. À l'issue de ce test, si l'hypothèse est rejetée, toutes les autres hypothèses le sont également et la procédure s'arrête. Si l'hypothèse n'est pas rejetée, la deuxième hypothèse ayant la plus grande p-valeur est alors testée et ainsi de suite jusqu'à ce qu'une hypothèse soit rejetée. Par exemple, Hochberg en 1988 a proposé une procédure séquentielle ascendante, nommée procédure de Hochberg (Hochberg, 1988). Cette procédure, contrôle le FWER sous certaines formes de dépendance (corrélations entre les hypothèses) (Simes, 1986). La procédure de Hochberg est similaire à la procédure de Holm où toutes les hypothèses sont testées avec les mêmes seuils dans les deux procédures, c'est-à-dire $\alpha_i^* = \frac{\alpha}{m+1-i}$ $(i = 1, \ldots, m)$. Cependant, dans cette procédure, les hypothèses sont triées dans l'ordre décroissant tandis que dans la procédure de Holm les hypothèses sont triées différemment (ordre croissant). De plus, les procédures ascendantes étant moins conservatrices et donc plus puissantes que les procédures descendantes, la procédure de Hochberg présentée ci-dessus peut être considérée comme une procédure de Holm améliorée.

Procédures adaptatives : Dans ces procédures, les données sont exploitées *a posteriori* afin d'augmenter la puissance de détection des procédures. Ces procédures tentent pour cela d'utiliser les données dans le but d'obtenir de l'information sur la loi de probabilité de la statistique de test ou encore sur les structures de dépendances présentes dans les données (Schweder and Spjøtvoll, 1982; Finner and Gontscharuk, 2009; Sarkar et al., 2012; Guo and Sarkar, 2020). L'estimation de la proportion d'hypothèses nulles vraies, notées π_0 (où $\pi_0 = m_0/m$) est très utilisée par les procédures issues de cette classe. L'estimation de ce paramètre s'explique

notamment par son implication directe dans le niveau auquel le critère d'erreur est contrôlé. Dans l'exemple de la procédure de Bonferroni où l'on souhaite contrôler le FWER au niveau α :

$$\begin{aligned} FWER &= \mathbb{P}(FP > 0) = \mathbb{P}(\bigcup_{i=1}^{m_0} \{p_i^* \leq \alpha\}) \leq \sum_{i=1}^{m_O} \mathbb{P}(p_i^* \leq \alpha) \\ FWER &\leq \sum_{i=1}^{m_O} \frac{\alpha}{m} \leq \frac{\alpha \times m_0}{m} \leq \alpha \times \pi_0 \leq \alpha \end{aligned} \tag{2.3}$$

où p_i^* correspond à la p-valeur ajustée pour le test i.

Pour gagner en puissance, différentes procédures estimant ce paramètre ont été proposées.

2.2.5 . Procédures de tests multiples contrôlant le FDR

Parmi les procédures basées sur le contrôle du FDR, nous en avons étudié deux, elles sont présentées ci-dessous :

Benjamini et Hochberg : La procédure de Benjamini et Hochberg (que nous nommerons BH par la suite) a été proposée en 1995 en même temps que le FDR (Benjamini and Hochberg, 1995). C'est une procédure séquentielle ascendante consistant à rejeter toutes les k hypothèses nulles correspondant aux k plus petites p-valeurs où $k = max(i \geq 0 : p_{(i)} \leq \frac{i\alpha}{m})$, $p_{(i)}$ étant les p-valeurs ordonnées.

Les auteurs ont introduit la procédure BH comme une procédure contrôlant le FDR sous l'hypothèse que les tests effectués sont indépendants. Il existe des procédures de tests multiples prenant en compte ces corrélations mais ces procédures ont tendance à être moins puissantes. Il a cependant été mis en évidence que sous certaines formes de dépendances connues telles que la dépendance PRDS (*Positive Regression Dependent on Subset of null statistics* en anglais) (Benjamini and Yekutieli, 2001) ou encore en présence de dépendances faibles (Storey, 2002), la procédure BH assurait tout de même le contrôle du FDR. Par conséquent, dans les données génomiques telles que les GWAS où des structures de corrélations telles que les DL entre les SNPs sont présentes, la procédure BH est considérée comme valide.

De plus, il a été démontré que lorsque les statistiques de test sont PRDS, la procédure BH contrôle en réalité le FDR au niveau $\pi_0\alpha$ où la proportion d'hypothèses nulles vraies n'est pas estimée (comme c'est le cas dans la procédure de Bonferroni décrite précédemment). Par conséquent, le FDR obtenu avec la procédure BH est plus petit que le niveau initialement fixé (c'est-à-dire FDR $\leq \pi_0\alpha$) (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001).

Qvalue : Afin de gagner en puissance, de nombreuses procédures adaptatives ont été développées afin d'estimer la proportion d'hypothèses nulles vraies π_0 (Storey, 2002; Dalmasso et al., 2005; Benjamini et al., 2006; Liang and Nettleton, 2012). L'une des plus utilisées est la procédure qvalue qui est basée sur l'estimation de la quantité $lim_{\lambda \to 1} \hat{\pi}_0(\lambda)$ par une spline cubique de la fonction suivante : $\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda\}}{m(1-\lambda)}$ (Storey, 2002; Storey and Tibshirani, 2003).

CHAPITRE 3
TESTS MULTIPLES PONDÉRÉS

Le chapitre précédent a permis d'introduire le principe du test d'hypothèse simple ainsi que celui des tests multiples. Dans ce chapitre, nous présentons les stratégies de tests multiples avec pondérations. Nous commencerons par une introduction de ces stratégies puis nous présenterons les procédures dont nous avons effectué l'évaluation. Enfin, nous présentons des procédures qui n'ont pas été intégrées à l'étude de simulations

Sommaire

3.1 . Principe

L'une des limites des approches de tests multiples standard est que toutes les hypothèses (qui correspondent aux SNPs dans le contexte des GWAS) sont considérées comme interchangeables. Cependant, les propriétés statistiques ou biologiques des tests individuels sont généralement différentes. Par exemple, pour une même taille d'effet, les tests statistiques réalisés pour des variants génétiques de faible MAF ont tendance à être moins puissants que pour des variants plus fréquents.

L'utilisation de seuils spécifiques dans les procédures séquentielles ou encore l'estimation de la proportion d'hypothèses nulles dans les procédures de tests multiples adaptatives permettent d'augmenter le nombre de découvertes réalisées. Cependant, l'exploitation et l'intégration de connaissances a priori dans les méthodes de tests multiples classiques peuvent être une solution afin d'obtenir une plus grande puissance de détection des variants associés (Roeder and Wasserman, 2009; Gui et al., 2012). De nombreuses recherches ont été effectuées pour exploiter l'information issue de covariables par le biais de pondérations des p-valeurs, ce qui est un bon moyen afin d'augmenter la puissance de détection tout en contrôlant le niveau du critère d'erreur choisi.

Le principe des procédures de tests multiples pondérés consiste à multiplier les seuils par les poids (que l'on note par la suite w), ou de manière équivalente, à multiplier les p-valeurs (ou les statistiques de test) par les poids inverses. De nombreuses procédures pondérées ont été proposées (Benjamini and Hochberg, 1997; Rubin et al., 2006; Genovese et al., 2006; Roquain and Wiel, 2008b,a; Kang et al., 2009; Scott et al., 2015; Ignatiadis et al., 2016; Lei and Fithian, 2018; Boca and Leek, 2018; Li et al., 2019; Zhang et al., 2019; Zhang and Chen, 2020).

Toutes ces méthodes pondèrent les hypothèses testées selon leur probabilité a priori d'être associé ou non avec le phénotype étudié. Ainsi, la puissance augmente pour certaines hypothèses individuelles tandis qu'elle diminue pour d'autres tout en maintenant le contrôle du critère d'erreur choisi. Afin de garantir le contrôle de celui-ci, il est nécessaire de contraindre les pondérations afin que les poids soient équilibrés. On impose alors que le poids moyen pour toutes les hypothèses soit égal à 1 ($(\sum_{i=1}^{m} w_i)/m = 1$) et que chacun des poids soit strictement positif ($w_i > 0$).

Cette stratégie peut être utilisée afin de mettre en avant certains variants qui sont difficilement détectés par les méthodes classiques tels que ceux ayant de faibles fréquences alléliques. Il existe différentes approches pour définir les poids :

Poids externes : La première méthode est la plus simple et la plus intuitive. Elle consiste à exploiter les connaissances a priori issues de recherches antérieures pour définir des poids externes (Genovese et al., 2006; Roeder et al., 2006; Hu et al., 2010).

Poids optimaux : La seconde approche consiste à exploiter les données afin de définir des pondérations. Cette stratégie repose alors sur des procédures adaptatives dont les poids sont estimés à partir des données afin d'optimiser certains critères, le plus souvent la puissance globale. (Wasserman and Roeder, 2006; Roeder et al., 2007; Roeder and Wasserman, 2009; Zhao and Zhang, 2014; Zhao and Fung, 2016; Durand, 2019). Dans cette dernière approche, différentes méthodes utilisant des covariables informatives ont été récemment introduites (Ignatiadis et al., 2016; Zhang and Chen, 2020).

On définit une covariable informative, notée $X = (x_i, \ldots, x_m)$, comme une variable continue ou catégorielle, indépendante des p-valeurs sous l'hypothèse nulle et informative sur la probabilité nulle ou la puissance statistique. Dans un contexte GWAS, la MAF, qui peut être estimée directement à partir de la matrice des génotypes, peut être utilisée comme covariable informative par les procédures adaptatives. Dans nos travaux, nous considérons ainsi la MAF comme covariable dans le but de favoriser la détection de SNPs rares ayant des effets importants par rapport à des SNPs plus fréquents ayant des effets moindres sur le phénotype étudié. Néanmoins, d'autres covariables peuvent être considérées, telles que la qualité du signal, la taille de l'échantillon ou encore la distance entre le variant génétique et la localisation génomique du phénotype dans l'analyse expression-QTL (Ignatiadis et al., 2016; Korthauer et al., 2019).

Nous présentons dans la suite de ce chapitre les différentes procédures de tests multiples pondérées contrôlant le FDR que nous avons étudiées, la plupart d'entre elles étant basées sur un calcul préliminaire des p-valeurs. Notons p_i les p-valeurs calculées pour les m hypothèses testées et P les variables aléatoires correspondantes. Un modèle de mélange à deux composantes est souvent utilisé afin de modéliser la distribution des p-valeurs. Les deux composantes correspondent aux hypothèses nulles et alternatives. Ainsi, la distribution marginale de chaque p-valeur peut être écrite de la manière suivante :

$$f(p) = \pi_0 f_0(p) + (1 - \pi_0) f_1(p) \tag{3.1}$$

où f_0 représente la densité sous l'hypothèse nulle, f_1 la densité sous l'hypothèse alternative, et $\pi_0 = Pr(H_i = 0)$ et $\pi_1 = Pr(H_i = 1)$ où H_i est la variable aléatoire telle que $H_i = 0$ si l'hypothèse nulle est vraie, $H_i = 1$ si l'hypothèse

alternative est vraie. Notons que si les statistiques de test sont des variables aléatoires continues, alors, sous l'hypothèse nulle, les p-valeurs suivent une distribution uniforme sur l'intervalle $[0, 1]$.

3.2 . Procédures évaluées

3.2.1 . Benjamini et Hochberg pondéré

La procédure BH pondérée (wBH) contrôlant le FDR a été introduite par Genovese et al. (2006). Cette procédure consiste à attribuer à chaque hypothèse nulle $H_{0,i}$, un poids positif tel que la somme des poids est égale au nombre d'hypothèses total ($\sum_{i=1}^{m} w_i = m$).

Une fois les poids attribués à chacune des hypothèses, la procédure BH est appliquée en remplaçant les p-valeurs p_i par les p-valeurs pondérées telles que :

$$\frac{p_i}{w_i}$$

Cette procédure wBH est directement dérivée de la procédure BH et s'appuie uniquement sur l'information externe pour définir les poids.

3.2.2 . False Discovery Rate Regression

La procédure False Discovery Rate Regression (FDRreg) introduite par Scott et al. (2015) est une procédure adaptative contrôlant le FDR dans laquelle la proportion d'hypothèses nulles vraies π_0 est estimée. Cependant, cette quantité π_0 est rendue dépendante de la covariable. Ainsi, nous avons :

$$\pi_0(x_i) = Pr(H_i = 0|X_i = x_i) \text{ et } FDR(x_i) = \mathbb{E}\left(\frac{FP}{max(R, 1)}|X_i = x_i\right)$$

avec $\pi_0(x_i)$ représentant les poids spécifiques à chaque hypothèse.

Ainsi, en notant z_i les statistiques de test, le modèle de mélange à deux composantes (Équation 3.1) peut s'écrire de cette façon :

$$f(z_i) = \pi_0(x_i)f_0(z_i) + (1 - \pi_0(x_i))f_1(z_i) \tag{3.2}$$

Dans cette approche, la densité sous l'hypothèse alternative, notée $f_1(z_i)$ représente un mélange de la loi sous l'hypothèse nulle (qui est supposée gaussienne) avec changements de positions.

Les proportions de mélange de $f_1(z_i)$ sont estimées via un algorithme récursif prédictif (Newton, 2002). Les paramètres du modèle dans l'équation 3.2, sont ensuite estimés par un algorithme EM (Espérance-Maximisation) supposant les proportions de mélange de la distribution alternative comme fixes. Une approche entièrement bayésienne basée sur la méthode de Monte-Carlo par chaînes de Markov (MCMC, pour *Markov Chain Monte Carlo* en anglais) est également proposée pour l'estimation conjointe de la proportion de mélange et les paramètres du modèle de régression.

3.2.3 . Science-Wise False Discovery Rate

La procédure Science-Wise False Discovery Rate (swfdr) introduite par Boca and Leek (2018) est similaire à la méthode FDRreg. En effet, cette méthode fait également dépendre π_0 et le FDR de la covariable. Cependant, tandis que FDRreg estime conjointement π_0 et le FDR en supposant que les statistiques de test sont normalement distribuées, la procédure swfdr estime d'abord la proportion d'hypothèses nulles vraies, puis un estimateur plug-in du FDR est utilisé.

Pour estimer les proportions $\pi_0(x_i)$, une approche similaire à la méthode qvalue (présentée précédemment Section 2.2.5) est proposée à la différence que :

$$\text{le rapport } \frac{\#\{p_i > \lambda\}}{m(1-\lambda)} \text{ est remplacé par } \frac{\hat{\mathbb{E}}(1_{P_i>\lambda}|X_i = x_i)}{(1-\lambda)}$$

où $\mathbb{E}(1_{P_i>\lambda}|X_i = x_i)$ est estimé à partir d'un modèle de régression logistique.

3.2.4 . Covariate Adaptive Multiple Testing

La procédure CAMT (Covariate Adaptive Multiple Testing) introduite par Zhang and Chen (2020) est également basée sur le modèle de mélange (Équation 3.2) avec des proportions dépendantes de la covariable. Cependant, cette procédure repose sur la version locale du FDR, le local fdr ($lfdr$) qui a été introduit par Efron et al. (2001). Le lfdr est défini comme la probabilité a posteriori qu'une hypothèse soit nulle compte tenu d'une p-valeur spécifique :

$$lfdr(p_i) = Pr(H = 0|P = p_i) = \frac{\pi_0 f_0(p_i)}{f(p_i)}$$

À partir de cette définition, le FDR peut être déduit de la relation $FDR = \mathbb{E}(lfdr|P \in \Gamma)$ où Γ est une région de rejet définie pour les p-valeurs (Efron et al., 2001; Dalmasso et al., 2007). De plus, la règle de décision optimale peut s'écrire de la façon suivante :

$$lfdr(p_i) \leq t \Leftrightarrow \frac{f_{1,i}(p_i)}{f_0(p_i)} \geq \frac{(1-t)\pi_0(x_i)}{t(1-\pi_0(x_i))}$$

Le principe de la procédure CAMT est de remplacer le rapport $\frac{f_{1i}}{f_0}$ dans la règle de décision optimale par une fonction de substitution $h_i(p) = (1-k_i)p_i^{-k_i}$. Ensuite, les paramètres $\pi_0(x_i)$ et k_i sont estimés à partir d'un algorithme EM afin de trouver le seuil optimal t permettant de contrôler le FDR au niveau souhaité.

Les auteurs ont également étendu la procédure au contrôle du FWER (Zhou et al., 2021). Au sein de la procédure, comme dans les procédures FDRreg et swfdr, les proportions d'hypothèses nulles vraies et la FDR sont rendus dépendants de la covariable. Cependant, l'objectif de CAMT est d'obtenir des poids optimaux, c'est-à-dire des pondérations permettant d'obtenir une puissance de détection maximale.

3.2.5 . Independent Hypothesis Weighting

La procédure IHW (pour *Independent Hypothesis Weighting* en anglais) a été introduite par Ignatiadis et al. (2016) aussi bien dans le cadre du contrôle du FDR que du FWER. Dans cette procédure, l'objectif est de trouver des poids optimaux maximisant la puissance globale. L'idée est de répartir les hypothèses en G groupes en fonction des valeurs ordonnées de la covariable. Ensuite, des poids positifs sont attribués à chaque groupe g afin de maximiser le nombre de rejets.

Pour éviter un surajustement, les auteurs ont introduit une approche de splitting consistant à diviser aléatoirement l'ensemble des m hypothèses en k plis indépendamment des p-valeurs et des covariables. Pour chaque pli, un problème d'optimisation est appliqué aux hypothèses des $k-1$ plis restants afin d'en déduire des poids $\tilde{\omega}_g$ $(g = 1, ..., G)$ qui maximisent la puissance globale. Ensuite, les hypothèses du pli retenu se trouvant dans le groupe g se voient attribuer le poids $\tilde{\omega}_g$.

Pour rendre le problème d'optimisation convexe, les auteurs ont proposé d'utiliser l'estimateur de Grenander au lieu de la fonction de répartition empirique. De plus, pour résoudre le problème d'optimisation, ils ont ajouté un paramètre de régularisation λ tel que $\sum_{g=2}^{G} \|w_g - w_{g-1}\| \leq \lambda$ où $\lambda > 0$. Ce paramètre de régularisation permet que les poids des groupes successifs ne soient pas trop différents.

Enfin, une fois les poids estimés, une procédure wBH standard est appliquée.

3.3 . Autres procédures

3.3.1 . SABHA

La procédure SABHA (pour *Structure-Adaptive Benjamini–Hochberg Algorithm* en anglais) introduite par Li and Barber (2018) est une procédure adaptative prenant en compte la structure de la liste des probabilités a priori d'être sous H_0 des hypothèses. Parmi les différentes structures, les auteurs en ont étudié principalement deux : la première est une structure ordonnée dans laquelle on suppose que

les hypothèses alternatives vraies sont susceptibles d'apparaître au début de la liste d'hypothèses. La seconde est une structure groupée dans laquelle on suppose que les hypothèses alternatives vraies sont regroupées par groupe dans la liste d'hypothèses. Cependant, la procédure SABHA peut être adaptée pour d'autres types de structure.

Un seuil, fixé arbitrairement à 0.5 par les auteurs, est utilisé dans la procédure afin que seule l'information des p-valeurs supérieures à ce seuil soit exploitée pour déterminer les poids. En comparaison, ce seuil joue exactement le même rôle que le paramètre λ dans la procédure de Storey (Storey, 2002). Une fois les p-valeurs censurées, SABHA estime ensuite les poids par l'algorithme ADMM (pour *Alternating Direction Method of Multipliers* en anglais) qui permet de résoudre des problèmes d'optimisation convexes. Les pondérations représentent les probabilités des hypothèses d'être sous H_0. Les poids correspondent ainsi à une version locale de π_0 présente dans la procédure de Storey. La procédure BH pondérée avec les p-valeurs censurées est ensuite appliquée avec ces poids.

Pour garantir au mieux le contrôle du FDR et éviter le surajustement des données, les auteurs ont contraint les poids à l'aide de la complexité de Rademacher. Cependant, celle-ci peut entrainer une perte du contrôle du FDR avec un niveau légèrement supérieur à celui initialement fixé.

L'intégration de cette procédure dans l'étude de simulation a été envisagée. Nous nous sommes appuyés sur les codes disponibles afin de pouvoir l'incorporer dans notre évaluation. Néanmoins, une perte du contrôle du FDR dans les résultats préliminaires a été obtenue. En outre, le temps d'exécution de la procédure était extrêmement long. Nous avons donc choisi de ne pas l'inclure dans notre étude de simulation.

3.3.2 . AdaPT

La procédure AdaPT (pour *ADAptive P-value Thresholding* en anglais) introduite par Lei and Fithian (2018) est une procédure itérative utilisant un système de masquage de p-valeurs dans le but contrôler le FDR.

À chaque itération de la procédure AdaPT, pour un seuil noté $s_t(x)$, le $\hat{FDP}_i$ des hypothèses pour lesquelles $p_i \geq 1 - s_t(x_i)$ est estimé de la manière suivante :

$$R_t = |\{i : p_i \leq s_t(x_i)\}|$$
$$A_t = |\{i : p_i \geq 1 - s_t(x_i)\}|$$
$$\widehat{FDP}_t = \frac{1 + A_t}{R_t V 1}$$

où R_t correspond nombre de rejets et A_t correspond au nombre de p-valeurs pour lesquelles $p_i \geq 1 - s_t(x_i)$.

Le FDP estimé des p-valeurs non censurées est comparé au seuil α initialement fixé. S'il est inférieur à α, toutes les hypothèses H_i pour lesquelles $p_i \leq s_t(x_i)$ sont rejetées et la procédure s'interrompt. Dans le cas contraire, un nouveau seuil est fixé et ainsi de suite tant que $\hat{FDP}_i \geq \alpha$. Un choix optimal pour le seuil est l'utilisation de surface de niveau du taux de faux positif local dans lequel π_0 dépend d'une covariable (qui peut être multidimensionnelle).

Dans le cas où les p-valeurs sous l'hypothèse nulle sont uniformes et indépendantes, la procédure AdaPT contrôle le FDR. Cependant, en présence de dépendance entre les p-valeurs, une perte du contrôle du critère d'erreur est obtenue.

La stratégie de masquage des données dans AdaPT requiert cependant la mise en place d'un grand nombre d'itérations dans la procédure d'optimisation, ce qui peut s'avérer coûteux en temps de calcul. C'est notamment pour cette raison que nous n'avons pas intégré la procédure AdaPT dans notre étude de simulation.

3.3.3 . AdaFDR

La procédure AdaFDR introduite par Zhang et al. (2019) est une procédure intermédiaire entre la procédure AdaPT et IHW. En effet, tout comme AdaPT, AdaFDR est une procédure adaptative et itérative dans laquelle π_0 dépend de la covariable et utilise une stratégie de division des p-valeurs pour contrôler le FDR. De plus, AdaFDR permet d'exploiter plusieurs covariables.

L'algorithme naïf de la procédure AdaFDR consiste à estimer le paramètre $\hat{\pi}_0(x_i)$ par l'algorithme EM afin de pondérer les hypothèses telles que $w_i = \frac{1}{\hat{\pi}_0(x_i)}$, puis à effectuer des ajustements locaux dans le seuil de p-valeur afin de maximiser la puissance globale.

Afin d'éviter un surajustement des données, la procédure AdaFDR tout comme la procédure IHW, utilise une approche de fractionnement des hypothèses. AdaFDR divise aléatoirement les hypothèses en seulement deux sous-ensembles (de taille égale). Pour chacun d'eux l'algorithme naïf est appliqué. Le seuil de p-valeurs appris par l'un des groupes est ensuite utilisé dans le second et le nombre de rejets est calculé.

Bien qu'AdaFDR permette l'incorporation d'une covariable multidimensionnelle tout en ayant une part d'optimisation pour maximiser la puissance de détection, cette procédure présente deux inconvénients.

Le premier inconvénient est que la procédure AdaFDR requiert un nombre d'hypothèses relativement important, notamment d'hypothèses alternatives, en raison du fractionnement des données. Les auteurs recommandent un nombre d'hypothèses total idéalement supérieur à 10 000 et plus d'une centaine d'hypothèses alternatives. Les auteurs ont proposé une version adaptée au cas où l'une de ces deux contraintes n'est pas respectée, ou encore lorsque beaucoup de p-valeurs égales à 1 sont présentes dans les données. Cette procédure, nommée AdaFDR-fast, diffère de la procédure classique par l'absence de l'étape d'optimisation et par conséquent ne maximise pas la puissance de détection.

Le second inconvénient de la procédure AdaFDR est la possibilité d'obtenir des résultats très différents lorsqu'on applique plusieurs fois la procédure sur un même jeu de données en raison de l'étape de fractionnement aléatoire. Pour contrecarrer ce problème et obtenir des résultats reproductibles, les auteurs recommandent de fixer la graine aléatoire. Cependant, en dépit de la variabilité des résultats, ceux-ci restent tout de même valables étant donné le contrôle du FDR par la procédure.

Au cours de la thèse, nous avons appliqué les deux versions de la procédure proposée par Zhang et al. (2019) sur les données simulées. Cependant, les résultats obtenus, ont montré un manque de contrôle du FDR, quelle que soit la version utilisée. C'est pour ces raisons que l'intégration de ces procédures dans l'évaluation et la comparaison des procédures a été abandonnée.

CHAPITRE 4
UNE NOUVELLE PROCÉDURE, wBHa

Dans ce chapitre, nous présentons deux versions de la procédure développée au cours de la thèse. Le principe général de la méthode est expliqué, suivi d'une description détaillée des différentes étapes de notre méthode.

Sommaire

4.1 . Principe général de la méthode

Notre objectif est de favoriser la détection des variants rares tout en maintenant une bonne puissance de détection globale et en conservant l'information individuelle. Pour cela, nous nous sommes inspirés des travaux de Dalmasso et al. (2008b) pour construire des poids qui dépendent de la fréquence allélique des marqueurs.

Le principe de la méthode que nous proposons est de définir les poids comme une fonction explicite de la covariable :

$$w(x_i, a) = \frac{m}{\sum_{j=1}^{m} \frac{1}{x_j^a}} \times \frac{1}{x_i^a} \text{ où } x_i = \text{MAF}_i \tag{4.1}$$

Nous avons cependant ajouté un paramètre libre (noté a) dont l'estimation permet d'introduire une part d'optimisation pour maximiser la puissance. L'étape d'optimisation de la puissance a été ajoutée dans la procédure dans le but de maintenir une bonne puissance globale même en l'absence de variants rares causaux. L'utilisation des MAF dans notre fonction permet de favoriser les hypothèses ayant de petites fréquences alléliques. De cette façon, les pondérations dépendent non seulement de la fréquence allélique des SNPs testés, mais également du paramètre a et permettent de favoriser la détection des variants rares tout en gardant une bonne puissance globale. L'approche développée afin d'estimer le paramètre a est présentée dans la section suivante.

4.2 . Algorithme naïf

L'algorithme naïf consiste à rechercher le a-optimal conduisant au nombre maximal de rejets (noté R) à partir d'une grille de valeurs allant de 0 à 10 par pas de 0.1 (valeurs choisies arbitrairement). Dans le cas où il existe plusieurs valeurs de a maximisant le nombre de rejets, **une stratégie de sélection** doit être mise en place. En effet, la fonction de poids que nous proposons intègre un paramètre unique a qui doit être le même pour toutes les hypothèses. Différentes stratégies présentées dans la Section 4.4 peuvent être envisagées.

Cependant, la procédure wBHa naïve conduit à une perte du contrôle du FDR qui résulte du surajustement. Pour contrôler le critère, l'utilisation d'une **approche de rééchantillonnage** des données peut être une solution. Une recherche de la stratégie la plus adaptée à la procédure a été réalisée.

4.3 . Stratégie de rééchantillonnage

Les stratégies de rééchantillonnage consistent à construire de nouveaux échantillons ou sous-échantillons à partir d'un échantillon initial. Ces sous-échantillons sont construits en respectant les caractéristiques de la distribution de l'échantillon original. Le paramètre a est estimé à partir de ces nouveaux échantillons ou sous-échantillons et non plus sur l'ensemble des données, réduisant par conséquent le surajustement. Plusieurs techniques de rééchantillonnage des données ont été explorées dans le développement de wBHa.

4.3.1 . Validation croisée k-fold et leave-one-out

Nous avons essayé une stratégie inspirée de la validation croisée k-fold (CV k-fold pour *Cross Validation* k*-fold* en anglais) (Stone, 1974; Geisser, 1975). Avec celle-ci, les hypothèses sont divisées aléatoirement en k sous-échantillons (plis) de taille égale. Une fois les hypothèses réparties en k-fold, la stratégie consiste à définir les pondérations des hypothèses et notamment à estimer le paramètre a-optimal d'un pli à partir des plis restants (autrement dit, la réunion des $k-1$ autres plis). Pour ce faire, à chaque itération de la validation croisée k-fold, l'algorithme naïf est appliqué sur les $k-1$ autres plis et le paramètre a-optimal est estimé à partir de ce nouveau sous-échantillon. Le processus de validation croisée est ainsi répété k fois jusqu'à ce que chaque pli soit retiré des données au moins une fois.

Avec cette stratégie de rééchantillonnage sans remplacement, les hypothèses appartiennent à un seul et unique pli, ce qui permet alors d'écarter à chaque itération une partie des hypothèses. Pour récapituler et illustrer le principe de cette méthode, nous présentons Figure 4.1 un exemple d'application avec un échantillon contenant neuf hypothèses et avec $k = 3$.

À l'issue des k itérations, une série de k valeurs de $a_{opt} = a_1, ..., a_k$ est obtenue. Pour résumer cette série de valeurs en une valeur unique, différentes approches ont été envisagées telles que l'utilisation de la valeur minimale, maximale, la médiane, la moyenne ou encore la valeur la plus proche de 1. La valeur moyenne étant celle donnant le meilleur résultat en termes de puissance et de FDR, c'est cet indicateur de position qui a été retenu pour synthétiser cette liste de valeur. En résumé, la valeur optimale a est obtenue en calculant la valeur moyenne de toutes les valeurs a_k.

Le choix du nombre de plis à une importance dans la validation croisée k-fold puisqu'il peut entraîner une augmentation du biais et/ou de la variance (Breiman and Spector, 1992; Kohavi, 1995; Jung, 2018). Bien qu'en pratique le nombre de plis utilisé soit souvent fixé à 5 ou 10 selon la taille de l'échantillon, d'autres valeurs peuvent être utilisées à condition qu'elles soient inférieures à la taille de l'ensemble

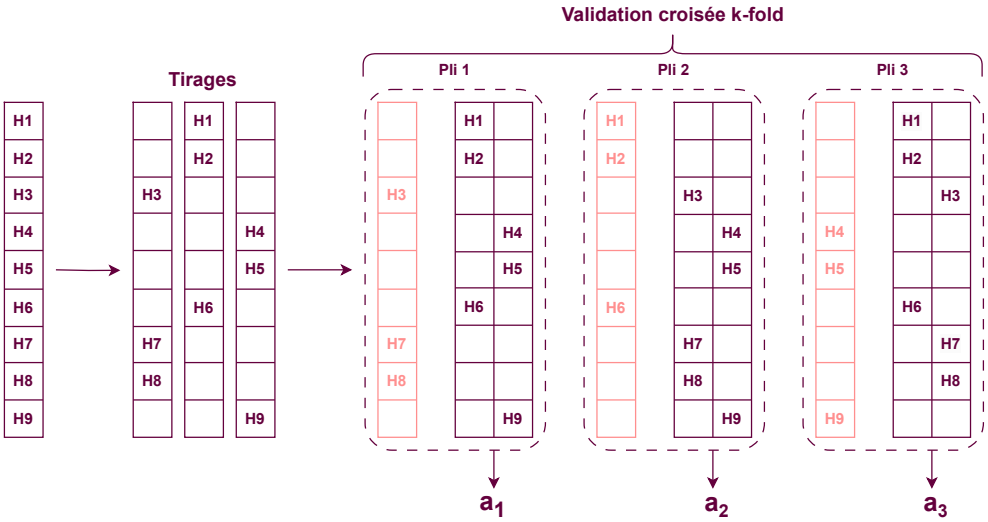

Figure 4.1 – Exemple d'application de la validation croisée k-fold pour un échantillon de 9 hypothèses et un nombre de plis égal à 3 ($k = 3$). (Réalisé avec diagrams.net)

de l'échantillon. Le cas particulier où k est égal à la taille des données, c'est-à-dire le nombre d'hypothèses total, représente la stratégie LOOCV (*Leave-One-Out Cross-Validation* en anglais). Avec cette stratégie, chaque pli contient une seule hypothèse. Le nombre de plis étant considérablement plus grand que le nombre habituellement utilisé, cette stratégie est alors coûteuse en temps de calcul. De ce fait, celle-ci n'est pas adaptée dans le cas des études d'association pangénomiques où le nombre d'hypothèses testées est important.

