UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Elodie PERSYN

Mémoire présenté en vue de l'obtention du
grade de Docteur de l'Université de Nantes
sous le sceau de l'Université Bretagne Loire

École doctorale : Biologie-Santé

Discipline : Recherche clinique, innovation technologique, santé publique
Spécialité : Biologie des organismes
Unité de recherche : l'institut du thorax, INSERM UMR 1087 / CNRS UMR 6291 / Université de Nantes

Soutenue le 25 octobre 2017

Analyse d'association de variants génétiques rares dans une population démographiquement stable

JURY

Président du jury	**David CAUSEUR**, Professeur, Agrocampus Ouest, Rennes
Rapporteurs :	**Laurent ABEL**, Directeur de recherche, Institut Imagine, Paris **David CAUSEUR**, Professeur, Agrocampus Ouest, Rennes
Examinateurs :	**Anne Louise LEUTENEGGER,** Chargée de recherche, Inserm UMR 946, Paris **Ndeye Coumba NDIAYE**, Maître de conférences, Université de Lorraine, Nancy **Jean-Jacques SCHOTT**, Directeur de recherche, l'institut du thorax, Nantes
Directeur de Thèse :	**Richard REDON**, Directeur de recherche, l'institut du thorax, Nantes
Co-encadrants de Thèse :	**Lise BELLANGER**, Maître de conférences, Laboratoire de Mathématiques Jean Leray, Université de Nantes **Christian DINA**, Ingénieur de recherche, l'institut du thorax, Nantes

REMERCIEMENTS

Ce travail de thèse a été financé par la région des Pays de la Loire dans le cadre du projet VACARME, mis en place par Hervé Le Marec et Richard Redon.

Je tiens tout d'abord à remercier mon directeur de thèse Richard Redon, ainsi que mes co-encadrants Lise Bellanger et Christian Dina pour leur supervision. J'ai eu la chance de pouvoir bénéficier de l'expérience de trois personnes formidables. Parmi leurs nombreuses qualités, je remercie Christian Dina pour son regard et ses nombreuses idées en épidémiologie génétique, Lise Bellanger pour m'avoir transmis sa rigueur dans la formalisation des méthodes statistiques et enfin Richard Redon pour son sens de l'organisation m'ayant permis la bonne avancée de mes recherches. Grâce à eux, j'ai beaucoup appris tant sur le plan professionnel que sur le plan personnel au fil de ces trois années.

Je remercie les rapporteurs Laurent Abel et David Causeur pour leur lecture et leur analyse critique consciencieuse de ce manuscrit.

Je remercie les personnes du comité de suivi de thèse Véronique Sébille et Marie de Tayrac ayant veillé à la bonne progression de ce travail.

. . . . . .

Je n'aurais pu faire toutes ces analyses statistiques très coûteuses en temps de calcul sans le cluster « BiRD » mis à disposition par la plateforme de bioinformatique GenoBiRD de l'institut du thorax. Parmi les personnes de la plateforme, je remercie Audrey Bihouée, Eric Charpentier et Jean-François Guillaume qui se sont toujours montrés disponibles pour m'aider avec l'utilisation du cluster.

Pour les conseils en programmation, j'ai pu compter sur les bioinformaticiens Eric Charpentier et Pierre Lindenbaum qui permettent de résoudre les difficultés que j'ai rencontrées extrêmement rapidement.

Je suis aussi reconnaissante envers le CCIPL (Centre de Calcul Informatique des Pays de la Loire) pour m'avoir mis à disposition leurs ressources informatiques au début de ma thèse, lors de l'installation de « BiRD ». Je remercie particulièrement Guy Moebs, pour m'avoir montré l'utilisation du cluster.

. . . . . .

Pour l'analyse des données sur le syndrome de Brugada, je remercie le chef d'équipe Jean-Jacques Schott d'avoir permis leur utilisation pour le développement de méthodes statistiques. Je remercie Solena Le Scouarnec et Matilde Karakachoff pour leur traitement des données et de m'avoir montré les points délicats à connaître lors de l'analyse statistique.

Pour l'analyse des données sur la maladie d'Alzheimer d'apparition précoce, je remercie Camille le Clézio et Dominique Campion pour avoir fourni les données de séquençage et aidé à la compréhension de leur structure. Je remercie aussi Pierre Lindenbaum m'ayant aidé avec le traitement des données.

. . . . . .

Je remercie bien évidemment l'équipe de génétique cardiovasculaire et la plateforme GenoBiRD de l'institut du thorax qui m'ont chaleureusement accueillie. Je me sens très chanceuse d'avoir passé trois années et demie en leur compagnie. Je remercie en particulier le groupe de statistiques, appelé MAGES, pour le partage des idées et des connaissances sur l'analyse statistiques de données génétiques formé par : Christian, Matilde, Floriane, Sidwell, Clément, Joanna et anciennement Pierre-François.

Je remercie la formidable secrétaire de l'équipe Ophélie Tindilière, pour son efficacité exemplaire à traiter les démarches administratives les plus ardues. C'est bien sûr sans oublier ses chères consœurs secrétaires : Aurélie, Corinne et anciennement Marie-Pierre.

Je remercie Stéphanie Chatel, chef de projet, pour le suivi de l'avancée de la thèse dans le cadre de VACARME.

Je remercie enfin toutes les personnes de l'institut du thorax que je n'ai pu citer, mais qui ont participé à rendre le lieu de travail si plaisant.

. . . . . .

Je remercie plus personnellement les doctorantes, Lindzy, Joanna et Marine, pour leur amitié et leur soutien. Elles ont été pour moi le facteur environnemental bénéfique qui a rendu ces années de thèse agréables.

Je remercie tous les autres doctorants ayant partagé mon bureau au cours des années : Marta, Antoine, Xavier et Pauline.

Je remercie tous les stagiaires de Master avec qui j'ai pu créer des liens et qui ont participé à la bonne ambiance : Tatiana, Pauline, Nelly, Manon, Aurélien, Anne-Sophie, Solène, Thomas, Ulysse, Mélanie, Laurabelle, Céline et Léa.

. . . . . .

Enfin je remercie ma famille et mes proches pour leur encouragement et leur confiance.

TABLE DES MATIÈRES

TABLE DES FIGURES

TABLE DES TABLEAUX

LISTE DES PUBLICATIONS

Persyn E, Karakachoff M, Le Scouarnec S, Le Clézio C, Campion D, Consortium FE, et al. **DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease**. Wang K, editor. PLOS ONE. 2017 Jul 24;12(7):e0179364.

Persyn E, Redon R, Bellanger L, Dina C. **The impact of a fine-scale population stratification on rare variant association test results.** En rédaction.

Le Scouarnec S, Karakachoff M, Gourraud J-B, Lindenbaum P, Bonnaud S, Portero V, Duboscq-Bidot L, Daumy X, Simonet F, Teusan R, Baron E, Violleau J, **Persyn E**, Bellanger L, Barc J, Chatel S, Martins R, Mabo P, Sacher F, Haïssaguerre M, Kyndt F, Schmitt S, Bézieau S, Le Marec H, Dina C, Schott J-J, Probst V, Redon R. **Testing the burden of rare variation in arrhythmia-susceptibility genes provides new insights into molecular diagnosis for Brugada syndrome.** Hum Mol Genet. 2015 May 15;24(10):2757–63.

AVANT-PROPOS

Ce travail de thèse a été effectué à l'institut du thorax, au sein de l'équipe « Génétique cardiovasculaire » de Jean-Jacques SCHOTT, sous la direction de Richard REDON et l'encadrement de Lise BELLANGER et Christian DINA. Depuis 2009, l'activité de recherche de l'équipe est tournée autour de l'identification de nouvelles variations génétiques impliquées dans les maladies cardiovasculaires, telles que les arythmies cardiaques (e.g. syndrome de Brugada, syndrome du QT long, repolarisation précoce), les valvulopathies (e.g. prolapsus valvulaire mitral, rétrécissement aortique calcifié), les dyslipidémies et les anévrismes intracrâniens. Cette recherche s'appuie sur des méthodes pluridisciplinaires de diagnostic clinique, biologie moléculaire, bioinformatique et de biostatistiques permettant l'analyse de données issues des technologies de séquençage et de génotypage à haut-débit.

Cette thèse est financée par la région des Pays de la Loire dans le cadre du projet de recherche VACARME (VAincre les maladies CArdiovasculaires, Respiratoires et MÉtaboliques). VACARME a été lancé en 2013, utilisant la recherche translationnelle pour répondre aux enjeux de la médecine de précision en plein essor. Ceci permettrait l'identification de nouveaux biomarqueurs permettant le diagnostic des maladies et de nouvelles voies biologiques pour un traitement plus ciblé. Pour une amélioration de la détection de variants lors des analyses génétiques, le programme stratégique d'épidémiologie génétique se concentre sur l'étude et le développement d'outils statistiques.

L'analyse d'association de variants génétiques rares est maintenant réalisable avec le développement des technologies de séquençage. Depuis 2007 jusqu'à maintenant, de nombreuses méthodes statistiques ont été mises au point, du fait de la complexité d'étudier des variations génétiques très peu observées dans la population. Cette thèse, axée sur ces méthodes statistiques, permet de présenter les problématiques liées à ce type d'analyse.

INTRODUCTION

I- VERS UNE MEILLEURE CONNAISSANCE DU GÉNOME HUMAIN ET DE SA VARIABILITÉ

La génétique naît vers la fin du XIXème siècle suite aux travaux pionniers de Gregory Mendel publiés dans les années 1860 sur la transmission des caractères. Ses travaux oubliés et enfin reconnus qu'à partir des années 1900 ont permis la compréhension actuelle des premiers principes de l'hérédité.

Suite à ces travaux, de nombreuses hypothèses tentent d'expliquer le mécanisme moléculaire lié à la transmission des caractères. Ce n'est qu'en 1944 que la molécule d'ADN est identifiée par Avery, MacLeod et McCarty, comme molécule support de l'hérédité. La structure en double hélice de l'ADN, élucidée par Watson et Crick en 1953, constitue une étape déterminante dans l'avancée des connaissances. Cependant les techniques d'étude des séquences nucléotidiques restent rudimentaires jusqu'à la fin des années 1970, et l'invention de nouveaux outils est nécessaire. Une grande avancée est permise par le développement de nouvelles techniques de séquençage mises au point par Frederick Sanger en 1977 et par Allan Maxam et Walter Gilbert en 1976–1977. L'amélioration des techniques avec notamment l'amplification de l'ADN par PCR (*polymerase chain reaction*) permet une grande avancée dans la connaissance du génome humain à partir des années 1980. Les techniques biomoléculaires permettant de manipuler la séquence d'ADN, sont alors appelées le génie génétique.

Les études individuelles de régions génétiques ne permettent pas de connaître le génome humain dans sa totalité. C'est en 1990, qu'est alors officiellement lancé le projet *Human Genome Project* (HGP) ou projet Génome Humain. Celui-ci a pour objectif de séquencer en intégralité le génome humain. Cette entreprise de grande ampleur est mise en place pour une durée de 15 ans avec un budget de 3 milliards de dollars. Une première ébauche de la séquence est publiée en 2001 et la séquence complète est terminée en 2004. De ce premier génome de référence, il est établi qu'environ 20 000 - 25 000 gènes sont présents dans le génome humain [1]. Les exons, c'est-à-dire les régions transcrites des gènes, ne représentent qu'un faible pourcentage du génome humain. Une compagnie privée, appelée *Celera*

Genomics, obtient lors de la même période la séquence brute de l'ADN mais celle-ci ne sera pas rendue publique.

Suite au projet HGP, qui a permis d'obtenir la séquence nucléotidique du génome humain, un nouveau projet, appelé *Encyclopedia of DNA Elements* (ENCODE), est lancé en 2003 [2]. Celui-ci a pour objectif d'identifier tous les éléments fonctionnels du génome humain, c'est-à-dire tous les éléments agissant à l'échelle des protéines et des ARN, ainsi que les éléments de régulation.

Afin de mieux expliquer la diversité des caractères entre individus et entre populations, de nouveaux projets voient le jour afin de mieux comprendre la diversité génétique des populations humaines. La meilleure connaissance des variabilités génétiques est permise grâce au recueil de l'information génétique de groupes d'individus provenant de diverses populations. Le projet international HapMap[3], débutant en 2002, a pour objectif de créer une « *haplotype map* », c'est-à-dire de cartographier les variations fréquentes de la séquence d'ADN dans le génome humain. Une carte des « *single nucleotide polymorphisms* » (SNP) - marqueurs génétiques de structure simple - est publiée en 2005 et compte plus d'un million de SNP [4]. Le projet *1000 Genomes*, démarrant en 2008, a recueilli l'information génétique de milliers d'individus provenant de différentes populations afin de mieux comprendre les variabilités génétiques. Cette base de données très volumineuse est beaucoup utilisée dans le cadre des études génétiques.

La mise en place de tels projets a été facilitée par la réduction considérable des temps et des coûts de séquençage. De nouvelles techniques de séquençage à haut débit appelées *Next-Generation Sequencing* (NGS), en 2004, permettent une réduction considérable des temps de séquençage. Le génome humain à 1000$ a été enfin possible dès 2014 avec la société Illumina.

Cette avancée des technologies permet d'étudier la génétique des populations à des échelles géographiques de plus en plus fine. En France, la population de l'ouest de la France a été étudiée par Matilde Karakachoff et al. (2015) dans notre laboratoire, à partir du génotypage de 1684 individus [5]. Depuis, quelques années, les projets de génétique de population se précisent en France. Ceux dans lesquels est impliquée l'équipe de génétique de l'institut du

thorax sont présentés dans la Figure 1. Le projet *The French Exome Project* (FREX)[1], a pour objectif de constituer une base de données pour un panel de référence d'exomes de la population Française. Les données d'exome d'environ 600 individus répartis sur l'ensemble du territoire, permettraient d'étudier la génétique de la population française ainsi que de fournir une population de référence pour les études d'association cas-témoins. Suite à l'évolution des techniques et des tendances, le projet *FranceGenRef*[2] vise le séquençage de milliers de génomes complets. Ces projets sont en collaboration avec le projet « VAincre les maladies CArdiaques, Respiratoires et MÉtaboliques » (VACARME)[3] [6]. Ce projet lancé en 2013, a permis la collecte d'information génétique pour des donneurs de sang de la région Pays de la Loire. Cette biocollection « Population de RÉférence du Grand Ouest » (PREGO) rassemble 5000 personnes nées dans les Pays de la Loire et le Morbihan. Les échantillons d'ADN de l'ensemble des participants sont génotypés, une centaine d'entre eux sont sélectionnés pour le séquençage d'exome en vue du projet FREX, enfin environ 350 sont sélectionnés pour le séquençage du génome entier pour le projet *FranceGenRef*.

Ces données permettent une meilleure connaissance du génome humain et de sa variabilité, et sont nécessaires aux études génétiques de populations et de maladies. Le recueil de données de plus en plus volumineuses, décrites comme *Big Data*, représente des enjeux technologiques importants. D'après les estimations de Stephens et al. (2015) [7], 20 des plus grandes institutions génomiques ont collectivement une capacité de stockage de 100PB (soit 10^5 TB). La centralisation et le partage de toutes ces données est un pas nécessaire pour des études génétiques efficaces. Par exemple, la base de données *Exome Aggregation Consortium* (ExAC) [8], mise en ligne en 2014, a été permise grâce à la coalition de chercheurs, dans un but d'agrégation et d'harmonisation des données issues de projets de séquençage d'exome.

[1] Le projet « The French Exome Project », porté par Emmanuelle Genin (Brest), est financé dans le cadre de l'infrastructure France Génomique. Les partenaires participant à l'organisation des collectes sont :

- Inserm UMR 1078, UBO, EFS, CHU Brest
- Inserm UMR 1087, CNRS UMR 6291, l'institut du Thorax, CHU Nantes et Université de Nantes
- Inserm UMR 1079, CNR-MAJ, CHU de Rouen, Université de Rouen, IRIB, Normandie Université, Rouen, France
- Inserm UMR 1167, Institut Pasteur de Lille, Université Lille-Nord de France
- Inserm UMR 897, Université de Bordeaux
- CHU Dijon

Les exomes des participants sont séquencés au Centre National de Génotypage (CNG) à Evry.

[2] Le projet « *FranceGenRef* », porté par Jean-François Deleuze du CNG à Evry, est financé dans le cadre du laboratoire d'excellence « GENomique MEDicale » (GenMed).

[3] Le projet VACARME, porté par Hervé Le Marec et Richard Redon (Nantes), est financé par la région des Pays de la Loire.

Cette base de données permet de fournir une liste exhaustive des variations génétiques et de leurs fréquences dans diverses populations.

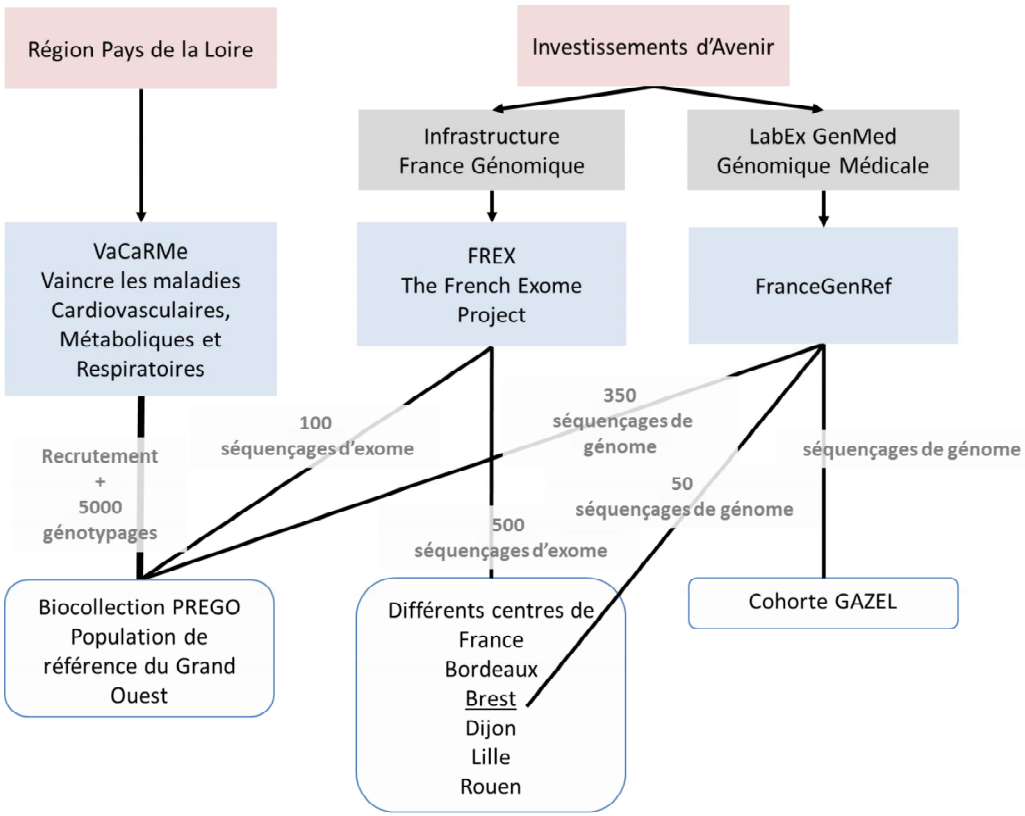


Figure 1. Les différents projets pour la constitution de panels de référence en génétique.
En rouge : les financements. En gris : infrastructures. En bleu : noms des projets. En blanc : la provenance des échantillons d'ADN. Sont aussi indiqués les nombres de génotypage, de séquençage d'exome ou de génome, dans le cadre des différents projets.

II- IDENTIFICATION DE FACTEURS DE RISQUE GÉNÉTIQUES POUR DES PATHOLOGIES

Les études génétiques ont permis de grandes avancées dans le domaine de la santé, en identifiant des facteurs génétiques de risque pour de nombreuses pathologies. Les pratiques ont beaucoup évolué avec l'amélioration des technologies et des connaissances sur le génome humain.