Lors du développement de la méthode, plusieurs valeurs du nombre de plis ont été testées ($k = \{5, 10, 50\}$), mais aucune n'a permis d'obtenir un contrôle strict du FDR.

4.3.2 . Bagging

Nous avons essayé une méthode inspirée de la stratégie Bagging (Bootstrap Aggregating), également appelée Bagging, introduite par Breiman en 1996 (Breiman, 1996; González et al., 2020). Elle repose sur la stratégie Bootstrap (Efron, 1979; Efron and Tibshirani, 1986) et permet de réduire la variance obtenue et de limiter le surajustement. Pour ce faire, K sous-échantillons sont générés en échantillonnant aléatoirement avec remise m^* hypothèses parmi l'échantillon initial contenant m hypothèses testées. Puis l'algorithme naïf est appliqué pour chacun d'eux.

Avec cette stratégie, une hypothèse peut apparaître en un ou plusieurs exemplaires dans un sous-échantillon, elle peut aussi, ne pas être choisie. Nous avons essayé différentes valeurs de m^* : m et $\frac{m}{K}$ comme illustré par les Figures 4.2 et 4.3.

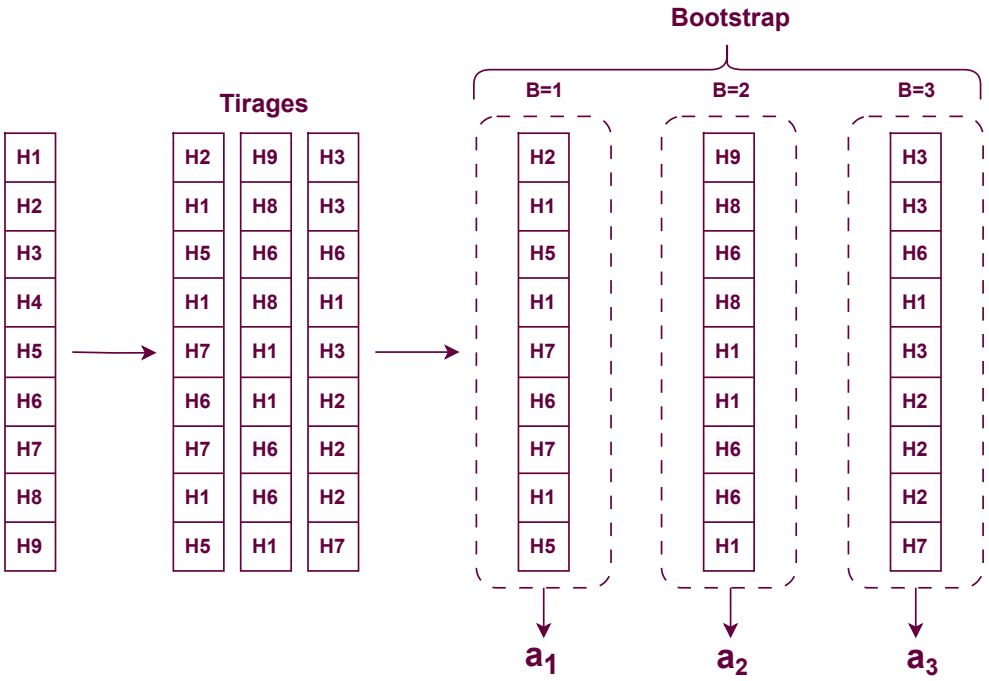

Figure 4.2 – Exemple d'application avec $m^* = m$ pour un échantillon de 9 hypothèses et un nombre d'échantillons Bootstrap égal à 3 ($K = 3$). (Réalisé avec diagrams.net)

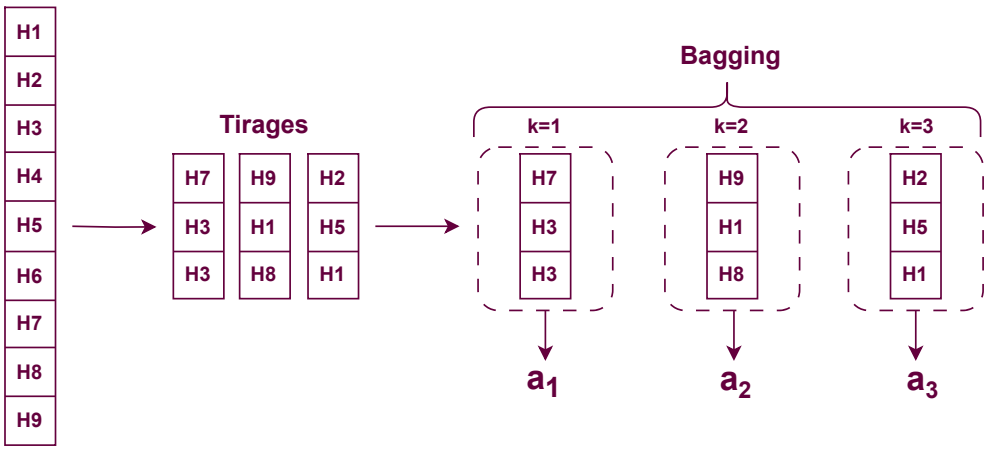

Figure 4.3 – Exemple d'application avec $m^* = m/K$ pour un échantillon de 9 hypothèses et un nombre de sous-échantillons égal à 3 ($K = 3$). (Réalisé avec diagrams.net)

Ainsi, à l'issue des K itérations, un vecteur de K valeurs de $a_{opt} = a_1, ..., a_K$ est obtenu. Pour résumer cette série de valeurs en une valeur unique, différentes méthodes ont été appliquées. La valeur moyenne donnait le meilleur résultat pour

toutes les valeurs de m^* que nous avons essayées. C'est donc ce résumé qui a été retenu pour synthétiser cet ensemble de valeur.

Tout comme pour la validation croisée k-fold, le choix de la valeur des paramètres à une influence sur les résultats obtenus. Alors que le Bagging standard génère des échantillons Bootstrap de taille égale à celle de l'échantillon initial, des résultats optimaux sont souvent obtenus avec des ratios d'échantillonnage inférieurs au choix standard (Martínez-Muñoz and Suárez, 2010).

Différentes valeurs de nombre d'échantillons Bootstrap ont été testées lors du développement de la méthode avec $m^* = m$ ($K = \{2, 10, 20, 200, 500, 1000\}$), cependant, les résultats ont montré un manque de contrôle du FDR par la procédure wBHa.

En nous inspirant de cette stratégie, nous avons également essayé d'utiliser des échantillons Bootstrap plus grands que la taille de l'échantillon initial. Ainsi, K nouveaux échantillons obtenus par tirage aléatoire avec remise de taille $m^* = m \times 10$ sont créés. Cependant, nous avons obtenu des résultats peu concluants (FDR non contrôlé) en dépit de plusieurs valeurs de K testées ($K = \{1, 10, 20\}$). Cette stratégie ne semble donc pas adaptée à notre contexte.

En choisissant $m^* = m/K$, la méthode devient moins coûteuse en temps de calcul qu'avec la stratégie m^*=m. De plus, avec cette stratégie, les tailles des sous-échantillons, et donc les tirages aléatoires sont plus petits qu'avec la stratégie $m^* = m$ et plus nombreux qu'avec la stratégie de la validation croisée k-fold. La diversité de ceux-ci se retrouve augmentée. La différentiation des sous-échantillons par rapport à l'échantillon initial est plus poussée. Les résultats que nous avons obtenus avec cette stratégie indiquent un bon contrôle du FDR.

Nous avons donc utilisé $m^* = m/K$ pour la suite.

4.4 . Stratégie de sélection du a-optimal

L'algorithme naïf appliqué à chacun des échantillons Bootstrap consiste à sélectionner la valeur de a conduisant au nombre maximal de rejets. Cependant, il est possible d'obtenir différentes valeurs de a conduisant au même nombre de rejets. Plusieurs approches, présentées ci-après, ont été envisagées pour sélectionner une valeur unique de a.

4.4.1 . Indicateurs de positions

Les indicateurs de positions comme la moyenne, la médiane, le minimum et le maximum ont été utilisés. Les résultats obtenus avec ces différents résumés numériques montrent une puissance globale élevée, mais un manque de contrôle du FDR. La difficulté principale réside dans l'équilibre entre une puissance compétitive et un FDR contrôlé puisque l'augmentation de l'un entraîne inévitablement l'augmentation de l'autre.

4.4.2 . Valeur la plus proche de 1

Nous avons également employé une stratégie sélectionnant le a le plus proche de la valeur 1 pour résumer la série statistique. En effet, lorsque le paramètre a est égal à 1, appliquer la procédure wBHa revient à utiliser la procédure wBH classique. Choisir le a le plus proche de 1 conduit donc à limiter l'influence du paramètre a. Cependant, les résultats obtenus ont montré un FDR élevé.

4.4.3 . Fonction de lissage

Nous avons par ailleurs effectué un lissage de la fonction $R = \zeta(a)$ en utilisant des splines cubiques, c'est-à-dire des fonctions définies par morceaux par des polynômes de degrés 3. La valeur de a retenue est alors celle qui maximise la fonction spline.

La puissance obtenue avec cette version de wBHa est élevée tout en garantissant le contrôle du FDR. Cette version de la procédure a donc été conservée et évaluée au travers de l'étude de simulation présentée Chapitre 5. Dans cette version 1 de wBHa, l'algorithme naïf est appliqué au travers de la stratégie de rééchantillonnage inspirée du Bagging où le paramètre K est initialement fixé à 60. D'autre part, des régressions spline cubiques sont utilisées dans wBHa afin de sélectionner la valeur unique maximisant le nombre de rejets dans chacune des itérations du Bagging. L'algorithme de la version 1 de wBHa est présenté dans l'Algorithme 1.

Bien que cette version de wBHa soit compétitive comparativement aux procédures existantes actuelles, elle présente l'inconvénient de ne pas être reproductible dans certains cas. Nous avons donc proposé une alternative en introduisant une version 2 de wBHa, décrite dans la section suivante.

4.4.4 . Définition d'intervalles

Étant donné que le nombre de rejets est recherché à partir d'une grille de valeurs allant de 0 à 10 par des pas de 0.1, il arrive que les valeurs maximisant R soient des valeurs successives définissant ainsi des intervalles.

Algorithme 1 : Algorithme d'optimisation de a de la procédure wBHa version 1

Input : Un m-tuple de p-valeurs $P = (p_1, ...p_m)$ et de covariables $X = (x_i, ..., x_m)$, un niveau nominal $\alpha \in (0, 1)$ du FDR et un nombre de plis (nombre d'échantillons Bootstrap) $K = 60$.

Output : a optimal

for $k_i = 1, ..., K$ **do**

  Réchantillonnage (Tirage) avec remise de $\frac{m}{K}$ hypothèses;

  for $a = 0, 0.1, 0.2, ..., 10$ **do**

    Application de la procédure wBH au niveau α avec $w(x_i, a) = \frac{m}{\sum_{j=1}^{m} \frac{1}{x_j^a}} \times \frac{1}{x_i^a}$;

    Calcul et sauvegarde du nombre de rejets R;

  end

  Interpolation spline cubique des R en fonction de a **then** choisir a qui maximise R

end

a optimal obtenu en calculant la moyenne des K valeurs ;

En présence d'un seul et unique intervalle, la valeur maximale est sélectionnée. En présence de plusieurs intervalles, nous avons choisi de sélectionner la valeur maximale de l'intervalle le plus long, s'il y en a un, ou la valeur maximale de l'intervalle le plus proche de 1 si plusieurs intervalles ont la même longueur.

Cette version 2 de wBHa est plus stable que la précédente et permet d'obtenir une puissance élevée tout en contrôlant le FDR. L'algorithme de la version 2 de wBHa est présentée dans l'Algorithme 2.

4.4.5 . Comparaison des différentes stratégies

Afin d'illustrer la différence entre toutes les stratégies de sélection du a maximisant R que l'on a présentées, prenons des exemples. Soit L la liste de valeurs de a obtenues :

— **Exemple 1 :** $L_1 = \{1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 8.3, 8.4, 8.5, 8.6, 8.7, 8.8, 8.9, 9.0, 9.1\}$

— **Exemple 2 :** $L_2 = \{1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 8.3, 8.4, 8.5, 8.6, 8.7, 8.8, 8.9, 9.0, 9.1\}$

— **Exemple 3 :** $L_3 = \{1.7, 1.8, 1.9, 2.0, 2.1, 2.2, 2.3, 2.4, 2.5, 8.3, 8.4, 8.5, 8.6, 8.7, 8.8, 8.9, 9.0, 9.1, 9.2\}$

Le tableau 4.1 présente les résultats obtenus dans les trois exemples pour chacune des stratégies présentées dans cette section. Dans ce tableau, nous observons que selon la stratégie adoptée, la valeur sélectionnée peut être très différente.

4.5 . Développement logiciel

La procédure wBHa version 2 a été implémentée dans un package R, appelé wBHa, disponible dans un dépôt GitHub au lien suivant https://github.com/obryludivine/wBHa. Le package wBHa, a été documenté et contient l'ensemble des données de simulation et les données réelles qui ont été utilisés pour l'évaluation de notre procédure et celles existantes. Un script permettant de reproduire les graphiques présentés dans le manuscrit est également disponible dans la documentation du package. Un second dépôt GitHub dans lequel les scripts permettant de reproduire les données simulées a été créé et est disponible dans le lien suivant : https://github.com/obryludivine/wBHa_simulation.

Algorithme 2 : Algorithme d'optimisation de a de la procédure 2 de wBHa

Input : Un m-tuple de p-valeurs $P = (p_1, ...p_m)$ et de covariables $X = (x_i, ..., x_m)$, un niveau nominal $\alpha \in (0, 1)$ du FDR et un nombre de plis (nombre d'échantillons Bootstrap) $K = 100$.

Output : a optimal

for $k_i = 1, ..., K$ **do**

  Réchantillonnage (Tirage) avec remise de $\frac{m}{K}$ hypothèses;

  for $a = 0, 0.1, 0.2, ..., 10$ **do**

    Application de la procédure wBH au niveau α avec $w(x_i, a) = \frac{m}{\sum_{j=1}^{m} \frac{1}{x_j^a}} \times \frac{1}{x_i^a}$;

    Calcul et sauvegarde du nombre rejets R;

  end

  Enregistrement des valeurs de a maximisant R dans un L-tuple ordonné ($L \geq 1$) $A = (a_1, ..., a_L)$;

  if $L > 1$ **then**

    Calcul des différences successives dans A ;

    Définition des bornes d'intervalles ayant des différences supérieures à 0.1 ;

    Regroupement des L valeurs de A dans les v intervalles ainsi définis ;

    if $v = 1$ **then**

      Sauvegarde de la valeur maximale du vecteur A ;

    else

      Calcul de la longueur de chaque intervalle ;

      if *un intervalle est plus long que les autres* **then**

        Sauvegarde de la valeur maximale du plus long intervalle;

      else

        Sauvegarde de la valeur maximale de l'intervalle le plus proche de 1 ;

      end

    end

  end

end

a optimal obtenu en calculant la moyenne des K valeurs ;

	Exemple 1	Exemple 2	Exemple 3
Min	1.7	1.7	1.7
Max	9.1	9.1	9.2
Médiane	5.4	2.6	8.3
Moyenne	5.4	5.25	5.6
La plus proche de 1	1.7	1.7	1.7
Splines cubiques	8.34	8.34	8.34
Intervalles	2.5	2.6	9.2

Table 4.1 – Exemple illustrant les différences obtenues selon la stratégie de sélection du a maximisant le nombre de rejets employée dans la procédure wBHa pour 3 exemples distincts.

CHAPITRE 5
ÉTUDE DE SIMULATIONS

L'objectif de ce chapitre est de présenter le plan de simulation que nous avons mis en place pour évaluer des différentes procédures de tests multiples présentées dans les chapitres précédents. Différents types de simulation ont été considérés dans cette étude : tout d'abord des simulations pour lesquelles les données ont été intégralement simulées (que nous appellerons "données complètement simulées") et des simulations pour lesquelles les données ont été générées à partir d'un jeu de données existant (que nous appellerons "données semi-simulées"). Nous présentons également un jeu de données réelles auquel les différentes procédures ont été appliquées.

Sommaire

5.1 . Données complètement simulées

5.1.1 . Génotypes

Le codage de la matrice des génotypes dépend du modèle génétique supposé (modèles présentés dans la Section 1.3.3). Pour un SNP bi-allélique, d'allèles A et a :

— **Modèle dominant :** Si l'on suppose que l'allèle A est dominant par rapport à l'allèle a, alors on utilise le codage suivant : 0 pour le génotype aa et 1 pour les génotypes aA et AA.

— **Modèle récessif :** Si l'on suppose que l'allèle A est récessif par rapport à l'allèle a, alors on utilise le codage suivant : 0 pour les génotypes aa et Aa, et 1 pour le génotype AA.

— **Modèle additif :** Si l'on suppose un effet co-dominant avec un effet intermédiaire pour le génotype hétérozygote Aa (modèle additif) et que l'allèle A est l'allèle alternatif, alors on utilise le codage suivant : 0 pour le génotype homozygote aa, 1 pour le génotype hétérozygote Aa et 2 pour le génotype homozygote AA.

Le choix du modèle (et donc du codage) a une importance puisqu'il a une influence sur la puissance de détection des SNPs associés. En effet, selon le modèle supposé, les individus seront repartis en deux ou trois groupes de génotypes.

Bien que le modèle additif conduise à une perte de puissance lorsque le véritable mode d'hérédité est récessif, il permet de maintenir une bonne puissance globale quelle que soit la réalité Lettre et al. (2007). Ce modèle est donc plus largement utilisé dans les études d'association. C'est pourquoi nous avons considéré un modèle génétique additif pour coder la matrice de génotypes G, composée de n lignes correspondant aux individus (n fixé à 2000) et m colonnes correspondant aux SNPs ($m \in \{8000, 14000, 20000\}$). Pour résumer, G est une matrice de génotypes de taille $n \times m$ avec $G_{ij} \in \{0, 1, 2\}$ et $i = (1, \ldots, m)$ et $j = (1, \ldots, n)$.

Pour simuler la structure de dépendance entre les SNPs, notre modèle de simulation est inspiré de celui de Dehman et al. (2015) et de Stanislas et al. (2017) à partir des travaux de Wu et al. (2009). Le génotype complet de chaque individu i a été généré à partir d'une distribution normale multivariée de dimensions m :

$$G_i^* \sim \mathcal{N}_m(0, \Sigma) \tag{5.1}$$

Dans ce modèle, la matrice Σ est une matrice diagonale par bloc telle que dans chaque bloc, toutes les variables sont équicorrélées au niveau ρ. Pour illustrer cela, prenons le cas d'une matrice de covariance avec une taille de bloc égale à 2 ($B = 2$) avec un coefficient de corrélation égale à ρ, on obtient la matrice Σ ci-dessous :

$$\Sigma = \begin{pmatrix} 1 & \rho & 0 & 0 & 0 & 0 \\ \rho & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \rho & 0 & 0 \\ 0 & 0 & \rho & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \rho \\ 0 & 0 & 0 & 0 & \rho & 1 \end{pmatrix} \text{ où } \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \text{ est une matrice carrée.}$$

Dans nos simulations, la taille des blocs est fixée à 10 ($B = 10$). Nous avons considéré différentes valeurs de ρ dans notre étude de simulation qui sont les suivantes : 0 (cas indépendant), 0.10, 0.20, 0.35, 0.5 et 0.75.

Pour obtenir les génotypes, les variables continues obtenues à partir de la relation 5.1 doivent être discrétisées. Pour cela, on utilise l'équation de Hardy-Weinberg suivante :

$$p^2 + q^2 + 2pq = 1 \tag{5.2}$$

dans laquelle pour chacun des SNPs, p représente la fréquence d'un des deux allèles possibles et $q = 1 - p$. Les fréquences génotypiques sont obtenues à partir des fréquences alléliques :

— $\mathbb{P}(G = 2) = p^2$
— $\mathbb{P}(G = 1) = 2pq$
— $\mathbb{P}(G = 0) = q^2$

Nous avons fixé (arbitrairement) p comme étant la MAF de sorte que $p \leq q$. Nous avons finalement pour chaque SNP :

— $G_{ij} = 2$ si $G^*_{ij} < q_{p^2,N(0,1)}$,
— $G_{ij} = 1$ si $q_{p^2} < G^*_{ij} < q_{(1-p)^2,N(0,1)}$,
— $G_{ij} = 0$ si $q_{(1-p)^2,N(0,1)} < G^*_{ij}$,
où $q_{.,N(0,1)}$ est la fonction quantile de la loi normale.

Nous proposons deux manières de générer les MAF selon la nature du variant : pour les m_o variants non causaux, elles ont été générées à partir d'une distribution uniforme entre 0.01 et 0.5 ($U[0.01, 0.5]$). Quant aux m_1 variants causaux, ils ont

été divisés en 4 sous-ensembles distincts dans lesquels les MAF ont été générées à partir des distributions suivantes :

— Groupe 1 (SNPs rares) : $U[0.01, 0.05]$
— Groupe 2 (SNPs moyennement rares) : $U[0.05, 0.15]$
— Groupe 3 (SNPs moyennement communs) : $U[0.15, 0.25]$
— Groupe 4 (SNPs communs) : $U[0.30, 0.40]$

Le nombre de SNPs contenus dans chacun de ses sous-groupes a été obtenu par la division euclidienne de m_1 par 4. Le reste de cette division a été ajouté au groupe 4. Nous avons considéré différentes valeurs de m_1 dans notre étude de simulation telles que $m_1 \in \{5, 10, 15, 20, 25, 50, 100, 150\}$.

5.1.2 . Phénotypes

Pour simuler les phénotypes des individus, nous avons considéré deux types de variables possibles : les caractères quantitatifs puis qualitatifs. Ces deux cas correspondent respectivement aux études de traits (phénotypes) quantitatifs et aux études cas-témoins. Un modèle linéaire est utilisé dans le cas de phénotypes quantitatifs, tandis qu'un modèle logistique est utilisé dans le cas de phénotypes binaires.

Phénotypes quantitatifs

Nous décrivons tout d'abord le modèle sur lequel reposent les phénotypes quantitatifs générés dans notre étude de simulation, à savoir, le modèle linéaire. Pour chaque individu i où $i = (1, \ldots, m)$, la valeur de son phénotype Y_i est déterminée à partir du modèle suivant :

$$Y_i = \sum_{j=1}^{m} G_{ij}\beta_j + \epsilon_i \text{ où } j = (1, \ldots, m) \tag{5.3}$$

dans lequel :
— G_{ij} correspond au génotype du SNP j pour l'individu i.

— Le paramètre β_j représente la taille d'effet (*effect size* en anglais) du SNP j sur le phénotype de l'individu i, autrement dit cette valeur reflète la taille d'effet du génotype du SNP j sur le phénotype. Nous présentons, Section 5.1.2, les différentes tailles d'effets considérées dans notre étude de simulation.

— Le paramètre ϵ_i représente l'erreur résiduelle du modèle. Ces résidus ont été générés à partir d'une loi normale $\mathcal{N}(0, \sigma^2)$.

Pour calibrer la force de l'association, σ^2 a été fixé en fonction du coefficient de détermination R^2 comme dans les travaux de Stanislas et al. (2017) :

$$\sigma_i^2 = \frac{(R^2 - 1)\sum(G_{ij}\beta_j - \bar{Y}_i)^2}{R^2(2 - n)} \tag{5.4}$$

Ce coefficient exprime la proportion des variations de Y_i expliquées par le modèle. Pour les données complètement simulées, nous avons fixé la valeur du coefficient de détermination R^2 à 0.2.

Phénotypes binaires

Pour les études cas-témoins, les phénotypes ont été générés à partir d'un modèle logistique. Ainsi, pour chaque individu i où $i = (1, \ldots, m)$, la valeur de son phénotype Y_i est obtenue à partir du modèle suivant :

$$\mathbb{P}(Y_i = 1|G_{ij}) = \frac{e^{\beta_0 + G_{ij}\beta_j}}{1 + e^{\beta_0 + G_{ij}\beta_j}} \text{ for } j = (1, \ldots, m) \tag{5.5}$$

dans lequel :

— G_{ij} et β_j, comme dans le cadre du modèle linéaire, représentent respectivement le génotype du SNP j et la force de son effet sur le phénotype de l'individu i. Nous détaillons, Section 5.1.2, les différentes tailles d'effets considérées dans notre étude de simulation.

— Le paramètre β_0 représente l'ordonnée à l'origine du modèle, elle est définie par la valeur moyenne attendue de Y lorsque tous les $G = 0$. Dans notre étude de simulation, la valeur de β_0 a été fixé de manière à obtenir des proportions équilibrées de cas et de témoins dans les échantillons.

Tailles d'effets

Nous détaillons dans cette section les tailles d'effet que nous avons considérées (β_j).

Les variants non causaux (m_0) sont les marqueurs considérés comme non associés avec le phénotype étudié. De ce fait, nous avons fixé leurs tailles d'effets à zéro, c'est-à-dire $\beta_{j_{m_0}} = 0$ quel que soit le type de leurs modèles considérés.

Pour les variants causaux, nous avons défini quatre sous-groupes selon les valeurs de MAF. Nous avons par conséquent appliqué différentes tailles d'effets afin de considérer plusieurs scénarios. Nous en avons élaboré trois distincts dont le premier est celui de référence (scénario 1) où les variants causaux rares ont un effet plus

important que les variants communs. Ce scénario représente ainsi le contexte dans lequel nous nous sommes placés pour le développement des procédures wBHa que nous avons présenté Chapitre 4. Pour une évaluation juste des différentes méthodes, nous avons également considéré deux autres scénarios où les variants rares ne sont plus favorisés. Dans le scénario 2, les tailles d'effets du scénario 1 sont inversées, en d'autres termes, les variants communs du scénario 2 ont un effet plus important que les variants moins fréquents. Dans le troisième et dernier scénario, les effets ne sont pas différenciés entre les sous-groupes de variants causaux, autrement dit les effets appliqués sont identiques et tous les $\beta_{j_{m_1}}$ sont égaux.

Pour chacun des scénarios, nous avons donc appliqué des valeurs distinctes de tailles d'effets pour chacun des sous-groupes de variants causaux. Nous avons par ailleurs différencié les valeurs attribuées aux sous-groupes de variants causaux pour un scénario en fonction du type de phénotype utilisé. Dans le cas de phénotypes binaires, nous avons utilisé le rapport des cotes, autrement appelé *Odds Ratio* (OR). Ce rapport mesure l'effet d'une variable explicative (ici le génotype d'un SNP j) sur une variable à expliquer binaire (ici le phénotype étudié).

L'odds (cotes) pour un génotype donné est défini comme le rapport de la probabilité de présenter le phénotype étudié ($Y = 1$) sachant le génotype sur la probabilité de ne pas présenter le phénotype ($Y = 0$) sachant le génotype. Pour un génotype $G = 1$, son odds est défini par :

$$odds(G = 1) = \mathbb{P}(Y = 1|G = 1)/\mathbb{P}(Y = 0|G = 1)$$

L'OR correspond au rapport des cotes qui est défini par :

$$OR = \frac{odds(G = 1)}{odds(G = 2)} = \frac{\mathbb{P}(Y = 1|G = 1)/\mathbb{P}(Y = 0|G = 1)}{\mathbb{P}(Y = 1|G = 2)/\mathbb{P}(Y = 0|G = 2)}$$

Lorsque la probabilité de présenter le phénotype d'intérêt ($Y = 1$) est faible, l'OR s'interprète comme le risque relatif qui correspond donc au rapport de la probabilité de présenter le phénotype $Y = 1$ en présence du génotype à risque par rapport à la probabilité de présenter ce phénotype en l'absence de celui-ci. L'OR est une mesure positive ($OR \in [0, +\infty[$), pouvant s'interpréter de différentes manières selon sa valeur :

— Si $OR = 1$, cela signifie que le génotype n'a pas d'effet particulier sur le phénotype, autrement dit, porter ce génotype n'augmente ni ne diminue pas le risque de présenter le phénotype étudié.

— Si $OR > 1$, cela indique que la présence de ce génotype augmente le risque de présenter le phénotype étudié, on définira la version de ce SNP comme un facteur de risque.

— Si $OR < 1$, cela indique que la présence de ce génotype diminue le risque de présenter le phénotype étudié, on définira la version de ce SNP comme un facteur de protection.

Dans un modèle logistique (Équation 5.5), l'OR peut être exprimé de la manière suivante :

$$OR = e^{\beta_j} \Leftrightarrow log(OR) = \beta_j \tag{5.6}$$

Ainsi, dans notre étude de simulation, nous avons exprimé les tailles d'effets des différents sous-groupes de SNPs causaux dans chaque scénario à partir d'OR.

Pour ce qui est des phénotypes quantitatifs, nous avons attribué les valeurs présentées dans le Tableau 5.1 qui résume les β considérés dans notre étude de simulation en fonction de la nature du phénotype ainsi que du scénario choisi.

		SNPs non causaux	**SNPs causaux Rares**	**SNPs causaux Médium Rares**	**SNPs causaux Moyens**	**SNPs causaux Communs**
Traits Quantitatifs	Scé. 1	0	4	3	2	1
	Scé. 2	0	1	2	3	4
	Scé. 3	0	2	2	2	2
Traits Binaires	Scé. 1	0	log(2.2)	log(1.8)	log(1.5)	log(1.3)
	Scé. 2	0	log(1.3)	log(1.5)	log(1.8)	log(2.2)
	Scé. 3	0	log(1.5)	log(1.5)	log(1.5)	log(1.5)

Table 5.1 – Tailles d'effets (β) des SNPs pour les phénotypes quantitatifs et binaires dans trois scénarios.

5.1.3 . Grands nombres d'hypothèses testées

Pour évaluer les différentes procédures avec un nombre d'hypothèses testées plus réaliste, nous avons simulé des données avec des valeurs m et m_1 telles que : $m \in \{100000,\ 200000,\ 500000\}$ et $m_1 \in \{100, 150, 250\}$ et avec $R^2 = 0.5$. Cependant, en raison d'un temps de calcul excessif pour générer les données, nous n'avons considéré que le cas indépendant avec des traits quantitatifs.

5.2 . Données semi-simulées

Pour avoir des simulations avec une structure de corrélation plus complexe, nous avons simulé des données basées sur un jeu de données réelles issues d'une étude sur l'infection au VIH (Dalmasso et al., 2008a).

Dans cette étude, 307 851 SNPs mesurés pour 605 individus (dont 531 hommes et 74 femmes) ont été analysés afin d'identifier de nouveaux variants génétiques associées aux taux d'ARN viral plasmatique et d'ADN proviral. Pour des raisons de temps de calcul trop longs, nous avons limité notre étude de simulation à la matrice génotypique correspondant au chromosome 6 qui a été largement rapporté dans la littérature. De plus, la Figure 5.1 présentant le Manhattan plot produit à l'issue des tests d'associations entre la charge virale plasmatique et les génotypes à l'aide d'un modèle linéaire, confirme que la majorité des variants associés sont localisés sur le chromosome 6.

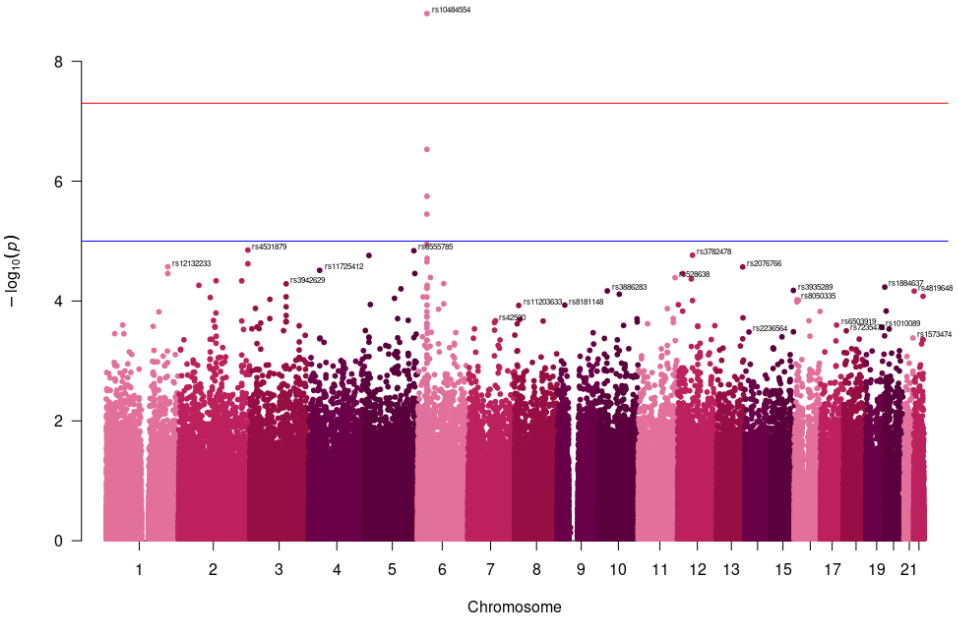

Figure 5.1 – Manhattan plot de l'étude d'association portant sur le VIH issue du travail de Dalmasso et al. (2008a). Les lignes bleues et rouges correspondent à deux seuils de significativité : $-log_{10}(10^{-5})$ et $-log_{10}(5.10^{-8})$ respectivement.