Les premières découvertes ont concerné des anomalies chromosomiques visibles au microscope. La première anomalie chromosomique a été identifiée en 1959, avec la découverte par microscopie d'un chromosome 21 surnuméraire pour la trisomie 21 [9].

Une carte de liaison a été développée en 1980 [10] avec des marqueurs génétiques appelés « polymorphismes de longueur des fragments de restriction » (RFLP, *restriction fragment length polymorphism*). Ceci a permis l'étude de familles présentant une prévalence anormalement élevée de personnes malades, avec « l'analyse de liaison ». Cette approche d'épidémiologie génétique, se basant sur les recombinaisons entre marqueurs génétique, permet de localiser, plus ou moins finement sur les chromosomes, les gènes liés à la maladie. Grâce à cette nouvelle technique, a été localisé pour la première fois, en 1983 [11], un gène impliqué dans une maladie. Il s'agit d'un gène, situé sur le chromosome 4, à transmission autosomique dominante pour la maladie d'Huntington. La cartographie de nouveaux gènes a surtout concerné dans un premier temps les maladies mendéliennes, c'est-à-dire expliquées par un seul locus génétique, et les maladies présentant plusieurs gènes majeurs. Des cartes de variations génétiques avec une meilleure résolution ont été élaborées pour localiser plus rapidement et finement les gènes responsables.

Les maladies dites complexes impliquent de nombreux facteurs génétiques et environnementaux. Pour ces maladies, de nombreux variants génétiques à faible risque sont facteurs de prédisposition. Les « études d'association génétiques » sont menées afin d'identifier ces variations génétiques à faible risque. Par opposition aux analyses de liaison, se concentrant sur des familles, les études d'association permettent de comparer, à l'échelle d'une population, les fréquences alléliques entre les personnes malades et les personnes témoins. Les études d'association génome-entier ont été rendues possibles grâce au génotypage de nombreuses personnes pour de nombreuses variations génétiques fréquentes appelées *Single nucleotide polymorphisms* (SNPs) et réparties sur tout le génome. La première étude a été réalisée, en 2005, pour la dégénérescence maculaire liée à l'âge [12]. Depuis de nombreuses études d'association génome-entier (GWAS, *genome-wide association studies*) ont été menées [13].

Avec le développement des nouvelles technologies de séquençage, il est maintenant possible de séquencer un grand nombre d'individus pour le génome entier, l'exome ou les parties

codantes de gènes candidats. Ainsi, en plus des variations génétiques fréquentes, les variations génétiques rares peuvent être également étudiées.

Chaque personne est unique et manifeste une même maladie de façon différente. L'étude génétique des maladies a pour but de mieux comprendre cette complexité. L'identification de biomarqueurs génétiques spécifiques est un enjeu de taille pour s'approcher d'une médecine personnalisée, avec l'individualisation sur-mesure des traitements médicaux.

III- VERS UNE MÉDECINE PERSONNALISÉE

La thèse s'inscrit dans le cadre du projet VACARME, où la recherche translationnelle sur les maladies cardiaques, respiratoires et métaboliques a pour objectif de développer une médecine personnalisée. La **médecine personnalisée** consiste à adapter sur-mesure le traitement d'un patient. Il s'agit de « donner au bon patient le bon traitement, chaque médicament étant donné à la bonne dose au bon moment » [14]. Le terme « **médecine de précision** » est parfois préféré, considérant que la médecine se veut déjà à la base personnalisée. Cependant une vraie médecine personnalisée semble impossible dans la mesure où il est difficile d'individualiser le choix d'un médicament. Le terme « **médecine stratifiée** » est alors employé lorsqu'un sous-groupe de patients est identifié et pour lequel il existe un traitement ciblé avec un meilleur rapport bénéfice-risque. Le développement de cette médecine a des enjeux humains, technologiques et économiques.

Avec cette médecine de précision, les patients bénéficient d'un diagnostic précoce et de traitements adaptés avec une meilleure prise en charge. Avec le diagnostic précoce des maladies grâce à l'utilisation de biomarqueurs, une meilleure prévention est possible. La prédiction de l'effet des traitements sur des catégories de patients, permet de choisir le bon traitement et de limiter la prise de médicaments inefficaces. Elle devrait alors permettre de diminuer les coûts de santé liés aux traitements et soins inadaptés.

Cette médecine de précision est permise grâce à l'avancée des technologies et des connaissances. En plus du développement des technologies de biologie moléculaire, ceci est rendu possible grâce aux développements d'outils permettant le recueil, le stockage et l'analyse d'un volume important de données biologiques à mettre en relation les unes avec les autres.

La médecine personnalisée s'est beaucoup développée depuis quelques années aux Etats-Unis, au Royaume-Uni et en Chine. En 2015, Barack Obama, l'ancien président des Etats-Unis a annoncé le lancement de la *Precision Medicine Initiative* avec un budget de 215 millions de dollars. La France, bien qu'en retard en termes de nombres de séquençages et d'analyses par an par rapport à ces pays, voit la filière de la médecine personnalisée en plein essor. Un plan de 670 millions d'euros a été annoncé en juin 2016, par l'ancienne ministre de la santé, Marisol Touraine, afin de développer une médecine personnalisée. Ce plan s'appuie sur le rapport « France Médecine Génomique 2025 » [15] d'Yves Lévy, le président de l'alliance nationale pour les sciences de la vie et de la santé (Aviesan). Ce rapport répond à la demande effectuée par l'ancien Premier ministre, Manuel Valls, en avril 2015, afin « d'examiner la mise en place, et la prospective sur 10 ans, des conditions de l'accès au diagnostic génétique dans notre pays ». L'un des objectifs de ce plan est de mettre en place une filière nationale de médecine génomique compétitive avec l'international, avec l'exportation d'un savoir-faire. Dans ce plan, est prévue l'instauration d'un parcours de soins permettant la médecine génomique pour les patients concernés. Enfin, cette filière doit être un « levier d'innovation scientifique et technologique, de valorisation industrielle et de croissance économique». Ce plan prend en compte les spécificités du système français et la dimension éthique pour répondre aux objectifs.

La génétique prend une place très importante dans la médecine personnalisée, d'où le terme parfois de « médecine génomique ». Les biomarqueurs génétiques permettent, tout d'abord, de mettre en place une prévention avec le diagnostic précoce des maladies. Ils permettent aussi de sélectionner les patients les plus répondeurs d'un traitement donné. Afin de développer une médecine de précision, il est alors important d'identifier de nouveaux biomarqueurs afin de comprendre plus finement le développement des maladies.

C'est dans ce cadre que le projet VACARME s'inscrit, avec le but d'identifier de nouveaux biomarqueurs génétiques, et/ou de nouvelles cibles thérapeutiques pour un ensemble de maladies d'intérêt.

IV- OBJECTIFS DE LA THÈSE

Depuis quelques années, la médecine de précision prend de l'ampleur avec le diagnostic précoce des maladies et le traitement ciblé via le dépistage de facteurs génétiques de risque. Identifier et expliquer les facteurs génétiques de risque pour des maladies complexes est un enjeu pour le domaine de la recherche en santé. Les approches ont beaucoup évolué avec le développement des technologies moléculaires et l'agrégation des connaissances dans les bases de données. Les études d'association génome-entier se sont d'abord concentrées sur les variants génétiques fréquents pour expliquer les maladies communes, en partant du postulat « *common disease - common variants* ». Elles ont permis de mettre en évidence de nombreux gènes impliqués dans de nombreuses maladies [16,17]. Cependant les loci génétiques identifiés ne permettent d'expliquer qu'une très faible partie de l'héritabilité des maladies [18]. Plusieurs hypothèses [19,20] ont été émises pour répondre à cette héritabilité manquante : (i) de nombreux variants fréquents avec un faible effet ne sont pas détectés avec la limite de puissance des tests statistiques ; (ii) les variants génétiques rares ont des effets plus forts ; (iii) d'autres types de variations génétiques et épigénétiques contribuent également ; (iv) les modèles statistiques considérés sont trop simples ; (v) les estimations de l'héritabilité sont biaisées. Dans le cadre de cette thèse, nous nous intéressons à l'étude des variants génétiques rares qui sont supposés montrer des effets plus forts que les variants fréquents, en se basant sur les hypothèses de sélection évolutive [21–27].

Les études d'association dans le cadre de variants rares présentent des enjeux supplémentaires en comparaison avec les variants fréquents [28–32]. Du fait de leur faible fréquence, tester individuellement les variants n'est pas possible. De nombreuses méthodes statistiques ont été développées avec la stratégie commune de tester des groupes de variants rares. La variété des méthodes statistiques existante s'explique par la difficulté de tester l'association entre un groupe hétérogène de variants et une maladie. Dans le cadre de cette thèse, nous visons dans un premier temps à identifier les enjeux et les principales stratégies pour tester l'association de groupes de variants rares. Afin de mieux les comprendre, nous souhaitons aussi évaluer la performance de certains tests, en termes de puissance et d'erreur de type I, à partir de la simulation de données génétiques selon différents scénarios. L'application à de vraies données est nécessaire afin de comparer les similarités entre les résultats obtenus.

À l'institut du thorax, l'équipe de génétique s'intéresse à l'identification de gènes impliqués dans divers maladies cardiovasculaires. Le gène *FLNA* a déjà été identifié comme jouant un rôle dans le développement du prolapsus valvulaire mitral [33]. Ce gène a déjà été de nombreuses fois décrit pour des pathologies variées, avec des localisations différentes des mutations [34]. Parmi les projets en cours, une étude d'association avec des variants génétiques rares de plusieurs gènes candidats est menée, a révélé que les variants rares présents chez les personnes atteintes sont principalement localisés dans certaines régions du gène. La détection de ces regroupements de variants rares chez les cas, est alors intéressante dans le cadre des études d'association. Peu de tests statistiques ont été proposés avec l'intégration des positions des variants [35–40]. C'est pourquoi, dans le cadre de cette thèse, nous avons développé un test statistique, appelé DoEstRare, pour « *Density-oriented Estimation for Rare variant positions* », [41], qui permet de détecter des variants à risque localisés dans des régions spécifiques des gènes. Nous avons comparé sa puissance et son comportement aux autres tests étudiés, avec la simulation de données et l'application à des données réelles.

Dans un deuxième temps, nous souhaitons étudier l'impact de la différence génétique entre populations, qu'on appelle stratification de population sur les résultats des tests d'association. La stratification de population est bien connue pour être un facteur de confusion lors de l'analyse des variants fréquents. Il a aussi été montré que cela pouvait aussi entraîner une inflation des p-values des tests de variants rares [42–47]. C'est pour cette raison que le choix des témoins pour les études génétiques est une étape importante dans l'analyse des variants rares. En effet, ces variations génétiques sont supposées être récentes et donc géographiquement localisées. Avec l'accumulation de données, des structures de population à échelle géographique fine sont observables à partir de l'analyse exploratoire de profils génétiques [5,48]. Différents projets présentés précédemment, dont l'institut du thorax fait partie, sont en cours et s'intéressent à l'histoire démographique de la population française à partir des données génétiques. Ces données seront utilisées comme références pour les études génétiques pour l'identification de gènes. Il est alors important de savoir si une structure à échelle géographique fine, telle à l'échelle de la population française ou d'une région

administrative, peut entraîner une augmentation du nombre de faux-positifs[4] avec l'étude des variants rares. Cependant, dans la littérature, les conséquences sont principalement décrites pour des structures de population mondiale ou continentale (population européenne), dont certaines pouvant se rapprocher de pays voisins [42–47]. C'est pourquoi nous souhaitons étudier l'impact d'une structure fine sur les résultats des tests que nous avons comparés. Pour cela, nous nous sommes basés sur l'analyse de données simulées faisant intervenir différents modèles démographiques, en essayant de refléter différentes échelles de géographie.

Afin d'éviter les faux positifs dans les résultats des analyses, il est important d'incorporer l'information de cette structure de population dans les tests statistiques. Différentes méthodes de correction peuvent être employées. Cependant il a été relevé dans la littérature que les méthodes de correction ne permettent pas dans certains cas de corriger cet effet de structure de population [43,49]. Nous voulons évaluer l'impact de la structure de population sur les résultats des tests d'association et choisir une méthode correction adaptée.

Afin de mieux comprendre les études d'association génétique, une première partie de ce manuscrit **Préalables de génétique** est dédiée à resituer les notions d'épidémiologie génétique. Cette thèse visant un public large de statisticiens et de généticiens, les notions de génétique sont rappelées ainsi que le déroulement des analyses statistiques dans le cadre général des études d'association génétique. Ceci permet d'amener la problématique du passage de l'analyse de variants fréquents à l'analyse de variants rares et les enjeux statistiques permettant d'identifier des signaux d'association pour des groupes de variants. La deuxième partie de cette thèse **Partie I : Les tests d'association pour les** variants génétiques rares se concentre sur la compréhension et la comparaison des principales stratégies qui ont été développées (ou appliquées) pour le test de groupes de variants rares. Les comparaisons se sont basées sur deux approches de simulation de données pour évaluer les erreurs de type I et la puissance des tests. Tout aussi important, une comparaison des comportements des tests lors de l'application à de vraies données permet de souligner les enjeux pratiques. Enfin dans une dernière partie **Partie II : Étude de la structure de population dans le cadre des tests d'association** pour variants rares, nous étudions l'impact de structures de population à échelle géographique fine sur les résultats des tests.

[4] Un gène faux-positif est un gène déclaré significativement associé à la maladie bien que dans la réalité, aucun lien n'existe.

PRÉALABLES DE GÉNÉTIQUE

I- NOTIONS DE GÉNÉTIQUE

I.1- LES VARIATIONS GÉNÉTIQUES

Le matériel génétique

Le matériel génétique de l'Homme comprend 23 paires de **chromosomes**, avec 22 paires de chromosomes autosomaux et une paire de chromosomes sexuels. Chaque chromosome est une macromolécule d'ADN (acide désoxyribonucléique) surenroulée. Cette molécule est constituée de deux brins avec une structure en double hélice. Chacun de ces brins consiste en un enchaînement de **nucléotides**. Un nucléotide comprend un ose (désoxyribose), un groupe phosphate et une base azotée pouvant être une adénine (A), une thymine (T), une cytosine (C) ou une guanine (G). L'ordre des nucléotides constitue la séquence porteuse de l'information génétique.

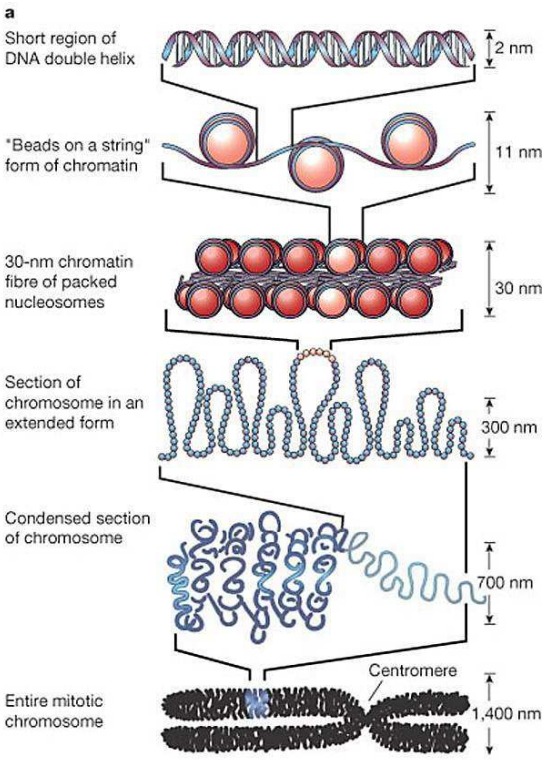


Figure 2. Organisation de l'ADN avec la structure de la chromatine.
Cette image est tirée de [50].

Les différents types de variations génétiques

Pour une meilleure connaissance du génome humain et de sa variabilité, de nombreuses variations génétiques ont été identifiées et répertoriées dans les bases de données.

Les variations ponctuelles affectent de courtes séquences nucléotidiques de l'ADN. Elles peuvent être des **substitutions**, des **insertions** ou des **délétions** de la séquence. Les insertions et les délétions sont parfois regroupées sous le terme de « ***indel*** ». Les substitutions d'un seul nucléotide dans la séquence d'ADN sont appelées « ***single nucleotide variants*** » (SNV). Les « ***single nucleotide polymorphisms*** » (SNP) sont des SNV présents chez plus de 1% de la population.

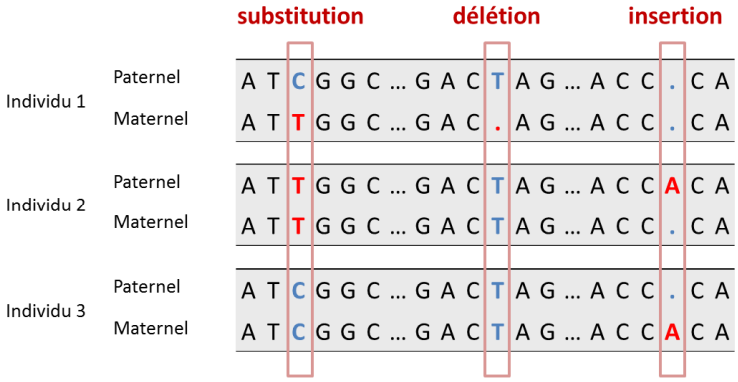


Figure 3. Les types de variations ponctuelles.
Les allèles de référence sont en bleu et les allèles alternatifs sont en rouge.

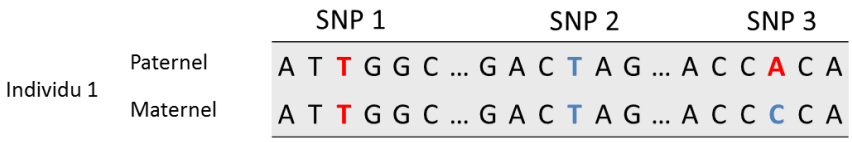
Les variations structurales ne sont pas étudiées dans le cadre de cette thèse. Elles concernent de longues séquences nucléotidiques. Celles-ci correspondent à des altérations génomiques qui impliquent des segments d'ADN de plus de 1kb, d'après la définition donnée dans la revue de Feuk et al. (2006) [51]. Elles comprennent les variations de nombres de copies (*copy-number variant*, CNV), les duplications, les inversions, les translocations, et les disomies uniparentales.

Les variations génétiques peuvent être annotées selon leurs conséquences. Par exemple, les variations se situant dans des gènes codant pour des protéines, peuvent avoir un impact plus ou moins important sur l'activité de la protéine. Dans le cadre des variations ponctuelles, les substitutions dites **synonymes** n'entraînent pas de changement d'acide aminé dans la

séquence protéique. Les substitutions **non-synonymes** sont appelées **faux-sens** lorsque l'acide aminé est remplacé par un autre, et **non-sens** lors de l'apparition d'un codon stop entraînant la production d'une protéine tronquée. Ces substitutions non-sens ont un impact plus grand sur la fonction de la protéine. Les insertions et les délétions ont aussi un fort impact car elles peuvent décaler le cadre le cadre de lecture de la séquence nucléotidique. Les conséquences de ces mutations sont moins prévisibles lorsqu'elles sont situées dans des régions non géniques ou dans des gènes ne codant pas pour des protéines.

Génotype et haplotype

Les différentes formes des variations génétiques sont appelées **allèles**. Chez l'être humain, les chromosomes sont présents en double exemplaire, avec un chromosome hérité du père et un chromosome hérité de la mère. La combinaison des deux allèles d'un individu, à un locus donné est appelée **génotype**. Soit le SNP 1 présenté dans la Figure 4, il présente deux allèles A et T dans la population. Les génotypes possibles dans la population sont alors AA, AT et TT. L'individu 1 présente le génotype Il est aussi possible de parler de génotype pour la combinaison d'allèles pour plusieurs loci, on y fera référence par le terme « **génotype multi-locus** » par la suite.



SNP ID	REF	ALT	Génotype individu 1
SNP 1	A	T	T/T
SNP 2	T	G	T/T
SNP 3	C	A	A/C

Figure 4. Génotype d'un individu

SNP ID est l'identifiant du SNP. REF est l'allèle de référence. ALT est l'allèle alternatif. Dans les bases de données est renseigné l'allèle de référence souvent correspondant à l'allèle majeur. Les autres variations de cet allèle sont les allèles alternatifs.