Les données du chromosome 6 ayant des valeurs manquantes, nous avons choisi un compromis entre la suppression et l'imputation des données. Nous avons tout d'abord supprimé les SNPs ayant plus de 10 valeurs manquantes. Les SNPs restants ont ensuite été imputés à l'aide de la méthode des K plus proches voisins, appelé également KNN (pour *k-Nearest Neighbors* en anglais). Le principe de cette méthode est de trouver pour chaque observation avec des données manquantes, les k observations complètes (c'est-à-dire sans données manquantes) les plus proches qui sont considérées comme les plus "représentatives" de l'observation. Pour cela, le paramètre k doit être préalablement fixé. Nous avons choisi une valeur de k égale à 1 afin que l'observation complète la plus proche soit admise comme la plus représentative. Pour déterminer l'observation la plus proche, la mesure de distance doit être également choisie au préalable. Nous avons choisi la distance euclidienne qui est la mesure la plus couramment utilisée. Une fois ces deux paramètres sélectionnés, l'algorithme consiste à attribuer à chaque valeur manquante la même valeur que celle de son voisin le plus proche. La Figure 5.2 illustre l'imputation à l'aide de la méthode KNN ($k = 1$) d'une matrice de génotypes codée par le modèle génétique additif pour 8 SNPs dont certains possèdent des données manquantes.

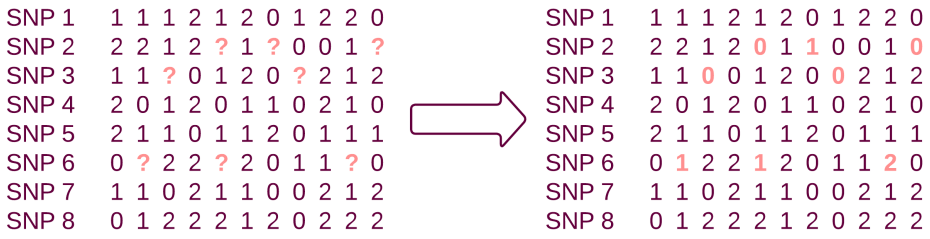

Figure 5.2 – Imputation d'une matrice de génotypes à l'aide la méthode KNN (avec $k = 1$) dans un échantillon de 8 SNPs dont certains possèdent des données manquantes (représentées par "?") pour 11 individus. Les SNPs 4, 7 et 1 sont respectivement les plus proches voisins des SNPs 2, 3 et 6 et ont été inférés à partir de ces SNPs. (Réalisé avec diagrams.net)

Une fois la matrice de génotypes filtrée et imputée, l'étape suivante consiste à générer les phénotypes. Pour ce faire, les SNPs causaux et non causaux doivent être tout d'abord identifiés. On commence pour cela par calculer les MAF des SNPs de la matrice de génotypes issue de l'imputation. Puis, les SNPs sont répartis en 4 groupes, selon leurs valeurs de MAF : le premier groupe correspond aux 558 SNPs ayant une MAF comprise entre 0.01 et 0.05, le second groupe correspond aux 4909 SNPs ayant une MAF comprise entre 0.05 et 0.15, le troisième groupe correspond aux 6674 SNPs ayant une MAF comprise entre 0.15 et 0.30 et le dernier groupe correspond aux 7840 SNPs ayant une MAF supérieure à 0.30. Une fois les variants

catégorisés, les variants causaux ont été tirés au hasard dans chaque groupe dans les mêmes proportions que pour les jeux de données complètement simulés. Les nombres de variants causaux considérés dans notre étude avec les données semi-simulées sont les suivants $m_1 \in \{20, 25, 50, 100, 150\}$.

Pour générer les phénotypes, il convient également de définir les tailles d'effets des SNPs. Comme pour les données complètement simulées, les tailles d'effets des SNPs non causaux ont été fixées à zéro ($\beta_{j_{m_0}} = 0$). Pour les tailles d'effets des SNPs causaux, nous nous sommes appuyés sur les valeurs réelles observées dans ce jeu de données. Pour cela, nous avons d'abord estimé le coefficient de régression de tous les SNPs significatifs lors de l'application de la méthode wBH au jeu de données d'origine au niveau $\alpha = 0.05$. Nous avons ensuite considéré les valeurs absolues des quatre quartiles de la distribution empirique de ces coefficients estimés pour définir la taille d'effet des variants causaux dans chacun des quatre groupes.

Nous avons considéré différents scénarios semblables à ceux pris en compte dans les données complètement simulées. Dans le scénario 1 les variants causaux rares ont un effet plus élevé que les variants communs, dans le scénario 2 les variants communs ont un effet plus élevé que les variants moins fréquents et dans le scénario 3 tous les effets sont également répartis entre les quatre groupes.

Une fois les génotypes triés et filtrés, les variants causaux et non causaux identifiés ainsi que leurs tailles d'effets définies, les phénotypes peuvent enfin être générés. Les phénotypes étudiés dans l'étude GWAS étant quantitatifs, nous avons choisi d'utiliser un modèle linéaire pour déterminer les phénotypes dans nos semi-simulations. Pour l'erreur résiduelle du modèle générée, nous avons fixé le paramètre R^2 à 0.8.

5.3 . Covariables et versions des packages

Une fois les phénotypes et les génotypes générés pour les données simulées (qu'elles soient complètement simulées ou basées sur un jeu de données réel), les p-valeurs sont calculées. À l'issue du calcul des p-valeurs, une procédure de tests multiples peut être appliquée. Nous avons donc appliqué les différentes procédures décrites dans les chapitres précédents à un niveau FDR de 5%. Pour chaque configuration, nous avons simulé 500 jeux de données.

5.3.1 . Pondérations et covariables

L'un des objectifs de la thèse étant de favoriser la détection des variants rares ayant des effets génétiques élevés, nous avons utilisé les fréquences alléliques comme covariable dans les procédures de tests multiples pondérées. En ce qui

concerne la procédure BH pondérée, nous avons fixé les poids de manière analogue à wBHa en considérant $w_i = \frac{m}{\sum_{j=1}^{m} \frac{1}{x_j}} \times \frac{1}{x_i}$ (où x_i est la MAF).

Cependant, comme mentionné précédemment, d'autres covariables informatives peuvent être prises en compte. À titre d'exemple, nous avons illustré l'utilisation de la méthode proposée pour favoriser les variants communs en remplaçant MAF par 1/MAF pour toutes les méthodes à l'exception de CAMT. Pour cette dernière méthode, nous avons utilisé $\log(1/MAF)$ afin d'éviter les problèmes de calcul dans l'algorithme EM dûs aux grandes valeurs de la covariable.

De plus, pour évaluer la robustesse de la méthode proposée à l'informativité de la covariable, nous avons simulé une covariable complètement non informative issue d'une distribution uniforme entre 0 et 1 ($U[0,1]$) lors de l'application des procédures pondérées.

5.3.2 . Versions des packages

Les packages, leurs versions et les fonctions utilisées dans toutes les analyses sont disponibles dans le Tableau 5.2.

Procédure	R package	Fonction	Version	Référence
BH	stats	p.adjust	4.2.1	Benjamini and Hochberg (1995)
qvalue	qvalue	qvalue	2.28.0	Storey and Tibshirani (2003)
FDRreg	FDRreg	FDRreg	0.2.1	Scott et al. (2015)
swfdr	swfdr	lm_qvalue	1.22.0	Boca and Leek (2018)
IHW	ihw	ihw	1.24.0	Ignatiadis et al. (2016)
CAMT	CAMT	camt.fdr	1.1	Zhang and Chen (2020)

Table 5.2 – Procédures incluses dans l'étude de comparaison.

5.4 . Critères d'évaluation

Nous présentons dans cette section les différents critères utilisés lors de l'évaluation des procédures.

5.4.1 . Puissance Globale

Afin d'évaluer la capacité des différentes procédures à détecter de vraies associations, pour chaque configuration, nous avons estimé la puissance moyenne $\mathbb{E}\left(\frac{VP}{m_1}\right)$ par la moyenne empirique (et son écart-type) du nombre de vraies découvertes sur les 500 jeux de données simulés divisé par m_1.

Cependant, pour les données corrélées, la définition des vrais et des faux positifs est ambiguë. En effet, dans le cas corrélés, un seul marqueur génétique influençant le phénotype peut conduire à l'obtention de plusieurs résultats significatifs pour les marqueurs corrélés (Benjamini et al., 2006; Siegmund et al., 2011; Brzyski et al., 2017). Pour résoudre ce problème, nous avons choisi d'estimer la puissance, et le FDR, dans les jeux de données corrélés en considérant des groupes de SNPs corrélés comme des unités d'intérêt. Les groupes ont été définis selon un seuil du coefficient de corrélation estimé de 0.8.

5.4.2 . Puissance dans le sous groupe des variants rares

Pour évaluer les performances des différentes méthodes spécifiquement dans le sous-groupe des variants rares, nous avons également estimé la puissance moyenne dans chaque sous-groupe $\mathbb{E}\left(\frac{VP_g}{m_{1g}}\right), g = (1, 2, 3, 4)$ (où $g = 1$ correspond au sous-groupe des variants causaux rares) par la moyenne empirique (et son écart-type) du nombre de vraies découvertes sur les 500 jeux de données simulés divisés par le nombre de variants causaux contenus dans chaque sous-groupe.

5.4.3 . Contrôle du FDR

Pour évaluer le contrôle FDR de chaque procédure, nous avons estimé le FDR par la moyenne empirique (et son écart-type) de la proportion de fausses découvertes observées sur les 500 jeux de données simulés.

5.5 . Données réelles

Pour illustrer les résultats obtenus avec les données simulées, nous avons appliqué les différentes procédures sur un jeu de données public issu de l'étude menée par Liu et al. (2017) et disponible dans la base de données Gene Expression Omnibus (GEO) (GSE90102). Cette étude porte sur la maladie de Crohn (MC) qui est une Maladie Inflammatoire Chronique de l'Intestin (MICI, IBD pour *Inflammatory Bowel Disease* en anglais) (Ananthakrishnan, 2015). Cette pathologie non contagieuse touche tous les segments du tube digestif (dont l'intestin) et se manifeste par des douleurs au niveau de l'abdomen par des diarrhées survenant par crises de durée et de fréquences variables. Les causes de cette maladie sont encore mal connues. Pourtant, certains facteurs de risques contribuent au développement de cette pathologie tels que des prédispositions génétiques ou environnementales (Abraham and Cho, 2009; Veauthier and Hornecker, 2018). En effet, il a été montré que les gènes NOD2 et ATG16L1, le tabagisme, l'usage de contraceptifs oraux, d'antibiotiques ou encore d'anti-inflammatoires non stéroïdiens peut augmenter le risque de développer cette maladie. Il n'existe pas actuellement de traitement curatif mais ceux existant permettent de soulager la douleur des patients afin d'améliorer la qualité de vie de ceux-ci.

En France, on estime que cette pathologie touche près d'une personne sur 1000, avec chaque année 8 nouveaux cas pour 100000 habitants (d'après l'assurance maladie française). Elle est le plus souvent diagnostiquée chez des patients âgés de 20 à 30 ans, mais peut être découverte à tout âge. De plus, la fréquence d'apparition de la maladie de Crohn a tendance à augmenter dans les pays industrialisés. D'après l'Association nationale française Afa Crohn RCH France, qui a fêté ses 40 ans en 2022, 120 000 personnes présentaient cette pathologie en France en 2022. De ce fait, cette maladie est un problème majeur de santé publique et les recherches la concernant sont essentielles afin de trouver de nouvelles cibles thérapeutiques.

L'un des objectifs de l'étude dont les données sont tirées était d'identifier de nouveaux variants génétiques potentiels influençant la maladie de Crohn. Pour cela, les auteurs se sont appuyés sur des résultats obtenus sur des patients occidentaux où le phénotype anormal des cellules de Paneth, qui sont des cellules tapissant la paroi de l'intestin grêle, est associé à la maladie de Crohn. En effet, il a été démontré que la morphologie de ces cellules était un biomarqueur chez des patients occidentaux présentant cette pathologie et qu'il en était de même concernant les gènes ATG16L1 et NOD2 (Wehkamp et al., 2004; Cadwell et al., 2008; Vandussen et al., 2014; Liu et al., 2018; Yang and Shen, 2021). La population japonaise étant génétiquement assez différente des populations occidentales, les auteurs ont émis l'hypothèse que de nouveaux variants génétiques associés au phénotype anormal des cellules de Paneth pourraient être découverts chez des patients japonais atteint de la MC.

Pour ce faire, les génotypes de 659 636 SNPs de 98 individus japonais atteints de la MC ont été récoltés ainsi que le phénotype des cellules de Paneth de ces individus. Avant de tester l'association entre ces génotypes et les phénotypes de ces individus qui représentait le pourcentage de cellules de Paneth anormales (phénotype quantitatif), l'étape de contrôle qualité des données est nécessaire. Pour notre analyse, nous n'avons considéré que les chromosomes autosomiques. Les contrôles de qualité standard ont été appliqués : nous avons supprimé tous les SNPs dont le call rate était inférieur à 95%, tous les SNPs présentant un écart significatif par rapport à l'équilibre de Hardy-Weinberg (p-valeur inférieure à 10^{-5}), et tous les SNPs avec une MAF inférieure à 0.01. La distribution des MAF de tous les SNPs restants après ces différentes étapes de contrôle qualité est présentée dans la Figure 5.3.

Après avoir appliqué ces filtres, l'association entre les génotypes des 607 720 SNPs restant et le phénotype des cellules de Paneth des 98 individus a pu être testée à l'aide du modèle du linéaire classique. Enfin, les différentes procédures de tests multiples présentées dans ce manuscrit ont été appliquées sur les p-valeurs issues de ces tests. Nous présenterons les résultats dans la Section 6.2 du Chapitre 6.

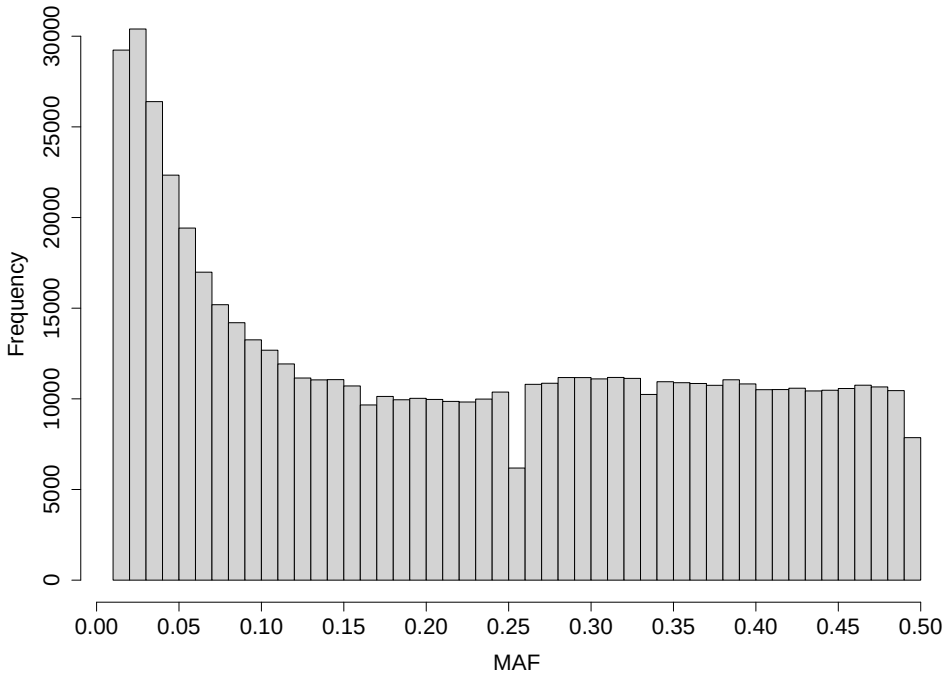

Figure 5.3 – Distribution des MAF des SNPs restant après l'étape de contrôle qualité des données sur la maladie de Crohn.

CHAPITRE 6
RÉSULTATS

Une fois le plan de simulations défini, les deux versions de notre procédure (wBHa version 1 et wBHa version 2) peuvent être évaluées au travers de celui-ci. Pour cela, elles ont toutes les deux été appliquées, ainsi que les méthodes existantes présentées dans ce manuscrit, sur nos données simulées. Nous avons comparé la puissance globale et la puissance des variants selon leurs fréquences alléliques ainsi que le contrôle du critère d'erreur FDR. Nous avons par ailleurs appliqué toutes ces procédures sur un jeu de données réel.

Sommaire

6.1 . Résultats de simulations

Nous présentons dans cette section les résultats obtenus lorsque la MAF a été utilisée comme covariable. Pour l'ensemble des simulations, nous détaillons les résultats obtenus concernant la puissance globale, la puissance des variants rares ainsi que le contrôle du FDR. Nous présentons par ailleurs les résultats obtenus pour des simulations ayant un grand nombre d'hypothèses testées. De plus, les résultats obtenus avec d'autres covariables que la MAF sont décrits dans cette section.

6.1.1 . Puissance Globale

Nous présentons dans cette section les résultats en termes de puissance globale, d'abord pour les données complètement simulées sans corrélations, puis pour celles où des corrélations ont été introduites. Nous présenterons par la suite les différences observées entre les deux versions de notre procédure. Enfin, nous décrivons les résultats observés avec les données semi-simulées.

La Figure 6.1 montre la puissance globale pour le scénario 1 (scénario de référence) avec des marqueurs indépendants. Les résultats pour les scénarios 2 et 3 sont disponibles en Annexe Figure S1 et Figure S2. Les résultats pour les marqueurs corrélés sont disponibles Figure 6.2 et en Annexe Figure S3 et Figure S4. Pour toutes les procédures, indépendamment des configurations, on observe que la puissance globale tend à diminuer avec le nombre total d'hypothèses testées m et qu'elle diminue également avec m_1. Ce résultat attendu s'explique par le fait que l'effet global se retrouve alors réparti sur un plus grand nombre de SNPs, rendant par conséquent l'effet individuel de chaque variant causal plus difficile à identifier. De plus, pour les jeux de données corrélés, on remarque que la puissance a tendance à augmenter avec ρ dans toutes les configurations.

Dans le scénario 1 avec des marqueurs indépendants, pour des petites et moyennes valeurs de m_1 ($m_1 \leq 25$ et $m_1 \leq 50$ pour les phénotypes quantitatifs et binaires, respectivement), wBHa et wBH ont tendance à être les procédures les plus puissantes, bien que pour les plus petites valeurs de m_1 dans le cas quantitatif ($m_1 \leq 15$), BH, qvalue et swfdr sont légèrement plus puissantes (Figure 6.1). Dans ces configurations, la procédure wBHa version 2 est plus puissante que la version 1. Pour des valeurs plus grandes de m_1, CAMT est la procédure la plus puissante pour les phénotypes quantitatifs (pour $m_1 \geq 50$) mais la moins puissante pour les phénotypes binaires pour toutes les configurations, où pour $m_1 \geq 100$ la procédure la plus puissante est IHW. À l'inverse, FDRreg a une bonne puissance globale pour les phénotypes binaires, mais est la procédure la moins puissante pour les phénotypes quantitatifs. Notons que si wBHa n'est pas toujours la procédure la plus puissante, elle a une assez bonne puissance globale dans toutes les configurations en comparaison des autres procédures.

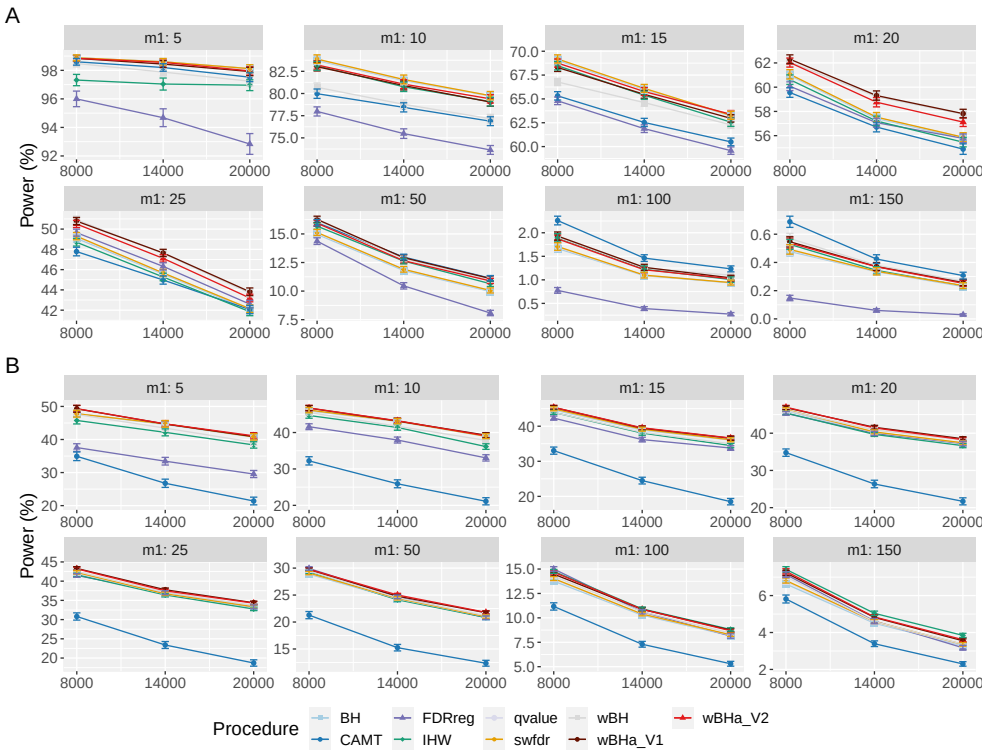

Figure 6.1 – Comparaison de la **puissance globale** dans le **scénario 1**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

Dans les scénarios 2 et 3 avec des marqueurs indépendants, pour des valeurs de m_1 petites et intermédiaires ($m_1 \leq 50$), wBHa, wBH, BH, qvalue et swfdr ont tendance à être les procédures les plus puissantes avec des résultats similaires (sauf pour les phénotypes binaires dans le scénario 3 où BH, qvalue et swfdr sont plus puissantes pour $m_1 \leq 50$) (Annexe Figure S1 et Figure S2). Pour les valeurs de m_1 plus grandes ($m_1 \geq 100$), IHW a tendance à être la procédure la plus puissante. Comme dans le scénario 1, CAMT donne de bons résultats avec les phénotypes quantitatifs, mais est la procédure la moins puissante pour les phénotypes binaires.

Pour les marqueurs corrélés (Figure 6.2, Annexe Figure S3 et Figure S4), wBHa fait partie des procédures les plus puissantes lorsque la valeur de m_1 est intermédiaire (dans le scénario 1) ou petite (dans les scénarios 2 et 3). Notons que dans le scénario 1 avec des traits binaires, wBHa version 2 a tendance à être la procédure la plus puissante dans toutes les configurations (Figure 6.2). De plus, on observe dans le cas de phénotype binaire lorsque $m_1 \geq 25$ pour $\rho = 0.75$ dans le scénario

de référence, que FDRreg devient la procédure la plus puissante. Dans les scénarios intermédiaires, FDRreg devient la plus puissante lorsque $m_1 = 150$ quelle que soit la valeur de ρ dans le scénario 2, et lorsque $m_1 \geq 100$ pour $\rho \geq 0.5$ dans le scénario 3.

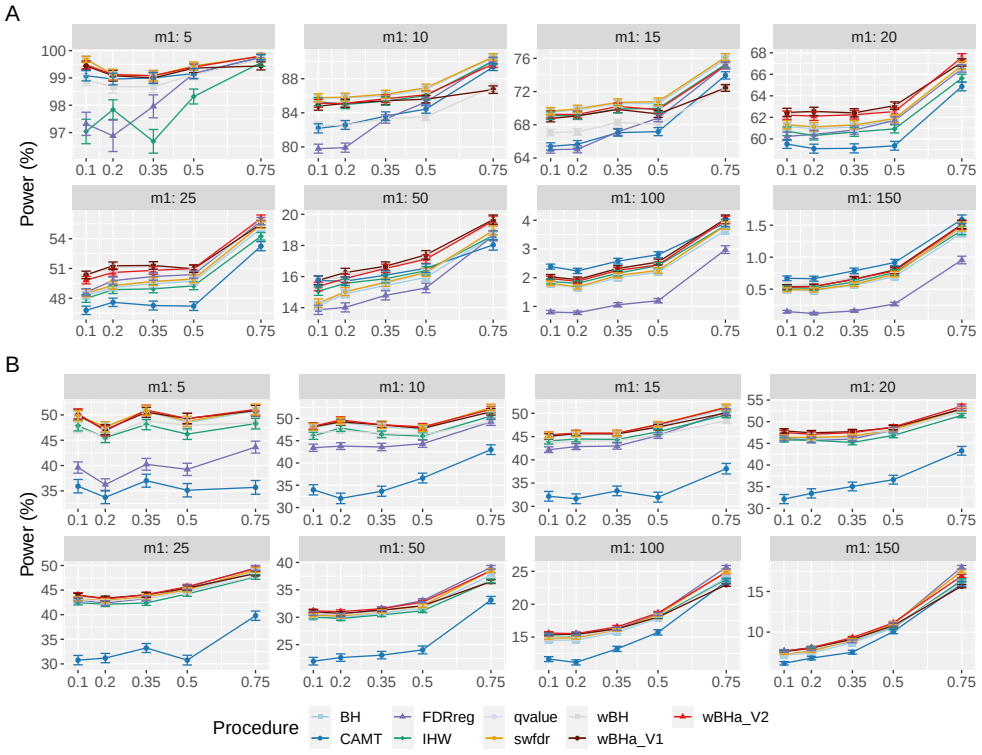

Figure 6.2 – Comparaison de la **puissance globale** dans le **scénario 1**, avec des variants **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

Lorsque l'on compare les deux versions de wBHa, nous avons des résultats hétérogènes selon le scénario dans lequel on se place et selon le type de phénotype étudié. Dans le scénario de référence (scénario 1), en présence de corrélation ou non, tandis que la procédure wBHa version 1 est plus puissante que la version 2 lorsque le phénotype est quantitatif (excepté pour $m_1 \leq 15$), nous observons l'inverse en présence de phénotype binaire. Dans les scénarios intermédiaires, en présence de corrélations ou non, la procédure version 2 est plus puissante que la version 1 dans la majorité des cas (excepté lorsque le nombre de variants causaux est faible en présence de phénotypes quantitatifs).

Les résultats obtenus avec les données semi-simulées sont similaires à ceux obtenus avec des données complètements simulées (Figure 6.3, Annexe Figure S5 et Figure S6). Dans tous les scénarios, pour des valeurs de m_1 petites et intermédiaires ($m_1 \leq 25$), wBHa (version 2) et wBH ont tendance à être les procédures les plus puissantes, tandis que CAMT et IHW sont plus puissantes pour de grandes valeurs de m_1 ($m_1 \geq 100$).

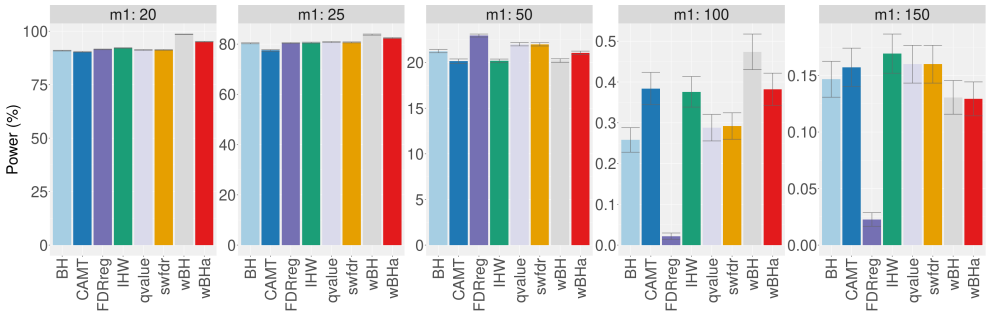

Figure 6.3 – Comparaison de la **puissance globale** dans le **scénario 1**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

6.1.2 . Puissance dans les sous-groupes

L'un de nos objectifs étant de favoriser la détection des variants de faibles fréquences, nous présentons dans cette section les résultats en termes de puissance dans le sous-groupe de variants rares, d'abord pour les simulations complètement simulées sans corrélations, puis pour celles où des corrélations ont été introduites. Nous présenterons par ailleurs les différences obtenues entre les deux versions de notre procédure. Enfin, nous décrirons les résultats observés avec les données semi-simulées.

La Figure 6.4 montre la puissance de détection des différentes procédures dans le sous-groupe des variants rares pour le scénario 1 (scénario de référence) avec des marqueurs indépendants. Les résultats pour les scénarios 2 et 3 sont disponibles en Annexe Figure S7 et Figure S8. Les résultats pour les marqueurs corrélés sont disponibles Figure 6.5 et en Annexe Figure S9 et Figure S10. Comme pour la puissance globale, la puissance dans le sous-groupe des variants rares tend à diminuer avec le nombre total d'hypothèses testées m et avec le nombre de SNPs causaux m_1 pour toutes les configurations. De plus, pour les marqueurs corrélés, la puissance tend à augmenter avec la valeur ρ.

Dans le scénario 1 avec des marqueurs indépendants (Figure 6.4), la procédure wBH, spécialement conçue pour ce contexte, est la procédure la plus puissante dans presque toutes les configurations. Cependant, notre procédure wBHa, quelle que soit sa version, montre une puissance assez importante par rapport aux autres procédures. Pour les grandes valeurs de m_1 ($m_1 \geq 50$), CAMT a tendance à être la procédure la plus puissante pour les phénotypes quantitatifs, mais elle est la moins puissante pour les phénotypes binaires. À l'inverse, FDRreg est performante pour les phénotypes binaires, mais elle est la moins puissante pour les phénotypes quantitatifs.

Il est intéressant de noter que dans les scénarios intermédiaires (scénarios 2 et 3), notre procédure et wBH ont tendance à être les plus puissantes dans le sous-groupe des variants rares pour toutes les configurations (Annexe Figure S7 et Figure S8). Cependant, comme attendu, la puissance de détection des variants rares de toutes les procédures est faible dans le scénario 2, où les effets les plus faibles sont attribués aux variants rares.

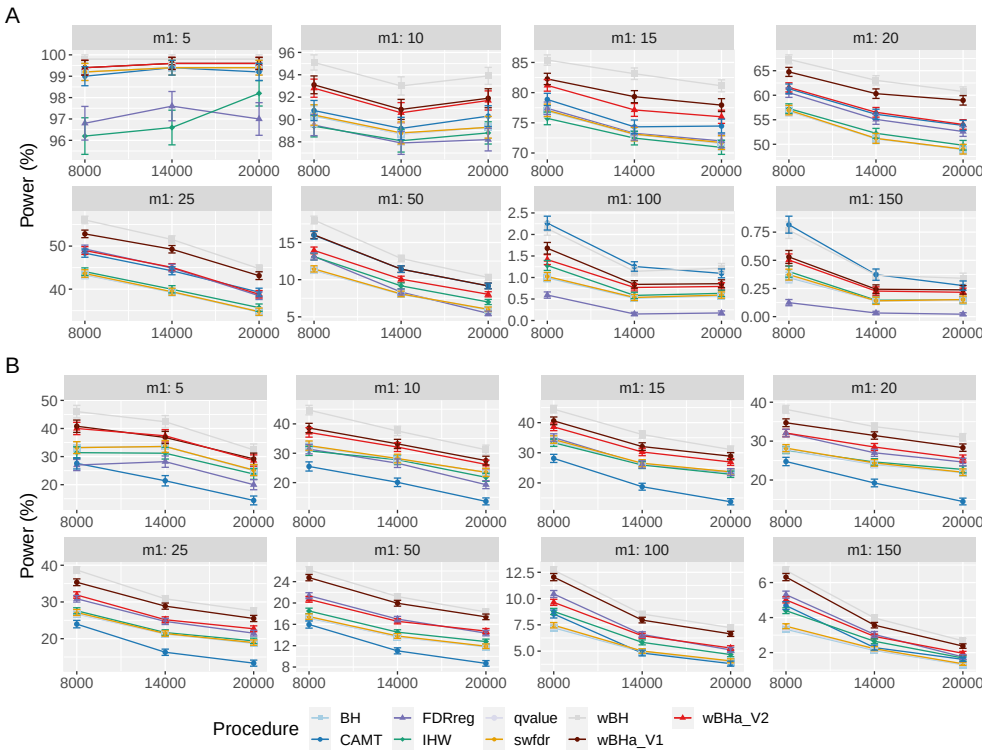

Figure 6.4 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 1**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

Dans le cas corrélé, pour tous les scénarios, nous avons obtenu des résultats similaires à ceux obtenus dans le cas indépendant (Figure 6.5, Annexe Figure S9 et Figure S10). Ainsi, wBHa et wBH ont tendance à être les procédures les plus puissantes dans le sous-groupe des variants rares pour toutes les configurations, excepté pour $m_1 \geq 50$ pour les traits quantitatifs dans le scénario 1.