Nous parlerons aussi par la suite de SNV **bi-allélique** si celui-ci présente deux allèles dans la population ou de SNV **multi-allélique** dans le cas de plus de 2 allèles. Dans les bases de données, est de plus défini un **allèle de référence**, correspondant à l'allèle du génome de référence ; les autres allèles sont dits **allèles alternatifs**.

Un **haplotype** est la combinaison des allèles à différents loci et situés sur un même chromosome. Reprenant, l'exemple de la figure, l'haplotype du chromosome hérité du père, pour les SNP 1 à 3, est TTA et celui hérité de la mère est TTC.

Fréquences alléliques

Pour un SNV donné, l'allèle le moins fréquent dans la population est appelé **allèle mineur** et l'allèle le plus fréquent, l'**allèle majeur**. La **fréquence de l'allèle mineur** (« *minor allele frequency* » : en abrégé MAF) est le critère permettant de définir si celui-ci est **rare** ou **fréquent**. L'estimation de la MAF dans une population pour un SNV présent sur un chromosome autosome (non sexuels) est :

$$\widehat{MAF} = \frac{\text{Nombre d'allèles mineurs dans la population}}{2 \times \text{Nombre d'individus}} \qquad \text{(I.1.1)}$$

et pour un SNV présent sur le chromosome X vaut :

$$\widehat{MAF} = \frac{\text{Nombre d'allèles mineurs dans la population}}{2 \times \text{Nombre d'individus féminins} + \text{Nombre d'individus masculins}} \qquad \text{(I.1.2)}$$

Un variant est communément défini comme rare si la MAF est inférieure à 1%. Cependant cette définition est arbitraire et est variable selon les pathologies étudiées.

I.2- ÉLÉMENTS DE STRUCTURE GÉNÉTIQUE

Le déséquilibre de liaison entre variants génétiques

La notion de **déséquilibre de liaison** (« *linkage disequilibrium* », LD) entre variants génétiques est importante pour les analyses statistiques. Il s'agit de l'association préférentielle entre les allèles de deux loci, certaines combinaisons d'allèles sur les chromosomes étant plus fréquentes que ce qui est attendu en cas d'indépendance. En considérant A_1 et B_1 les allèles d'un premier locus et A_2 et B_2 les allèles d'un deuxième locus [52], la mesure du LD D est :

$$D = |P(A_1A_2) - P(A_1)P(A_2)| \qquad \text{(I.2.1)}$$

Cette mesure du déséquilibre de liaison correspond à l'écart entre la fréquence observée de l'haplotype A_1A_2 ($P(A_1A_2)$) et la fréquence attendue en cas d'indépendance correspondant au produit des fréquences alléliques de A_1 et A_2 ($P(A_1)P(A_2)$). Si la présence d'un allèle est indépendante de celle de l'autre allèle, la mesure de D vaut 0. Il est alors dit qu'il n'y a aucun LD entre les variants génétiques. Une autre mesure du LD est le coefficient de corrélation r et peut s'exprimer en fonction de D de la façon suivante :

$$r = \frac{D}{\sqrt{P(A_1)P(B_1)P(A_2)P(B_2)}} \qquad \text{(I.2.2)}$$

Le coefficient de corrélation de Pearson est le coefficient le plus utilisé pour décrire le LD entre les SNV. La Figure 5 permet d'observer le LD par paires de SNP dans une région génétique.

Le LD est influencé par de nombreux facteurs tels que le taux de recombinaison, le taux de mutations, la dérive génétique, la sélection, et etc. Il est lié à la proximité des SNV sur le génome. Beaucoup moins d'évènements de recombinaison, pour une durée fixée, ont lieu entre deux SNV proches qu'entre deux SNV éloignés. C'est pourquoi les SNV proches ont plus de probabilité d'être transmis ensemble et présentent un plus fort LD.

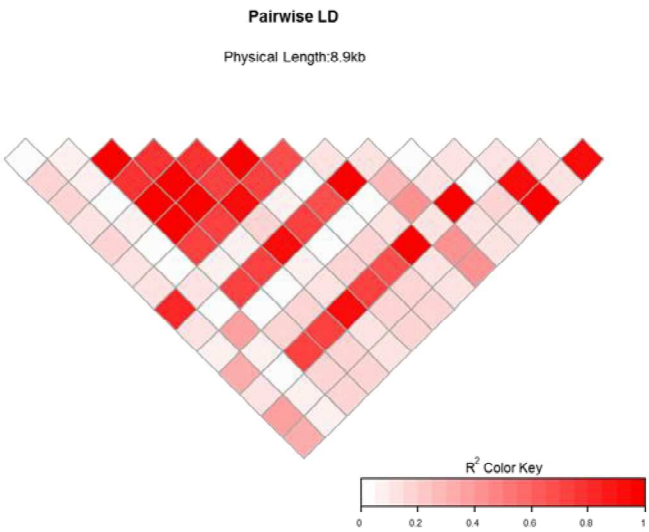


Figure 5. Graphe de LD par paire de SNP.
Ce graphe correspond à l'exemple du Package R LDheatmap. Le LD est mesuré par le coefficient de corrélation de Pearson.

Allèles IBS ou IBD entre individus

Deux allèles sont **identiques par état** (*identical by state*, en abrégé IBS) si la séquence nucléotidique est la même pour deux individus. Si ces deux allèles IBS sont hérités d'un même ancêtre commun, on dit que ces deux allèles sont **identiques par descendance** (*identity by descent*, IBD). Ces notions d'allèles IBS ou IBD sont employées dans le cadre de l'étude de la similarité génétique entre individus.

La Figure 6 permet d'illustrer la différence entre IBS et IBD. Par exemple, nous pouvons voir que les parents ont un allèle en commun pour le locus génétique donné, le nombre d'allèles IBS est alors IBS=1 pour ce locus entre les parents. Cependant cet allèle n'est pas hérité d'un même ancêtre commun, il y alors IBD=0. Quant aux deux enfants, ils présentent leur deux allèles IBS, alors IBS=2. Ces deux allèles sont aussi IBD, alors IBD=2.

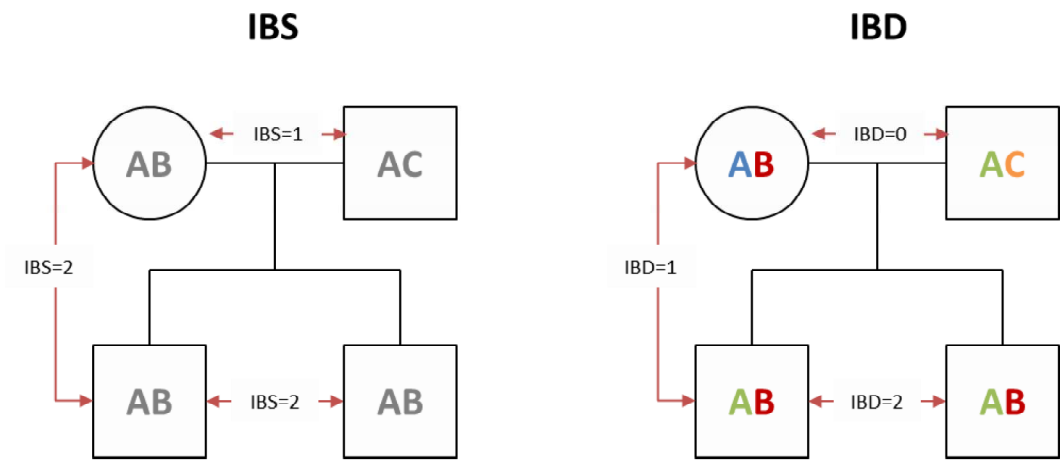


Figure 6. Notions d'états IBS et IBD.

Dans les arbres généalogiques, les hommes sont symbolisés par des carrés et les femmes par des cercles. IBS ici correspond au nombre d'allèles IBS entre les deux individus pour le locus représenté. IBS varie alors entre 0, 1 et 2. Il en est de même pour l'IBD. Pour la figure à droite, les allèles de même couleur sont IBD.

L'équilibre d'Hardy-Weinberg

L'**équilibre d'Hardy Weinberg** est une théorie de génétique des populations selon laquelle, dans le cadre d'une population « idéale », il y a équilibre des fréquences alléliques et génotypiques. Les conditions sous lesquelles s'appliquent le postulat sont : (1) population de taille infinie ; (2) panmixie (équiprobabilité et rencontre aléatoire des gamètes) ; (3) absence de migration, de sélection des individus et de mutation ; (4) les générations ne sont pas chevauchantes. Sous ces conditions, il est possible de déterminer les fréquences génotypiques à partir des fréquences alléliques. Soit A et B les allèles d'un locus donné, $p(A)$ et $p(B)$ leurs fréquences alléliques respectives, et $p(AA)$, $p(AB)$, $p(BB)$ les fréquences génotypiques dans la population idéale. La relation entre les fréquences génotypiques et les fréquences alléliques, dans le cadre d'un équilibre d'Hardy-Weinberg est :

$$\begin{cases} p(AA) = p(A)^2 \\ p(AB) = 2p(A)p(B) \\ p(BB) = p(B)^2 \end{cases} \qquad \text{(I.2.3)}$$

Une déviation de l'équilibre d'Hardy-Weinberg indique qu'une ou plusieurs conditions ne sont pas satisfaites. Cela peut résulter notamment d'une stratification de population.

L'indice de fixation F_{ST} entre populations

Afin de décrire la structure génétique due à la présence de plusieurs sous-populations, Wright (1951, 1965) [53,54] a décrit trois indicateurs, appelés statistiques F : $F_{ST,}$ F_{IT} et F_{IS}, en lien avec la population totale (T), les sous-populations (S) et les individus (I). Le F_{ST} est le plus couramment employé dans les études de génétique des populations et correspond, d'après la définition de Wright (1965) [54], à « *the correlation between random gametes within subdivisions, relative to gametes of the total population* ». Plusieurs méthodes d'estimation ont été proposées, et les interprétations données varient selon les auteurs. Selon la revue de Holsinger et Weir (2009) [55], le F_{ST} est directement lié à la variance des fréquences alléliques entre les populations, et réciproquement au degré de ressemblance entre les individus à l'intérieur des sous-populations. Il est perçu comme la proportion de diversité génétique due aux différences de fréquences alléliques entre les populations, ou comme la corrélation entre les allèles à l'intérieur des populations relativement à la population totale.

Pour le calcul du F_{ST} entre deux populations, nous avons utilisé la méthode de Weir and Cockerham (1984) [56], prenant en compte les tailles d'échantillon pour une estimation non biaisée. Le calcul du F_{ST} est détaillé dans l'Annexe II.

En pratique, le F_{ST} est utilisé comme un indicateur de différenciation génétique entre plusieurs populations (souvent entre 2 populations). Il est calculé à partir de l'information génotypique pour un échantillon d'individus provenant de différentes populations et pour l'ensemble des variants génétiques considérés fréquents. Lorsque le F_{ST} est égal à 0, les populations ne présentent pas de différence allélique. Plus le F_{ST} est grand plus la différenciation entre les populations est grande. Afin de donner un ordre de grandeur, l'étude de Nelis et al. (2009) [57] a estimé les F_{ST} suivants : Européens - Africains 0.153 ; Européens - Japonais 0.111 ; Européens - Chinois 0.110 ; Africains - Chinois 0.190 ; Africains -Japonais 0.192 ; Chinois - Japonais 0.007. Le F_{ST} entre des pays Européens voisins est souvent proche de 0.001.

II- LES ÉTUDES D'ASSOCIATION GÉNÉTIQUES POUR LES VARIANTS FRÉQUENTS

II.1- PRINCIPE

Les **études d'association génétique** sont menées afin de trouver de nouveaux loci génétiques impliqués dans des **traits phénotypiques complexes** [13]. Ces traits phénotypiques peuvent être des traits quantitatifs comme la taille d'un individu ou une maladie commune. Par la suite nous décrirons les études d'association uniquement dans le cadre de pathologies. Les maladies communes sont dites complexes, car on suppose qu'elles résultent de la combinaison de nombreux facteurs génétiques et environnementaux. L'enjeu est alors d'expliquer la totalité de l'héritabilité de la maladie, c'est-à-dire la part de la variance phénotypique expliquée par la composante génétique.

En opposition aux analyses de liaison qui sont des approches familiales, les études d'association peuvent être aussi des approches populationnelles. Les analyses de liaison permettent d'identifier des variants génétiques partagés par les membres atteints d'une même famille. Quant aux études d'association, nous faisons référence à celles permettant d'identifier des facteurs de risque à l'échelle de la population. Le principe est de tester statistiquement l'association entre les variants génétiques et la maladie dans un cadre d'**étude cas-témoins**. Dans la population sont échantillonnés des individus atteints (**cas**) et des individus non atteints (**témoins** ou **contrôles**). Les tests statistiques détectent des différences de fréquences alléliques entre cas et témoins.

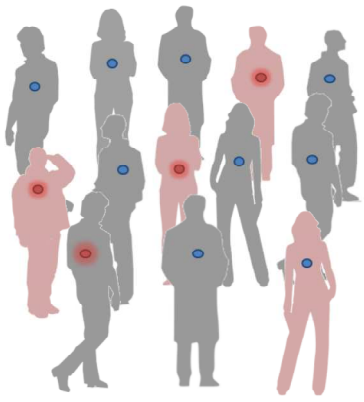


Figure 7. Phénotype et génotype dans la population.
Les personnes dans la population étudiée sont soit saines (gris) soit atteintes (rouge) pour la maladie. Ces personnes présentent un allèle génétique à risque (rouge) ou non (bleu). Les études d'association ont pour objectif d'identifier les variants génétiques à risque qui sont plus fréquents chez les personnes atteintes que les personnes saines.

Dans les parties suivantes, seront synthétisées les différentes étapes des études d'association menées classiquement pour les variants fréquents. Celles-ci sont proches de celles intervenant dans l'analyse de variants génétiques rares.

II.2- DESIGN D'UNE ÉTUDE

Recrutement des cas et des témoins

Étant donné la complexité de ces maladies, de **grandes tailles d'échantillon**, allant jusqu'à plusieurs milliers d'individus, sont nécessaires afin d'avoir une puissance suffisante de détecter des signaux d'association. En effet ces études reposent sur l'hypothèse que les variants génétiques sous-jacents ont une pénétrance variable. Cela veut dire que les personnes présentant une variation génétique à risque ne vont pas toutes présenter la maladie. La maladie résulte en partie de la combinaison de nombreuses variations génétiques augmentant ou diminuant le risque d'être atteint. On parle de facteurs de **susceptibilité** d'une maladie.

Dans ce genre d'étude, la **précision du diagnostic** des patients est importante. Certaines maladies sont liées à l'âge et peuvent se déclarer de manière plus ou moins précoce. Par exemple, une différence est effectuée entre la maladie d'Alzheimer d'apparition précoce (*early-onset Alzheimer's disease*) survenant chez les moins de 65 ans et la maladie d'Alzheimer d'apparition tardive (*late-onset Alzheimer's disease*). Parmi les témoins, certaines personnes être déclarées saines au moment du diagnostic et présenter la maladie tardivement. Les symptômes peuvent différer selon les patients ainsi que la sévérité de la maladie et peuvent être liés à différentes causes génétiques. C'est pourquoi il est important de bien définir l'âge des cas et des témoins, et les critères d'inclusion des patients atteints en fonction de leurs symptômes.

Les cas et les témoins doivent avoir un fond génétique de base similaire, c'est-à-dire être de même origine géographique. En effet il a déjà été observé des différences de fréquences alléliques à l'échelle mondiale ainsi qu'à l'échelle d'un pays. La **stratification de population** est bien connue pour être un facteur de confusion dans les études d'association génétique [13,58].

Le choix des cas et des témoins est alors une étape importante afin d'éviter tout biais statistique et d'en tirer des conclusions erronées.

Régions génétiques d'intérêt

Selon les connaissances biologiques accumulées sur la pathologie étudiée, les analyses se concentrent sur des gènes candidats pour des raisons de coûts, ou s'effectuent à l'échelle du génome entier. Dans le cadre d'**études d'association génome-entier** (*Genome-wide association study : GWAS*), les données sont recueillies pour des SNV répartis sur tout le génome. On compterait sur le génome humain environ 10 millions de SNP [3]. Les études d'association génome-entier « classiques à ce jour», i.e. menées à partir du génotypage de variants fréquents, testent 500 000 à 1 million de SNPs choisis afin de représenter l'ensemble des SNP du génome. On parle alors de « **tag SNP** » pour un SNP situé dans une région génétique en fort déséquilibre de liaison et représentant le groupe de SNP voisins. En raison de la corrélation due au déséquilibre de liaison, les SNP identifiés comme associés avec la maladie ne sont pas avec certitude les SNP réellement impliqués dans la maladie. Le SNP causal de la région génétique est dans certains cas un SNP à proximité du SNP identifié (Figure 8). Il s'agit alors de la détection d'une association indirecte entre un SNP non génotypé et la maladie.

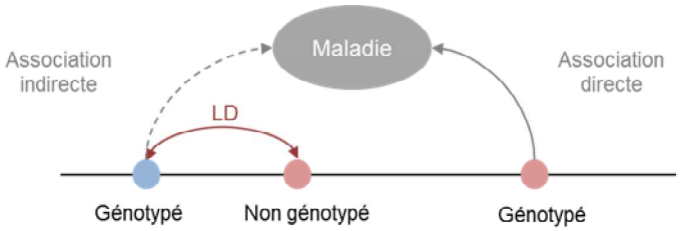


Figure 8. Le test d'association de SNP par des méthodes directes et indirectes.
Les variants impliqués dans la maladie sont en rouge et les variants non impliqués sont en bleu. Il est possible de détecter soit une association directe entre la maladie et un variant à risque génotypé, soit une association indirecte avec un variant neutre génotypé en fort LD avec le variant à risque non génotypé.

II.3- PRÉTRAITEMENT DES DONNÉES

Les SNP sont communément génotypés avec la technique des puces à ADN. Les puces ou « *array* » classiques de génotypage incluent environ 500 000 SNP et sont vendues par plaque de 96. Avec cette technique, l'étape d'identification des génotypes, ou « *calling* », passe par traitement d'intensités de fluorescence. Différents algorithmes peuvent être employés à cette étape, et certains génotypes ne peuvent être identifiés et sont signalés comme manquant. En

fonction des pourcentages de génotypes identifiés, ou « *call rate* », des variants génétiques ou des individus sont retirés de l'analyse. D'autres critères sont aussi utilisés pour filtrer les variants et les individus.

Filtre des variants

Un contrôle qualité des variants génétiques est nécessaire afin d'éviter tout faux-positif dans l'analyse. Les variants rares avec une MAF≤1-5% sont exclus de ces analyses car sont plus souvent sujets à des erreurs de génotypage et les tests statistiques utilisés classiquement ne sont pas assez puissants pour détecter des signaux d'association.

Les SNP présentant un taux de données manquantes élevé sont aussi retirés. Le seuil du « *call rate* » est en général de 95%.

Les variants présentant un grand déséquilibre d'Hardy-Weinberg dans les données, notamment chez les témoins, sont retirés car ceci serait dû à une mauvaise qualité de génotypage. Les seuils de p-value pour le test de Hardy Weinberg sont très variables d'une étude à l'autre, et peuvent se situer entre 0.001 et 1e-7. Ce filtre doit être effectué en considérant la structure de population potentiellement présente dans les données. En effet une structure de population ne satisfait pas les conditions d'équilibre d'Hardy-Weinberg et peut donc induire des p-values faibles pour ce test. Il est alors important d'identifier les sous-populations et de procéder à ce test à l'intérieur de chaque sous-population.

Filtre des individus

Un contrôle qualité est aussi effectué sur les individus. La concordance entre le sexe fourni par le génotypage et le sexe reporté par le patient est vérifiée, afin d'éviter toute erreur d'échantillon d'ADN.