Lorsque l'on compare les résultats des deux versions de wBHa dans le cas où la MAF est utilisée comme covariable, l'ensemble des figures présentées dans cette section montrent distinctement que wBHa version 1 est plus puissante dans le sous-groupe de variants rares que sa version 2 (quel que soit le scénario utilisé, le type de phénotype présent qu'il y ait ou non la présence de corrélations entre les SNPs).

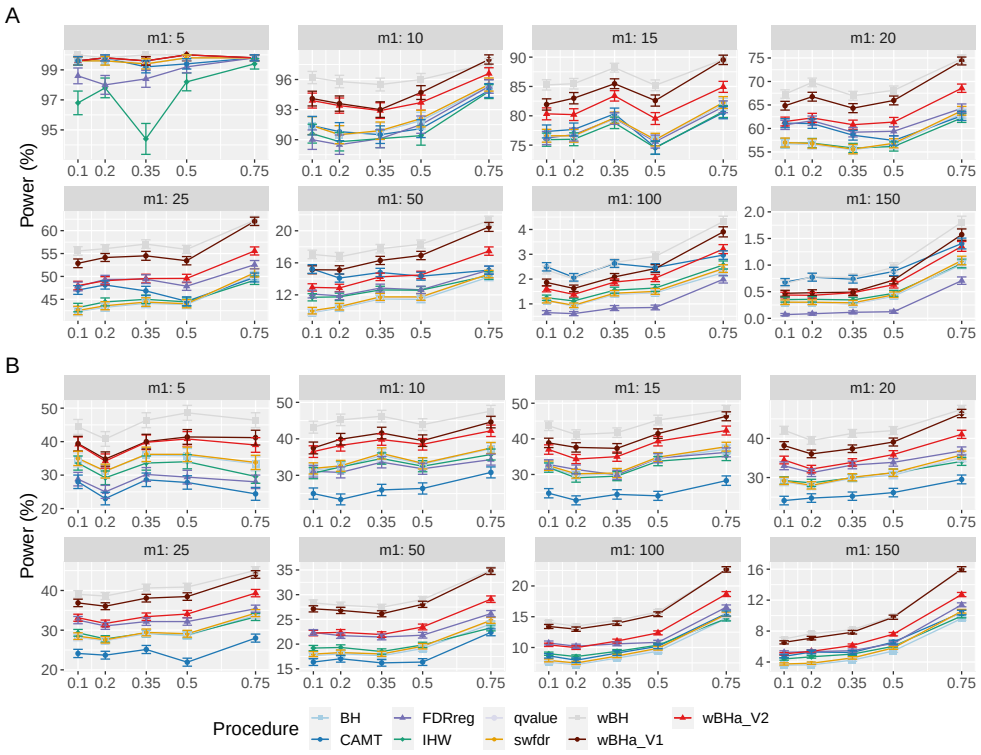

Figure 6.5 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 1**, avec des marqueurs **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

Lorsque l'on considère des données semi-simulées (Figure 6.6, Annexe Figure S11 et Figure S12), on remarque que notre procédure wBHa (version 2) est parmi les procédures les plus puissantes dans tous les contextes, la plus puissante étant wBH.

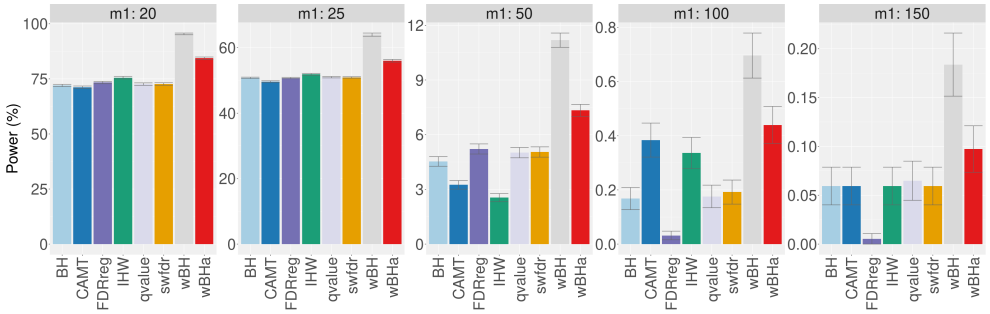

Figure 6.6 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 1**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

6.1.3 . Contrôle du FDR

Nous présentons dans cette section les résultats en termes de contrôle du FDR d'abord pour les données complètement simulées sans corrélations, puis pour les données avec corrélations (complètement simulées et semi-simulées).

La Figure 6.7 montre le FDR estimé pour toutes les procédures dans le scénario 1 avec des marqueurs indépendants. Ces résultats indiquent un bon contrôle du FDR pour toutes les procédures dans toutes les configurations (excepté pour FDRreg avec des phénotypes binaires). En effet, pour toutes les procédures sauf FDRreg, le FDR estimé est inférieur à 0.05 ou légèrement supérieur à ce seuil. Cela peut s'expliquer par le fait que dans nos configurations simulées, le nombre de rejets a tendance à être faible, entraînant une assez grande variabilité de la proportion de fausses découvertes. Ainsi, même avec la procédure BH (pour laquelle le contrôle du FDR a été théoriquement prouvé pour des tests indépendants), le FDR estimé est légèrement supérieur au seuil de 0.05 dans certains cas. Des résultats similaires ont été obtenus pour les scénarios 2 et 3 (Annexe Figure S13 et Figure S14 respectivement).

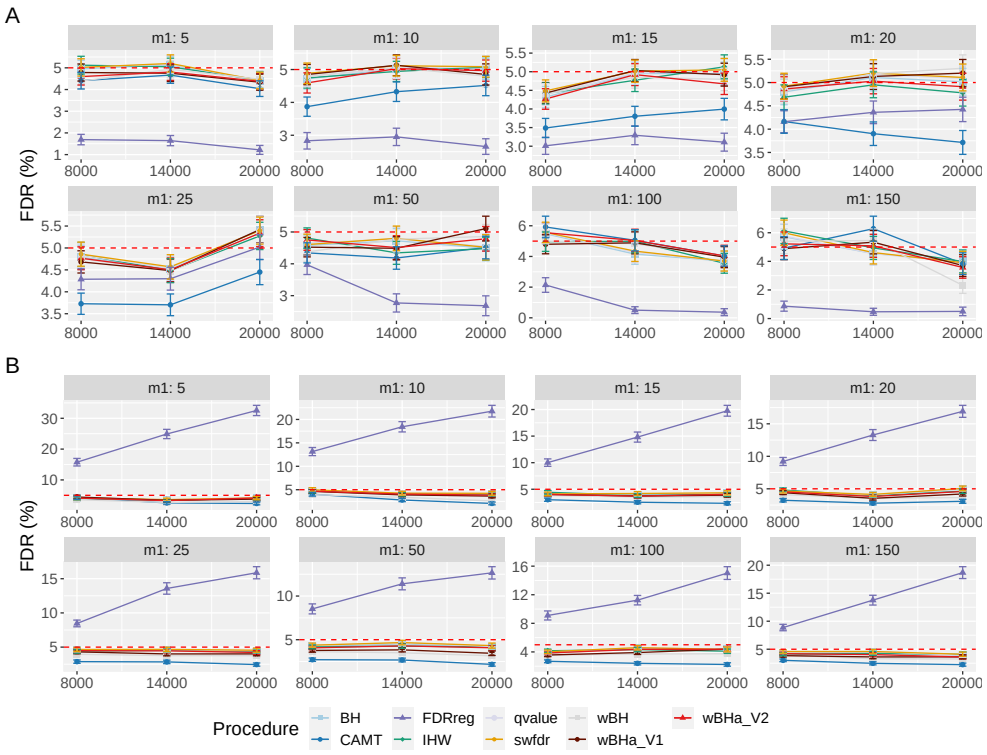

Figure 6.7 – Comparaison du **FDR** dans le **scénario 1**, avec des variants **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. La ligne pointillée rouge correspond au niveau cible de FDR (5 %). Les barres verticales illustrent les erreurs standard.

Dans le cas corrélé (Figure 6.8), le FDR estimé augmente avec ρ dans toutes les configurations. Des résultats similaires ont été obtenus dans les scénarios 2 et 3 (Annexe Figure S15 et Figure S16). De plus, le FDR estimé dans le cas des simulations basées sur le jeu de données réel (Figure 6.9, Annexe Figure S17 et Figure S18) a tendance à être élevé pour toutes les procédures. Ces résultats illustrent ainsi la difficulté de définir et de contrôler le FDR lorsque les hypothèses testées sont corrélées. Les FDR estimés étant similaires d'une procédure à l'autre, les comparaisons de puissance restent cependant pertinentes.

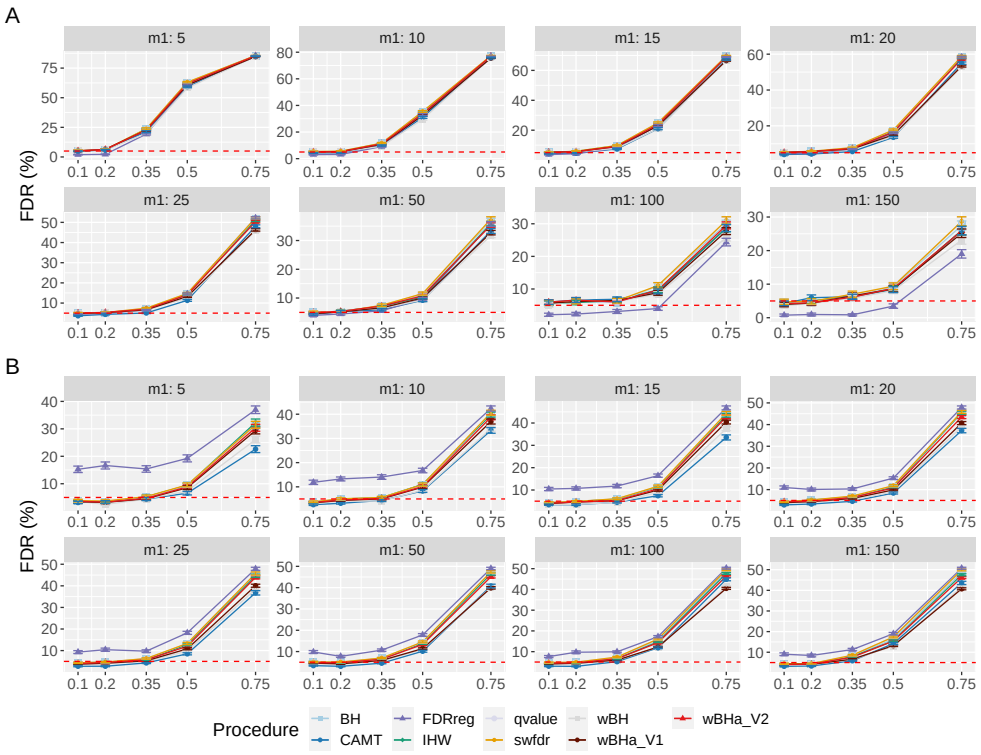

Figure 6.8 – Comparaison du **FDR** dans le **scénario 1**, avec des marqueurs **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. La ligne pointillée rouge correspond au niveau cible de FDR (5 %). Les barres verticales illustrent les erreurs standard.

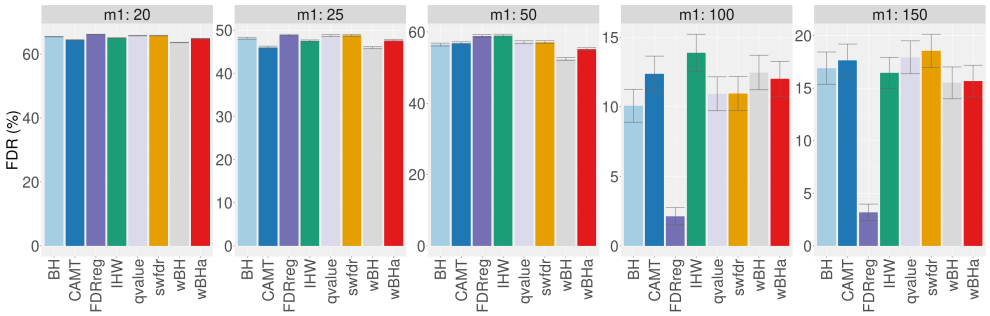

Figure 6.9 – Comparaison du **FDR** dans le **scénario 1**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

6.1.4 . Grand nombre d'hypothèses testées

Afin d'évaluer les procédures avec des simulations se rapprochant le plus possible de la réalité, nous avons également réalisé des simulations avec un nombre d'hypothèses testées plus important. Comme pour les simulations complètement simulées ou basées sur des données réelles, les procédures ont été comparées en termes de puissance globale, puissance dans le sous-groupe de variants rares ainsi que le contrôle du FDR. La procédure wBHa version 1 n'a cependant pas été incluse dans cette évaluation qui a été mise en œuvre au cours du développement de la version 2.

Lorsque l'on considère des valeurs plus élevées de m et m_1 pour les phénotypes quantitatifs, wBHa (version 2) fait partie des trois procédures les plus puissantes dans le scénario 1 tout en conservant une bonne puissance globale dans les scénarios 2 et 3 (Annexe Figure S19). wBHa est également l'une des trois procédures les plus puissantes dans le sous-groupe de variants rares dans le scénario 1, tandis que dans les scénarios 2 et 3, CAMT, IHW et swfdr ont tendance à être légèrement plus puissantes, en particulier pour les grandes valeurs de m_1 (Annexe Figure S20). La Figure S21 indique un assez bon contrôle du FDR pour toutes les procédures.

6.1.5 . Autres covariables

Pour illustrer l'utilisation de wBHa dans un contexte différent, nous avons également évalué les différentes procédures lorsque l'on cherche à favoriser les variants communs. Pour cela, nous avons remplacé la MAF par $1/\text{MAF}$ pour toutes les procédures sauf pour la procédure CAMT qui ne donnait pas toujours de résultats. En effet, les fréquences étant transformées par leur inverse, les faibles MAF conduisent à de très grandes valeurs de $1/\text{MAF}$. C'est notamment en raison de la présence de ces grandes variables que la procédure CAMT ne peut, parfois, pas appliquer l'algorithme EM provoquant par conséquent des erreurs de calcul et l'arrêt de la procédure immédiat. Plutôt que de retirer la procédure CAMT de la comparaison, nous avons choisi de modifier légèrement la covariable en utilisant $log(1/\text{MAF})$ plutôt que $1/\text{MAF}$.

Pour vérifier que cette modification n'a pas eu d'impact majeur sur les résultats, nous avons comparé les résultats obtenus avec la covariable $1/\text{MAF}$ et ceux obtenus avec $log(1/\text{MAF})$. Pour CAMT, cette comparaison a été réalisée uniquement à partir des jeux de données pour lesquels il était possible d'avoir des résultats, soit des jeux de données ayant des nombres de simulations variant de 136 à 500 par

configuration. Les Figures 6.10 et Figure 6.11 (ainsi que les Figure S22, Figure S23, Figure S24 et Figure S25 disponibles en annexe) montrent peu de différences entre les deux covariables pour CAMT.

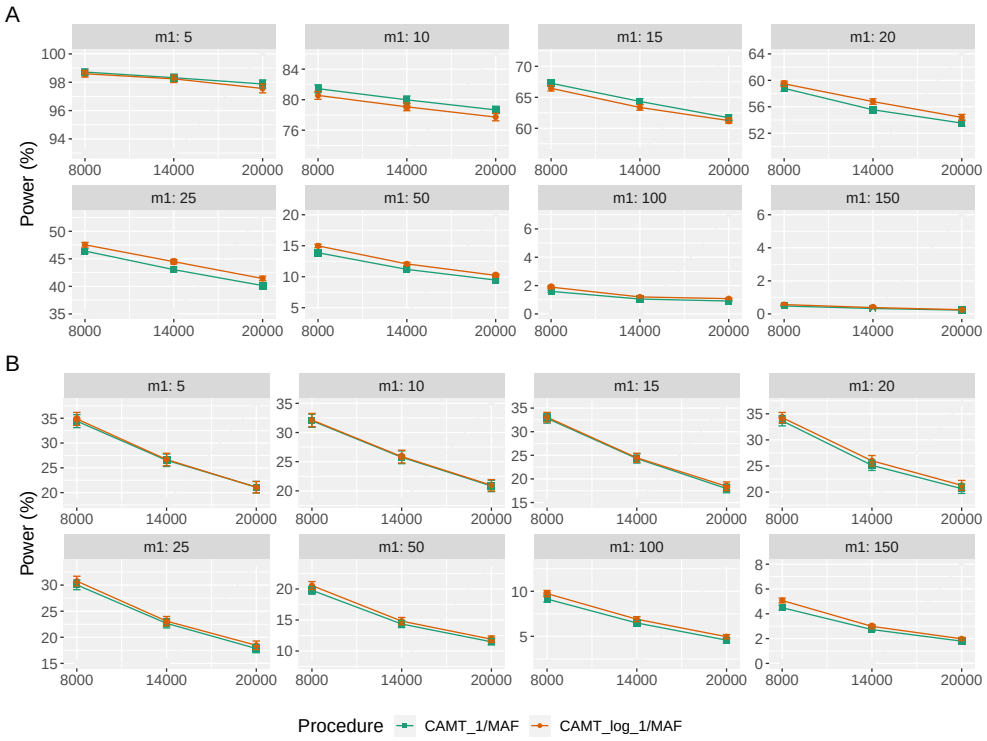

Figure 6.10 – Comparaison de la **puissance globale** de la procédure CAMT pour différentes covariables dans le **scénario 1**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

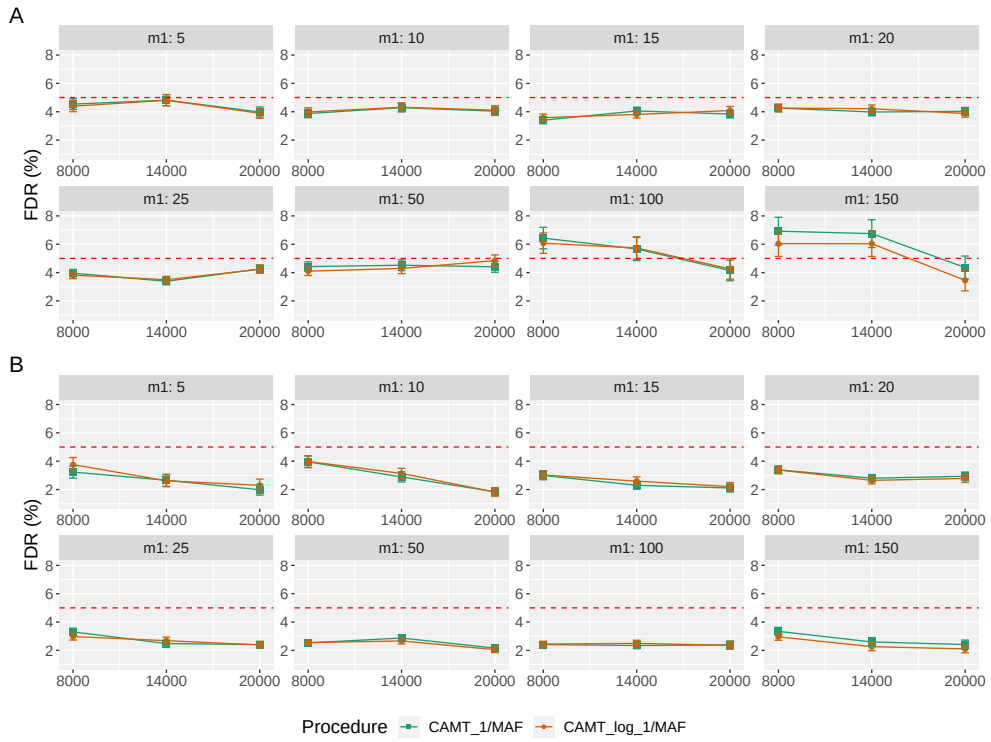

Figure 6.11 – Comparaison du **FDR** de la procédure CAMT pour différentes covariables dans le **scénario 1**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

Lorsque la MAF est remplacée par $1/\text{MAF}$ (ou $log(1/\text{MAF})$ pour CAMT) dans le but de favoriser les variants communs, notre procédure wBHa est l'une des procédures les plus puissantes pour détecter les variants communs dans tous les scénarios (Figure 6.12, Annexe Figure S26 et Figure S27) tout en conservant une puissance globale similaire à celle de la procédure BH non pondérée (Figure 6.13, Annexe Figure S28 et Figure S29). Le FDR est contrôlé par toutes les procédures sauf FDRreg pour les phénotypes binaires (Figure 6.14, Annexe Figure S30 et Figure S31).

Lorsque l'on compare les deux versions de wBHa, on remarque que les résultats sont relativement hétérogènes, comme lors de l'utilisation de la MAF comme covariable. En effet, dans le cas du scénario de référence (scénario 1), quel que soit le type de phénotype, la version 2 est plus puissante que la version 1 lorsqu'on a des valeurs de m_1 grandes ou intermédiaires ($m_1 \geq 20$ et $m_1 \geq 10$ pour le cas quantitatif et binaire respectivement). Dans les scénarios 2 et 3, la version 2 est

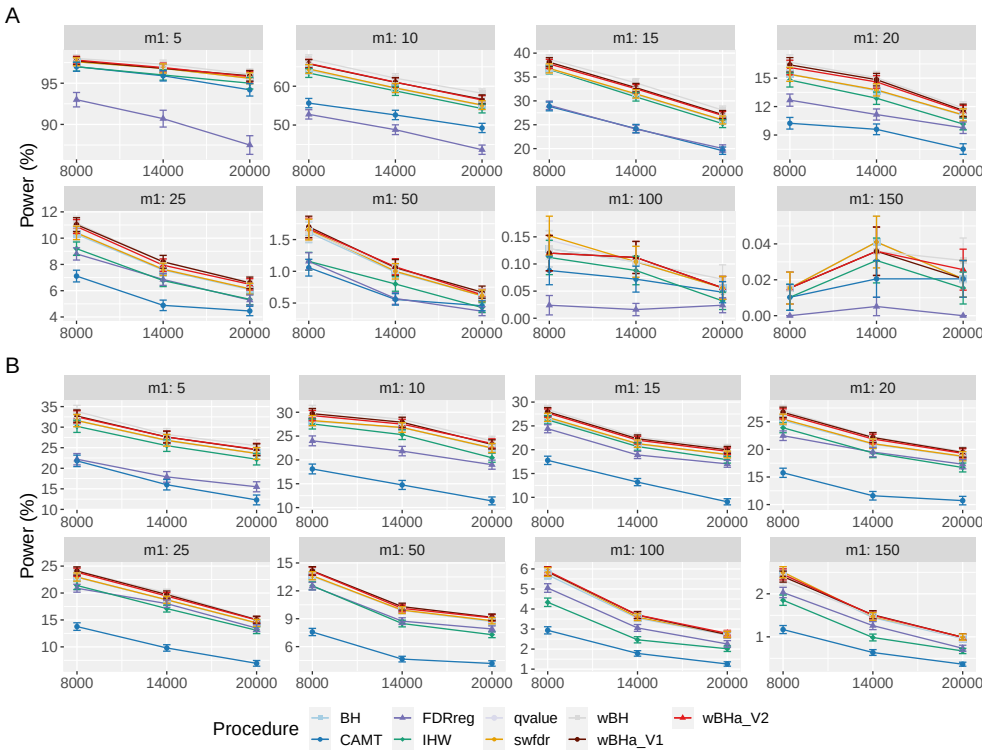

Figure 6.12 – Comparaison de la puissance dans le sous-groupe de **variants communs** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 1** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

globalement plus puissante que la version 1 (excepté pour des grandes valeurs de m_1 dans le cas linéaire, pour des petites et grandes valeurs de m_1 dans le cas binaire). Cependant, wBHa version 1 est la plus puissante dans le sous groupe de variants communs.

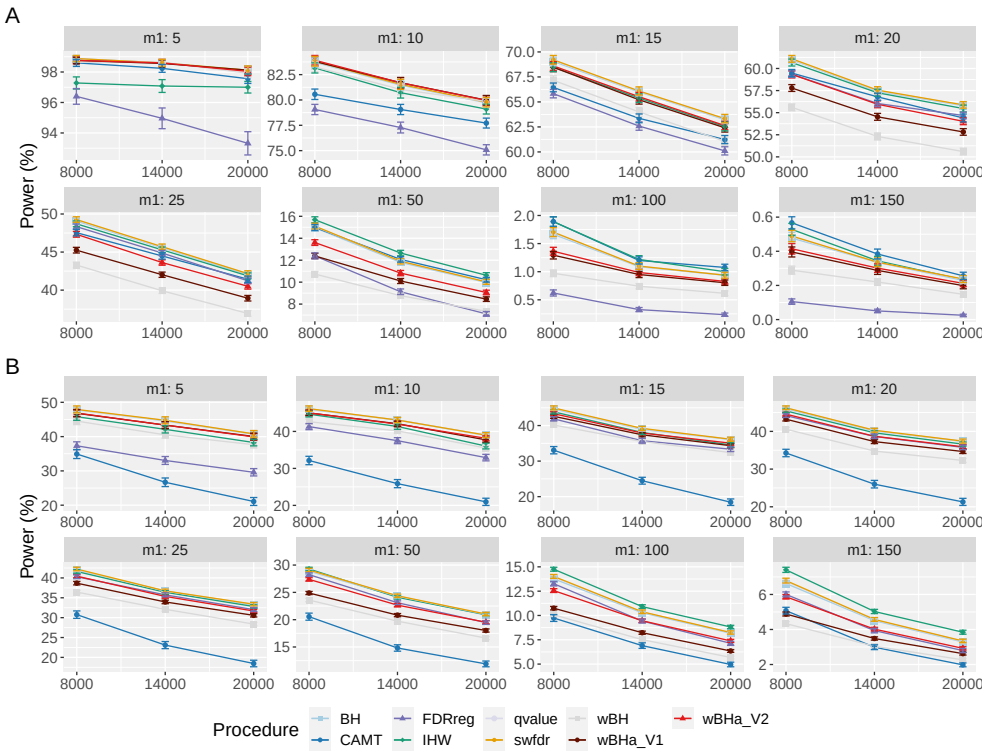

Figure 6.13 – Comparaison de la **puissance globale** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 1** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

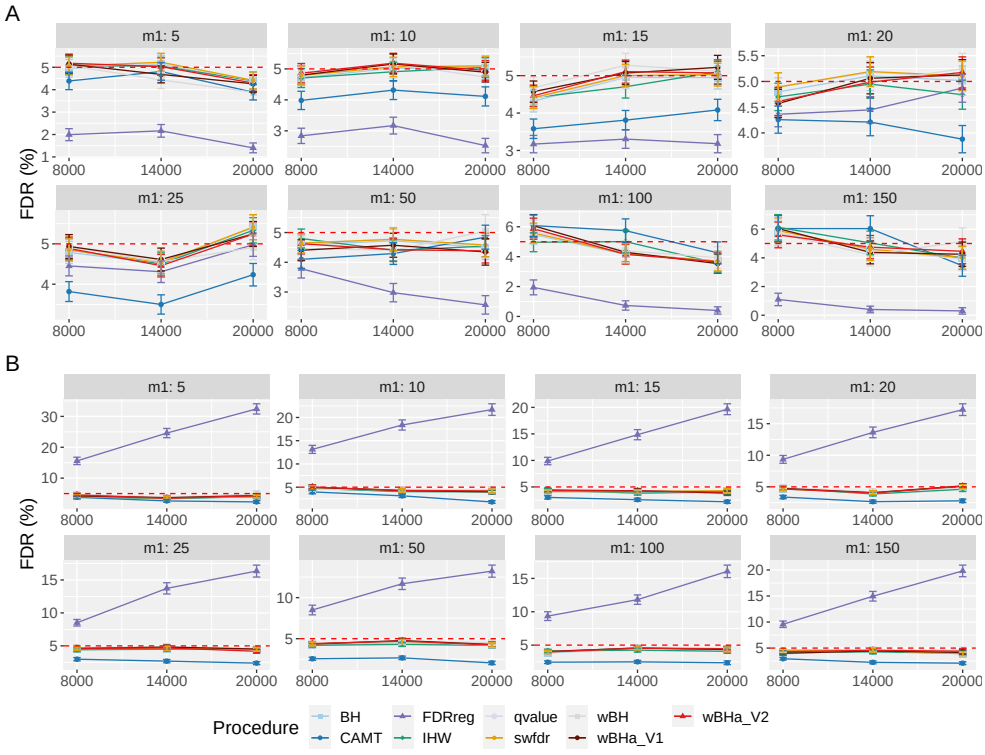

Figure 6.14 – Comparaison du **FDR** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 1** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

Lorsqu'on utilise une covariable non informative, les différentes procédures restent valables puisque le FDR est contrôlé au niveau souhaité (Figure 6.15, Annexe Figure S32 et Figure S33). Cependant, toutes les procédures pondérées ont tendance à avoir une puissance globale inférieure à la procédure BH non pondérée, bien que wBHa version 2 soit celle avec la plus petite perte (ce qui n'est pas le cas pour wBHa version 1) (Figure 6.16, Annexe Figure S34 et Figure S35).

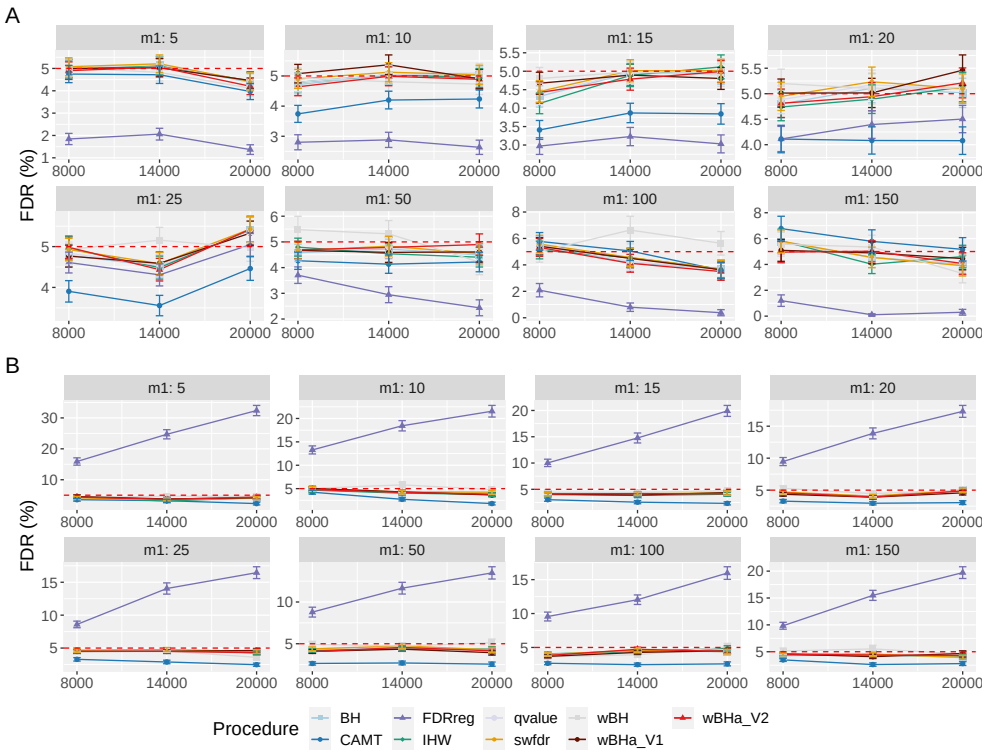

Figure 6.15 – Comparaison du **FDR** lors de l'utilisation de covariable **non informative** dans le **scénario 1** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

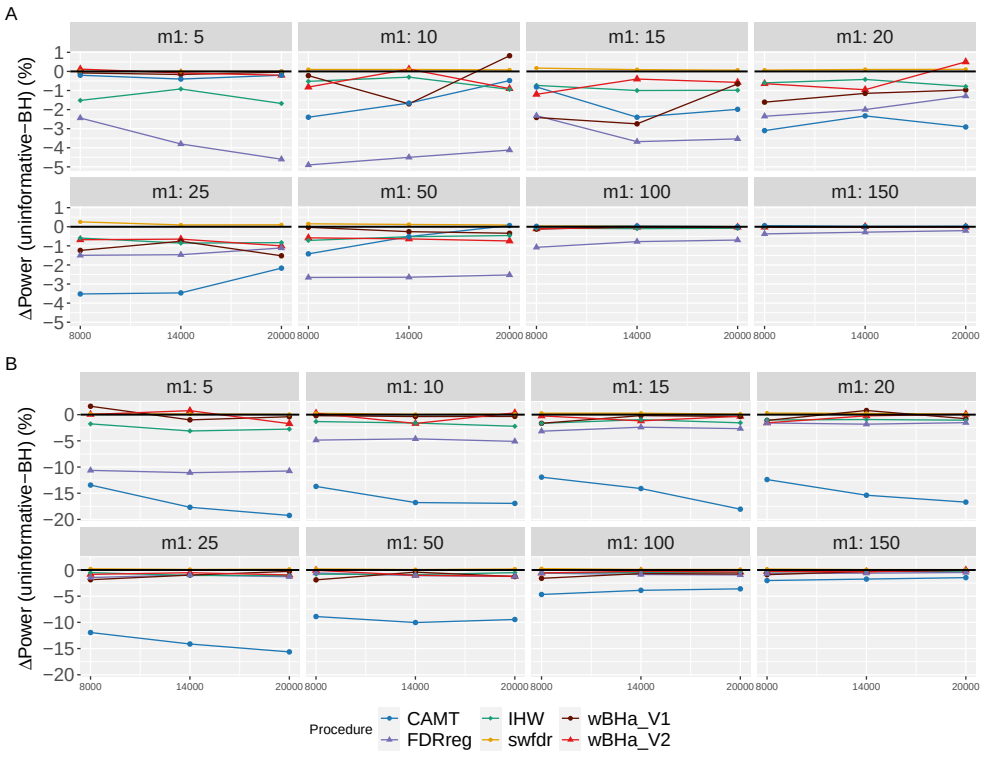

Figure 6.16 – **Différence de puissance globale** entre l'utilisation de covariable non informative et la procédure BH dans le **scénario 1** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement.

6.2 . Analyse de données réelles

Nous présentons tout d'abord les résultats obtenus avec toutes les procédures (exceptée wBHa version 1) lors de l'analyse des données réelles qui sont décrites Section 5.5. Puis, nous discutons de la variabilité des deux versions de wBHa.

6.2.1 . Puissances globales dans le sous-groupe de variants rares

La Figure 6.17 montre le nombre total de rejets pour chaque procédure pour différentes catégories de MAF. Les procédures CAMT et IHW, qui ont tendance à être les procédures les plus puissantes dans notre étude de simulation pour les phénotypes quantitatifs ont conduit, aux plus grands nombres totaux d'hypothèses nulles rejetées ($R = 111$ et $R = 109$, respectivement). La procédure wBHa version 2 a identifié 106 résultats significatifs. Les procédures BH, qvalue et swfdr ont conduit au plus petit nombre de rejets où $R = 43$ qui, est le même pour ces trois procédures. De plus, notons que la procédure wBHa version 2 a produit le plus grand nombre de rejets pour les SNPs ayant une MAF inférieure à 0.02.

Bien que wBHa version 2 ne soit pas la procédure la plus puissante, elle a néanmoins permis d'identifier 6 SNPs spécifiques (Figure 6.18) qui n'ont pas pu être sélectionnés par les autres procédures. Il est intéressant de noter que deux de ces SNPs, rs3772479 et rs2270569, sont situés respectivement dans les gènes FHIT et KIF9, qui ont été signalés comment ayant un rôle important dans les maladies inflammatoires de l'intestin (IBD pour *Inflammatory Bowel Disease* en anglais) (la maladie de Crohn étant une IBD) (Skopelitou et al., 2003; Xu and Qiao, 2006; Wierzbicki et al., 2009; Wang et al., 2018).

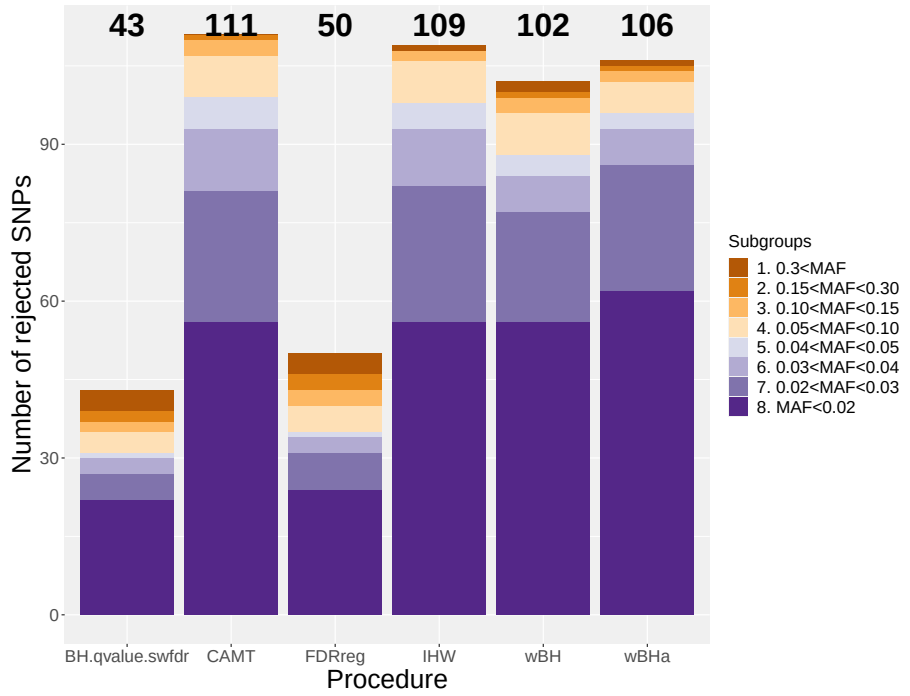

Figure 6.17 – Nombre de SNPs rejetés pour les sous-groupes de SNP pour chaque procédure.

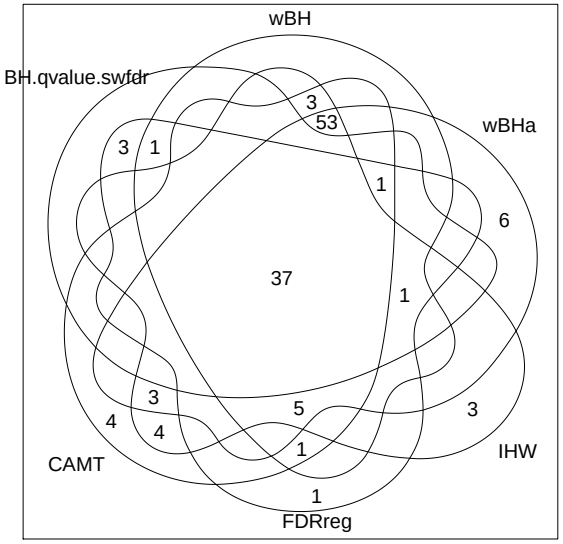

Figure 6.18 – Diagramme de Venn des SNP sélectionnés pour toutes les procédures.

6.2.2 . Reproductibilité

Comme évoqué précédemment Section 4.4.3, un problème de variabilité dans la procédure wBHa version 1 a été détecté, ce qui a conduit au développement d'une seconde procédure : wBHa version 2. En effet, en répétant plusieurs fois l'analyse d'un même jeu de données avec wBHa version 1, il est possible d'obtenir des résultats parfois très différents. Bien que ce ne soit pas le cas avec la plupart des jeux de données simulés, nous avons observé ce phénomène en analysant les données issues de l'étude de Liu et al. (2017). Nous avons donc comparé les deux versions de wBHa en répétant 500 fois l'analyse de ce même jeu de données.

La Figure 6.19 présente les deux boites à moustaches (ou *boxplot* en anglais) permettant de visualiser et de comparer la répartition du nombre de rejets pour chacune des versions de wBHa (1 et 2). On remarque que la procédure wBHa version 1 conduit à une variabilité plus grande que la version 2.

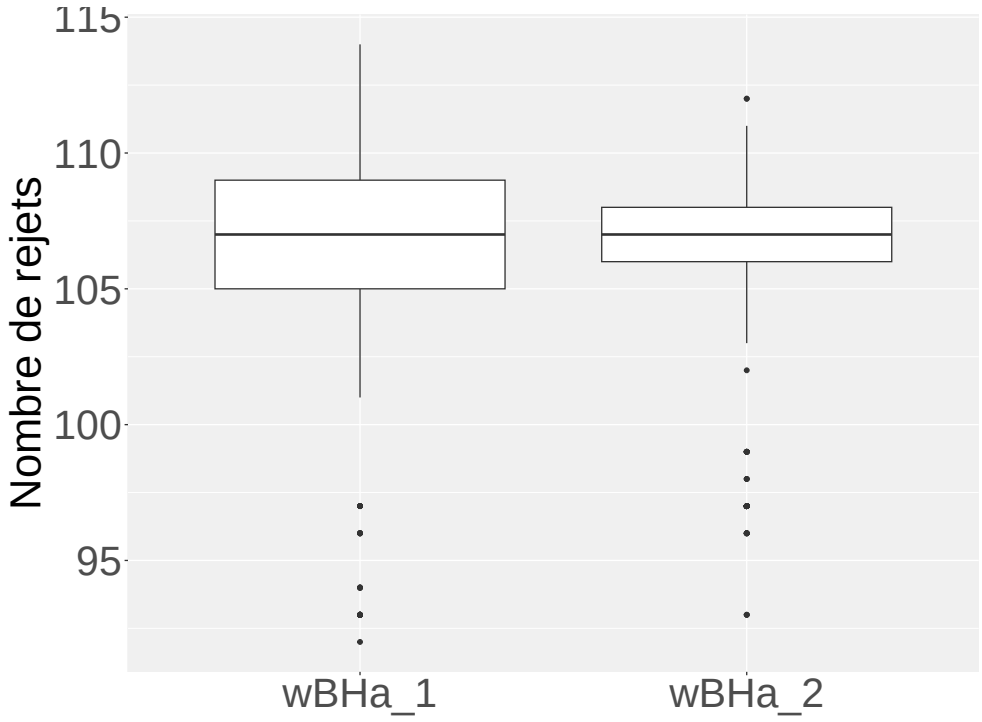

Figure 6.19 – Comparaison du **nombre de rejets** des différentes versions de wBHa sur les données sur la maladie de Crohn (pour 500 itérations).

CHAPITRE 7
DISCUSSION, CONCLUSION ET PERSPECTIVES

Sommaire

7.1 . Discussion

Dans cette thèse, nous avons évalué des procédures de tests multiples pondérés récentes dans le contexte des études d'association pangénomiques. Nous avons également introduit deux versions d'une nouvelle procédure appelée wBHa qui vise à prioriser la détection de marqueurs génétiques ayant une faible MAF tout en laissant la procédure adapter une fonction de pondération afin de maximiser la puissance de détection globale. Une évaluation des différentes procédures et de celle développée a été réalisée au travers d'une étude de simulation et d'une application sur un jeu de données réelles.

Pour les jeux de données dans lesquels les SNPs ont été simulés sous l'hypothèse d'indépendance des marqueurs, wBHa, quelle que soit sa version, a obtenu de bons résultats par rapport aux autres procédures, avec une assez bonne puissance globale dans toutes les configurations simulées. Comme l'a noté Korthauer et al. (2019), nous avons constaté qu'IHW et CAMT sont globalement plus puissantes lorsque la proportion d'hypothèses alternatives augmente. Cependant, la proportion d'hypothèses alternatives est difficile à estimer et l'utilisation de ces deux procédures dans un contexte dans lequel seuls quelques marqueurs sont associés au phénotype peut conduire à une diminution de la puissance globale. Le fait que wBHa tende à être la procédure la plus puissante dans les scénarios 2 et 3 pour des proportions plus faibles de variants causaux que dans le scénario 1 peut s'expliquer par la difficulté croissante à détecter les variants causaux lorsque leur taille d'effet est plus petite.

Lorsque l'on considère la puissance de détection des associations au sein du sous-groupe de variants rares, la procédure wBH standard est la plus puissante dans tous les scénarios. Cependant, wBHa est non seulement puissante dans le sous-groupe des variants rares, mais elle a également une bonne puissance globale. Ces résultats démontrent l'importance du paramètre d'optimisation a dans la procédure wBHa.

En ce qui concerne le contrôle du FDR, la plupart des procédures semblent contrôler correctement le critère d'erreur pour les jeux de données indépendants, bien que dans certains cas, le FDR estimé pour toutes les procédures (y compris BH) soit légèrement supérieur au seuil. Ceci peut s'expliquer par le faible nombre de rejets, entraînant une grande variabilité de la proportion de fausses découvertes. Cependant, FDRreg ne semble pas contrôler le FDR dans les études cas-témoin, ce qui est cohérent avec des résultats similaires obtenus par Boca and Leek (2018), Korthauer et al. (2019) et Zhang and Chen (2020) pour certaines configurations.

Pour les jeux de données avec corrélations (entièrement simulés ou basés sur un jeu de données réel), nous avons obtenu des résultats similaires en termes de puissance. Ainsi, wBHa a montré de bonnes performances par rapport aux autres

procédures. Comme attendu, une perte de contrôle du FDR est observée avec toutes les procédures lorsque les corrélations augmentent, et la difficulté de définir les vrais et les faux positifs reste un défi. Cependant, en pratique, l'influence des corrélations sur le contrôle du FDR peut être limitée en utilisant des méthodes telles que l'élagage (LD pruning) (Purcell et al., 2007).

Lorsque la MAF est utilisée comme covariable dans le scénario de référence (scénario 1), wBHa version 1 était plus puissante que wBHa version 2 dans la majorité des cas pour des phénotypes quantitatifs. Cependant, dans des configurations pour lesquelles wBHa tend à être plus puissante que des autres procédures, la version 2 tend à être plus puissante que la version 1. Ces résultats sont également observés dans les scénarios intermédiaires. On obtient des résultats similaires lorsque la covariable (1/MAF) est utilisée.

Ainsi, la puissance globale de wBHa version 2 a tendance à être plus grande que la puissance globale de wBHa version 1. Cependant, la puissance de détection des variants rares de wBHa version 1 est généralement plus grande que la puissance de détection des variants rares de wBHa version 2.

En résumé, bien que wBHa ne soit pas la procédure la plus puissante dans toutes les configurations, elle a montré de bonnes performances par rapport aux autres procédures, non seulement en termes de puissance globale, mais également en ce qui concerne la puissance de détection dans le sous-groupe de variants rares. En particulier, dans les scénarios 2 et 3 pour lesquels les variants rares ont des tailles d'effet modérées ou faibles, wBHa s'est avérée puissante dans le sous-groupe des SNPs rares, montrant ainsi son intérêt. Ainsi, wBHa permet la détection de variants rares tout en ayant une puissance globale similaire à celle des autres procédures, quelle que soit la taille d'effet des variants rares.

Pour illustrer les résultats obtenus avec les données simulées, nous avons appliqué les différentes procédures à un jeu de données réel public portant sur la maladie de Crohn. Les résultats obtenus sont cohérents avec ceux obtenus avec les données simulées. Les procédures avec pondérations ont une puissance plus grande que les procédures sans pondérations, ce qui se traduit par un plus grand nombre d'hypothèses nulles rejetées. De plus, wBHa a identifié six variants rares spécifiques qui n'ont pas pu être sélectionnés par les autres procédures. Parmi eux, deux marqueurs sont situés dans les gènes FHIT et KIF9, qui ont été rapportés comme étant impliqués dans les IBD, suggérant qu'ils pourraient correspondre à de véritables associations. Ces résultats soulignent l'intérêt de wBHa dans les applications réelles.

Au regard des deux versions de notre méthode, nous pouvons conclure que bien que wBHa version 2 ait une meilleure puissance globale, wBHa version 1 favorise davantage la détection des variants d'intérêt. Cependant, wBHa version 1 peut dans certains cas ne pas être reproductible. Bien que wBHa version 1 puisse être considéré comme valide dans la mesure où le risque d'erreur FDR est contrôlé (Zhang et al., 2019), l'utilisation pratique peut s'avérer délicate.

Aussi, nous pensons que prendre comme seuls critères d'évaluations les puissances et le contrôle du FDR n'est pas satisfaisant et la reproductibilité des résultats a son importance. Dans la mesure où wBHa version 1 n'est pas reproductible, surtout lorsque le nombre d'hypothèses nulles rejetées est relativement faible, cette version de wBHa peut être utilisée dès que l'on s'attend à un nombre important de marqueurs associés au phénotype. Mais, de manière générale, il nous semble préférable d'utiliser wBHa version 2 qui est plus robuste. Une étude plus poussée de la reproductibilité des différentes procédures serait cependant intéressante à mettre en place.

Les procédures de tests multiples pondérées adaptatives basées sur des covariables informatives sont ainsi très prometteuses dans le contexte des études d'association pangénomiques. Notre nouvelle procédure wBHa, qui a montré de bonnes performances dans tous les contextes, semble être un bon choix pour prioriser les variants rares sans perte de puissance globale.

7.2 . Conclusion et Perspectives

Dans cette thèse, nous avons développé une procédure de tests multiples pondérée originale. Deux versions de cette procédure sont proposées, chacune ayant ses avantages et inconvénients. Elles ont toutes les deux un objectif commun : favoriser la détection des variants rares qui manquent de puissance avec les procédures existantes dans le cadre des GWAS.

Le principe de notre procédure est basé sur la construction de poids définis par une fonction dépendant à la fois de la fréquence allélique des variants, mais aussi d'un paramètre libre permettant de maximiser la puissance. Pour limiter le phénomène de surapprentissage, une approche de rééchantillonnage des données basée sur le Bagging est appliquée afin d'estimer le paramètre optimal. Un paramètre unique maximisant le nombre de rejets est sélectionné pour chacun des sous-échantillons issus du Bagging. C'est à cette étape que les deux versions proposées diffèrent. La version 1 utilise des régressions splines cubiques tandis que la version 2 utilise une stratégie prenant en compte le fait que des valeurs successives du paramètre peuvent être obtenus.

Nous avons comparé notre approche avec des procédures de tests multiples avec pondérations existantes. Pour ce faire, un plan de simulation adapté au contexte des GWAS a été construit et ajusté tout au long de la thèse. Dans ce plan de simulation, la variation d'un certain nombre de paramètres tels que le nombre de marqueurs étudiés, le type de phénotype étudié (binaire ou quantitatif), les MAF, les effets des variants causaux, la structure de dépendance entre les SNPs ont été pris en compte.

Les résultats préliminaires de notre méthode ont indiqué de bonnes performances en termes de puissance de détection et de contrôle du risque d'erreur, mais la présence d'un problème de reproductibilité pour la version 1. Afin de corriger ce problème de non-reproductibilité, le choix des paramètres de la méthode de ré-échantillonnage ainsi que la manière de déterminer le paramètre optimal final dans la procédure ont dû être reconsidérés. La difficulté principale dans l'élaboration de cette procédure était l'obtention d'une puissance suffisamment compétitive tout en garantissant le contrôle du critère d'erreur.

Nous avons choisi de publier nos travaux dans la revue PeerJ, ce qui s'inscrit dans une démarche de science ouverte et reproductible. Le principe de la science ouverte consiste à rendre disponible et à partager les résultats de la recherche autant que possible par le biais de publication d'article dans des revues en *open-access*, le partage des données ou encore l'évaluation par les pairs. La science ouverte à de nombreux avantages (Barker et al., 2022; Hagger, 2022; Neumann, 2022; Pulverer, 2023). Par exemple, les articles publiés dans des revues en libre accès ont une meilleure visibilité que des articles uniquement accessibles par un abonnement payant, les collaborations entre chercheurs sont également facilitées. La publication des commentaires des relecteurs ainsi que des différentes versions de l'article à chaque itération apporte des garanties sur la qualité du processus d'évaluation. La reproductibilité des données a également suivi l'essor de la science ouverte, elle repose sur le principe FAIR (Findable, Accessible, Interoperable and Reusable). Le principe FAIR a pour objectif de favoriser la découverte, l'accès, l'interopérabilité et la réutilisation de données partagées (Wilkinson et al., 2016; Denecker and Toffano-Nioche, 2021). L'ensemble des scripts développés durant la thèse pour la création des jeux de données, la comparaison et l'évaluation des procédures, la visualisation et l'illustration des résultats ont été déposés sur GitHub. De plus, un package R documenté a été implémenté et est disponible sur GitHub.

Nos travaux ont permis d'obtenir des résultats ouvrant de nouvelles perspectives sur la détection et l'analyse des variants rares dans les études d'association pangénomiques. Des perspectives d'amélioration de notre procédure peuvent cependant être considérées comme la prise en compte de plusieurs covariables. Prendre en compte des déséquilibres de liaisons entre les variants pourraient être également

une piste d'amélioration de notre procédure. La procédure a été développée dans le cadre de données issues de puces à ADN. Une extension de cette procédure permettant d'intégrer les spécificités des données issues de techniques de séquençage pourrait également être envisagée.

Contributions

Publication

Obry Ludivine and Dalmasso Cyril (2023). Weighted multiple testing procedures in Genome-Wide Association Studies. PeerJ, 11 :e15369.

Conférence

Journées Ouvertes en Biologie, Informatique et Mathématiques 2020 (JOBIM) : Weighted multiple testing procedures in Genome-Wide Association Studies (Poster)

Logiciel

Dépôts GitHub :

Package R : https://github.com/obryludivine/wBHa

Scripts : https://github.com/obryludivine/wBHa_simulation

Annexes

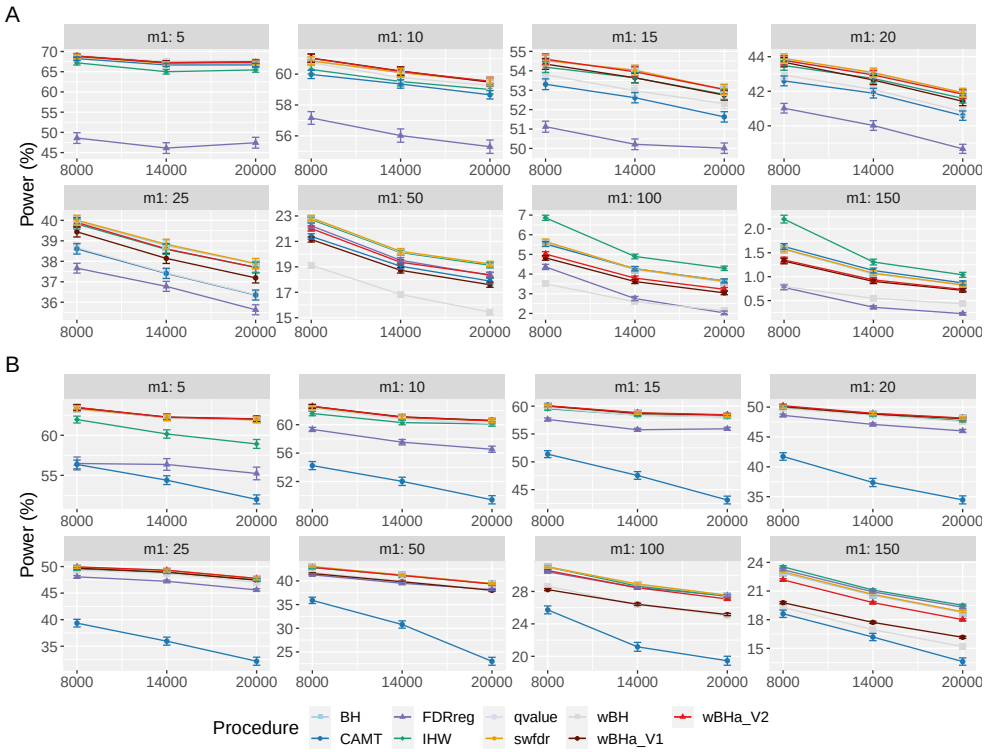

Figure S1 – Comparaison de la **puissance globale** dans le **scénario 2**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

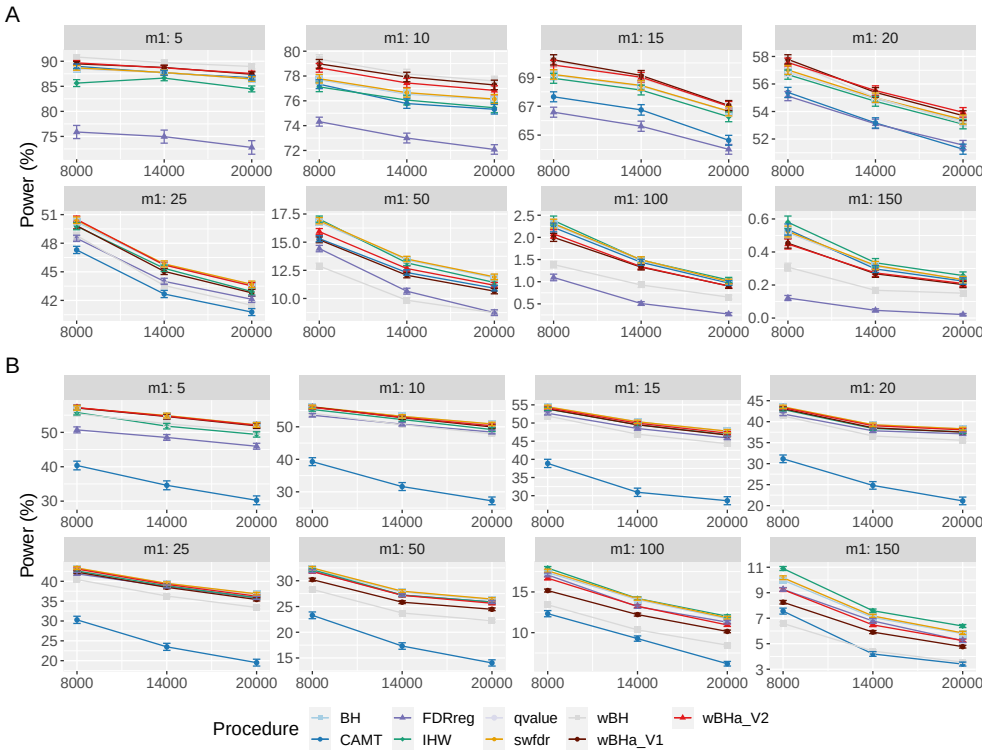

Figure S2 – Comparaison de la **puissance globale** dans le **scénario 3**, avec des variants **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

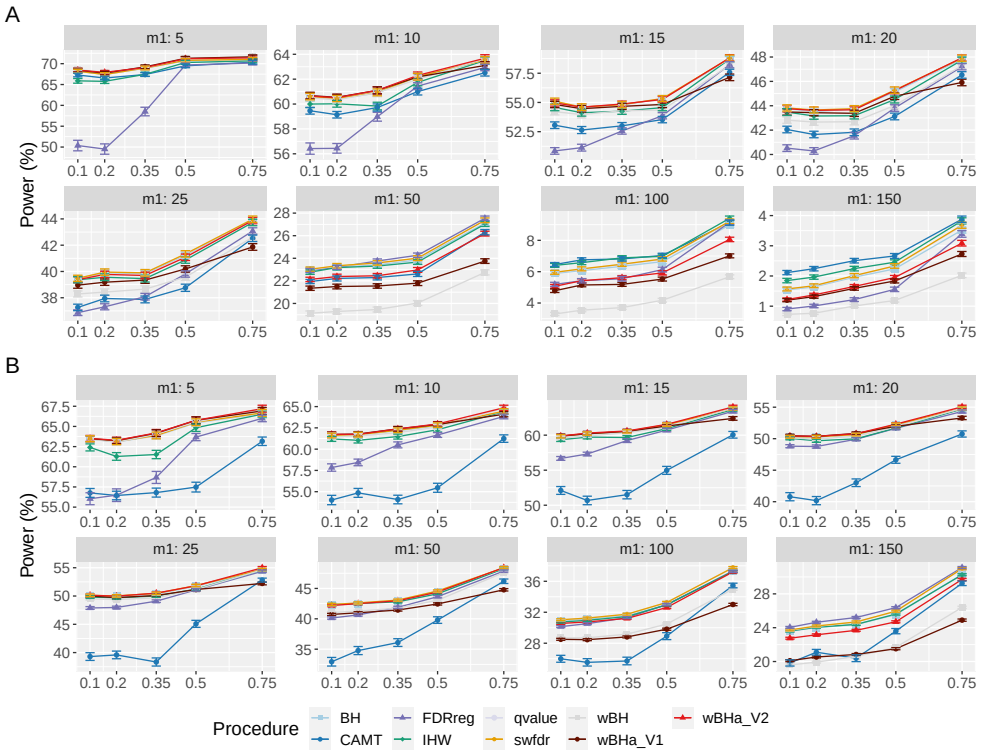

Figure S3 – Comparaison de la **puissance globale** dans le **scénario 2**, avec des variants **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

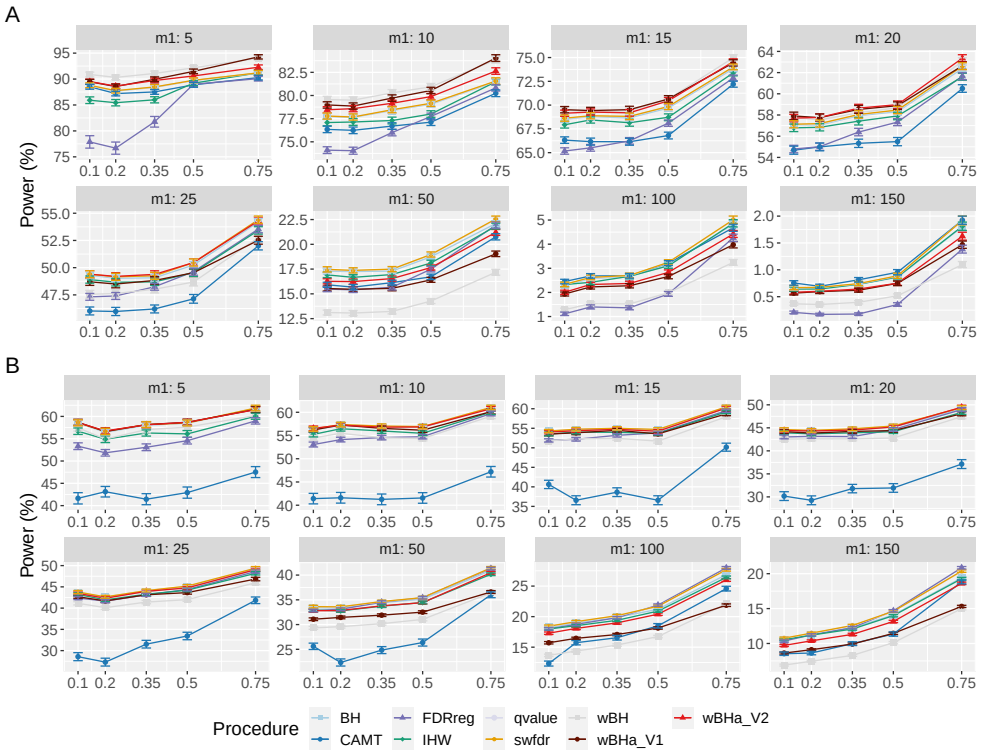

Figure S4 – Comparaison de la **puissance globale** dans le **scénario 3**, avec des variants **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

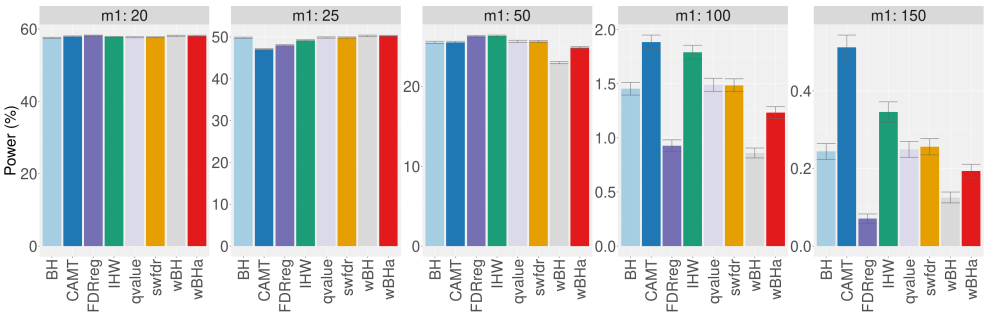

Figure S5 – Comparaison de la **puissance globale** dans le **scénario 2**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