Les études d'association supposent l'**indépendance des individus**. Au moment du design les cas et les témoins sont choisis afin de ne pas présenter un lien de parenté proche. Le degré d'apparenté est vérifié avec le calcul du taux d'allèles IBD pour chaque couple d'individus. Le taux d'allèles IBD peut être estimé à partir des taux d'allèles IBS et des fréquences alléliques avec le programme PLINK [59]. Pour chaque variant, il y a possibilité de 0, 1 ou 2

allèles IBS ou IBD entre deux individus. L'indicateur pour l'apparentement entre deux individus est le taux moyen d'allèles IBD sur le génome :

$$\hat{\pi} = 0.5 \times P(IBD = 1) + P(IBD = 2)$$

Cet indicateur est aussi appelé PIHAT et parmi les couples présentant un PIHAT supérieur à 1%-25%, un des deux individus est retiré de l'étude.

Enfin, les individus présentant des forts taux de données manquantes sont aussi retirés de l'analyse. En général les individus avec un *missing call rate* supérieur à 1%-5% sont retirés de l'analyse.

II.4- ANALYSE EXPLORATOIRE DE LA STRUCTURE DE POPULATION

Les populations cas et témoins doivent présenter des profils génétiques similaires afin d'éviter tout facteur de confusion. En effet si les cas et les témoins ne sont pas de la même origine géographique, les différences alléliques perçues ne sont pas liées à la maladie mais à la stratification de population. Pour illustration, la Figure 9 représente les différences de fréquences alléliques pour le SNP rs4988235 connu pour expliquer la répartition de la tolérance au lactose dans le monde [60]. On peut percevoir des différences de fréquences alléliques à l'échelle mondiale ainsi qu'à l'échelle européenne.

Une **stratification de population** entraîne une « inflation » des p-values des tests statistiques effectués sur les variants génétiques, i.e. des p-values plus faibles qu'attendu sous l'hypothèse nulle. Ceci se traduit par une augmentation du nombre de faux-positifs dans les résultats de l'étude. De manière générale, une analyse exploratoire multivariée est réalisée afin d'observer visuellement si la structure génétique est la même chez les cas et les témoins. La structure génétique observée peut être liée à une stratification de population en lien avec les origines géographiques différentes, ou peut être liée à des biais techniques. En effet il a déjà été remarqué des biais liés au génotypage des individus à différentes périodes. Les individus sont génotypés par plaque, c'est-à-dire par groupe de 96 en général. Cet effet plaque ou appelé « ***batch effect*** » dans la littérature peut aussi être observable sur les premières composantes d'une analyse multivariée sur les données de génotypage pour l'ensemble des échantillons.

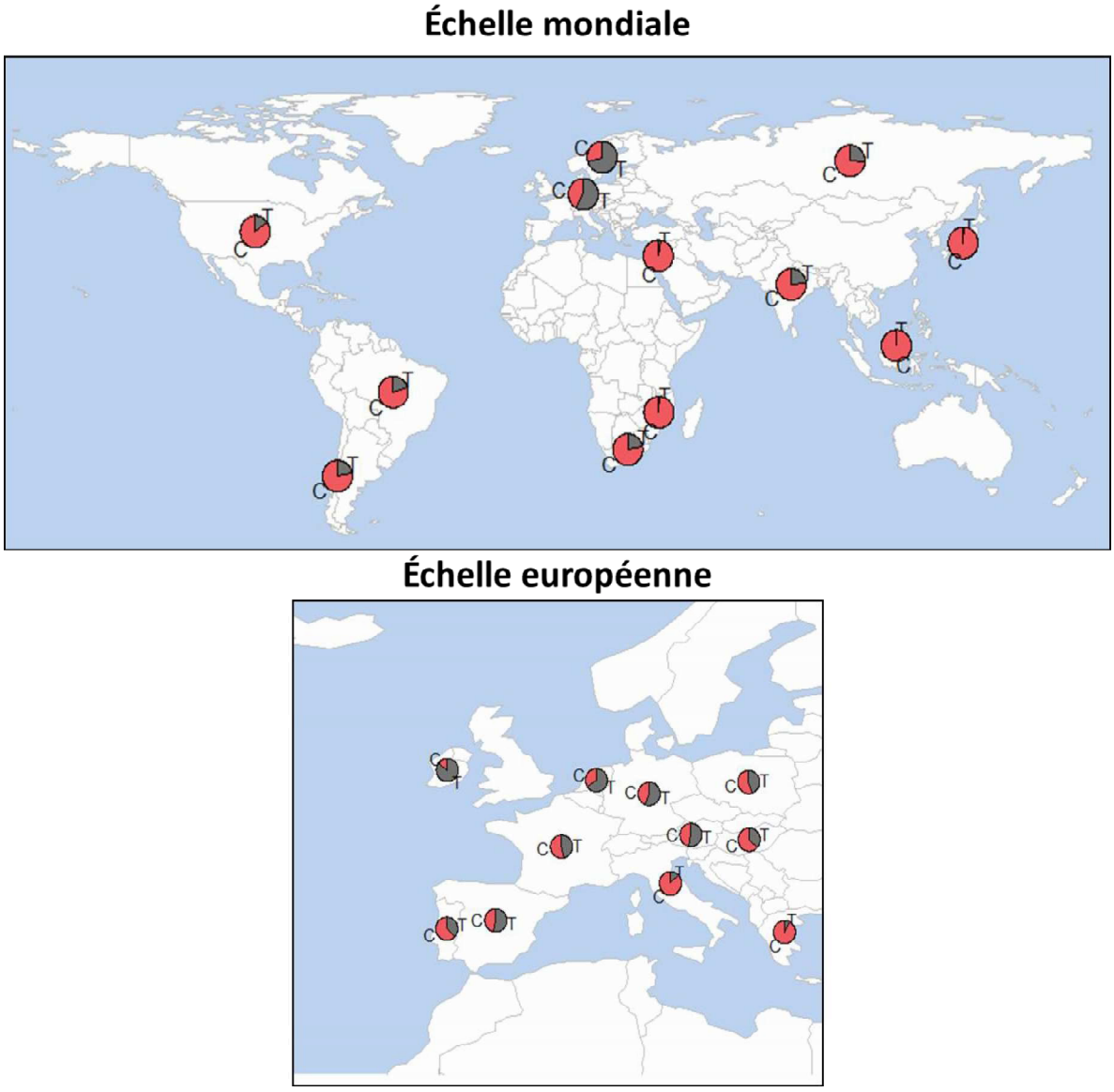


Figure 9. Fréquences alléliques pour le SNP rs4988235 associé à la tolérance au lactose.

Les fréquences alléliques sont issues de ALFRED (https://alfred.med.yale.edu/).

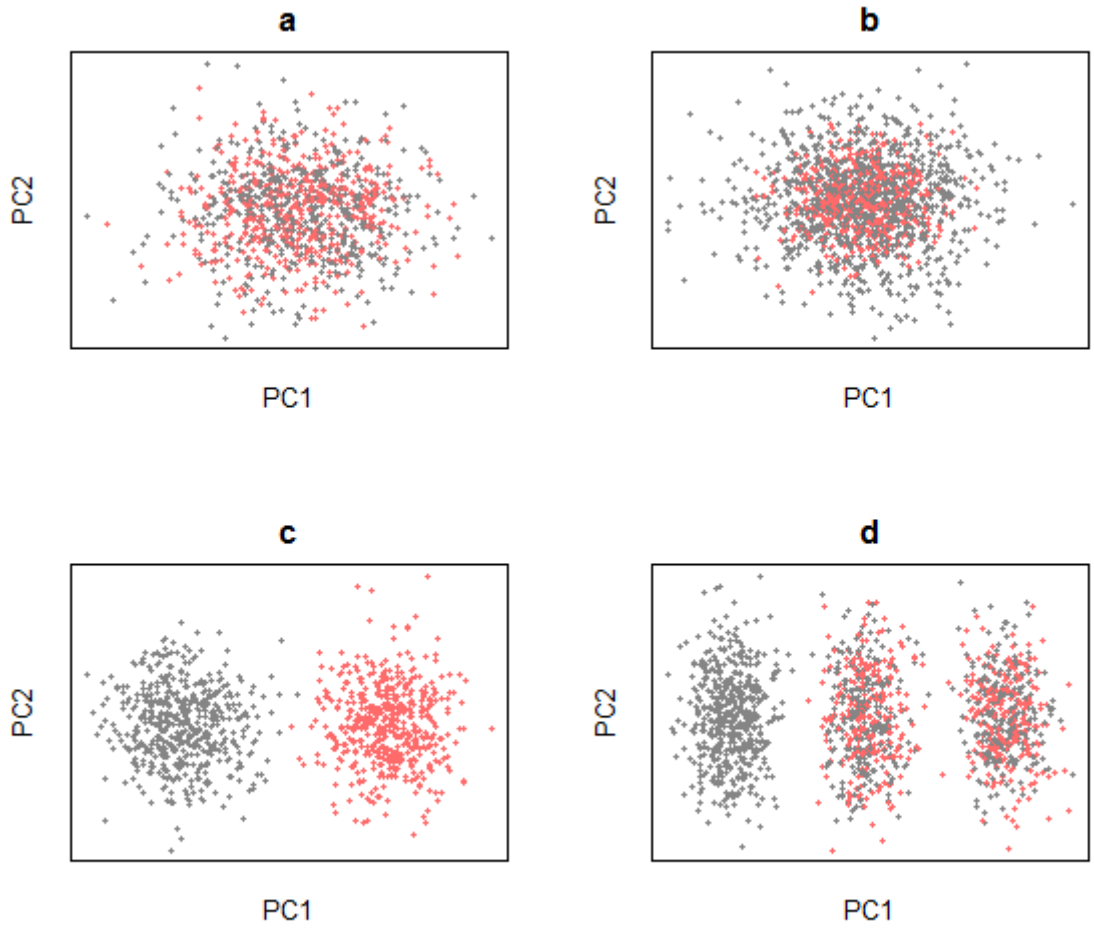


Figure 10. Différents cas de stratification de population.

Les cas sont en rouge et les témoins sont en gris. a. Les cas et les témoins proviennent de la même population. b. Les témoins présentent une variabilité génétique plus élevée. c. Les cas et les témoins proviennent de manière distincte de deux populations. d. Les cas et les témoins sont échantillonnés à partir de différentes populations.

L'analyse exploratoire multivariée est effectuée sur les variants considérés indépendants. En effet sur le génome, il y a des régions en fort déséquilibre de liaison et de longueurs variables. Ces blocs de LD ne présentent pas les mêmes nombres de variants. Ainsi prendre en compte tous les variants dans une analyse exploratoire multivariée donnerait plus de poids aux variants pour lesquels il existe de nombreux variants en LD. Dans un exemple purement imaginaire, si sur 1000 SNP choisis pour représenter le génome, 100 SNP (10% du nombre total) sont entièrement corrélés (r^2 =1), la première composante ne représentera pas les

différences moyennes sur le génome mais bien les différences pour un seul génotype représenté par ces 100 SNP. C'est pourquoi les SNP sont exclus afin de ne garder qu'un seul SNP par bloc de LD. Les méthodes exploratoires les plus couramment utilisées sont l'analyse en composante principale (ACP) et l'analyse de positionnement multidimensionnel (*multidimensional scaling*, en abrégé MDS). L'analyse MDS se base sur une matrice de dissimilarité entre individus. Cette mesure peut être la proportion d'allèles non IBS entre deux individus, pour les variants indépendants considérés dans l'analyse multivariée. Cette mesure est appelée « IBS distance » dans le guide du logiciel PLINK [59].

Selon la structure de population visible, plusieurs stratégies peuvent être considérées. Dans le cas idéal de la Figure 10.a, les populations cas et témoins présentent une structure commune. Une partie des cas et des témoins peuvent présenter une structure commune (Figure 10.b). Dans ce cas, on peut choisir de retirer les individus cas ou témoins s'éloignant de la structure commune. Si les cas et les témoins sont beaucoup trop différents car sont d'origines géographiques différentes (Figure 10.c), il n'y a pas possibilité de correction et l'étude n'est pas possible. Si parmi les cas et les témoins, on observe différentes sous-populations géographiques (Figure 10.d), les sous-populations n'étant présentes que chez les cas ou que chez les témoins doivent être écartées de l'analyse. Il est possible de conduire des analyses statistiques séparées sur les différentes sous-populations, en s'assurant que les effectifs de population le permettent, et combiner les résultats avec une méta-analyse. Il est aussi possible d'ajouter des covariables sur l'origine géographique dans le cadre d'un modèle de régression logistique (voir la partie II.5-Analyse d'association) pour tester l'effet génétique d'un variant.

Afin d'illustrer la démarche, la Figure 11 présente le cas concret de la GWAS pour le syndrome de Brugada (« *Brugada syndrome »,* BrS), et conduite par Bezzina et al. (2013) [61]. Une analyse MDS a été effectuée sur les cas et les témoins de l'étude. Les cas pour lesquels aucun témoin ne correspond ont été retirés et les individus ont été séparés en deux groupes homogènes pour une analyse stratifiée.

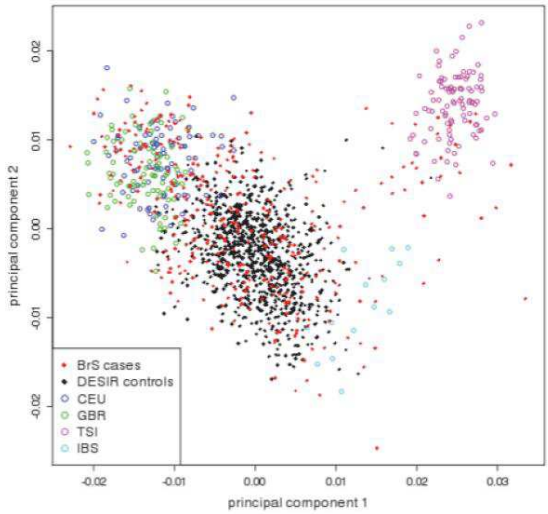


Figure 11. Analyse MDS sur les échantillons d'origine européenne pour l'étude du syndrome de Brugada.

II.5- ANALYSE D'ASSOCIATION

Lors des études d'association génome-entier classiques, les variants génétiques sont testés individuellement, on parle de **tests simple-marqueur**. Ainsi de nombreux tests statistiques sont effectués de manière indépendante. Plusieurs tests statistiques sont applicables pour tester l'association entre le variant génétique et le phénotype [62]. Il semble intuitif de tester l'indépendance entre le génotype et le phénotype à partir d'un tableau de contingence (Figure 12) par l'intermédiaire d'un test du chi-deux à 2 degrés de liberté. Il est aussi possible de supposer des modèles récessif ou dominant et d'obtenir une table de contingence 2x2 en regroupant des génotypes.

	Cas	Témoin
AA	N_{AA}^{A}	N_{AA}^{U}
AB	N_{AB}^{A}	N_{AB}^{U}
BB	N_{BB}^{A}	N_{BB}^{U}

Figure 12. Tableau de contingence pour le test d'indépendance entre le phénotype et le génotype.

Les modèles

L'association entre le phénotype et le génotype peut être testée avec la considération du modèle de régression logistique suivant :

$$logit\big(P(Y_i = 1)|X_{ij}, \boldsymbol{Z_i}\big) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta X_{ij} \qquad \text{(II.5.1)}$$

avec Y_i, le phénotype de l'individu i ; $\boldsymbol{Z_i'}$, le vecteur des covariables pour l'individu i; X_{ij}, le génotype de l'individu i pour le variant j ; α_0, le risque de base ; β, le coefficient de régression pour l'effet génétique ; $\boldsymbol{\alpha}$, le vecteur des coefficients de régression pour les effets des covariables.

Le génotype X_{ij} peut être codé de plusieurs façons. Le modèle génétique le plus fréquemment supposé est le modèle additif comme spécifié dans (II.5.2). Par exemple, si on considère un locus à 2 allèles A et B, les génotypes AA, AB et BB sont codés 0, 1 et 2. Dans ce modèle on considère que l'effet est proportionnel au nombre d'allèles en lien avec le phénotype. On peut aussi supposer un modèle génétique récessif ou dominant. Pour le modèle récessif, les génotypes AA, AB et BB sont codés 0, 0 et 1. Pour le modèle dominant les génotypes sont codés 0, 1 et 1. Dans le cadre d'un modèle génétique codominant, dans lequel chaque génotype a un effet qui lui est propre, le modèle de régression logistique est le suivant :

$$logit\left(\mathrm{P}\big(\mathrm{Y_i} = 1|X_{ij}, \boldsymbol{Z_i}\big)\right) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta_{AB}X_{ij_{AB}} + \beta_{BB}X_{ij_{BB}} \qquad \text{(II.5.2)}$$

avec $X_{ij_{AB}} = 1$ si l'individu présente le génotype AB pour le variant j, sinon $X_{ij_{AB}} = 0$. De même, $X_{ij_{BB}} = 1$ si l'individu présente le génotype BB pour le variant j, sinon $X_{ij_{BB}} = 0$. Ce modèle génétique suppose que chaque génotype présente un effet différent.

L'hypothèse nulle dans le cadre du premier modèle (II.5.2) est $H_0 : \beta = 0$. Lorsqu'on considère un modèle génétique codominant, l'hypothèse nulle est alors $H_0 : \beta_{AB} = \beta_{BB} = 0$. Le modèle génétique le plus choisi, bien que très discutable d'un point de vue biologique est le modèle additif. Il est préféré au modèle codominant pour des raisons de puissance statistique. En effet dans le cadre du modèle codominant, deux effets génétiques sont à estimer au lieu d'un seul pour le modèle additif.

L'avantage du modèle de régression logistique est la possibilité d'ajuster le modèle pour des facteurs non génétiques, comme le sexe, l'âge, l'origine géographique, etc., avec l'incorporation de covariables. Il est très fréquent d'incorporer les premières composantes de

l'analyse multivariée ACP ou MDS afin de prendre en compte la structure présente dans les données (voir la partie précédente **Analyse exploratoire de la structure de population**) [63].

Les tests statistiques

L'hypothèse nulle peut être testée avec le test de Wald, le test du rapport de vraisemblance ou un test du score de Rao [64].

Le test de Wald

Considérons le premier modèle de régression logistique (II.5.1). Le test de Wald compare l'estimation du paramètre testé $\hat{\beta}$ à la valeur sous l'hypothèse nulle $\beta = 0$. Le valeur $\hat{\beta}$ est estimée en maximisant la vraisemblance du modèle par rapport aux données. La statistique de test est ainsi

$$\frac{\hat{\beta}^2}{\mathrm{var}(\hat{\beta})}$$

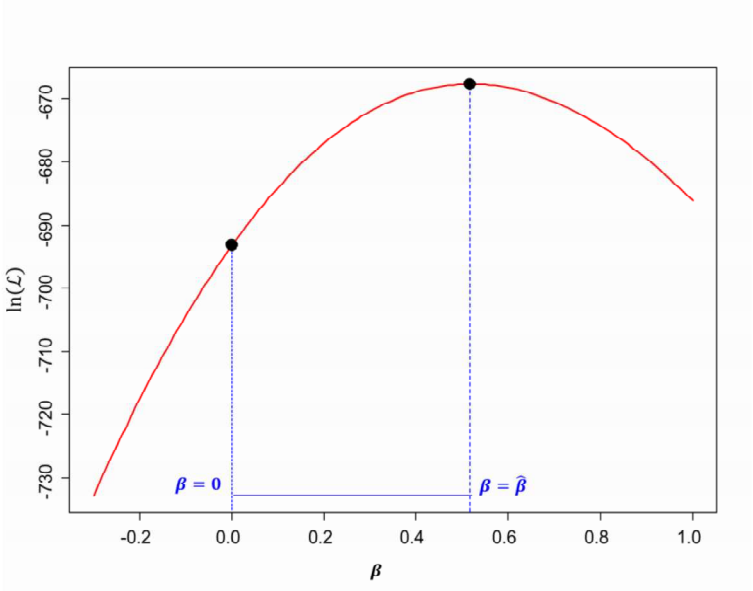


Figure 13. Le test de Wald

Le test du rapport de vraisemblance

Le **test du rapport de vraisemblance** compare les vraisemblances des deux modèles suivants :

$$\begin{aligned} \mathcal{M}_0 \quad & logit\left(P(Y_i = 1|X_{ij}, \boldsymbol{Z_i})\right) = \alpha_0 + \boldsymbol{Z_i'\alpha} \\ \mathcal{M}_1 \quad & logit\left(\mathrm{P}(\mathrm{Y_i} = 1|X_{ij}, \boldsymbol{Z_i})\right) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta X_{ij} \end{aligned} \qquad \text{(II.5.3)}$$

La vraisemblance d'un modèle correspond à la probabilité conditionnelle d'observer les données connaissant les paramètres. La statistique de test correspond à un logarithme du rapport des vraisemblances (*likelihood ratio*, LR) entre les modèles $\mathcal{M}_0$ et $\mathcal{M}_1$.

$$-2\ln(LR) = -2\ln\left(\frac{\mathcal{L}(\beta = 0)}{\mathcal{L}\left(\beta = \hat{\beta}\right)}\right)$$

avec $\mathcal{L}(\beta = 0)$ et $\mathcal{L}\left(\beta = \hat{\beta}\right)$ les estimations des vraisemblances pour les modèles respectifs $\mathcal{M}_0$ et $\mathcal{M}_1$. L'estimation $\hat{\beta}$ du coefficient de régression pour l'effet génétique est déterminée en maximisant la log-vraisemblance. La Figure 14 permet de représenter le principe du test du rapport de vraisemblance. Cette statistique de test suit un chi-deux à un degré de liberté.