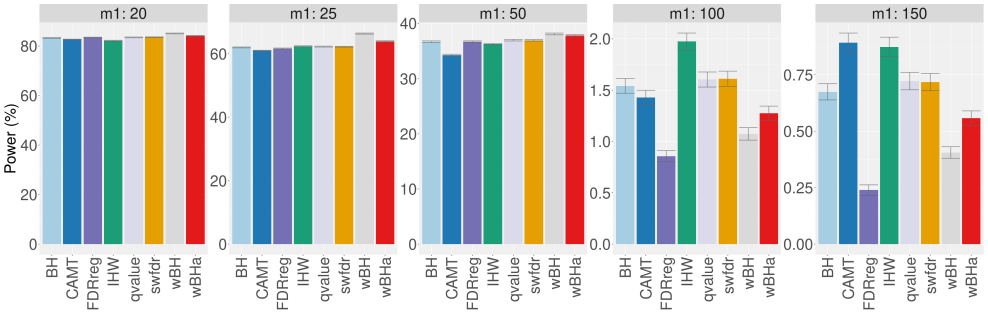

Figure S6 – Comparaison de la **puissance globale** dans le **scénario 3**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

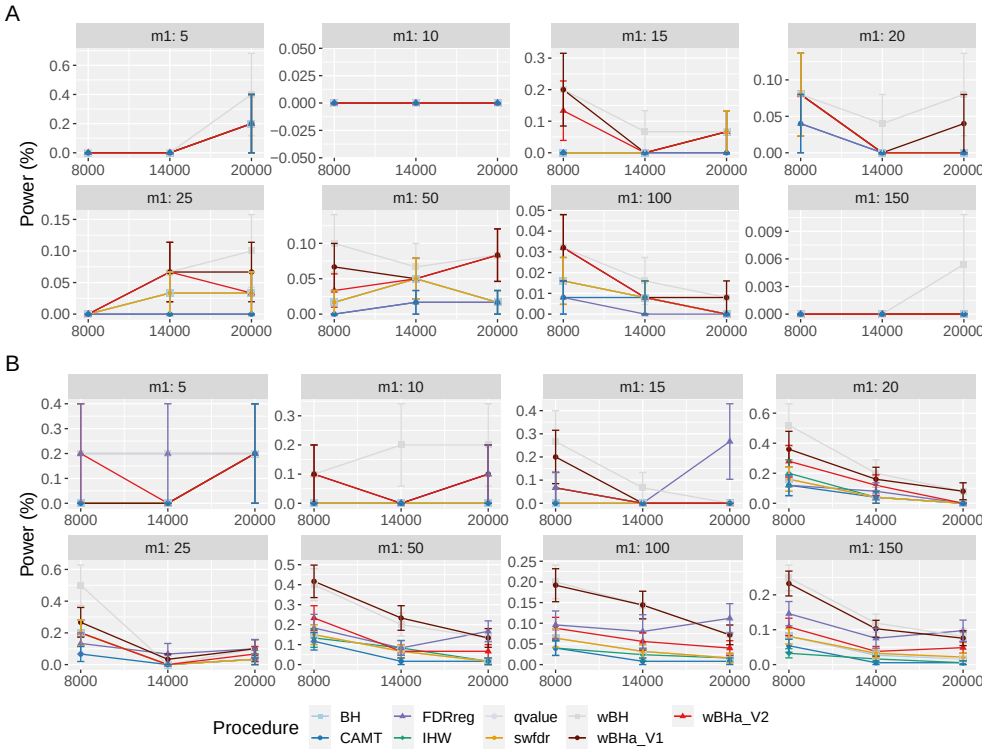

Figure S7 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 2**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

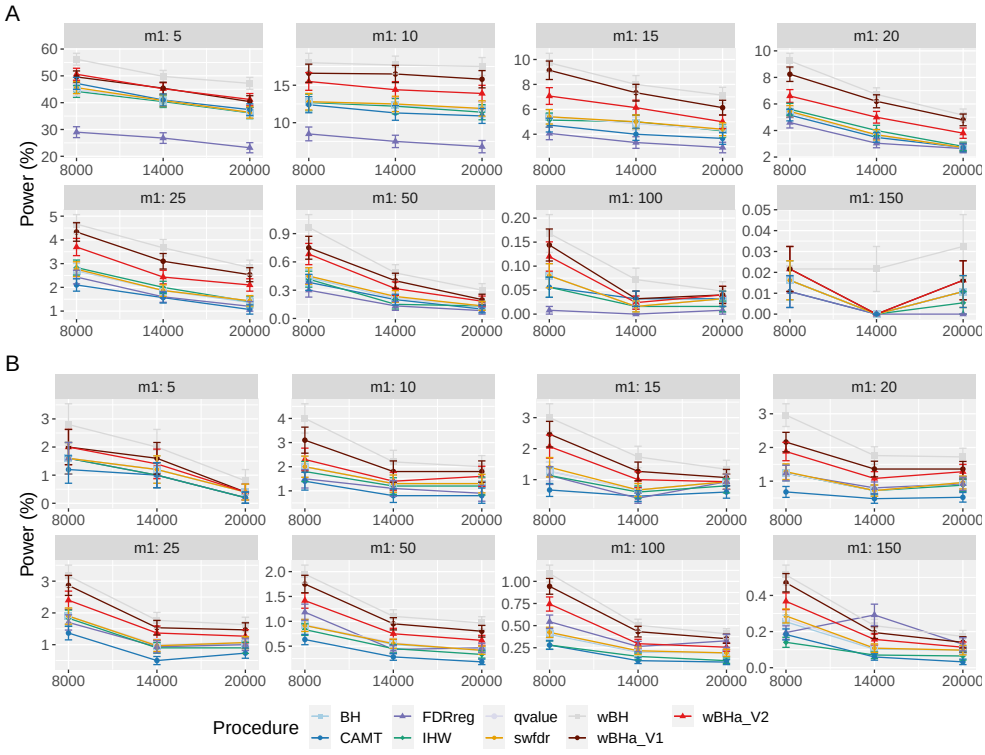

Figure S8 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 3**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

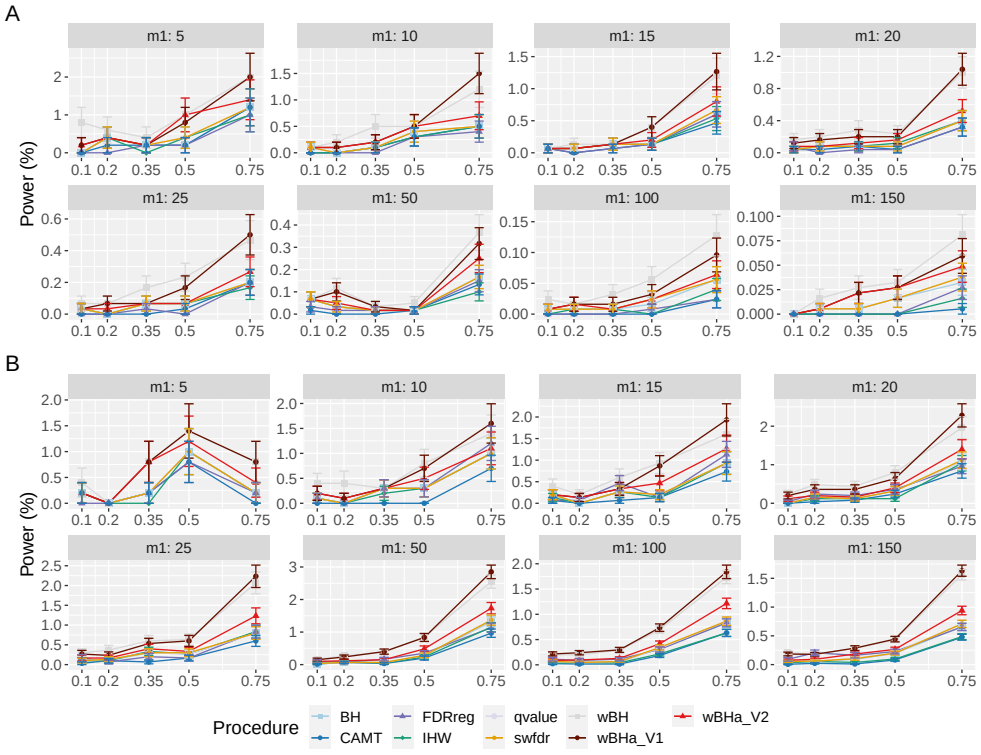

Figure S9 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 2**, avec des marqueurs **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

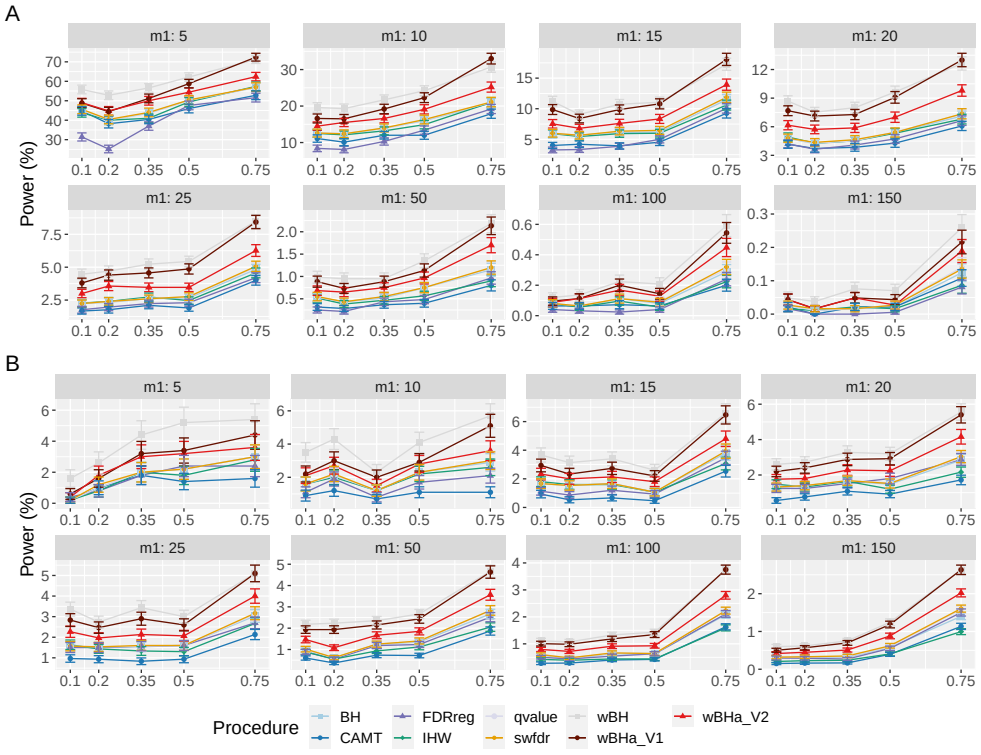

Figure S10 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 3**, avec des marqueurs **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

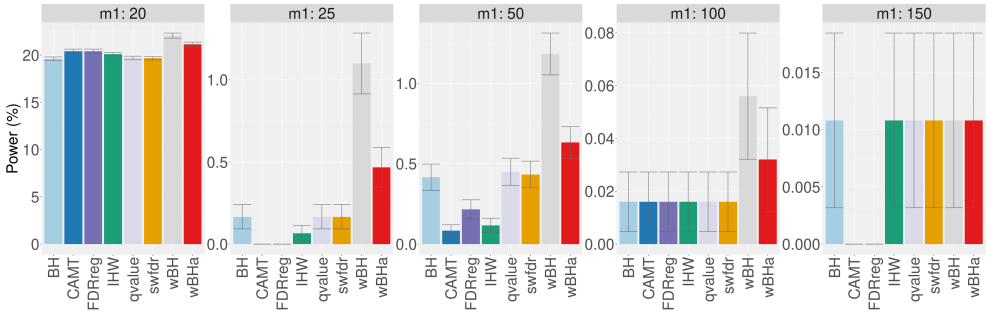

Figure S11 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 2**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

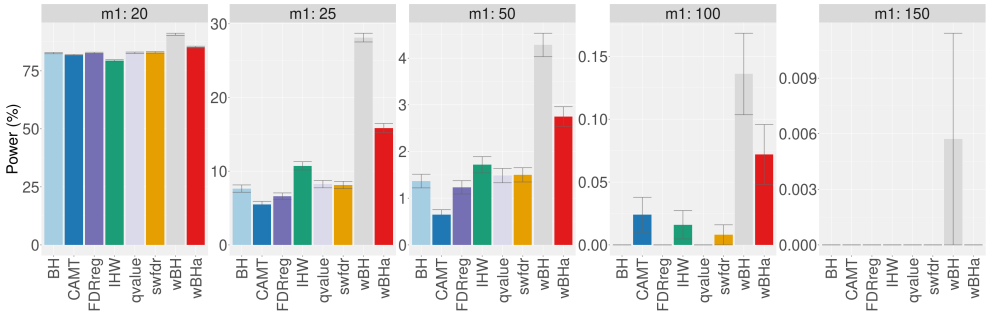

Figure S12 – Comparaison de la puissance dans le sous-groupe des **variants rares** dans le **scénario 3**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

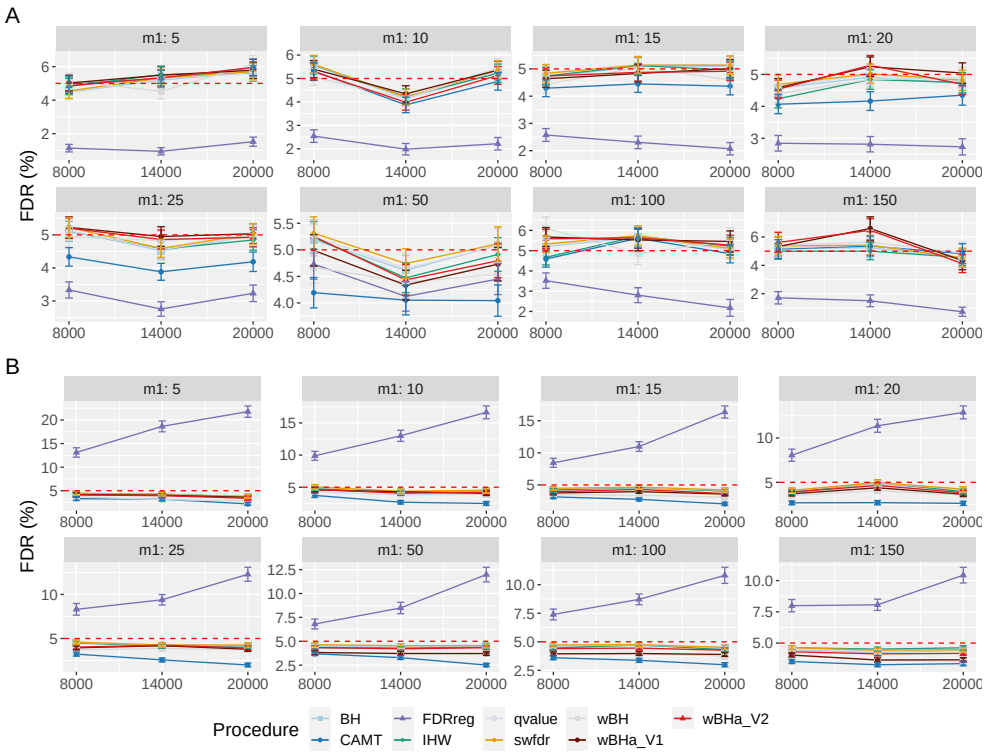

Figure S13 – Comparaison du **FDR** dans le **scénario 2**, avec des variants **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. La ligne pointillée rouge correspond au niveau cible de FDR (5 %). Les barres verticales illustrent les erreurs standard.

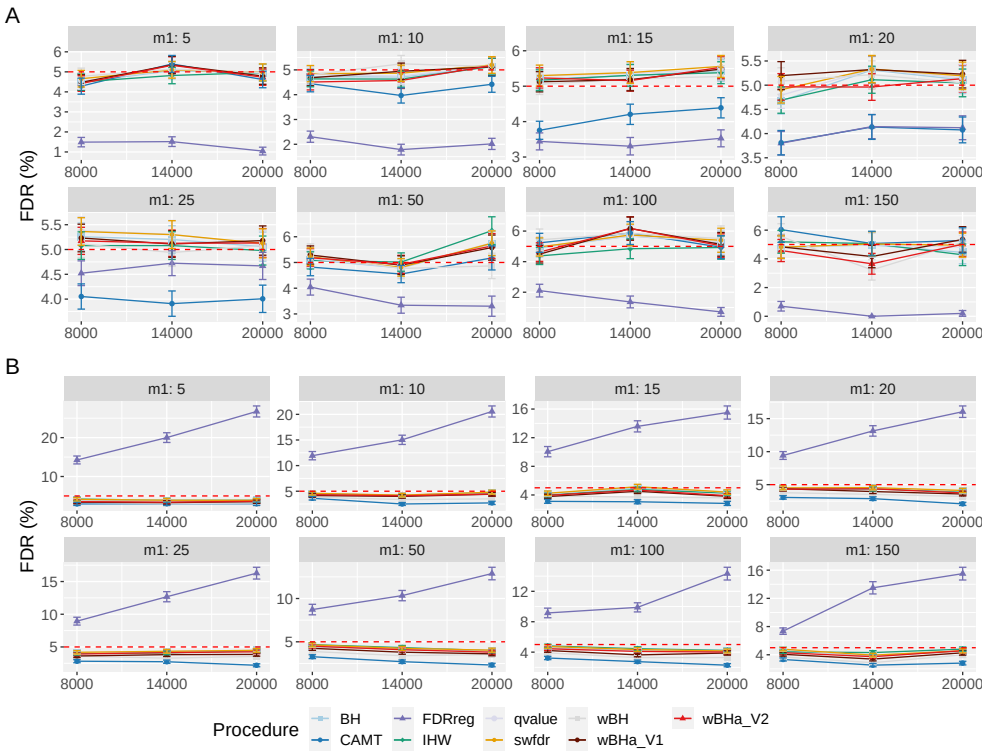

Figure S14 – Comparaison du **FDR** dans le **scénario 3**, avec des variants **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. La ligne pointillée rouge correspond au niveau cible de FDR (5 %). Les barres verticales illustrent les erreurs standard.

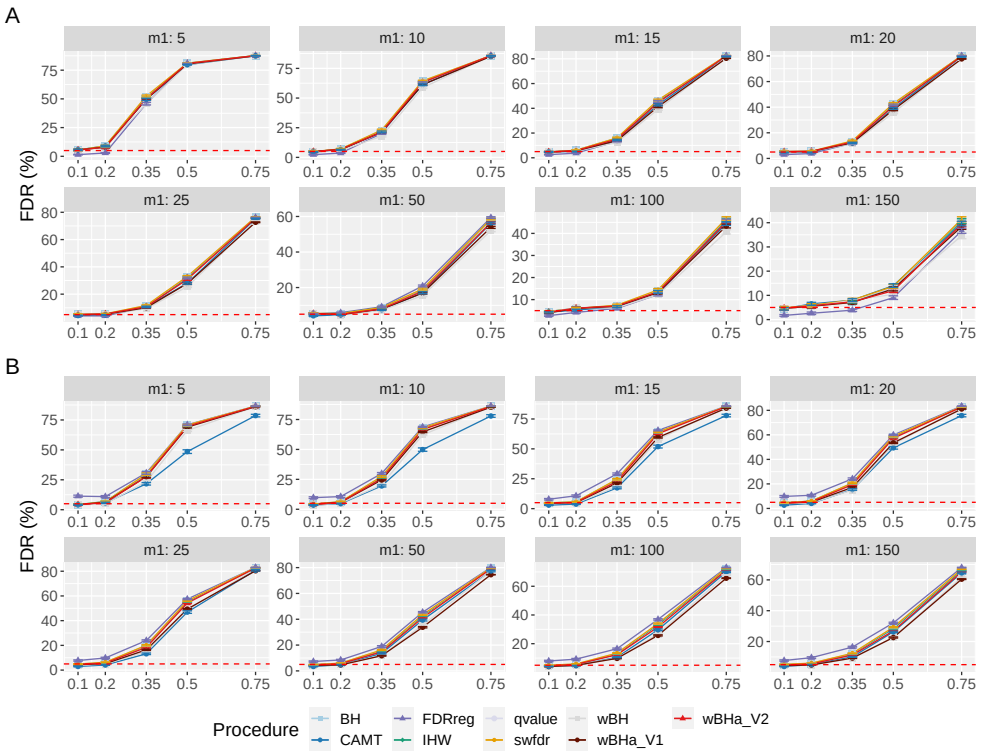

Figure S15 – Comparaison du **FDR** dans le **scénario 2**, avec des marqueurs **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. La ligne pointillée rouge correspond au niveau cible de FDR (5 %). Les barres verticales illustrent les erreurs standard.

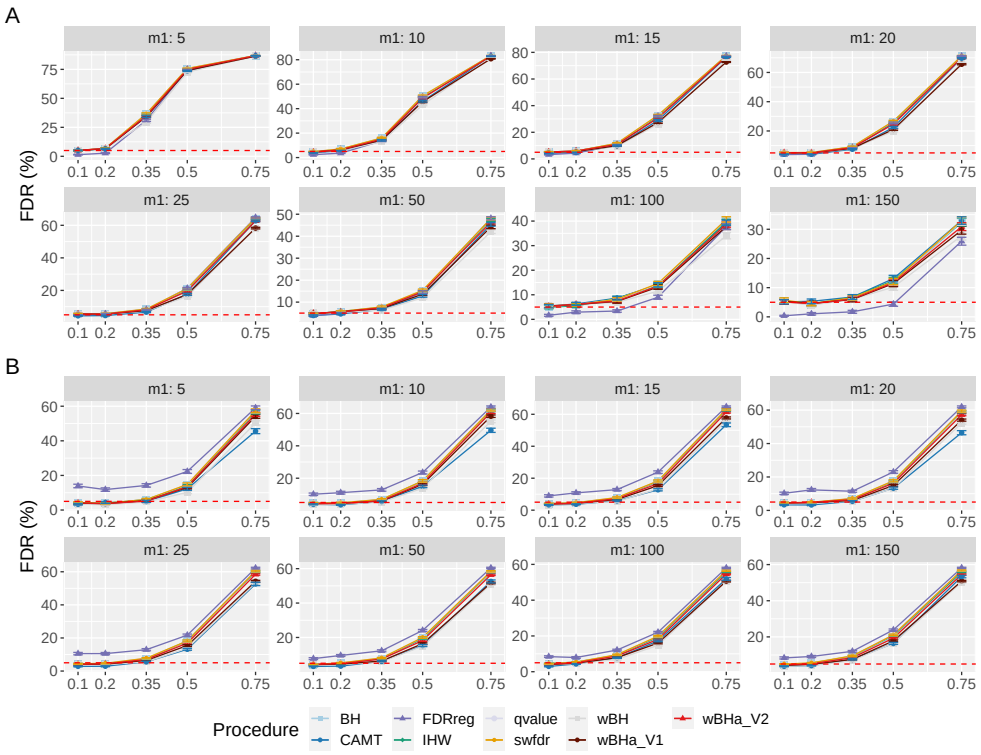

Figure S16 – Comparaison du **FDR** dans le **scénario 3**, avec des marqueurs **corrélés**, pour différentes valeurs de ρ et m_1 avec $m = 8000$. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. La ligne pointillée rouge correspond au niveau cible de FDR (5 %). Les barres verticales illustrent les erreurs standard.

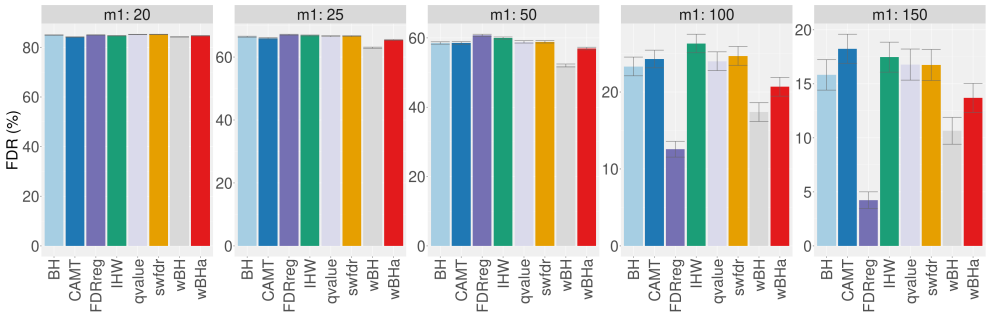

Figure S17 – Comparaison du **FDR** dans le **scénario 2**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

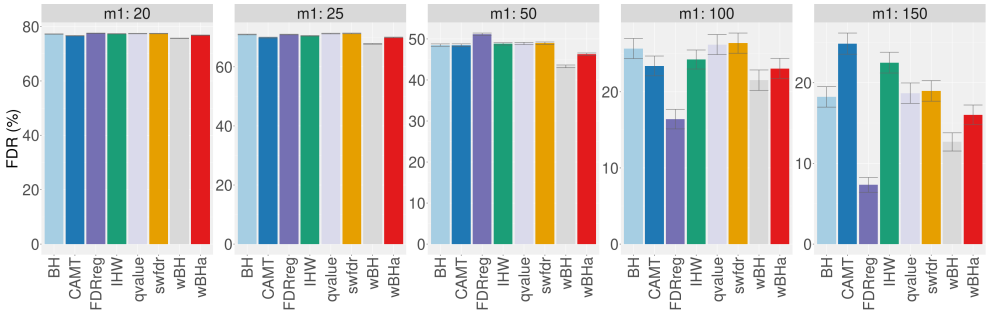

Figure S18 – Comparaison du **FDR** dans le **scénario 3**, avec des simulations basées sur des **données réelles**, pour différentes valeurs de m_1. Les barres verticales illustrent les erreurs standard.

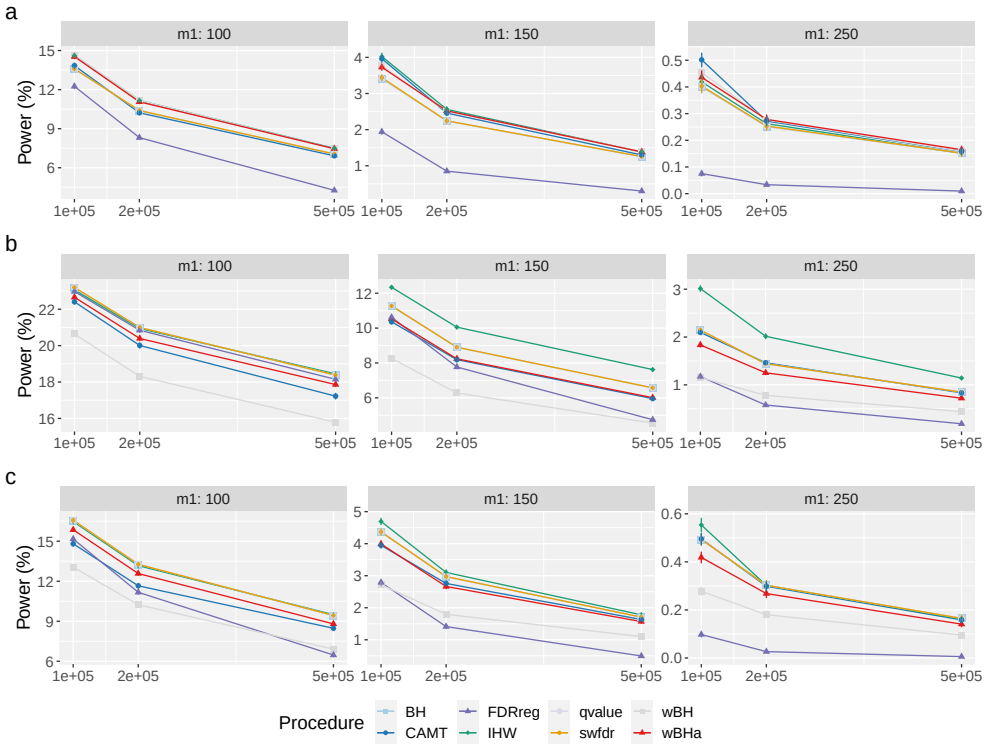

Figure S19 – Comparaison de la **puissance globale** avec des marqueurs indépendants pour de **grandes valeurs de *m*** et m_1 avec des phénotypes quantitatifs. Les panneaux a, b et c présentent les résultats pour le scénario 1, le scénario 2 et le scénario 3, respectivement. Les barres verticales illustrent les erreurs standard.

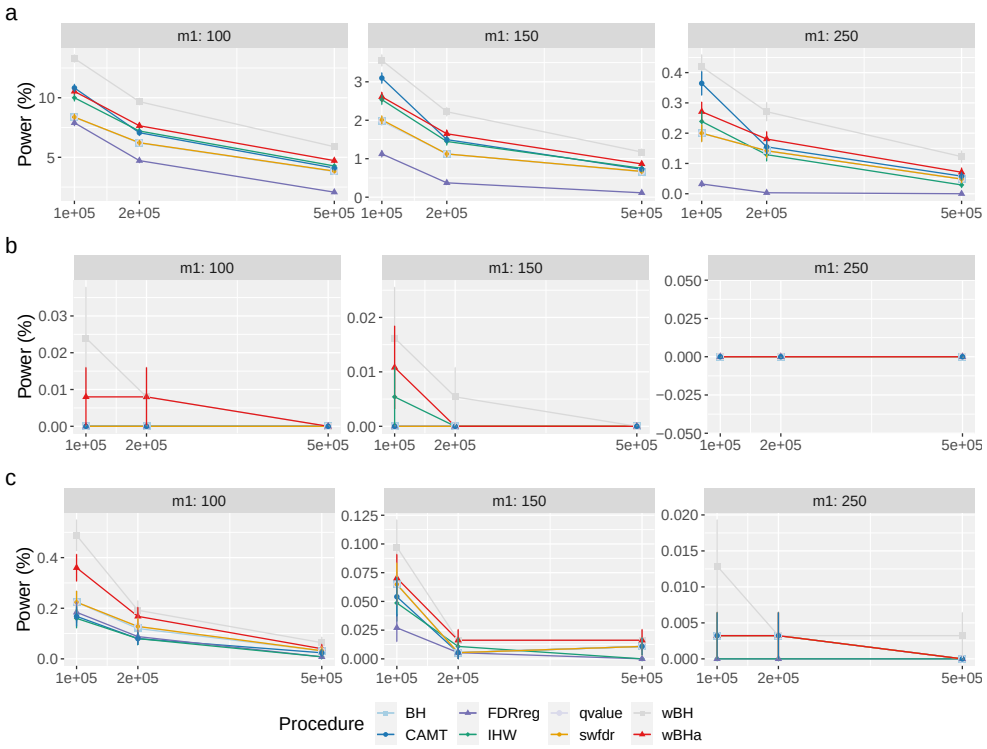

Figure S20 – Comparaison de la puissance dans le sous-groupe de **variants rares** avec des marqueurs indépendants pour de **grandes valeurs de *m*** et m_1 avec des phénotypes quantitatifs. Les panneaux a, b et c présentent les résultats pour le scénario 1, le scénario 2 et le scénario 3, respectivement. Les barres verticales illustrent les erreurs standard.