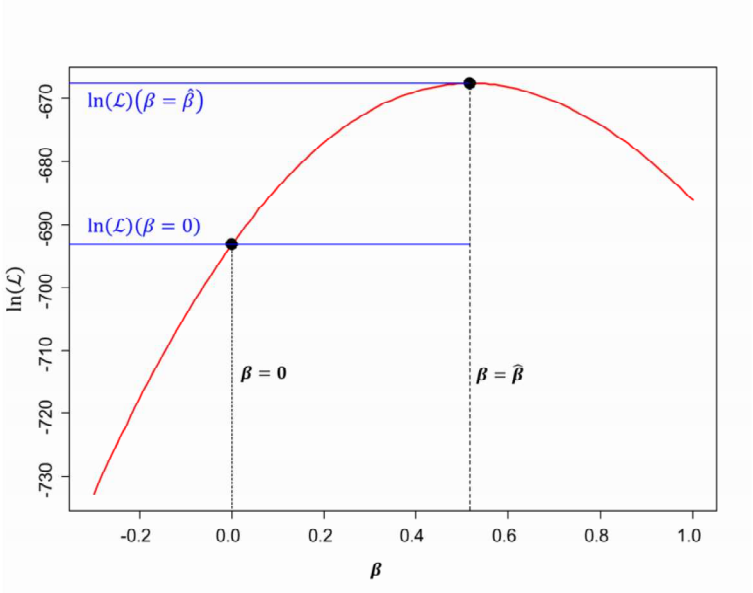


Figure 14. Le test du rapport de vraisemblance.

Le test du score

Le **test du score** calcule la pente de la courbe de log-vraisemblance pour le modèle en $\beta = 0$. Si cette pente est proche de 0, alors la valeur $\beta = 0$ est proche du maximum de vraisemblance. La Figure 15 permet de représenter le principe du test du score. La statistique pour le test du score est :

$$Q_j^2 = \frac{u_j^2}{v_j} \qquad \text{(II.5.4)}$$

avec $u_j = \left[\frac{\delta \ln(\mathcal{L})}{\delta\beta}(\beta = 0)\right]^2 = \sum_{i=1}^{N} X_{ij}(Y_i - \hat{\mu})$ et $v_j = E\left(-\frac{\delta^2 \ln(\mathcal{L})}{\delta\beta^2}(\beta = 0)\right) = \sum_{i=1}^{N}\left(X_i - \overline{X_{.j}}\right)^2 \hat{\mu}(1-\hat{\mu})$. v_j est la variance de u_j sous l'hypothèse nulle. La prise en compte des covariables est permise grâce à $\hat{\mu} = logit^{-1}(\alpha_0 + \boldsymbol{Z_i}'\boldsymbol{\alpha})$. La statistique de test Q_j^2 suit approximativement sous l'hypothèse nulle une loi du χ^2 à 1 degré de liberté.

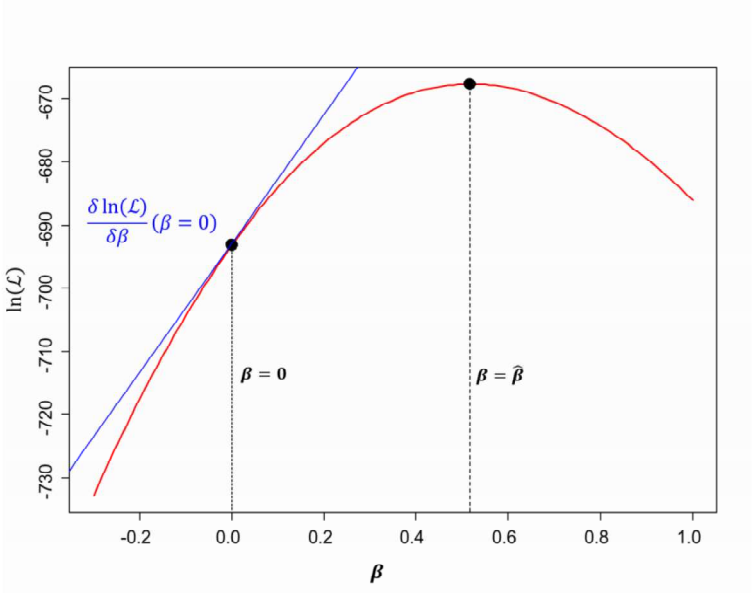


Figure 15. Le test du score.

II.6- INTERPRÉTATION DES RÉSULTATS

Lors des études GWAS, chaque SNP est testé individuellement. Les résultats présentent le niveau de significativité des SNP ainsi que l'orientation de l'effet sous la forme d'un *odd ratio* (OR).De cette liste de SNP, il faut déterminer quels sont les SNP associés à la maladie à partir d'un seuil de significativité. Dans ce contexte d'identification de variants associés

nécessitant des comparaisons multiples, une correction de type Bonferroni [65] s'impose pour limiter le nombre de faux-positifs. Le seuil de significativité est alors de 5% divisé par le nombre de tests réalisables pour l'ensemble du génome. Dans le contexte des GWAS, le seuil posé par la communauté scientifique est de 5e-08. Ceci correspond au seuil de 5% pour 1 million de tests indépendants effectués.

Les résultats sont souvent représentés par deux types de graphe : le ***quantile-quantile plot*** et le ***Manhattan plot.*** Pour ces graphiques, chaque point représente un variant génétique qui a été testé.

Le ***quantile-quantile plot*** (Q-Q plot) représente les logarithmes des p-values observées en fonction des logarithmes des p-values attendues en supposant une distribution uniforme sous l'hypothèse nulle. Un exemple de Q-Q plot selon trois scénarios de GWAS est représenté avec la Figure 16. Cette figure est issue de la revue écrite par Price et al. (2010) [66].La diagonale pour laquelle X=Y représente la situation sous l'hypothèse nulle. Ce graphe permet de savoir s'il y a des signaux d'association avec la maladie avec des variants génétiques présentant des p-values très faibles s'écartant de l'hypothèse nulle. Il permet aussi de visualiser s'il y a une inflation des p-values. En effet, on suppose que de nombreux variants ne sont pas associés à la maladie, et doivent ainsi se situer le long de la diagonale. Si de nombreuses p-values sont faibles et s'écartent de la diagonale, cela signifie qu'il y a une inflation due à une stratification de population. C'est pourquoi ce graphe est souvent réalisé afin de visualiser les éventuels biais statistiques présents dans les résultats.

Un indicateur appelé « ***genomic inflation factor*** » permet aussi de quantifier l'inflation des p-values [67]. A chaque p-value est indiquée la statistique correspondante observée pour la loi du χ^2 à 1 degré de liberté. Le « *genomic inflation factor* » est alors le ratio suivant :

$$\lambda_{50} = \frac{\underset{j}{\text{median}}\left(\chi_1^2(j)\right)}{0.455} \qquad \text{(II.6.1)}$$

avec $\text{median}_j\left(\chi_1^2(j)\right)$ la médiane des statistiques observées pour la loi du χ^2 à 1 degré de liberté, les variants indicés par j. La valeur 0.455 correspond à la valeur attendue sous H_0 de la médiane des statistiques. Si ce ratio est supérieur à 1, il y a une inflation des p-values. Le seuil de décision pour indiquer la présence d'une inflation des p-values est en général de 1.05.

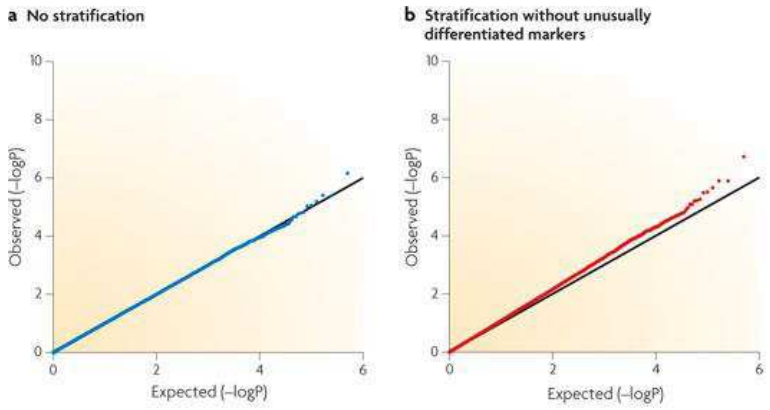


Figure 16. Q-Q plot selon trois scénarios de GWAS (figure de Price et al. (2010) [66]).
a) Pas de stratification de population. Les p-values suivent une distribution uniforme. b) Stratification de population sans signal d'association parasite. On observe une légère inflation des p-values, i.e. les p-values observées sont inférieures à celles attendues sous H_0. c) Stratification de population avec des signaux d'association parasites. On observe une très grande inflation pour un ensemble de marqueurs.

Le « ***Manhattan plot*** » représente les logarithmes des p-values en fonction des positions des variants génétiques. Un exemple de Manhattan plot est représenté par la Figure 17 pour l'étude du syndrome de Brugada [61]. Le seuil de significativité est situé à 5e-08, permettant de corriger les tests multiples par la méthode de Bonferroni. Un seuil de suggestivité est aussi parfois placé pour distinguer des variants d'intérêt proches du seuil de significativité. Dû au LD entre les variants proches sur le génome, un signal d'association est représenté par un ensemble de variants génétiques significatifs. Ce graphe représentant les résultats en fonction des positions, permet de savoir le nombre de signaux d'association identifiés. Sur le Manhattan plot pour le syndrome de Brugada, nous observons principalement deux signaux d'association.

D'autres outils peuvent être utilisés pour visualiser les signaux d'association de manière plus fine dans la région d'intérêt, comme LocusZoom [68].

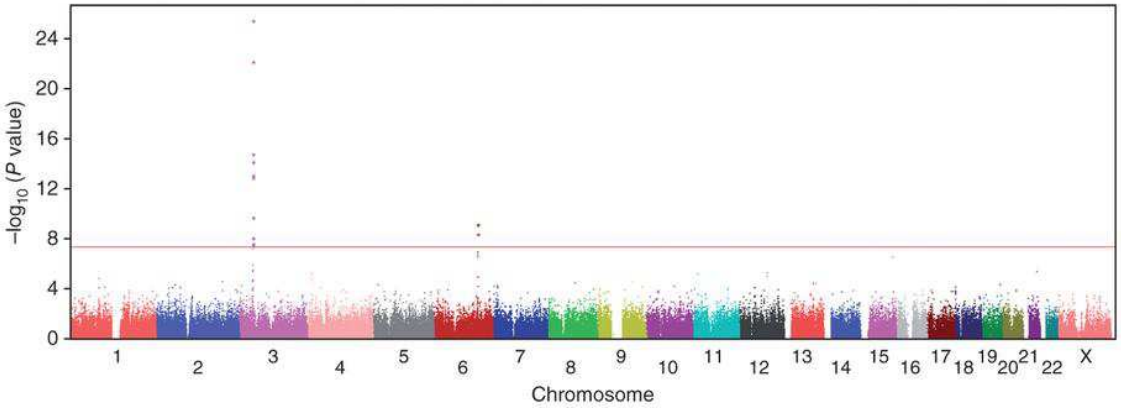


Figure 17. Manhattan plot des résultats de la GWAS sur le syndrome de Brugada de Bezzina et al. (2013) [61].

II.7- VALIDATION DES RÉSULTATS

Lors des études d'association, afin de valider les signaux d'association obtenus [69], une **phase de réplication** est souvent conduite et recommandée. Ceci consiste à étudier les régions intéressantes identifiées à partir de données complémentaires indépendantes suivant le même schéma d'analyse. Ceci permet d'éviter tout biais d'échantillonnage, et de confirmer les signaux d'association dans plusieurs jeux de données.

Après la validation des résultats statistiques par réplication, la localisation du variant causal pour la maladie reste à déterminer. L'**annotation des signaux d'association** à partir des bases de données permet de sélectionner les meilleures variations génétiques candidates pour expliquer la maladie. Une première étape est de lister les SNP en LD avec les SNP significatifs. En effet, comme il a été dit précédemment, les SNP génotypés peuvent être directement ou indirectement associés à la maladie par LD. Afin de faire le tri parmi la liste de SNP en fort LD, une annotation est effectuée sur les conséquences fonctionnelles. Par exemple, les variations génétiques non-synonymes, c'est-à-dire entraînant un changement de la séquence protéique, sont les plus faciles à repérer pour expliquer la maladie. Cependant les variations génétiques peuvent aussi être localisées dans des régions non codantes pour des protéines.

Les analyses *in silico* peuvent aider à raffiner les régions génétiques d'intérêt pour la maladie, mais l'étude des mécanismes biologiques impliqués reste nécessaire. C'est pourquoi des études fonctionnelles *in vitro* ou *in vivo* sont menées en laboratoire, et sont importantes pour

l’identification des variants causaux. La reproduction partielle du phénotype considéré en laboratoire avec la variation considérée est la validation finale de son implication avec la maladie. Avec une meilleure connaissance des voies biologiques impliquées, il est possible de mieux comprendre la maladie et de rechercher de nouvelles pistes de traitement.

III- DE L'ÉTUDE DES VARIANTS FRÉQUENTS À L'ÉTUDE DES VARIANTS GÉNÉTIQUES RARES

III.1- LES LIMITES DES GWAS

La mise en place des études d'association génome-entier ont constitué une avancée majeure dans la découverte de nouveaux variants génétiques à risque pour diverses pathologies. La première GWAS, en 2005, a porté sur la dégénérescence maculaire liée à l'âge [12]. Dans cette étude, 96 cas ont été comparés à 50 témoins pour 116 204 SNP répartis sur tout le génome. Ceci a constitué une avancée majeure en termes d'analyse avec l'identification d'un locus génétique situé sur le chromosome 1 (p-value de 4e-08). Depuis cette étude, de nombreuses GWAS ont été menées grâce à une rapide évolution des techniques. La capacité de détection des signaux d'association est meilleure grâce à l'augmentation des tailles d'échantillon. Les puces de génotypages utilisées sont plus denses et comprennent entre 500 000 et 1 million de marqueurs. Afin d'illustrer cette évolution en termes de chiffres, la Figure 18 montre le nombre de publications depuis 2005 à fin 2016, et a été réalisée à partir des données[5] du catalogue en ligne des GWAS [70]. Durant cette période, 2467 études d'association ont été publiées et répertoriées dans le catalogue GWAS[6].

L'une des attentes des études d'association était de pouvoir expliquer toute l'héritabilité des maladies et de pouvoir prédire les phénotypes étudiés. Elles ont permis en effet des milliers d'associations SNP-phénotype (Figure 19). Cependant il a été constaté, dès 2008, que les marqueurs génétiques identifiés avec les GWAS n'expliquent qu'une faible part de l'héritabilité [18]. Il a été pris pour exemple l'étude de la taille d'un individu, qui est un caractère avec une héritabilité très élevée (environ 80%-90%). Trois études ont été menées avec un grand nombre d'individus allant jusqu'à environ 30 000 personnes [71–73]. Néanmoins, il a été observé que les variants identifiés, n'expliquent qu'environ 5% de l'héritabilité de la taille d'un individu. Le terme « ***missing heritability*** » est souvent utilisé dans la littérature pour désigner cette part d'héritabilité manquante que les scientifiques cherchent à expliquer.

[5] Le fichier des données du catalogue GWAS est «gwas_catalog_v1.0-associations_e88_r2017-04-24.tsv » et a été téléchargé en 05/2017 à l'URL https://www.ebi.ac.uk/gwas/api/search/downloads/full .
[6] Il est possible que certaines études parues en 2016 ne soient pas encore répertoriées dans le catalogue des GWAS.

Suite à cette constatation, plusieurs hypothèses ont été émises pour expliquer l'héritabilité manquante[18,19,26,74] telles que :

- le manque de puissance statistique pour détecter des variants fréquents à faible risque
- d'autres types de variations génétiques ne sont pas étudiés
- une surestimation de l'héritabilité pour certains traits phénotypiques
- le modèle statistique pour expliquer la susceptibilité d'une maladie est trop simple

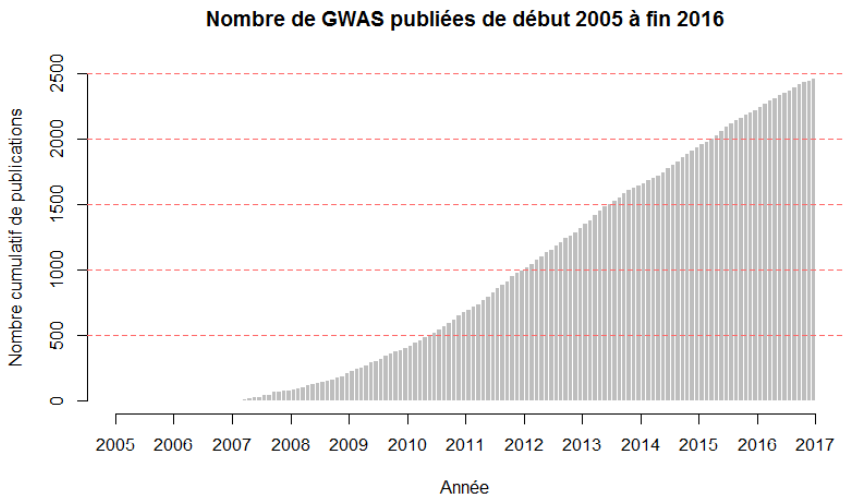


Figure 18. Évolution du nombre de GWAS publiées de début 2005 à fin 2016.
Cette figure est basée sur 1945 GWAS publiées et répertoriées dans le catalogue GWAS entre début 2005 et fin 2014.

Il a été émis l'hypothèse du **modèle dit infinitésimal** [26], avec des centaines voire des milliers de variations génétiques fréquentes expliquant les phénotypes. Ces SNP à faible pénétrance seraient très difficiles à détecter statistiquement à cause de la puissance limitée des tests statistiques et requerraient des tailles d'échantillonnage dans les populations beaucoup plus grandes. Cette hypothèse est renforcée par une étude de Yang et al. (2010) montrant que l'information génétique portée par un très grand nombre de variants fréquents expliquerait une grande partie de l'héritabilité de certains traits [75]. Actuellement, la plupart des variants fréquents identifiés par les études d'association génome-entier présentent des Odd-Ratios (OR) assez faibles. Les OR médians des signaux d'association pour des maladies très étudiées

sont indiqués dans le Tableau 1 d'après les données récoltées dans le catalogue des GWAS (25/06/2017).

Tableau 1. OR médian des signaux d'associations pour des maladies très étudiées, reportées dans le catalogue des GWAS (25/06/2017).

Maladie (*reported trait*)	**Nombre d'études**	**Nombre d'associations**	**OR médian >1 (nombre d'associations avec OR indiqué)**
Type 2 diabetes	43	406	1.270 (385)
Schizophrenia	28	908	1.075 (888)
Alzheimer's disease	21	65	1,3333 (43)
Parkinson's disease	20	174	1.728 (165)
Bipolar disorder	19	169	1.19 (137)

Les études d'association génome-entier se sont focalisées jusqu'à présent sur les variants fréquents. Une autre hypothèse pour expliquer l'héritabilité des maladies est l**'implication de variants rares à effet fort** [21–27]. En effet une hypothèse de la génétique des populations est que le processus de sélection diminuerait la fréquence des variants à haut risque. On passe alors d'une hypothèse communément appelée « *common disease-common variant* » (CD-CV) à une hypothèse dite « *common disease-rare variant* » (CD-RV). Avec l'avancée des connaissances, il est maintenant connu que les SNV rares contribuent également beaucoup au développement de maladies.

Figure 19. Diagramme du catalogue des GWAS et datant de décembre 2016.

Ce diagramme est téléchargé depuis l'URL ftp://ftp.ebi.ac.uk/pub/databases/gwas/timeseries/png/gwas-diagram_2016-Q4.png. Chaque point représente un SNP significativement associé à un trait pour le seuil de 5e-08. Chaque couleur représente une catégorie de trait étudié. L'ensemble des publications a permis d'identifier environ 33000 SNP significativement associés à des pathologies et autres traits phénotypiques divers.

La Figure 20, tirée de la figure de McCarthy et al. (2008) [74] et reprise par Manolio et al. (2009) [19] , permet d'expliquer le type d'étude menée pour identifier des variants génétiques en fonction de leur fréquence et de leur taille d'effet. Les GWAS classiques sont menées afin de détecter des variants fréquents avec des effets globalement faibles. Les variants rares et peu fréquents, sont supposés avoir des effets plus forts, et sont étudiés grâce au séquençage des individus à l'échelle du génome, de l'exome ou de gènes candidats pour la pathologie. En effet, le génotypage n'est pas adapté à l'identification de nouvelles mutations rares car suppose la connaissance de la localisation des variations. Ces variants rares ou peu fréquents à effet intermédiaire peuvent être analysés par approche familiale ou cas-témoin. Les variants très rares, à l'origine de maladies dites mendéliennes (ou mono-géniques) rares sont identifiables par des analyses de liaison réalisées sur des familles. Enfin les variants rares avec des effets très faibles sont très difficiles voire impossible à identifier.

D'autres variations génétiques telles que les CNVs qui sont des variations structurales du génome peuvent aussi expliquer la part non identifiée de l'héritabilité [76]. Des hypothèses ont aussi été émises sur l'implication des variations épigénétiques [77].

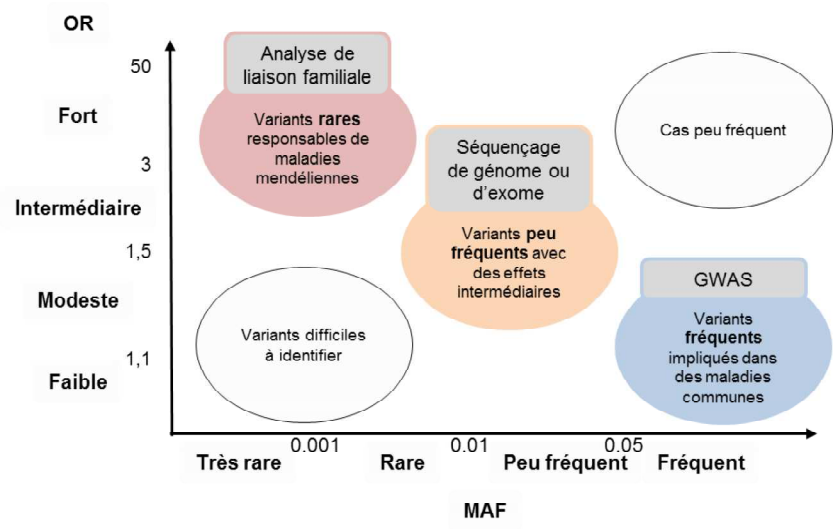


Figure 20. Faisabilité d'identification des variants génétiques en fonction de leur MAF et de leur taille d'effet.

Cette figure est basée de celle de la publication de McCarthy et al. (2008) [74]

Le modèle statistique, testant individuellement les variants génétiques, est certainement trop simple pour représenter la réalité. Cependant, des analyses basées sur des modèles complexes ne sont pas toujours réalisables. Par exemple, il n'est pas possible de construire un modèle de régression logistique avec tous les effets marginaux des variants génétiques car le nombre de paramètres à estimer dans le modèle serait trop élevé. Les études statistiques en génétique doivent faire le compromis entre la faisabilité et la pertinence des hypothèses envisagées. Une voie de recherche est l'étude des phénomènes d' « interaction » des variants génétiques, appelés aussi « épistasie » [78,79]. Ces interactions s'expliquent par la présence de nombreuses voies de signalisation métaboliques. La complexité de ces réseaux serait liée à des processus d'évolution permettant une redondance de l'information. Par exemple une même molécule peut être produite via différentes voies. C'est l'altération de ces différentes voies qui impacteraient fortement la susceptibilité de développer une maladie. D'autres études se concentrent sur les interactions gène-environnement et nécessitent la récupération de données supplémentaires concernant l'environnement des personnes [80,81].

Parmi ces différents axes d'étude génétique des maladies, nous nous focalisons dans cette thèse sur l'identification de variants génétiques rares par l'approche des études d'association, à l'échelle d'une population.

III.2- L'ÉTUDE DE VARIANTS GÉNÉTIQUES RARES

Afin d'expliquer en partie l'héritabilité manquante des maladies, des études se concentrent sur les variants génétiques rares. Ces études sont maintenant possibles grâce à l'amélioration des techniques de séquences pour des temps et des coûts diminués. En effet l'identification de nouvelles variations rares propres à la population étudiée n'est permise qu'avec un séquençage des régions génétiques. Dans la littérature, les études de maladie pour des variations rares se placent le plus souvent dans le contexte de séquençage d'exome ou le séquençage de gènes candidats.

Différentes approches peuvent être menées pour identifier des mutations rares responsables de maladie. Les approches familiales, supposent des mutations propres à la famille augmentant le risque de maladie. Les mutations rares sont filtrées selon des critères de comparaison entre les

personnes atteintes et non atteintes. Les études d'association pour les variants génétiques rares sont aussi menées afin de comparer à l'échelle d'une population générale des cas et des témoins non apparentés.

Comme nous avons pu le voir précédemment, de très nombreuses études GWAS ont été menées pour tester l'association entre des SNP et une maladie d'intérêt. Les procédures pour mener à bien une telle étude sont bien rodées. Des outils comme PLINK [59] permettent d'effectuer un grand nombre d'étapes d'analyses. Dans le cadre des études d'association pour variants rares, des adaptations sont nécessaires. La Figure 22 permet de présenter les principales étapes d'une étude d'association pour les variants fréquents et pour les variants rares et permet de souligner les principales différences.

Le séquençage des échantillons d'ADN et prétraitement bioinformatique des données

Les étapes de production et de prétraitement des données sont résumées dans la Figure 21.

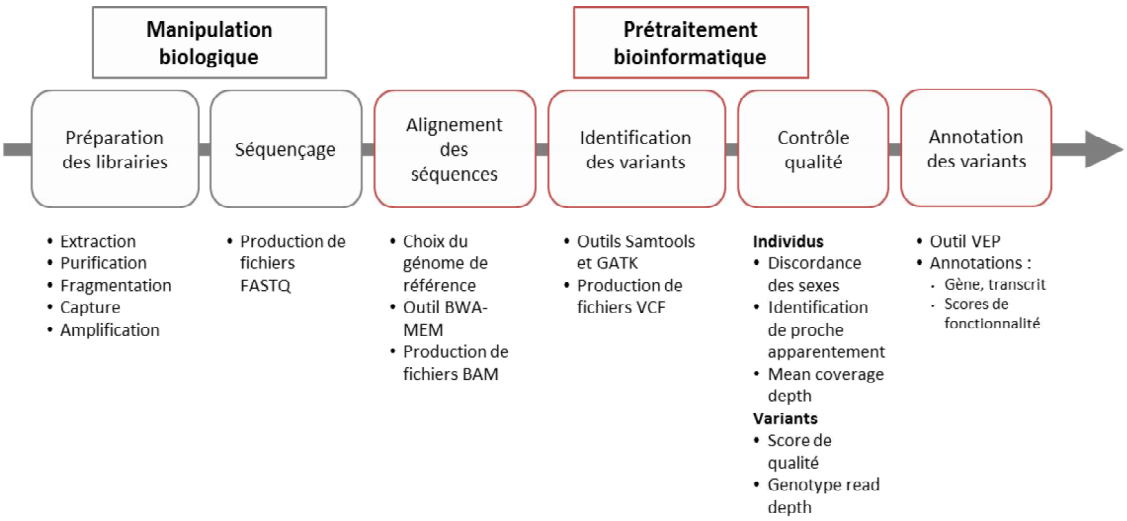


Figure 21. Étapes de production et de prétraitement des données pour l'analyse de variants rares.

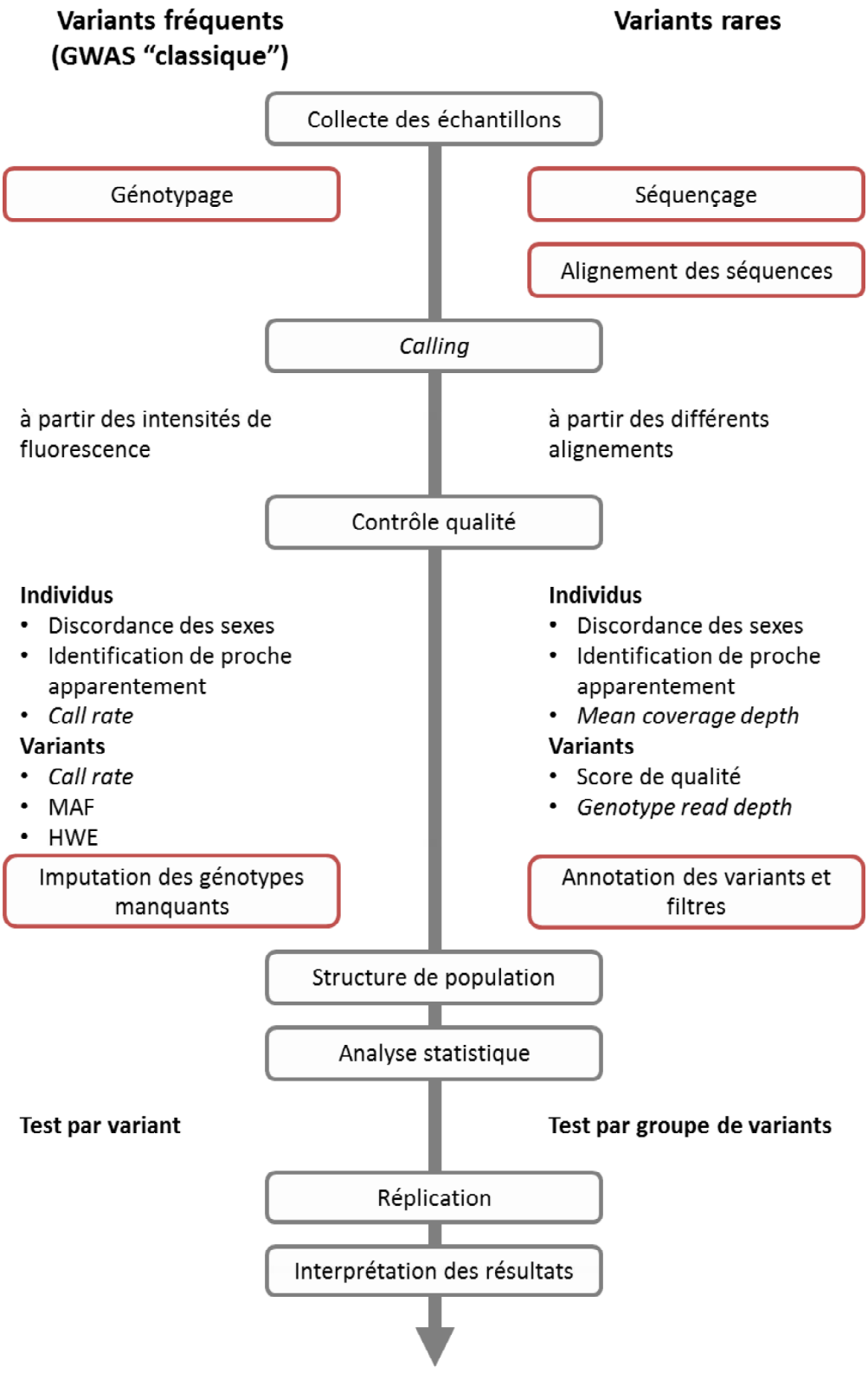


Figure 22. Étapes conceptuelles d'une étude d'association pour les variants fréquents et pour les variants rares.

Avant le séquençage, la préparation de la librairie, i.e. la préparation d'ADN à séquencer, est nécessaire. L'ADN du patient subit de nombreuses opérations, les étapes variant en fonction de l'ADN voulu. Selon l'étude, le séquençage peut concerner le génome entier, l'exome ou les régions codantes de gènes candidats. Pour résumer les nombreuses opérations, l'ADN est fragmenté en de nombreux morceaux, qui sont ensuite amplifiés. Dans le cadre de séquençage d'une partie spécifique du génome, une étape de capture des fragments voulus se déroule avant l'étape d'amplification. Les nombreux fragments d'ADN obtenus, appelés *reads* constituent la librairie et sont ensuite séquencés avec les nouvelles technologies de séquençage haut débit (*Next-Generation Sequencing*, NGS).

Du séquenceur sont récupérés des fichiers FASTQ '.fq' (un par échantillon), lesquels contiennent les séquences « *raw* » et les qualités par base de tous les *reads* séquencés. Ces fichiers sont très volumineux dans le cadre des séquençages de génome entier. Les nombreuses courtes séquences d'ADN sont ensuite alignées sur un génome de référence (e.g. GRCh38/hg38 ou GRCh37/hg19) grâce à un outil bio-informatique répandu BWA-MEM [82]. L'alignement consiste à faire correspondre les courtes séquences à des séquences du génome, et donc de pouvoir les localiser. Les différentes séquences alignées sont stockées dans des fichiers SAM '.sam' (*Sequence Alignment Map*). Ces fichiers étant très volumineux, car contenant de nombreuses informations comme l'identifiant, la séquence, la qualité d'alignement, les positions, etc, ils sont compressés dans des fichiers binaires appelés BAM '.bam' (*Binary Alignment Map*).

L'identification des variants génétiques ou étape de « *calling* » est communément réalisée avec les outils SAMtools [83] et GATK [84]. Un fichier VCF '.vcf' est généré par patient, et recensant tous les variants génétiques qu'ils présentent par rapport au génome de référence.

Le contrôle qualité est très important dans le cadre des variants génétiques rares. En effet il est important de pouvoir distinguer une mutation rare d'une erreur technique. Il est encore difficile à présent de détecter ces erreurs techniques dont la répartition est très hétérogène, i.e. certaines régions génétiques sont plus souvent sujettes aux erreurs que d'autres.

Dans le contexte des études d'association pour les variants génétiques rares, les variants sont annotés avant l'analyse pour connaître leur fréquence dans la ou les populations étudiées. Ceci permet aussi de savoir si ces variants sont présents dans les bases de données externes telles qu'ExAC qui regroupent des milliers d'exome. Si les fréquences diffèrent beaucoup des

bases de données, ces variants sont susceptibles d'être liés à des erreurs techniques. De plus, les variants rares sont testés par groupe (ce sera discuté plus loin, dans la partie sur les choix des tests statistiques). Ces groupes sont définis à partir des annotations sur l'appartenance à un gène ou un transcrit. Enfin, afin d'optimiser la composition des groupes, les variants rares peuvent être filtrés sur un ou plusieurs scores de fonctionnalité ou annotations. En effet, un filtre peut être mis en place pour ne garder que les variants les plus susceptibles d'avoir un effet sur la maladie. L'équipe de bioinformatique de l'institut du thorax utilise par exemple l'outil VEP (*Variant Effect Predictor*) [85] disponible sur le site d'Ensembl. Cet outil utilise par exemple les *sequence ontology terms* [86] décrivant les conséquences des mutations. Des scores de fonctionnalité comme SIFT [87] ou PolyPhen [88] sont aussi très utilisés pour filtrer les variants. Il existe bien d'autres scores de fonctionnalité pour prédire la fonctionnalité d'un variant [89,90]. Ionita-Laza et al. (2016) [91] ont développé un score combinant différents scores de fonctionnalités. Il est difficile de savoir quel filtre appliquer, car étant tout à fait arbitraire, c'est pourquoi lors des études, différents filtres sont employés pour essayer d'identifier des gènes significativement associés.

Choix de tests statistiques adaptés

Les tests simple-marqueur, utilisés couramment lors des GWAS sur les variants fréquents ne sont pas adaptés pour les variants rares. En effet la puissance de ces tests est limitée par le faible effectif de personnes porteuses d'une mutation rare à un locus donné. La stratégie est de cumuler l'information pour un groupe de variants génétiques rares, de façon à augmenter les effectifs. Ces groupes correspondent très souvent aux parties codantes des gènes. Il est en effet beaucoup plus facile de comprendre l'impact direct du changement de la séquence nucléotidique sur la séquence protéique. Beaucoup de méthodes statistiques ont été développées pour les variants rares. Une liste des publications de méthodes d'association pour variants rares est présentée dans le Tableau 2. Cette liste se veut exhaustive, mais il est possible que des méthodes aient pu être oubliées. Bien plus d'une cinquantaine de tests d'association pour variants rares ont été publiés, montrant la complexité de tester un groupe de variants. Les principaux points de difficulté sont les suivants :

- le groupe peut mélanger des variants causaux et des variants non causaux ;

- les variants causaux peuvent être soit à risque (augmentant la probabilité d'être malade), soit protecteurs (diminuant la probabilité) ;
- la répartition des variants causaux sur le gène peut être homogène ou localisée ;
- les effets des variants peuvent ne pas être additifs.

Différents cas de figure énoncés sont schématisés dans la Figure 23.

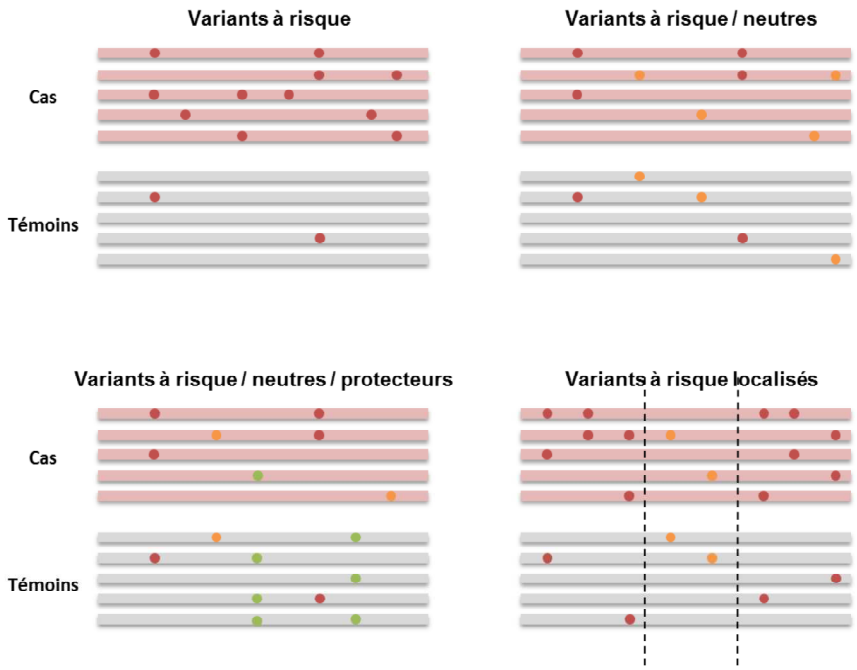


Figure 23. Différents scénarios génétiques pour la susceptibilité d'une maladie.
Les barres rouges et grises correspondent respectivement aux séquences d'ADN chez les cas et les témoins. Les points sont les variants rares et sont de couleur différente selon le type d'effet sur la susceptibilité de la maladie ; les points rouges, orange et verts sont respectivement les variants à risque, neutres et protecteurs.