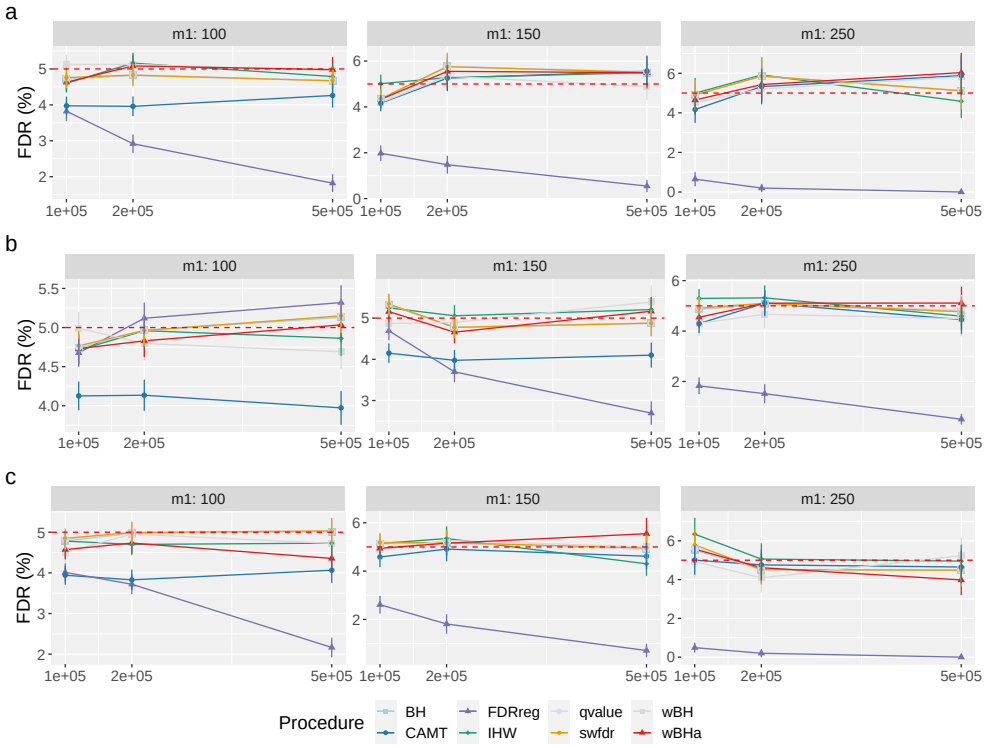

Figure S21 – Comparaison du **FDR** avec des marqueurs indépendants pour de **grandes valeurs de *m*** et m_1 avec des phénotypes quantitatifs. Les panneaux a, b et c présentent les résultats pour le scénario 1, le scénario 2 et le scénario 3, respectivement. Les barres verticales illustrent les erreurs standard.

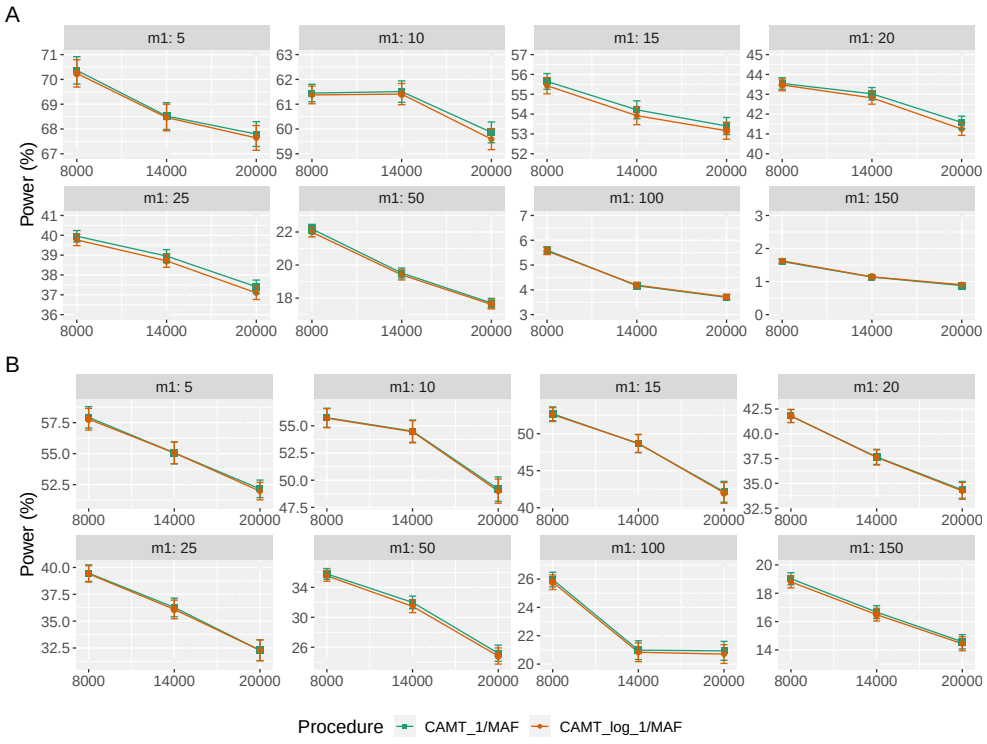

Figure S22 – Comparaison de la **puissance globale** de la procédure CAMT pour différentes covariables dans le **scénario 2**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

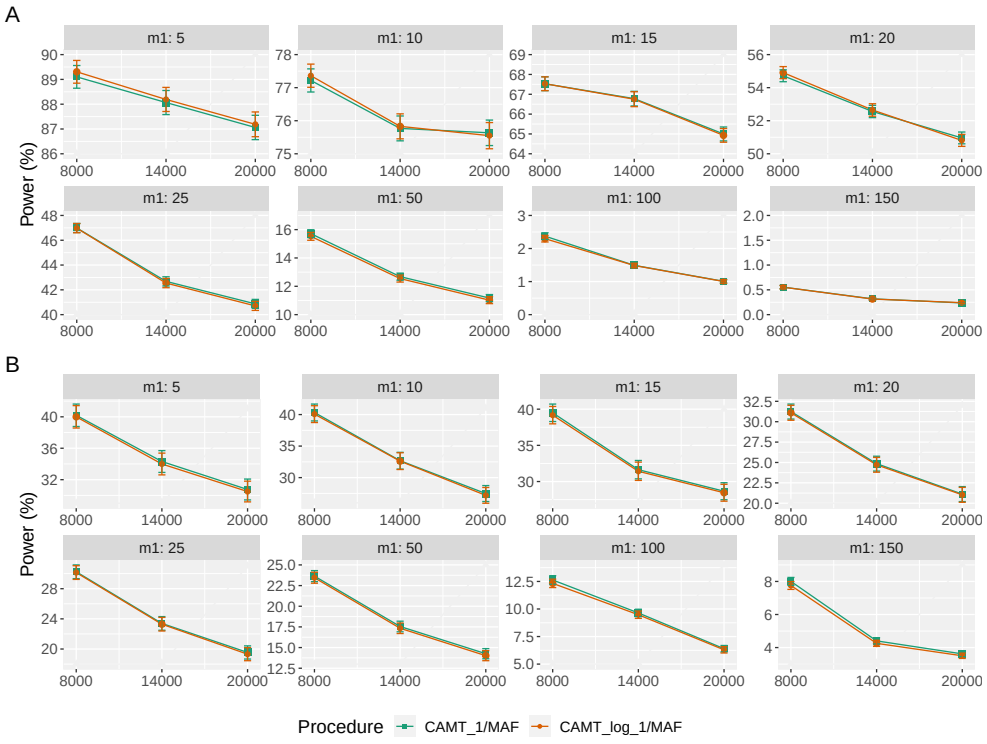

Figure S23 – Comparaison de la **puissance globale** de la procédure CAMT pour différentes covariables dans le **scénario 3**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

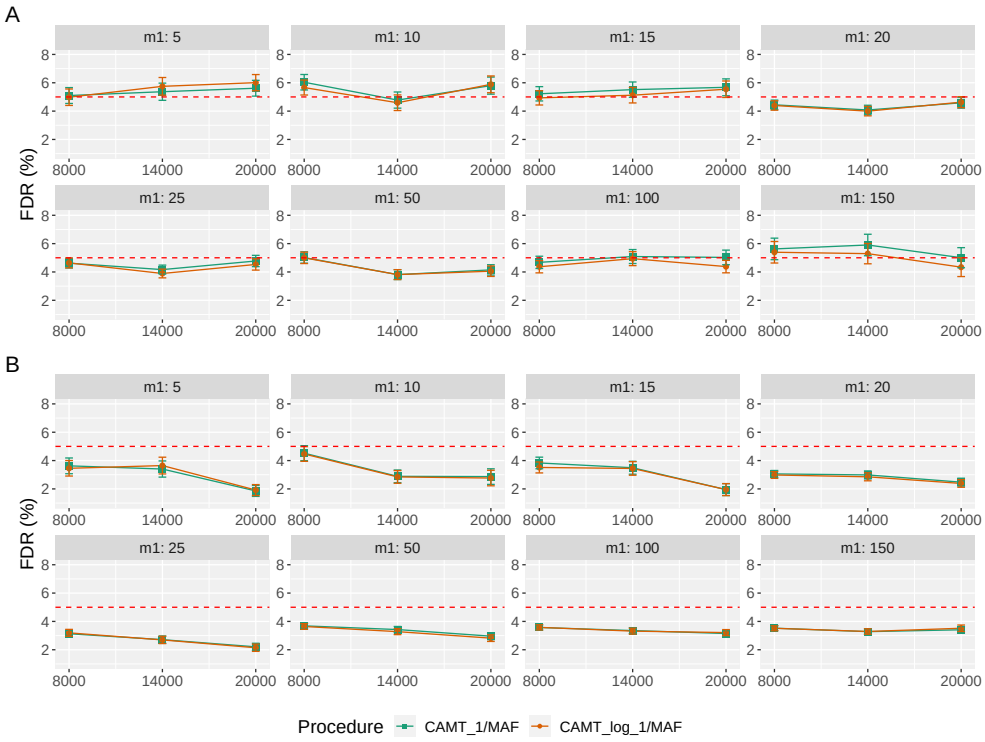

Figure S24 – Comparaison du **FDR** de la procédure CAMT pour différentes covariables dans le **scénario 2**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

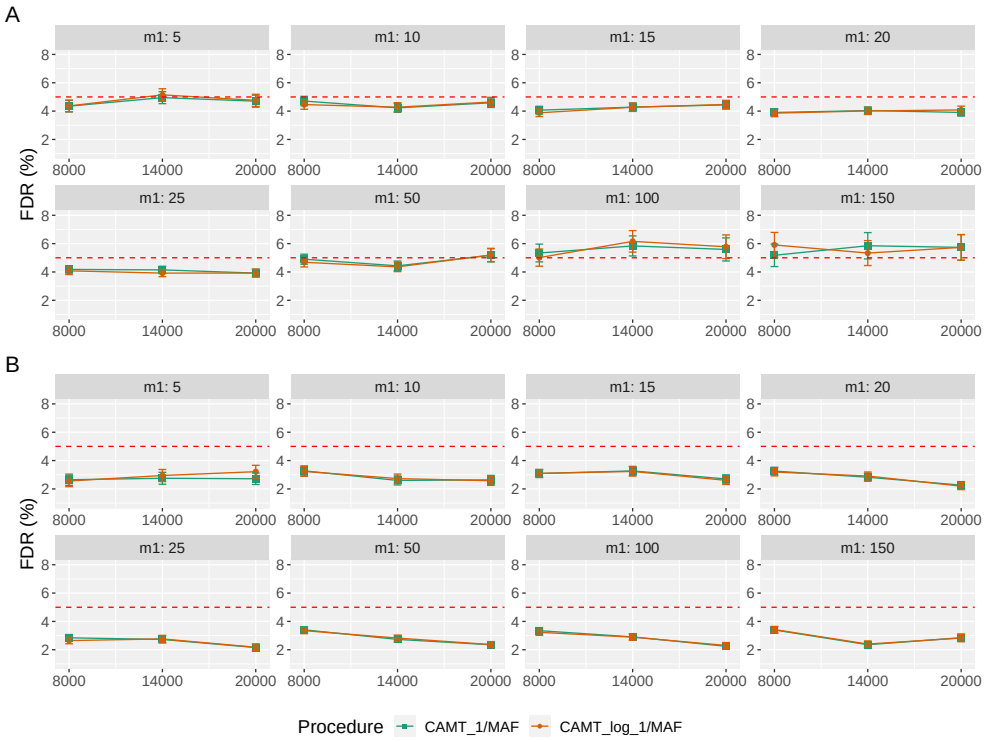

Figure S25 – Comparaison du **FDR** de la procédure CAMT pour différentes covariables dans le **scénario 3**, avec des marqueurs **indépendants** ($\rho = 0$), pour différentes valeurs de m et m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

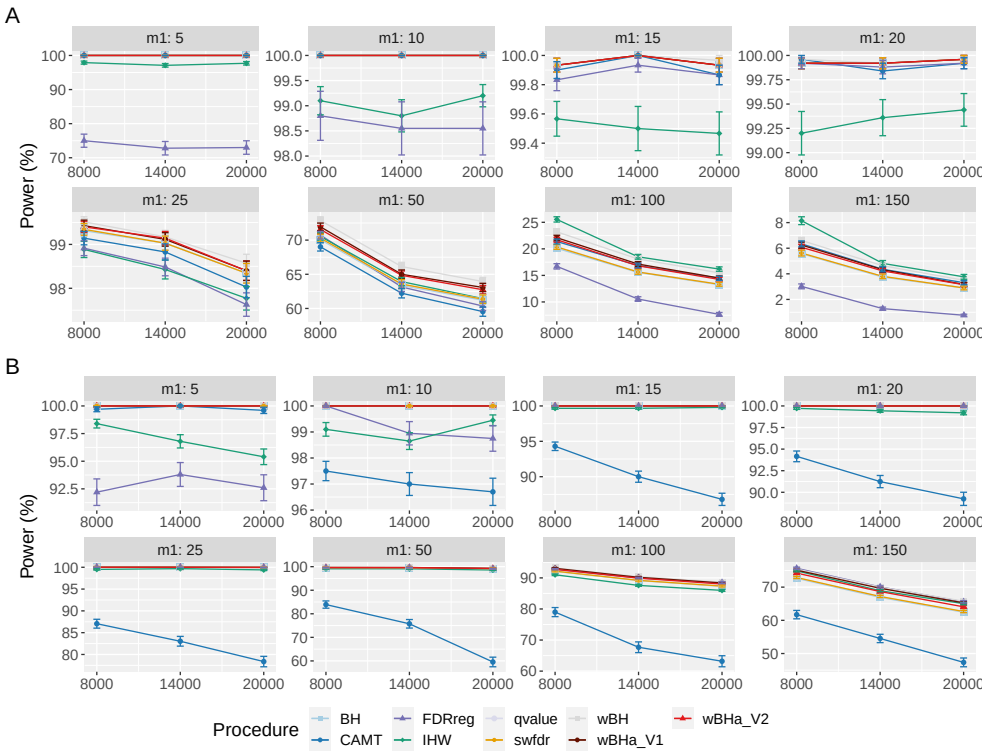

Figure S26 – Comparaison de la puissance dans le sous-groupe de **variants communs** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 2** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

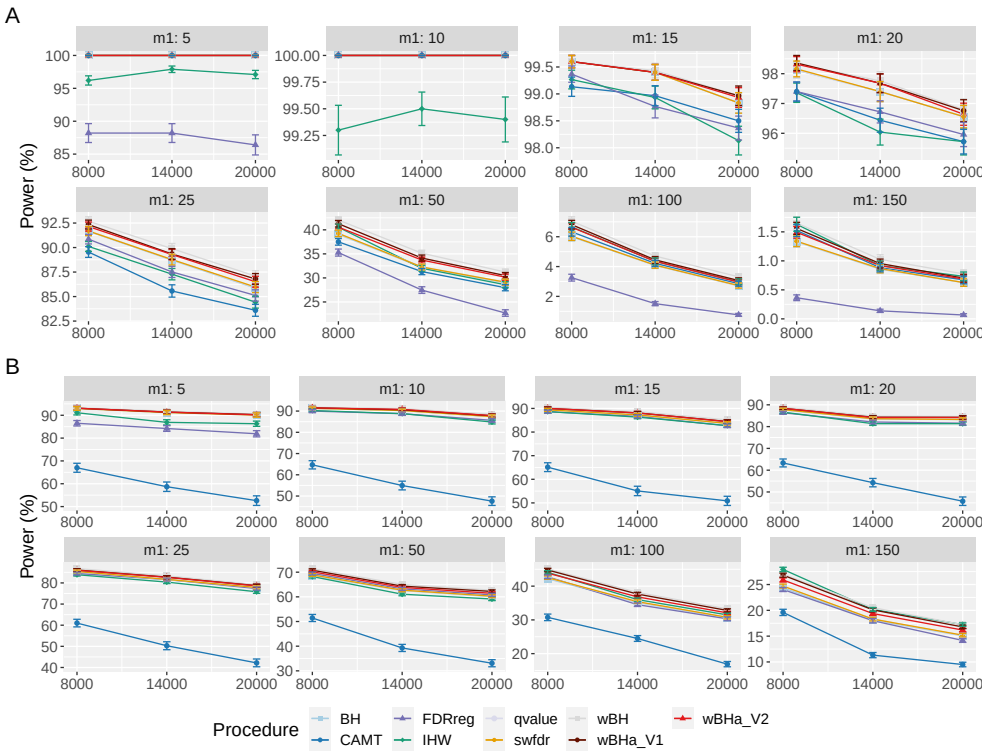

Figure S27 – Comparaison de la puissance dans le sous-groupe de **variants communs** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 3** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

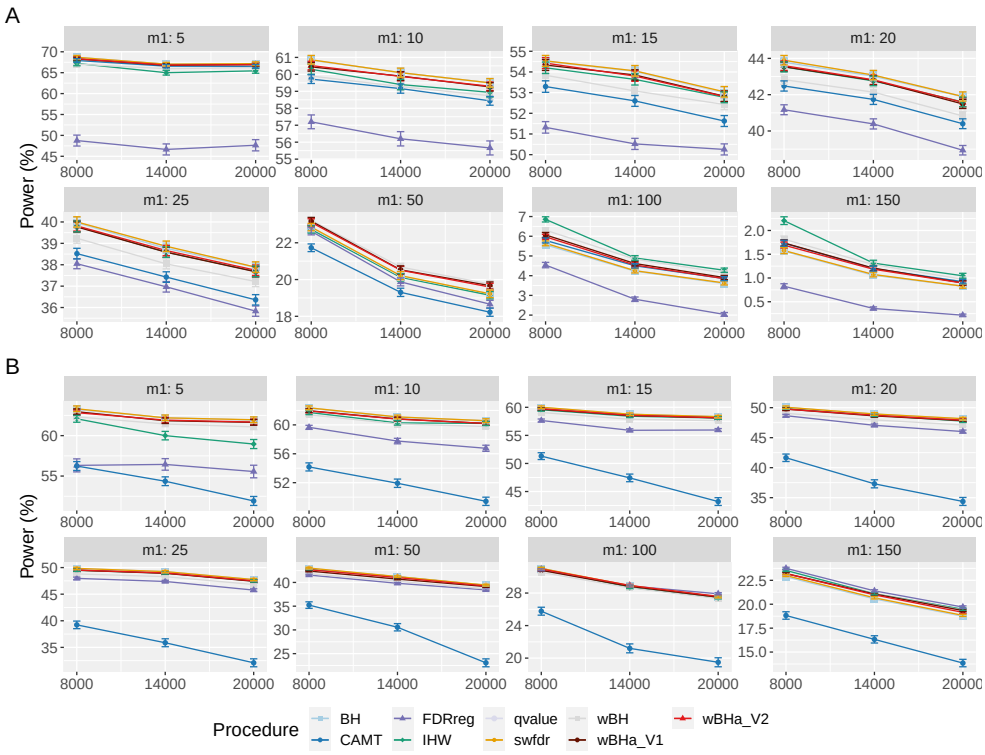

Figure S28 – Comparaison de la **puissance globale** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 2** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

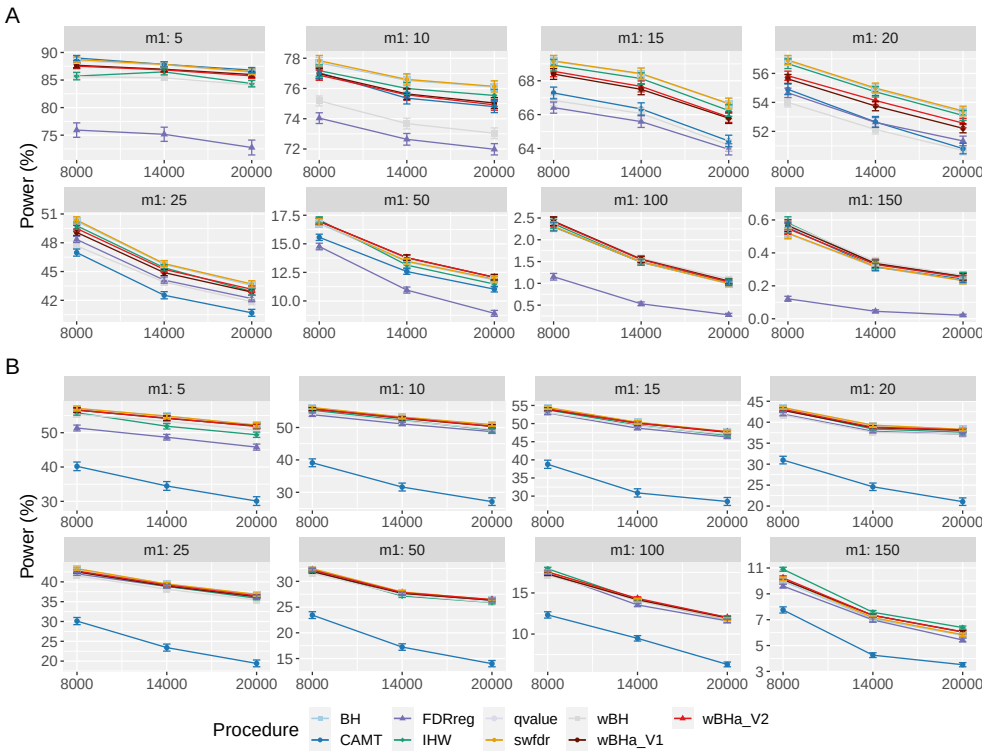

Figure S29 – Comparaison de la **puissance globale** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 3** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

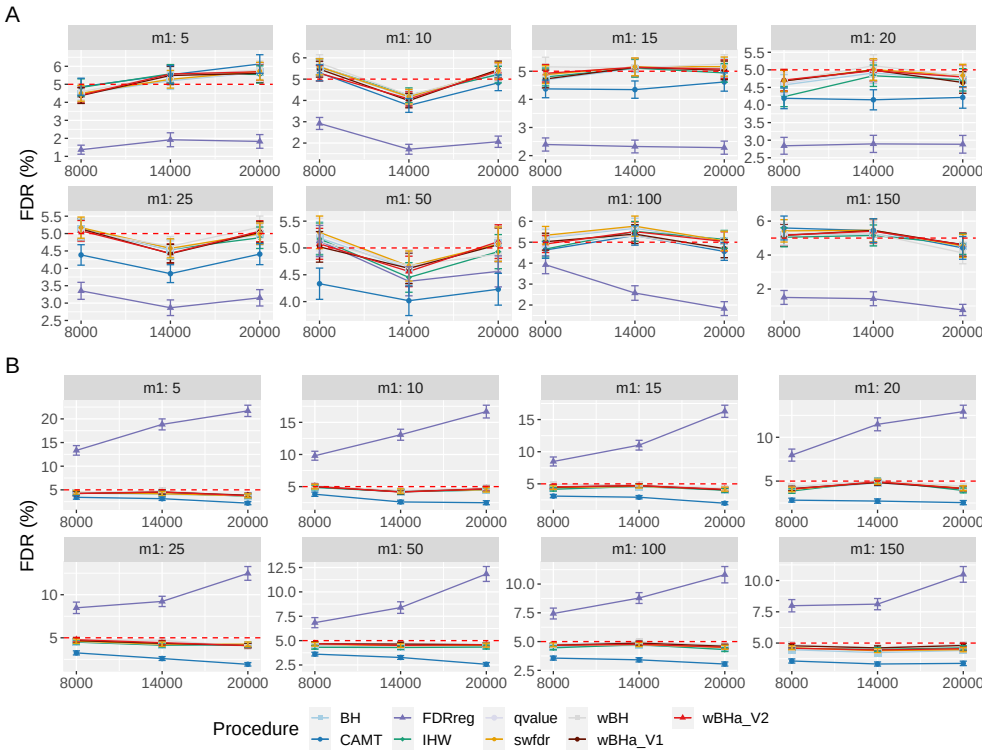

Figure S30 – Comparaison du **FDR** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 2** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

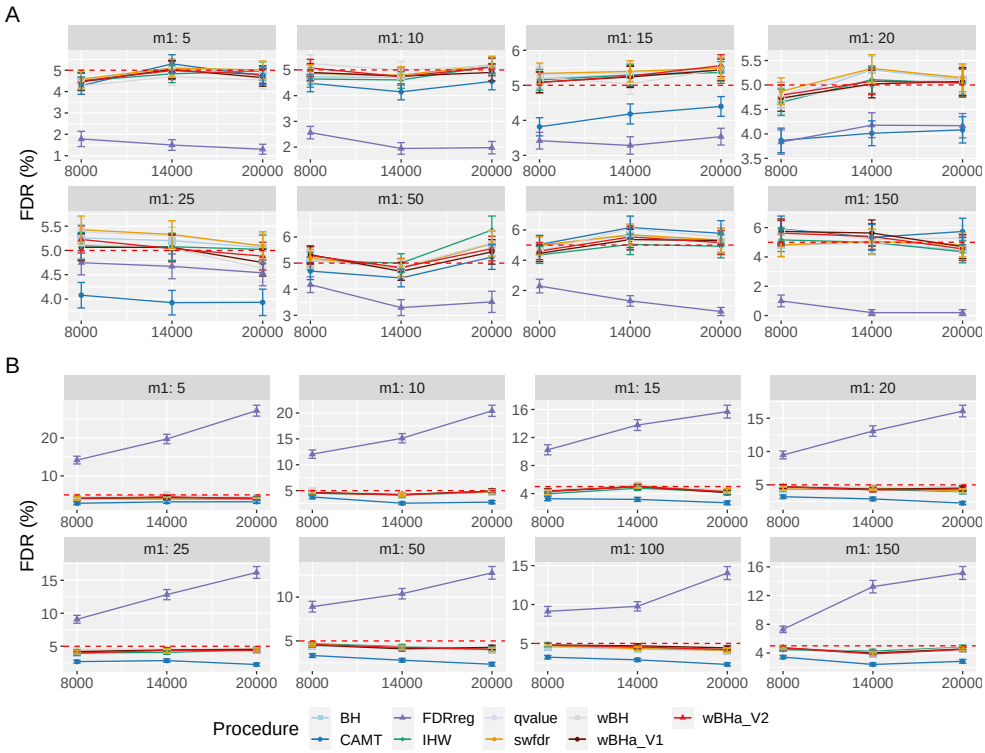

Figure S31 – Comparaison du **FDR** lors de l'utilisation de **1/MAF** comme covariable dans le **scénario 3** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

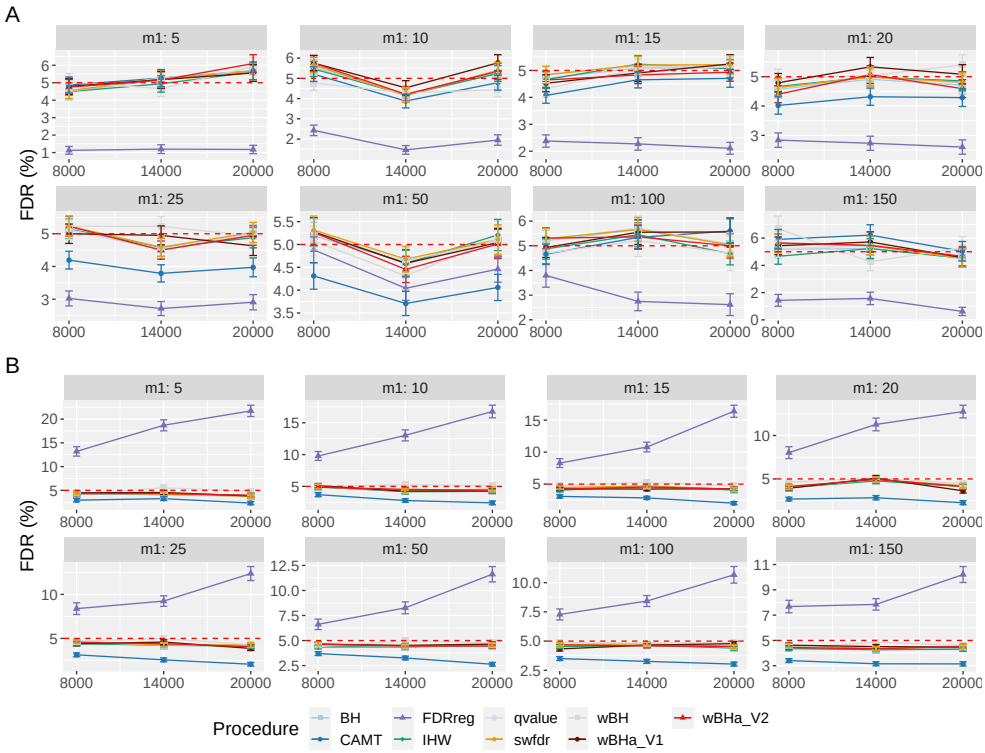

Figure S32 – Comparaison du **FDR** lors de l'utilisation de covariable **non informative** dans le **scénario 2** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

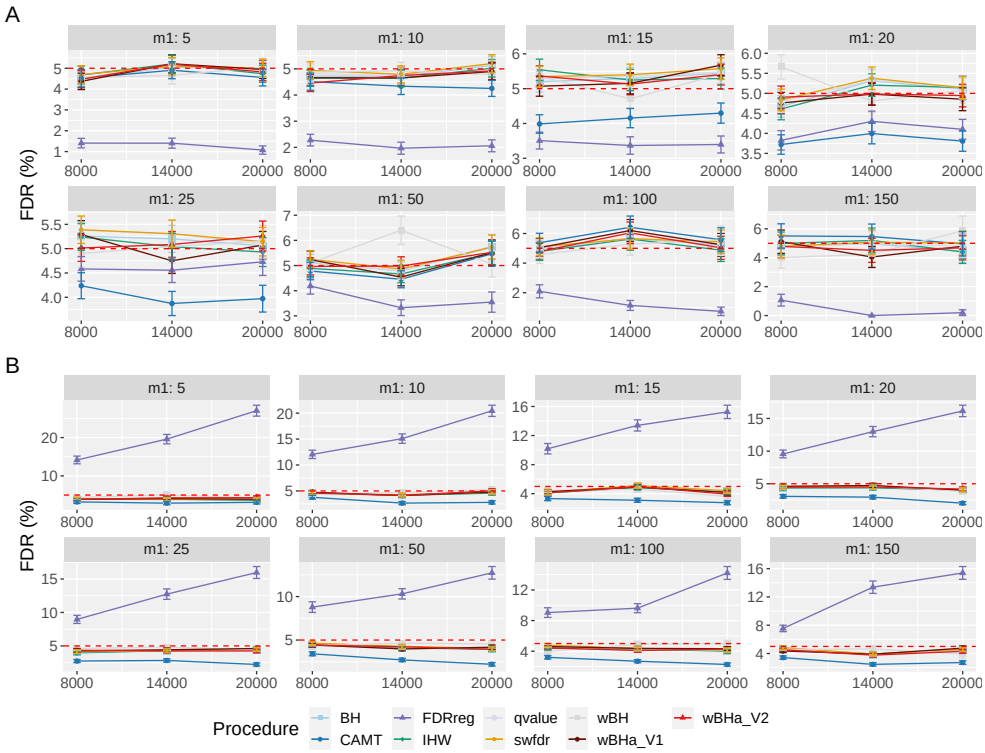

Figure S33 – Comparaison du **FDR** lors de l'utilisation de covariable **non informative** dans le **scénario 3** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement. Les barres verticales illustrent les erreurs standard.