Différentes stratégies sont utilisées pour tester un ensemble de variants génétiques rares [30]. On retrouve comme principales stratégies :

- *burden tests* : compare les proportions d'allèles rares pour l'ensemble de la région testée entre les cas et les témoins.
- *variance-component tests* : considèrent la présence de variants rares avec des effets très différents au sein du gène (à risque, neutres, protecteurs) et testent la variance des effets génétiques.

- *p-value combination tests* : combinent les p-values des tests simple-marqueurs.
- *combined tests* : combinent différentes stratégies.

Il existe bien sûr de nombreuses méthodes statistiques avec des stratégies autres que celles qui sont énoncées. Ces stratégies sont plus détaillées dans la partie **Principales catégories de test**.

En pratique, les tests les plus couramment sont les tests CMC [92] (617 citations) (peut être aussi appelé CAST [93] (203), ou encore le test de Morris and Zeggini 2010 [94] (239)), wSum, SKAT [95] (603) et SKAT-O [96] (173) (méthodes expliquées dans la partie **Présentation générale des tests étudiés**). Le *wordcloud* de la Figure 24 montre les tests les plus cités dans la littérature.

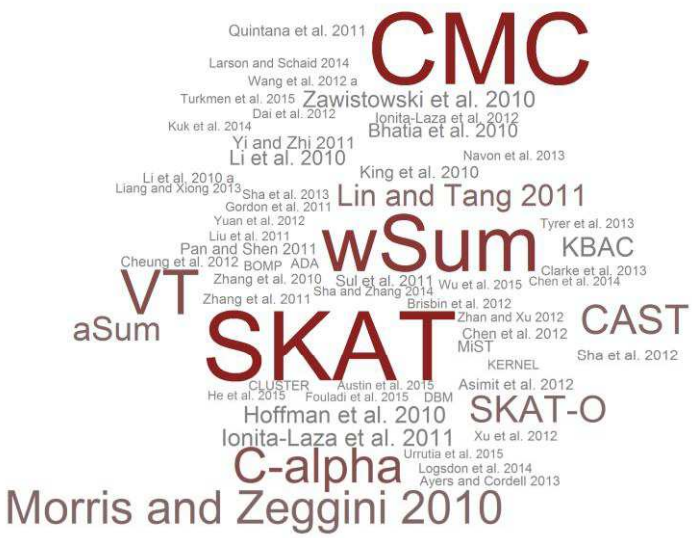


Figure 24. Wordcloud des publications de méthodes pour les variants rares.
La taille des mots est en relation avec le nombre de citations obtenues pour le 07/07/2017. Pour chaque méthode, les auteurs ainsi que l'année de publication est précisée. Les tests étudiés par la suite dans cette thèse sont écrit sous forme d'abbréviation.

Chaque test statistique repose sur des hypothèses biologiques différentes, et sont susceptibles de fournir des résultats différents. C'est pourquoi il est souvent recommandé de ne pas se limiter à un seul test, mais d'en utiliser plusieurs pour détecter le plus de signaux d'association.

Tableau 2. Liste des publications de méthodes adaptées ou développées pour les études d'association de variants génétiques rares avec des maladies complexes

Auteurs	Année	
Morgenthaler and Thilly	2007	[93]
Li and Leal	2008	[92]
Madsen and Browning	2009	[98]
Morris and Zeggini	2010	[94]
Li et al.	2010	[101]
Han and Pan	2010	[103]
Price et al.	2010	[105]
Bhatia et al.	2010	[107]
Liu and Leal	2010	[109]
Li et al.	2010	[110]
Zawistowski et al.	2010	[112]
Hoffmann et al.	2010	[114]
King et al.	2010	[116]
Zhang et al.	2010	[118]
Yi and Zhi	2011	[120]
Ionita-Laza et al.	2011	[122]
Sul et al.	2011	[124]
Neale et al.	2011	[126]
Pan and Shen	2011	[127]
Gordon et al.	2011	[129]
Wu et al.	2011	[95]
Zhang et al.	2011	[132]
Lin and Tang	2011	[134]
Quintana et al.	2011	[136]
Dai et al.	2012	[138]
Yuan et al.	2012	[140]
Liu et al.	2012	[142]
Asimit et al.	2012	[144]
Wang et al.	2012	[146]
Ionita-Laza et al.	2012	[35]
Sha et al.	2012	[149]
Wang and Fingert	2012	[151]
Sha et al.	2013	[153]
Lee et al.	2012	[96]
Cheung et al.	2012	[156]
Xu et al.	2012	[158]
Zhan and Xu	2012	[160]
Brisbin et al.	2012	[162]
Fier et al.	2012	[36]

Auteurs	Année	
Chen et al.	2012	[97]
Chen et al.	2013	[37]
Wang et al.	2013	[99]
Sun et al.	2013	[100]
Navon et al.	2013	[102]
Tyrer et al.	2013	[104]
Fang et al.	2013	[106]
Ayers and Cordell	2013	[108]
Schaid et al.	2013	[38]
Liang and Xiong	2013	[111]
Clarke et al.	2013	[113]
Cheng et al.	2014	[115]
Larson and Schaid	2014	[117]
Won et al.	2014	[119]
Kuk et al.	2014	[121]
Lin et al.	2014	[123]
Logsdon et al.	2014	[125]
Lin et al.	2014	[39]
Sun and Wang	2014	[128]
Sha and Zhang	2014	[130]
He et al.	2015	[131]
Chen et al.	2014	[133]
Turkmen et al.	2015	[135]
Fouladi et al.	2015	[137]
Wu et al.	2015	[139]
Austin et al.	2015	[141]
Lee et al.	2015	[143]
Zhou and Wang	2015	[145]
Greco et al.	2016	[147]
Urrutia et al.	2015	[148]
Hu et al.	2016	[150]
Fang et al.	2016	[152]
Sun et al.	2016	[154]
Li and Chen	2016	[155]
Hasegawa et al.	2016	[157]
Li et al.	2017	[159]
Sofer	2017	[161]
Sugasawa et al.	2017	[163]

PARTIE I : LES TESTS D'ASSOCIATION POUR LES VARIANTS GÉNÉTIQUES RARES

I- PRÉSENTATION GÉNÉRALE DES TESTS ÉTUDIÉS

Les tests d'association pour variants rares étudient l'association entre des groupes de variants rares et une maladie complexe. Ces groupes sont souvent définis à partir de l'unité biologique qu'est le gène. Nous utiliserons souvent par la suite le terme gène pour évoquer la région génétique d'intérêt. Les variants qui composent ces groupes sont hétérogènes : (1) par leur fréquence allélique, certains variants sont plus rares que d'autres ; (2) par leur taille d'effet, certains variants ont des conséquences plus importantes ; (3) par l'orientation de l'effet, certains allèles rares augmentent le risque de développer la maladie et d'autres le diminuent. Il est difficile de traduire statistiquement les hypothèses biologiques sous-jacentes en lien avec la maladie. C'est pourquoi de nombreuses méthodes existent, reposant sur des hypothèses nulles différentes. Les principales stratégies pour les tests d'association sont présentées dans le Tableau 3, la classification étant inspirée de celle de la revue de Lee et al. (2014) [30]. Nous expliquons plus en détails les différentes catégories de tests par la suite avec la description des tests étudiés.

Nous nous intéressons de plus à des tests incorporant les positions des variants dans l'analyse, car il a déjà été montré que des variants impliqués dans une maladie pouvaient être concentrés dans certaines parties spécifiques des gènes. L'équipe de génétique de l'institut du thorax a mis en évidence l'implication de mutations sur le gène *FLNA* dans le développement du prolapsus valvulaire mitral [33]. Ce gène est aussi impliqué dans de nombreuses autres pathologies variées. Les mutations à risque ne sont pas réparties de la même façon sur le gène selon la maladie [34]. Peu de tests prennent en compte les positions dans le calcul de la statistique de test [35–40]. C'est pourquoi nous avons développé un test, appelé DoEstRare pour « *Density-oriented Estimation for Rare variant positions* » [41], qui a pour but d'identifier des regroupements de variants rares à risque chez les cas.

Tableau 3. Tests d'association pour variants rares

Positions	Catégorie	Description de la stratégie	Méthodes
Non	*Burden tests*	Calcul d'un score génétique par individu correspondant à une variable binaire	CAST[93]
		Calcul d'un score génétique par individu correspondant à une somme pondérée des nombres d'allèles mineurs	wSum[98], VT[105], aSum[103]
	Variance-component tests	Teste une composante de la variance des effets génétiques	C-alpha[126], SKAT[95], SKAT-O[96] MiST [100]
	P-value combination tests	Combinaison des p-values issues de tests simple-marqueur	ADA[123]
	KBAC *test*	Analyse de génotypes multi-loci	KBAC[109]
Oui	*Sliding-window tests*	Une statistique est calculée par région génétique par le système de fenêtres glissantes.	BOMP[37]
	Kernel matrix tests	Une matrice noyau est utilisée pour incorporer les distances entre variants dans la statistique	CLUSTER[39], KERNEL[38], PODKAT[40]
	DBM *test*	Les distances physiques entre les variants rares sont calculées. La distribution des distances pondérées est comparée entre les cas et les témoins.	DBM[36]
	DoEstRare	Comparaison des densités de positions des variants et de leur fréquence moyenne chez les cas et les témoins.	DoEstRare [41]

Abréviations: ADA, *adaptive combination of P-values for rare variant association testing*; aSum, *data-adaptive sum test*; BOMP, *burden or mutation position*; CAST, *cohort allelic sum test*; CLUSTER, test de Lin (2014); DBM, *distance-based measure*; KBAC, *kernel-based adaptive cluster*; KERNEL, test de Schaid et al. (2013), PODKAT, *position-dependent kernel association test*; SKAT, *sequence kernel association test*; VT, *variable threshold*; wSum, *weighted sum test*

I.1- NOTATIONS

La structure des données analysées est présentée dans la Figure 25.

Soit **X**, la matrice des génotypes avec X_{ij} le génotype de l'individu $i \in \{1, \dots, N\}$ pour le variant $j \in \{1, \dots, P\}$. Les génotypes sont codés selon le modèle additif, i.e. le génotype correspond au nombre d'allèles mineurs et $X_{ij} \in \{0,1,2\}$. Dans le contexte de variants rares, le génotype homozygote pour l'allèle rare n'est quasiment jamais observé, et la matrice est principalement composée de 0 et rarement de 1.

Soit $\boldsymbol{Y}$, le vecteur des phénotypes avec $Y_i = 1$ si l'individu est atteint de la maladie et $Y_i = 0$ si l'individu est sain.

Soit **Z**, la matrice des covariables avec Z_{im} la valeur de la covariable $m \in \{1, \dots, M\}$ pour l'individu i.

Soit $\boldsymbol{l}$, le vecteur des positions des variants, avec l_j la position du variant j.

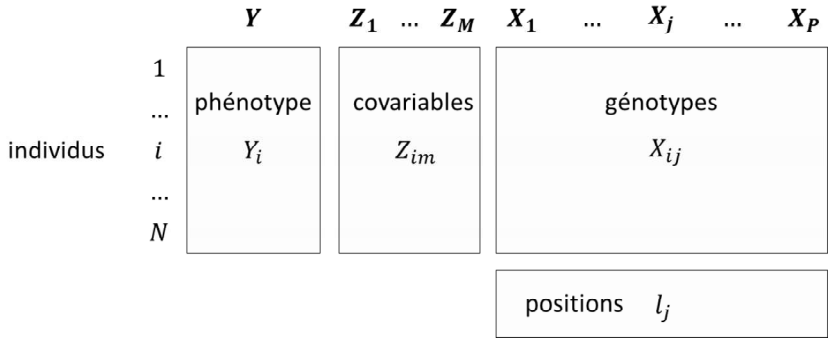


Figure 25. Structure des données et notations

Par la suite nous utiliserons A (« *Affected* ») pour désigner les cas et U (« *Unaffected* ») les témoins. Les effectifs des populations A et U sont alors respectivement $N^A = card(A)$ et $N^U = card(U)$. $N = N^A + N^U$ correspond au nombre d'individus dans la population d'étude totale.

Par la suite :

- m_j^A est le nombre de mutations rares pour le variant j chez les cas : $m_j^A = \sum_{i \in A} X_{ij}$
- m_j^U, est le nombre de mutations rares pour le variant j chez les témoins : $m_j^U = \sum_{i \in U} X_{ij}$

I.2- PRINCIPALES CATÉGORIES DE TEST

Tests « burden »

Parmi les méthodes statistiques, une catégorie de tests est très citée, les « *burden tests* », aussi appelés « *pooled tests* ». Le terme « *burden* » peut se traduire par fardeau génétique. La stratégie générale de ces tests est d'agréger les nombres d'allèles rares afin d'obtenir des effectifs plus importants.

Sum test

Une façon d'agréger les mutations est de résumer l'information génétique d'un individu par un score génétique. Dans le cadre du « *Sum test* », ou parfois appelé « *burden test* » dans certaines publications, le score d'un individu est le nombre d'allèles rares présents dans le gène. Un exemple est schématisé dans la Figure 26.

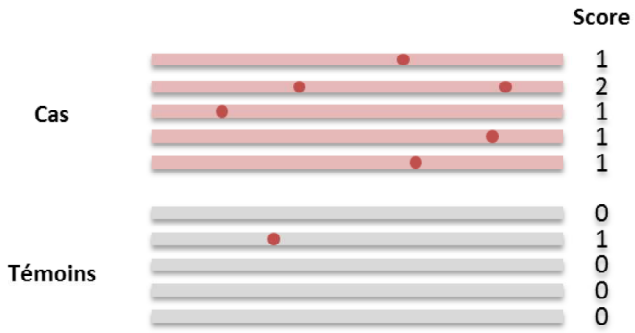


Figure 26. Score par individu dans le cadre du *Sum test*

En reprenant les notations, le score génétique S_i de l'individu i s'écrit alors de la façon suivante :

$$S_i = \sum_{j=1}^{P} X_{ij} \qquad \text{(I.2.1)}$$

L'association entre ce score génétique et la maladie étudiée peut être testée en se basant sur le modèle de régression logistique suivant :

$$logit\big(P(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i})\big) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta \sum_{j=1}^{P} X_{ij} \qquad \text{(I.2.2)}$$

avec α_0 le risque de base, β le coefficient de régression pour l'effet du score et $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)$ le vecteur des coefficients de régression pour les covariables. La composante génétique testée dans le modèle est signalée en rouge. Nous avons indiqué les covariables dans le modèle, mais les analyses effectuées dans cette Partie I n'incorporent pas d'information extérieure.

Dans le cadre d'un modèle de régression logistique, l'hypothèse nulle est $H_0 : \beta = 0$. La statistique du test de score peut être par exemple employée (voir la partie **Récapitulatif des burden tests et test du score p57**).

Pour ce test parfois appelé « *sum test* » ou RVT1 par Morris et Zeggini (2010) [94], il est supposé que les effets sont les mêmes pour tous les variants rares présents dans le gène et sont additifs. Cependant il est très improbable que tous les variants aient la même conséquence.

wSum

La méthode « *weighted sum*» (wSum) de Madsen et Browning (2009) [98], est similaire au test **Sum** (p52) avec un système de pondération des variants génétiques pour moduler les effets des variants selon leur degré de susceptibilité d'être lié à la maladie. Le score calculé pour l'individu i est alors

$$S_i = \sum_{j=1}^{P} w_j X_{ij} \qquad \text{(I.2.3)}$$

avec w_j le poids accordé au variant j. Le poids décrit par Madsen et Browning est une fonction de la fréquence allélique. Il s'écrit $w_j = \frac{1}{\sqrt{N.\widehat{MAF_j^U}\left(1-\widehat{MAF_j^U}\right)}}$ avec $\widehat{MAF_j^U}$, la MAF du variant j estimée à partir des témoins. Ce poids permet d'accorder plus de poids aux variants rares de façon à ce que le test ne soit pas totalement influencé par les variants fréquents.

Madsen et Browning (2009), utilise un test des rangs de Wilcoxon pour comparer les scores entre les cas et les témoins. La significativité de ce test peut être calculée au moyen d'une procédure de permutations des phénotypes. La statistique de test est supposée suivre approximativement une loi normale de moyenne μ et écart-type σ (estimés au préalable à l'aide de permutations). En se basant sur une loi approchée sous H_0, beaucoup moins de permutations sont nécessaires pour atteindre des p-values faibles, en comparaison avec la procédure classique de permutations.

En intégrant ce score dans un modèle de régression logistique, on obtient:

$$logit\big(P(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i})\big) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta \sum_{j=1}^{P} w_j X_{ij} \qquad \text{(I.2.4)}$$

L'association entre le score génétique est la maladie peut être aussi testée, en considérant le modèle de régression logistique avec pour hypothèse nulle $H_0 : \beta = 0$.

aSum

La méthode « *data-adaptive sum test* » (aSum) [103] proposé par Han et Pan en 2010, compare aussi des scores génétiques entre cas et témoins avec la prise en compte de variants protecteurs dans le système de pondération. En reprenant le score (I.2.4) présenté pour le test **wSum** (p53), les poids w_j sont égaux à 1 ou -1 selon si le variant est considéré comme à risque ou protecteur.

Afin de savoir si un variant est protecteur ou non, un test simple marqueur (test du score) est réalisé selon un modèle de régression logistique univarié suivant :

$$logit\big(P(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i})\big) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta_j X_{ij} \qquad \text{(I.2.5)}$$

Si le coefficient de corrélation $\beta_j < 0$ et la p-value est inférieure au seuil α alors le variant est considéré comme protecteur ($w_j = -1$) sinon il est considéré comme à risque ($w_j = 1$). Le seuil choisi est $\alpha = 0.10$ selon les recommandations énoncées par Basu et Pan (2011) [164].

Un test de score est ensuite appliqué pour évaluer l'association entre le score génétique S et le phénotype Y (voir Analyse d'association, p27). La significativité est évaluée par méthode de permutations des phénotypes.

VT

Le test "*variable threshold*" (VT), développé par Price et al. (2008), permet de faire une sélection adaptative des variants rares pris dans l'analyse. Au lieu de filtrer les variants rares sur un seuil de MAF fixe qui est arbitraire, plusieurs seuils de MAF sont considérés, d'où le terme d'approche avec un seuil variable. Les poids sont égaux à 0 ou 1 selon la MAF du variant rare. Le score génétique peut s'écrire de la façon suivante :

$$S_{t,i} = \sum_{j=1}^{P} w_j(t) X_{ij} \qquad \text{(I.2.6)}$$

avec $w_j(t) = I\left(MAF_j \leq t\right)$, t étant le seuil de MAF variable.

Price et al. (2008) calcule une statistique de test pour le seuil t, qui ressemble beaucoup à la statistique du test du score :

$$Q(t) = \frac{U(t)}{V^*(t)^{\frac{1}{2}}} \qquad \text{(I.2.7)}$$

avec $U(t) = \boldsymbol{S_t'}(\boldsymbol{Y} - \overline{\boldsymbol{Y}})$ et $V^*(t) = (\boldsymbol{S_t} - \overline{\boldsymbol{S}}_{\boldsymbol{t}})'(\boldsymbol{S_t} - \overline{\boldsymbol{S}}_{\boldsymbol{t}})$ avec S_t le vecteur des scores pour le seuil t pour l'ensemble des individus et $\bar{S}_t$ la moyenne de ce score. La statistique finale du test VT est $Q_{max} = \max_t Q(t)$.

CAST

La test « *cohort allelic sum test* » (CAST) [93], développé par Morgenthaler et Thilly (2007), est dit « *collapsing* » car le score génétique est transformé en variable binaire. Ce test permet de comparer la proportion de personnes porteuses d'au moins une mutation rare pour le gène chez les cas et les témoins. Le score génétique est alors binaire, il est égal à 1 si la personne est porteuse d'au moins une mutation rare dans le gène (Figure 27).