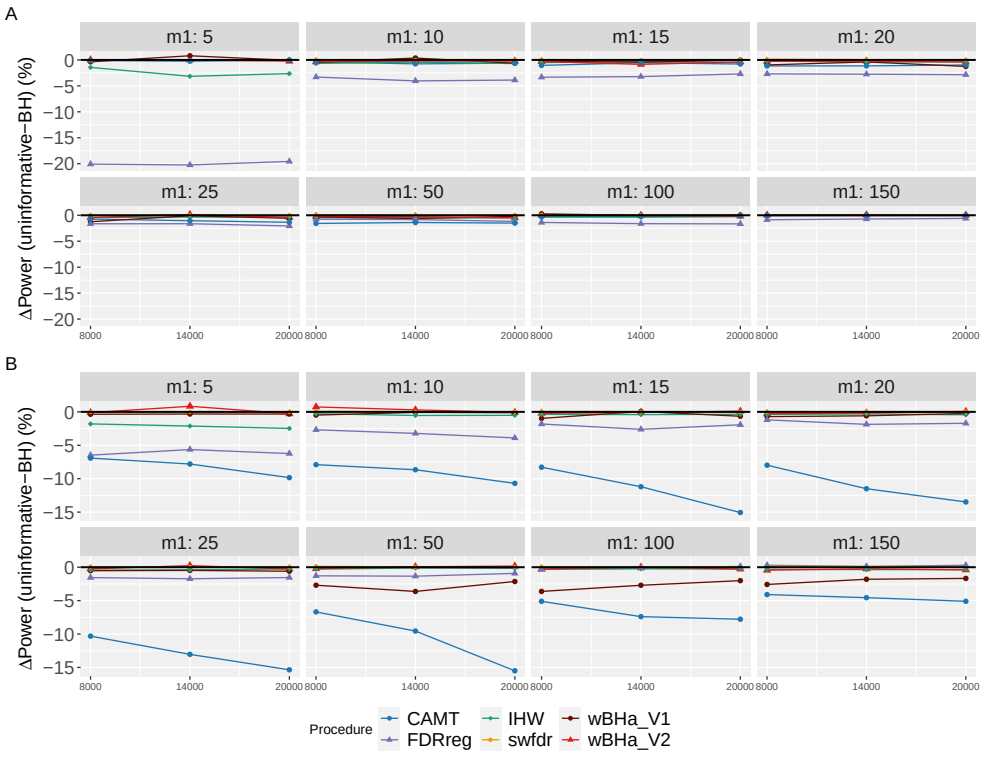

Figure S34 – **Différence de puissance globale** entre l'utilisation de covariable non informative et la procédure BH dans le **scénario 2** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement.

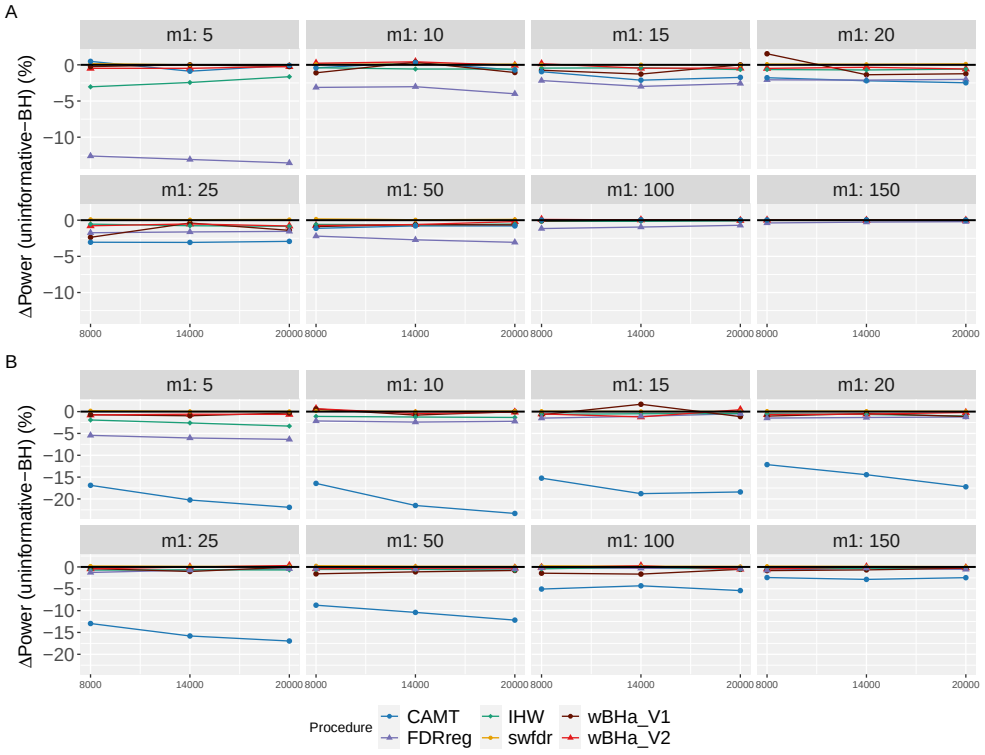

Figure S35 – **Différence de puissance globale** entre l'utilisation de covariable non informative et la procédure BH dans le **scénario 3** avec des marqueurs indépendants pour différentes valeurs de m et de m_1. Les panneaux A et B présentent les résultats pour les phénotypes quantitatifs et binaires, respectivement.

Bibliographie

Abraham, C. and Cho, J. H. (2009). Inflammatory bowel disease. *The New England journal of medicine*, 361 :2066–2078.

Ananthakrishnan, A. N. (2015). Epidemiology and risk factors for ibd. *Nature reviews. Gastroenterology & hepatology*, 12 :205–217.

Auer, P. L. and Lettre, G. (2015). Rare variant association studies : considerations, challenges and opportunities. *Genome medicine*, 7 :Article16.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7 :781–791.

Bandyopadhyay, B., Chanda, V., and Wang, Y. (2017). Finding the sources of missing heritability within rare variants through simulation. *Bioinformatics and biology insights*, 11 :1–5.

Barker, M., Hong, N. P. C., Katz, D. S., Lamprecht, A. L., Martinez-Ortiz, C., Psomopoulos, F., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., and Honeyman, T. (2022). Introducing the fair principles for research software. *Scientific Data*, 9 :622.

Barrett, J. C., Hansoul, S., and et al., D. L. N. (2008). Genome-wide association defines more than thirty distinct susceptibility loci for crohn's disease. *Nature genetics*, 40 :955–962.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society : Series B (Methodological)*, 57 :289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24 :407–418.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93 :491–507.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29 :1165–1188.

Boca, S. M. and Leek, J. T. (2018). A direct approach to estimating false discovery rates conditional on covariates. *PeerJ*, 2018 :e6035.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nature genetics*, 40 :695–701.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche e Commerciali Di Firenze*, 8.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24 :123–140.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review*, 60 :291–319.

Brinster, R., Köttgen, A., Tayo, B. O., Schumacher, M., and Sekula, P. (2018). Control procedures and estimators of the false discovery rate and their application in low-dimensional settings : An empirical investigation. *BMC Bioinformatics*, 19 :1–10.

Brookes, A. J. (1999). The essence of snps. *Gene*, 234 :177–186.

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., and Sabatti, C. (2017). Controlling the rate of gwas false discoveries. *Genetics*, 205 :61–75.

Bush, W. S. and Moore, J. H. (2012). Chapter 11 : Genome-wide association studies. *PLoS computational biology*, 8 :e1002822.

Cadwell, K., Liu, J. Y., Brown, S. L., Miyoshi, H., Loh, J., Lennerz, J. K., Kishi, C., Kc, W., Carrero, J. A., Hunt, S., Stone, C. D., Brunt, E. M., Xavier, R. J., Sleckman, B. P., Li, E., Mizushima, N., Stappenbeck, T. S., and IV, H. W. V. (2008). A unique role for autophagy and atg16l1 in paneth cells in murine and human intestine. *Nature*, 456 :259–263.

Collins, F. S., Brooks, L. D., and Chakravarti, A. (1998). A dna polymorphism discovery resource for research on human genetic variation. *Genome Research*, 8 :1229–1231.

Dalmasso, C., Bar-Hen, A., and Broët, P. (2007). A constrained polynomial regression procedure for estimating the local false discovery rate. *BMC Bioinformatics*, 8 :1–12.

Dalmasso, C., Broët, P., and Moreau, T. (2005). A simple procedure for estimating the false discovery rate. *BIOINFORMATICS ORIGINAL PAPER*, 21 :660–668.

Dalmasso, C., Carpentier, W., Meyer, L., Rouzioux, C., Goujard, C., Chaix, M. L., Lambotte, O., Avettand-Fenoel, V., Clerc, S. L., de Senneville, L. D., Deveau, C., Boufassa, F., Debré, P., Delfraissy, J. F., Broet, P., and Theodorou, I. (2008a). Distinct genetic loci control plasma hiv-rna and cellular hiv-dna levels in hiv-1 infection : the anrs genome wide association 01 study. *PloS one*, 3 :e3907.

Dalmasso, C., Génin, E., and Trégouet, D. A. (2008b). A weighted-holm procedure accounting for allele frequencies in genomewide association studies. *Genetics*, 180 :697–702.

Dehman, A., Ambroise, C., and Neuvial, P. (2015). Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16 :1–14.

Denecker, T. and Toffano-Nioche, C. (2021). FAIR_bioinfo : a turnkey training course and protocol for reproducible computational biology. *JOSE - Journal of Open Source Education*, 4 :68.

Dickhaus, T. (2014). *Simultaneous statistical inference : With applications in the life sciences*, volume 9783642451829. Springer-Verlag Berlin Heidelberg.

Dudbridge, F. and Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genetic epidemiology*, 32 :227–234.

Dudoit, S. and Van der Laan, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer New York.

Duggal, P., Gillanders, E. M., Holmes, T. N., and Bailey-Wilson, J. E. (2008). Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. *BMC Genomics*, 9 :516–518.

Durand, G. (2019). Adaptive p-value weighting with power optimality. *Electronic Journal of Statistics*, 13 :3336–3385.

Efron, B. (1979). Bootstrap methods : Another look at the jackknife. *The Annals of Statistics*, 7 :1–26.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102 :93–103.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1 :54–75.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96 :1151–1160.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., and Nadeau, J. H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics*, 11 :446–450.

Fang, A. M., Lee, A. Y., Kulkarni, M., Osborn, M. P., and Brantley, M. A. (2009). Polymorphisms in the vegfa and vegfr-2 genes and neovascular age-related macular degeneration. *Molecular Vision*, 15 :2710–2719.

Farcomeni, A. (2007). Some results on the control of the false discovery rate under dependence. *Scandinavian Journal of Statistics*, 34 :275–297.

Finner, H. and Gontscharuk, V. (2009). Controlling the familywise error rate with plug-in estimator for the proportion of true null hypotheses. *Journal of the Royal Statistical Society. Series B : Statistical Methodology*, 71 :1031–1048.

Foulkes, A. S. (2009). *Applied Statistical Genetics with : R For Population-based Association Studies*. Springer New York, NY.

Galwey, N. W. (2009). A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33 :559–568.

Gao, X., Starmer, J., and Martin, E. R. (2008). A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, 32 :361–369.

Geisser, S. (1975). The predictive sample reuse method with applications. *Source : Journal of the American Statistical Association*, 70 :320–328.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93 :509–524.

González, S., García, S., Ser, J. D., Rokach, L., and Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning : Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64 :205–237.

Gui, J., Tosteson, T. D., and Borsuk, M. (2012). Weighted multiple testing procedures for genomic studies. *BioData Mining*, 5 :Article4.

Guo, W. and Sarkar, S. (2020). Adaptive controls of fwer and fdr under block dependence. *Journal of Statistical Planning and Inference*, 208 :13–24.

Györffy, B., Gyorffy, A., and Tulassay, Z. (2005). The problem of multiple testing and solutions for genome-wide studies. *Orvosi Hetilap*, 146 :559–563.

Hagger, M. S. (2022). Developing an open science 'mindset'. *Health Psychology and Behavioral Medicine*, 10 :21.

HapMap, C. I., Altshuler, D. M., Gibbs, R. A., and et al., L. P. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467 :52–58.

HapMap, C. I., Belmont, J. W., Boudreau, A., and et al., S. M. L. (2005). A haplotype map of the human genome. *Nature*, 437 :1299–1320.

HapMap, C. I., Frazer, K. A., Ballinger, D. G., and et al., D. R. C. (2007). A second generation human haplotype map of over 3.1 million snps. *Nature*, 449 :851–861.

Hill, W. G. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and applied genetics.*, 38 :226–231.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75 :800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6 :65–70.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, 105 :1215–1227.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Methods*, 13 :577–580.

Janssens, A. C. J., Moonesinghe, R., Yang, Q., Steyerberg, E. W., Duijn, C. M. V., and Khoury, M. J. (2007). The impact of genotype frequencies on the clinical validity of genomic profiling for predicting common chronic diseases. *Genetics in medicine : official journal of the American College of Medical Genetics*, 9 :528–535.

Jung, Y. (2018). Multiple predicting k-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, 30 :197–215.

Kang, G., Ye, K., Liu, N., Allison, D. B., and Gao, G. (2009). Weighted multiple hypothesis testing procedures. *Statistical applications in genetics and molecular biology*, 8 :Article23.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, pages 1137–1145. Morgan Kaufmann Publishers Inc.

Korte, A. and Farlow, A. (2013). The advantages and limitations of trait analysis with gwas : A review. *Plant Methods*, 9 :1–9.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20 :1–21.

Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis : Study designs and statistical tests. *American Journal of Human Genetics*, 95 :5–23.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics (Oxford, England)*, 13 :762–775.

Lei, L. and Fithian, W. (2018). Adapt : an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 80 :649–679.

Lettre, G., Lange, C., and Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genetic Epidemiology*, 31 :358–362.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations ; heterotic models. *Genetics*, 49 :49–67.

Li, A. and Barber, R. F. (2018). Multiple testing with the structure-adaptive benjamini–hochberg algorithm. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 81 :45–74.

Li, B. and 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 :311–321.

Li, M.-X., Yeung, J. M. Y., Cherny, S. S., and Sham, P. C. (2012). Evaluating the effective numbers of independent tests and significant p-value

thresholds in commercial genotyping arrays and public imputation reference datasets. *Human Genetics*, 131 :747–756.

Li, Z., Li, X., Liu, Y., Shen, J., Chen, H., Zhou, H., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *American journal of human genetics*, 104 :802–814.

Liang, K. and Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 74 :163–182.

Liu, D. J. and Leal, S. M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genetics*, 6 :1–14.

Liu, T. C., Kern, J. T., VanDussen, K. L., Xiong, S., Kaiko, G. E., Wilen, C. B., Rajala, M. W., Caruso, R., Holtzman, M. J., Gao, F., McGovern, D. P., Nunez, G., Head, R. D., and Stappenbeck, T. S. (2018). Interaction between smoking and atg16l1t300a triggers paneth cell defects in crohn's disease. *The Journal of clinical investigation*, 128 :5110–5122.

Liu, T. C., Naito, T., Liu, Z., Vandussen, K. L., Haritunians, T., Li, D., Endo, K., Kawai, Y., Nagasaki, M., Kinouchi, Y., McGovern, D. P., Shimosegawa, T., Kakuta, Y., and Stappenbeck, T. S. (2017). Lrrk2 but not atg16l1 is associated with paneth cell defect in japanese crohn's disease patients. *JCI Insight*, 2 :e91917.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5 :e1000384.

Maher, B. (2008). Personal genomes : The case of the missing heritability. *Nature*, 456 :18–21.

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., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., MacKay, T. F., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461 :747–753.

Marouli, E., Graff, M., Medina-Gomez, C., Lo, K. S., Wood, A. R., Kjaer, T. R., Fine, R. S., Lu, Y., Schurmann, C., Highland, H. M., Rüeger, S.,

Thorleifsson, G., Justice, A. E., Lamparter, D., Stirrups, K. E., Turcot, V., Young, K. L., Winkler, T. W., Esko, T., Karaderi, T., Locke, A. E., Masca, N. G., Ng, M. C., Mudgal, P., Rivas, M. A., Vedantam, S., Mahajan, A., Guo, X., Abecasis, G., Aben, K. K., Adair, L. S., Alam, D. S., Albrecht, E., Allin, K. H., Allison, M., Amouyel, P., Appel, E. V., Arveiler, D., Asselbergs, F. W., Auer, P. L., Balkau, B., Banas, B., Bang, L. E., Benn, M., Bergmann, S., Bielak, L. F., Blüher, M., Boeing, H., Boerwinkle, E., Böger, C. A., Bonnycastle, L. L., Bork-Jensen, J., Bots, M. L., Bottinger, E. P., Bowden, D. W., Brandslund, I., Breen, G., Brilliant, M. H., Broer, L., Burt, A. A., Butterworth, A. S., Carey, D. J., Caulfield, M. J., Chambers, J. C., Chasman, D. I., Chen, Y. D. I., Chowdhury, R., Christensen, C., Chu, A. Y., Cocca, M., Collins, F. S., Cook, J. P., Corley, J., Galbany, J. C., Cox, A. J., Cuellar-Partida, G., Danesh, J., Davies, G., Bakker, P. I. D., Borst, G. J. D., Denus, S. D., Groot, M. C. D., Mutsert, R. D., Deary, I. J., Dedoussis, G., Demerath, E. W., Hollander, A. I. D., Dennis, J. G., Angelantonio, E. D., Drenos, F., Du, M., Dunning, A. M., Easton, D. F., Ebeling, T., Edwards, T. L., Ellinor, P. T., Elliott, P., Evangelou, E., Farmaki, A. E., Faul, J. D., Feitosa, M. F., Feng, S., Ferrannini, E., Ferrario, M. M., Ferrieres, J., Florez, J. C., Ford, I., Fornage, M., Franks, P. W., Frikke-Schmidt, R., Galesloot, T. E., Gan, W., Gandin, I., Gasparini, P., Giedraitis, V., Giri, A., Girotto, G., Gordon, S. D., Gordon-Larsen, P., Gorski, M., Grarup, N., Grove, M. L., Gudnason, V., Gustafsson, S., Hansen, T., Harris, K. M., Harris, T. B., Hattersley, A. T., Hayward, C., He, L., Heid, I. M., Heikkilä, K., Øyvind Helgeland, Hernesniemi, J., Hewitt, A. W., Hocking, L. J., Hollensted, M., Holmen, O. L., Hovingh, G. K., Howson, J. M., Hoyng, C. B., Huang, P. L., Hveem, K., Ikram, M. A., Ingelsson, E., Jackson, A. U., Jansson, J. H., Jarvik, G. P., Jensen, G. B., Jhun, M. A., Jia, Y., Jiang, X., Johansson, S., Jørgensen, M. E., Jørgensen, T., Jousilahti, P., Jukema, J. W., Kahali, B., Kahn, R. S., Kähönen, M., Kamstrup, P. R., Kanoni, S., Kaprio, J., Karaleftheri, M., Kardia, S. L., Karpe, F., Kee, F., Keeman, R., Kiemeney, L. A., Kitajima, H., Kluivers, K. B., Kocher, T., Komulainen, P., Kontto, J., Kooner, J. S., Kooperberg, C., Kovacs, P., Kriebel, J., Kuivaniemi, H., Küry, S., Kuusisto, J., Bianca, M. L., Laakso, M., Lakka, T. A., Lange, E. M., Lange, L. A., Langefeld, C. D., Langenberg, C., Larson, E. B., Lee, I. T., Lehtimäki, T., Lewis, C. E., Li, H., Li, J., Li-Gao, R., Lin, H., Lin, L. A., Lin, X., Lind, L., Lindström, J., Linneberg, A., Liu, Y., Liu, Y., Lophatananon, A., Luan, J., Lubitz, S. A., Lyytikäinen, L. P., MacKey, D. A., Madden, P. A., Manning, A. K., Männistö, S., Marenne, G., Marten, J., Martin, N. G., Mazul, A. L., Meidtner, K., Metspalu, A., Mitchell, P., Mohlke, K. L., Mook-Kanamori, D. O., Morgan, A., Morris, A. D., Morris, A. P., Müller-Nurasyid, M., Munroe, P. B., Nalls, M. A., Nauck, M., Nelson, C. P., Neville, M., Niel-

sen, S. F., Nikus, K., Njølstad, P. R., Nordestgaard, B. G., Ntalla, I., O'Connel, J. R., Oksa, H., Loohuis, L. M., Ophoff, R. A., Owen, K. R., Packard, C. J., Padmanabhan, S., Palmer, C. N., Pasterkamp, G., Patel, A. P., Pattie, A., Pedersen, O., Peissig, P. L., Peloso, G. M., Pennell, C. E., Perola, M., Perry, J. A., Perry, J. R., Person, T. N., Pirie, A., Polasek, O., Posthuma, D., Raitakari, O. T., Rasheed, A., Rauramaa, R., Reilly, D. F., Reiner, A. P., Renström, F., Ridker, P. M., Rioux, J. D., Robertson, N., Robino, A., Rolandsson, O., Rudan, I., Ruth, K. S., Saleheen, D., Salomaa, V., Samani, N. J., Sandow, K., Sapkota, Y., Sattar, N., Schmidt, M. K., Schreiner, P. J., Schulze, M. B., Scott, R. A., Segura-Lepe, M. P., Shah, S., Sim, X., Sivapalaratnam, S., Small, K. S., Smith, A. V., Smith, J. A., Southam, L., Spector, T. D., Speliotes, E. K., Starr, J. M., Steinthorsdottir, V., Stringham, H. M., Stumvoll, M., Surendran, P., Hart't, L. M., Tansey, K. E., Tardif, J. C., Taylor, K. D., Teumer, A., Thompson, D. J., Thorsteinsdottir, U., Thuesen, B. H., Tönjes, A., Tromp, G., Trompet, S., Tsafantakis, E., Tuomilehto, J., Tybjaerg-Hansen, A., Tyrer, J. P., Uher, R., Uitterlinden, A. G., Ulivi, S., Laan, S. W. V. D., Leij, A. R. V. D., Duijn, C. M. V., Schoor, N. M. V., Setten, J. V., Varbo, A., Varga, T. V., Varma, R., Edwards, D. R. V., Vermeulen, S. H., Vestergaard, H., Vitart, V., Vogt, T. F., Vozzi, D., Walker, M., Wang, F., Wang, C. A., Wang, S., Wang, Y., Wareham, N. J., Warren, H. R., Wessel, J., Willems, S. M., Wilson, J. G., Witte, D. R., Woods, M. O., Wu, Y., Yaghootkar, H., Yao, J., Yao, P., Yerges-Armstrong, L. M., Young, R., Zeggini, E., Zhan, X., Zhang, W., Zhao, J. H., Zhao, W., Zheng, H., Zhou, W., Rotter, J. I., Boehnke, M., Kathiresan, S., McCarthy, M. I., Willer, C. J., Stefansson, K., Borecki, I. B., Liu, D. J., North, K. E., Heard-Costa, N. L., Pers, T. H., Lindgren, C. M., Oxvig, C., Kutalik, Z., Rivadeneira, F., Loos, R. J., Frayling, T. M., Hirschhorn, J. N., Deloukas, P., and Lettre, G. (2017). Rare and low-frequency coding variants alter human adult height. *Nature*, 542 :186–190.

Martínez-Muñoz, G. and Suárez, A. (2010). Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition*, 43 :143–152.

Matsuo, H., Yamamoto, K., Nakaoka, H., Nakayama, A., Sakiyama, M., Chiba, T., Takahashi, A., Nakamura, T., Nakashima, H., Takada, Y., Danjoh, I., Shimizu, S., Abe, J., Kawamura, Y., Terashige, S., Ogata, H., Tatsukawa, S., Yin, G., Okada, R., Morita, E., Naito, M., Tokumasu, A., Onoue, H., Iwaya, K., Ito, T., Takada, T., Inoue, K., Kato, Y., Nakamura, Y., Sakurai, Y., Suzuki, H., Kanai, Y., Hosoya, T., Hamajima, N., Inoue, I., Kubo, M., Ichida, K., Ooyama, H., Shimizu, T., and Shinomiya, N. (2016). Genome-wide association study of clinically defined gout iden-

tifies multiple risk loci and its association with clinical subtypes. *Annals of the Rheumatic Diseases*, 75 :652–659.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits : consensus, uncertainty and challenges. *Nature reviews. Genetics*, 9 :356–369.

Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases : a Cohort Allelic Sums Test (CAST). *Mutation research*, 615 :28–56.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7 :e1001322.

Neumann, J. (2022). Fair data infrastructure. *Advances in biochemical engineering/biotechnology*, 182 :195–207.

Neuvial, P. (2008). Asymptotic properties of false discovery rate controlling procedures under independence. *Electronic Journal of Statistics*, 2 :1065–1110.

Newton, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankhya : The Indian Journal of Statistics, Series A*, 64 :306–322.

Otani, T., Noma, H., Nishino, J., and Matsui, S. (2018). Re-assessment of multiple testing strategies for more efficient genome-wide association studies. *European Journal of Human Genetics*, 26 :1038–1048.

Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, 12 :465–474.

Owen, A. B. (2005). Variance of the number of false discoveries. *Journal of the Royal Statistical Society. Series B : Statistical Methodology*, 67 :411–426.

Panagiotou, O. A., Evangelou, E., and Ioannidis, J. P. (2010). Genome-wide significant associations for variants with minor allele frequency of 5% or less—an overview : A huge review. *American Journal of Epidemiology*, 172 :869–889.

Patnala, R., Clements, J., and Batra, J. (2013). Candidate gene association studies : a comprehensive guide to useful in silico tools. *BMC genetics*, 14 :Article39.

Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic epidemiology*, 32 :381–385.

Pounds, S. B. (2006). Estimation and control of multiple testing error rates for microarray studies. *Briefings in bioinformatics*, 7 :25–36.

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., and Sunyaev, S. R. (2010a). Pooled association tests for rare variants in exon-resequencing studies. *American Journal of Human Genetics*, 86 :838–838.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010b). New approaches to population stratification in genome-wide association studies. *Nature reviews. Genetics*, 11 :459–463.

Project, C. G., Altshuler, D. L., Durbin, R. M., and et al., G. R. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467 :1061–1073.

Pulverer, B. (2023). Open access for open science. *EMBO reports*, 24 :e57638.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., Bakker, P. I. D., Daly, M. J., and Sham, P. C. (2007). Plink : A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81 :559–575.

Qiu, X., Klebanov, L., and Yakovlev, A. (2005). Correlation between gene expression levels and limitations of the empirical bayes methodology for finding differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, 4 :Article34.

Riancho, J. A. (2012). Genome-wide association studies (gwas) in complex diseases : advantages and limitations. *Reumatologia clinica*, 8 :56–57.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273 :1516–1517.

Robbins, R. B. (1918). Some applications of mathematics to breeding problems III. *Genetics*, 3 :375–389.

Roeder, K., Bacanu, S. A., Wasserman, L., and Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. *American journal of human genetics*, 78 :243–252.

Roeder, K., Devlin, B., and Wasserman, L. (2007). Improving power in Genome-Wide Association Studies : weights tip the scale. *Genetic epidemiology*, 31 :741–747.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 24 :398–413.

Roquain, E. and Wiel, M. V. D. (2008a). Multi-weighting for fdr control. *arXiv :hal-00306219v1*.

Roquain, E. and Wiel, M. V. D. (2008b). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3 :678–711.

Rubin, D., Dudoit, S., and Laan, M. V. D. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Statistical applications in genetics and molecular biology*, 5 :Article19.

Sabatti, C., Service, S., and Freimer, N. (2003). False discovery rate in linkage and association genome screens for complex disorders. *Genetics*, 164 :829–833.

Sarkar, S. K. (2006). False discovery and false nondiscovery rates in single-step multiple testing procedures. *Annals of Statistics*, 34 :394–415.

Sarkar, S. K., Guo, W., and Finner, H. (2012). On adaptive procedures controlling the familywise error rate. *Journal of Statistical Planning and Inference*, 142 :65–78.

Schweder, T. and Spjøtvoll, E. (1982). Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69 :493–502.

Scott, J. G., Kelly, R. C., Smith, M. A., Zhou, P., and Kass, R. E. (2015). False discovery rate regression : an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110 :459–471.

Siegmund, D., Yakir, B., and Zhang, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. *Annals of Applied Statistics*, 5 :645–668.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73 :751–754.

Skopelitou, A. S., Katsanos, K. H., Michail, M., Mitselou, A., and Tsianos, E. V. (2003). Immunohistochemical expression of fhit gene product

in inflammatory bowel disease : significance and correlation with clinicopathological data. *European journal of gastroenterology & hepatology*, 15 :665–673.

Sladek, R., Rocheleau, G., and et al., J. R. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445 :881–885.

Stanislas, V., Dalmasso, C., and Ambroise, C. (2017). Eigen-epistasis for detecting gene-gene interactions. *BMC bioinformatics*, 18 :Article54.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Source : Journal of the Royal Statistical Society. Series B (Methodological)*, 36 :111–147.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64 :479–498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100 :9440–9445.

Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic epidemiology*, 37 :334–344.

Tabor, H. K., Risch, N. J., and Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits : practical considerations. *Nature reviews. Genetics*, 3 :391–397.

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, 20 :467–484.

Tamhane, A. C. and Dunnett, C. W. (1999). Stepwise multiple test procedures with biometric applications. *Journal of Statistical Planning and Inference*, 82 :55–68.

Thomas, D. C. (2004). *Statistical methods in genetic epidemiology*. Oxford University Press.

Turner, S., Armstrong, L. L., Bradford, Y., and et al., C. S. C. (2011). Quality control procedures for genome wide association studies. *Current protocols in human genetics*, 68 :1.19.1–1.19.18.

Vandussen, K. L., Liu, T. C., Li, D., Towfic, F., Modiano, N., Winter, R., Haritunians, T., Taylor, K. D., Dhall, D., Targan, S. R., Xavier, R. J., McGovern, D. P., and Stappenbeck, T. S. (2014). Genetic variants synthesize to produce paneth cell phenotypes that define subtypes of crohn's disease. *Gastroenterology*, 146 :200–209.

Veauthier, B. and Hornecker, J. R. (2018). Crohn's disease : Diagnosis and management. *American Family Physician*, 98 :661–669.

Walters, E. (2016). The p-value and the problem of multiple testing. *Reproductive BioMedicine Online*, 32 :348–349.

Wang, S., Hou, Y., Chen, W., Wang, J., Xie, W., Zhang, X., and Zeng, L. (2018). Kif9-as1, linc01272 and dio3os lncrnas as novel biomarkers for inflammatory bowel disease. *Molecular Medicine Reports*, 17 :2195–2202.

Wasserman, L. and Roeder, K. (2006). Weighted hypothesis testing. *arXiv :math/0604172v1*.

Wehkamp, J., Harder, J., Weichenthal, M., Schwab, M., Schäffeler, E., Schlee, M., Herrlinger, K. R., Stallmach, A., Noack, F., Fritz, P., Schröder, J. M., Bevins, C. L., Fellermann, K., and Stange, E. F. (2004). Nod2 (card15) mutations in crohn's disease are associated with diminished mucosal α-defensin expression. *Gut*, 53 :1658–1664.

Wierzbicki, P., Adrych, K., Kartanowicz, D., Wypych, J., Stanislawowski, M., Zwolinska-Wcislo, M., Celinski, K., Skrodzka, D., Godlewski, J., Korybalski, B., Smoczynski, M., and Kmiec, Z. (2009). Overexpression of the fragile histidine triad (fhit) gene in inflammatory bowel disease. *Journal of physiology and pharmacology : an official journal of the Polish Physiological Society*, 60 Suppl 4 :57–62.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J., Groth, P., Goble, C., Grethe, J. S., Heringa, J., t Hoen, P. A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S. A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., Lei, J. V. D., Mulligen, E. V., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific Data 2016 3 :1*, 3 :160018.

Won, S., Kim, W., Lee, S., Lee, Y., Sung, J., and Park, T. (2015). Family-based association analysis : A fast and efficient method of multivariate association analysis with multiple variants. *BMC Bioinformatics*, 16 :46.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the Sequence Kernel Association Test. *American Journal of Human Genetics*, 89 :82–93.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics (Oxford, England)*, 25 :714–721.

Xu, C., Tachmazidou, I., Walter, K., Ciampi, A., Zeggini, E., and Greenwood, C. M. (2014). Estimating genome-wide significance for whole-genome sequencing studies. *Genetic Epidemiology*, 38 :281–290.

Xu, C. M. and Qiao, C. H. (2006). Loss of fragile histidine triad protein expression in inflammatory bowel disease. *World Journal of Gastroenterology : WJG*, 12 :7355–7360.

Yang, E. and Shen, J. (2021). The roles and functions of paneth cells in crohn's disease : A critical review. *Cell Proliferation*, 54 :e12958.

Zhang, M. J., Xia, F., and Zou, J. (2019). Fast and covariate-adaptive method amplifies detection power in large-scale multiple hypothesis testing. *Nature Communications*, 10 :1–11.

Zhang, X. and Chen, J. (2020). Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *Journal of the American Statistical Association*, 117 :1–31.

Zhao, H. and Fung, W. K. (2016). A powerful FDR control procedure for multiple hypotheses. *Computational Statistics & Data Analysis*, 98 :60–70.

Zhao, H. and Zhang, J. (2014). Weighted p-value procedures for controlling FDR of grouped hypotheses. *Journal of Statistical Planning and Inference*, 151-152 :90–106.

Zhou, H., Zhang, X., and Chen, J. (2021). Covariate adaptive familywise error rate control for genome-wide association studies. *Biometrika*, 108 :915–931.

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R., and Lander, E. S. (2014). Searching for missing heritability : Designing rare variant association studies. *Proceedings of the National Academy of Sciences of the United States of America*, 111 :455–464.