Le score génétique C s'écrit alors de la manière suivante :

$$C_i = \mathrm{I}\left(\sum_{j=1}^{P} X_{ij} > 0\right) = I(S_i > 0) \qquad \text{(I.2.8)}$$

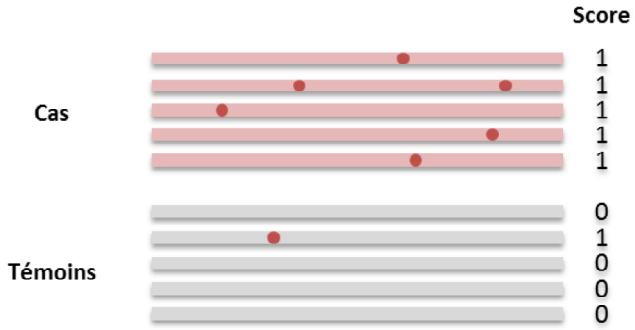


Figure 27. Score « burden » pour le test CAST

L'association peut être testée de différentes façons. Elle peut être testée au moyen d'un modèle de régression logistique (test du score ou test du rapport de vraisemblance). Le modèle s'écrit alors :

$$logit\left(P(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i})\right) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta * \mathrm{I}\left(\sum_{j=1}^{P} X_{ij} > 0\right) \qquad \text{(I.2.9)}$$

Mais le test le plus couramment utilisé, car adapté aux petits effectifs, est le test exact de Fisher. Ce test s'effectue sur la table de contingence 2x2 suivante :

Tableau 4. Table de contingence 2 x 2 pour le test exact de Fisher dans le cadre du test CAST

	Cas	Témoin
Porteur d'un variant rare	$\sum_{i=1}^{N^A} C_i$	$\sum_{i=1}^{N^U} C_i$
Non porteur d'un variant rare	$N^A - \sum_{i=1}^{N^A} C_i$	$N^U - \sum_{i=1}^{N^U} C_i$

Le test « *Combined Multivariate and Collapsing* » (CMC), développé par Li et Leal (2008) reprend le concept de « *collapsing* » du test CAST [92]. Dans ce test, les variants fréquents et les variants rares du gène sont pris en compte. L'étape de « *collapsing* » est réalisée sur les variants rares avec la création d'une nouvelle variable $C_i, i \in \{1, \ldots, N\}$ identique à celle de

CAST. Cette variable C et les variables correspondant aux génotypes pour les variants fréquents sont testées par une approche multivariée. Le test décrit est le test de Hotelling.

Dans la littérature, il est très fréquent que le test CMC s'effectue uniquement sur les variants rares. Il est alors identique au test CAST. Le test RVT2 développé par Morris et Zeggini (2010) [94] consiste en un test du rapport de la vraisemblance pour le modèle présenté dans l'équation (I.2.9). Par la suite nous nous référons au test CAST lors de ce type d'hypothèse nulle.

Récapitulatif des *burden tests* et test du score

Les différents tests présentés précédemment calculent tous un score génétique S_i par individu i (voir

Tableau 5). Afin de tester l'association entre ce score génétique et la maladie, nous pouvons considérer le modèle de régression logistique suivant :

$$logit\left(P(Y_i = 1|S_i, \boldsymbol{Z_i})\right) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta S_i \qquad \text{(I.2.10)}$$

Tableau 5. Récapitulatif du score génétique utilisé pour chaque « *burden test* »

Test	Score génétique	Détail
Sum	$S_i = \sum_{j=1}^{P} w_j X_{ij}$	$w_j = 1$
wSum		$w_j = \frac{1}{\sqrt{N.\widehat{MAF_j^U}\left(1-\widehat{MAF_j^U}\right)}}$
aSum		$w_j = \begin{cases} -1 & \text{si variant protecteur} \\ 1 & \text{si variant à risque} \end{cases}$
VT	$S_{VT_i}(t) = \sum_{j=1}^{P} w_j(t) X_{ij}$	$w_j = I(MAF_j \leq t)$
CAST	$S_{CAST_i} = I\left(\sum_{j=1}^{P} X_{ij} > 0\right)$	

L'hypothèse nulle pour tester l'effet de ce score génétique sur la probabilité d'être atteint est $H_0: \beta = 0$. Cette hypothèse nulle peut être testée avec le test du score [165,166]. La statistique de score U et sa variance V sont :

$$U = \boldsymbol{S}'(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}})$$

$$V = \frac{1}{N-1}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}})'(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}}) * (\boldsymbol{S} - \overline{\boldsymbol{S}})'(\boldsymbol{S} - \overline{\boldsymbol{S}})$$

avec $\hat{\mu} = logit^{-1}(\widehat{\alpha_0} + \boldsymbol{Z}_i'\widehat{\boldsymbol{\alpha}})$ et and $\overline{\boldsymbol{S}} = \left(\sum_{i=1}^{N} \frac{S_i}{N}\right)_N$. La statistique de test vaut alors $Q = \frac{U^2}{V}$. Dans le cas où il n'y a pas de covariable $\mathbf{Z}$, $\widehat{\boldsymbol{\mu}}_{\boldsymbol{N}} = \left(\sum_{i=1}^{N} \frac{Y_i}{N}\right)_N$ correspond à la proportion de cas dans les données, ou encore à la moyenne de la variable Y.

Variance-component tests

Des tests appelés « *variance component tests* » permettent de considérer des situations où les variants rares ont des effets très hétérogènes. En effet beaucoup de tests « *burden* » ne sont pas adaptés à la présence de variants avec des effets opposés. Les variants dits protecteurs diminuent le risque d'être atteint de la maladie, tandis que les variants à risque augment ce risque.

C-alpha

Le test C-alpha [126], proposé par Neale et al. en 2011, est une approche permettant de détecter des mélanges d'effets (neutre, protecteur et à risque) pour un groupe de variants rares.

Pour un variant rare j observé m_j fois, on suppose que, m_j^A, le nombre de mutations rares chez les cas pour un variant j donné, suit une loi binomiale de paramètres m_j, p_j. Sous H_0, $p_j = p_0 \; \forall j \in \{1, \dots, P\}$, p_0 étant la probabilité qu'une mutation rare soit présente aléatoirement chez les cas ou chez les témoins ($= 0.5$ lors d'un nombre équilibré de cas et de témoins). L'hypothèse alternative est qu'il existe un mélange de distributions au sein du groupe de variants rares, certains sont neutres ($p_j = p_0$), certains à risque ($p_j > p_0$) et certains protecteurs ($p_j < p_0$).

Le test C-alpha repose sur le principe qu'un mélange de distributions au sein du groupe conduit à une augmentation de la variance. La statistique de test T compare ainsi les variances observées et attendues pour chaque variant j sous H_0.

$$CALPHA = \sum_{j=1}^{P} \left[\left(m_j^A - m_j p_0 \right)^2 - m_j p_0 (1 - p_0) \right]$$

La distribution de la statistique de test sous H_0est déterminée par méthode de permutations des phénotypes, permettant de calculer la probabilité critique.

SKAT

Le test « *sequence kernel association test* » (SKAT) [95], proposé par Wu et al. en 2011, est une généralisation du test C-alpha. Il a été développé par Wu et al. (2011) pour tester l'effet joint de multiples variants rares et fréquents présents au sein d'une région sur un phénotype.

Modèle linéaire

Avant de considérer le modèle général proposé par la méthode SKAT, il est plus simple d'introduire modèle linéaire. Soit le modèle de régression logistique suivant :

$$logit(P(Y_i = 1 | \boldsymbol{X_i}, \boldsymbol{Z_i})) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \boldsymbol{X_i'\beta} \qquad \text{(I.2.11)}$$

qui peut aussi s'écrire de cette façon

$$logit(P(Y_i = 1 | \boldsymbol{X_i}, \boldsymbol{Z_i})) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \sum_{j=1}^{P} \beta_j w_j X_{ij} \qquad \text{(I.2.12)}$$

avec α_0, l'intercepte ; $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)$ le vecteur des coefficients de régression pour les covariables $\boldsymbol{Z_m}, m \in \{1, \dots, M\}$; $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)$ le vecteur des coefficients de régression pour les effets génétiques des P variants ; et $w_j, j \in \{1, \dots, P\}$ les poids des variants.

Dans ce modèle, pour augmenter la puissance du test, Wu et al. (2011) supposent que les effets génétiques sont aléatoires. Les coefficients β_j suivent une distribution arbitraire de moyenne $E(\beta_j) = 0$ et de variance $V(\beta_j) = w_j \tau$. De plus les effets génétiques sont considérés non-corrélés deux à deux avec $\forall\, j \neq j'\ corr(\beta_j, \beta_{j'}) = 0$.

L'hypothèse nulle pour le test SKAT est H_0: $\boldsymbol{\beta} = 0$, i.e. il n'y aucun effet génétique quel que soit le variant. Car les effets génétiques sont supposés aléatoires et de moyenne nulle, l'hypothèse nulle est alors équivalente à $H_0 : \tau = 0$. SKAT permet alors de tester la variance des effets génétiques.

La statistique de test est :

$$Q = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}})'\mathbf{K}(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{N}) \qquad \text{(I.2.13)}$$

où $\mathbf{K} = \mathbf{XWWX}'$ est une matrice de dimension $N \times N$ avec $\mathbf{W} = diag(w_1, \dots, w_P)$. La matrice diagonale $\mathbf{W}$ permet de pondérer les variants. La prise en compte des covariables est permise grâce à $\hat{\mu} = logit^{-1}(\alpha_0 + \boldsymbol{Z_i\alpha})$.

La pondération est une fonction de la MAF et permet d'accorder plus de poids aux variants rares dans l'analyse, comme pour la pondération de Madsen et Browning (2009) [98] dans le cadre du test wSum. Le poids pour le variant j est $w_j = Beta(\widehat{MAF_j}; 1{,}25)$ avec $\widehat{MAF_j}$, la MAF du variant j estimée à partir de l'ensemble des individus. Les paramètres de la loi sont choisis par Wu et al. (2011) afin d'augmenter le poids des variants rares tout en gardant des coefficients raisonnables pour des variants avec une MAF comprise entre 1% et 5%.

Sous H_0, Q suit un mélange de distributions du chi-deux, qui peut être approché par la méthode Davies [167].

On peut aussi noter que le modèle de régression linéaire généralisé sans système de pondération s'apparente au test envisagé plus tôt par Pan (2009)[168], et s'appelant SSU (*sum of squared scores*). Ce test a été développé pour prendre en compte des variants avec des effets différents, ce qui n'est pas le cas avec le test Sum. La statistique de test s'écrit dans le cadre du SSU

$$\begin{aligned} SSU &= (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}})'\mathbf{XX}'(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{N}) \\ &= \boldsymbol{UU}' \end{aligned}$$

avec $\boldsymbol{U} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}})'\mathbf{X} = \left(U_1, \dots, U_j, \dots, U_P\right)'$, le vecteur des statistiques du score pour le test simple marqueur de chaque variant $j \in \{1, \dots, P\}$.

Modèle général

Le modèle présenté précédemment est un modèle linéaire. Le test SKAT permet de considérer des modèles plus généraux de la forme :

$$logit(P(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i})) = \alpha_0 + \boldsymbol{Z_i'\alpha} + f(\boldsymbol{X_i}) \quad \text{(I.2.14)}$$

avec $f(\boldsymbol{X_i})$ une fonction des génotypes de l'individu i. Ce modèle permettant d'expliquer la probabilité d'être atteint en fonction des génotypes est flexible et est déterminé par le choix de la matrice noyau **K** (*kernel*).

K est le noyau qui détermine la forme du modèle général. Cette matrice noyau de dimension $N \times N$ peut être considérée comme une matrice de similarité génétique entre individus. Les auteurs proposent par exemple la matrice de similarité génétiques IBS, avec chaque élément de la matrice correspondant au nombre d'allèles rares partagés entre les individus.

Nous avons choisi d'étudier le test SKAT pour le modèle linéaire, étant le plus utilisé (en pratique il semble être exclusivement utilisé) et le plus facile à comprendre.

Combinaison de stratégies

SKAT-O

Il a été montré que dans le cas où la majorité des mutations au sein d'un gène étaient à risque, SKAT pouvait se révéler moins puissant que les tests *burden* [164]. C'est pourquoi Lee et al., en 2012, ont proposé une amélioration de SKAT, SKAT-O [169] (O : *optimal*), afin d'augmenter la puissance de détecter une association à l'aide d'une combinaison d'un *burden test* et d'un *variance component test*.

SKAT-O est basé sur le modèle de régression logistique que SKAT. Il fait intervenir un paramètre supplémentaire ρ, qui détermine la structure de corrélation entre les effets génétiques : $\forall j\ corr(\beta_j, \beta_{j'}) = \rho$. Plusieurs valeurs de ce paramètre sont testées afin de maximiser la puissance.

La statistique de test devient :

$$Q_\rho = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}})' \mathbf{K}_{\boldsymbol{\rho}} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}}) \quad \text{(I.2.15)}$$

avec $\mathbf{K_\rho} = \mathbf{XWR_\rho WX'}$ pour le modèle de régression linéaire où $\mathbf{R_\rho} = (1-\rho)\boldsymbol{I_P} + \rho\mathbf{1_P}\mathbf{1_P}'$ et $\mathbf{W} = diag(w_1, \dots, w_P)$ et $\hat{\mu} = logit^{-1}(\alpha_0 + \boldsymbol{Z_i\alpha})$.

Plus simplement, $\boldsymbol{R_\rho}$ est une matrice dont les valeurs sur la diagonale sont égales à 1 et en dehors de la diagonale sont égales à ρ.

Voici dans le tableau ci-dessous deux cas particuliers de la statistique de test (I.2.15) :

Tableau 6. Cas particuliers de la statistique de test de SKAT-O.

Condition	Statistique	Test
$\rho = 0$	$Q_0 = \sum_{j=1}^{P} w_j^2 \left[\sum_{i=1}^{N}(Y_i - \bar{Y})X_{ij}\right]^2$	*linear SKAT*
$\rho = 1$	$Q_1 = \left[\sum_{j=1}^{P} w_j^2 \sum_{i=1}^{N}(Y_i - \bar{Y})X_{ij}\right]^2$	*burden test*

SKAT-O est bien une combinaison entre SKAT et un test *burden*. En application ρ est déterminé parmi la série de valeurs allant de 0 à 1 avec un pas de 0.1 telle que la probabilité critique soit la plus faible. La statistique finale est, en effet, le minimum des p-values obtenues pour chaque valeur de ρ. La distribution sous H_0 de la statistique est approchée par la méthode décrite par Lee et al. (2012).

MiST

Le test « *mixed effects test* » (MiST)[100], développé par Sun et al. (2013), a pour objectif de fournir un modèle général pour tester l'association entre des variants rares et un phénotype. Dans le cadre d'un modèle de régression logistique, ce test repose, comme pour le test SKAT, sur le modèle général suivant :

$$logit(P(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i})) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \sum_{j=1}^{P} \beta_j w_j X_{ij} \qquad \text{(I.2.16)}$$

avec α_0, l'intercept, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_M)$ le vecteur des coefficients de régression pour les covariables $\boldsymbol{Z_m}, m \in \{1, \dots, M\}$ et $\boldsymbol{\beta} = (\beta_1, \dots, \beta_P)$ le vecteur des coefficients de régression pour les P variants présents au sein du gène.

Avec la méthode MiST, les effets génétiques sont décomposés en une composante fixe et une composante aléatoire. On considère pour la composante fixe que les variants partageant les mêmes caractéristiques ont le même effet. Soit **B** la matrice de dimension $P \times K$, avec B_{jk} la caractéristique k du variant j. Les effets génétiques peuvent être écrits de cette façon :

$$\beta_j = \boldsymbol{B_j'\pi} + \delta_j \qquad \text{(I.2.17)}$$

avec $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ le vecteur des coefficients de régression pour les effets génétiques fixes et δ_j la composante génétique aléatoire qui suit une distribution de moyenne 0 et de variance τ^2. En comparaison avec le test SKAT, la matrice de caractéristiques **B** peut par exemple correspondre au vecteur des poids de SKAT dépendant de MAF.

Le modèle de régression logistique (I.2.16) peut se réécrire en incorporant (I.2.17) :

$$logit(P(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i})) = \alpha_0 + \boldsymbol{Z_i'\alpha} + (\boldsymbol{X_i'B})\boldsymbol{\pi} + \boldsymbol{X_i'\delta} \qquad \text{(I.2.18)}$$

Voici deux cas particuliers du modèle (I.2.18) :

Condition	**Modèle**	**Test**
$\boldsymbol{\delta = 0}$ $\boldsymbol{B} = (w_1, \dots, w_P)'$	$logit(P(Y_i = 1)) = \alpha_0 + \boldsymbol{Z_i\alpha} + \pi \sum_{j=1}^{P} w_j X_{ij}$	*burden test (*wSum*)*
$\boldsymbol{\pi = 0}$	$logit(P(Y_i = 1)) = \alpha_0 + \boldsymbol{Z_i\alpha} + \boldsymbol{X_i'\delta}$	*linear* SKAT

L'hypothèse nulle correspond à l'absence d'association entre le phénotype et les variants génétiques. Elle s'écrit $H_0: \boldsymbol{\pi = 0}\ et\ \boldsymbol{\tau^2 = 0}$. Les statistiques de test pour chacune des composantes génétiques, fixe et aléatoire, sont calculées indépendamment et ensuite combinées.

Pour la composante fixe, la statistique du score est calculée, préférée au rapport de vraisemblance. La statistique pour l'hypothèse nulle $H_0: \boldsymbol{\pi = 0}$, en supposant que $\boldsymbol{\delta = 0}$ est :

$$U_\pi = (\mathbf{XB})'(\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}) \quad \text{(I.2.19)}$$

avec $\widehat{\mu_i} = logit^{-1}(\alpha_0 + Z_i\alpha)$.

Pour la composante aléatoire des effets génétiques, une statistique du score quadratique, similaire à SKAT, est calculée pour tester la composante de la variance τ^2 des effets génétiques. La statistique correspondant à l'hypothèse nulle H_0: $\boldsymbol{\tau}^2 = \mathbf{0}$ est :

$$Q_{\tau^2} = (\boldsymbol{Y} - \hat{\mu})'\mathbf{XX}'(\boldsymbol{Y} - \hat{\mu}) \quad \text{(I.2.20)}$$

avec $\hat{\mu}_i = logit^{-1}(\alpha_0 + Z_i\alpha + (X_i'\mathrm{B})\pi)$.

Les statistiques de test U_π et Q_{τ^2} sont indépendantes et donc peuvent être combinées. La p-value de chacune des deux statistiques est calculée grâce à une approximation de la distribution de la statistique sous l'hypothèse nulle avec la méthode de Liu et al. (2009) [170]. Les p-values sont ensuite combinées par la méthode de Fisher (p64) pour obtenir la significativité du test général.

Tests de combinaison des p-values

Certains tests ont pour stratégie de tester les p-values des tests simple-marqueur. Dans cette étude nous avons choisi d'utiliser le test ADA. Afin d'en expliquer le principe de manière plus claire, la méthode de Fisher est aussi décrite bien que non considérée dans la comparaison des tests.

Méthode de Fisher

La méthode de Fisher [171] est une méthode combinant l'information de tests indépendants. Elle peut être employée pour tester un groupe de variants génétiques en combinant les p-values obtenues pour chaque test simple-marqueur. Soient $p_j, j \in \{1, \dots, P\}$ les p-values obtenues pour les variants rares du gène, la statistique de la méthode de Fisher est la suivante :

$$S = -2\sum_{j=1}^{P} \log(p_j) \quad \text{(I.2.21)}$$

Sous l'hypothèse nulle, les p-values sont considérées comme les réalisations d'une variable aléatoire suivant une loi uniforme $U(0,1)$. La méthode de Fisher permet de tester s'il y a un

effet global des variants rares sur la susceptibilité de la maladie. Lorsque les tests sont mutuellement indépendants, la statistique de test S suit de manière asymptotique une loi du χ^2 à $2P$ degrés de liberté.

ADA

Le test ADA [123], dont le nom complet est « *adaptive combination of P-values for rare variant association testing* »,proposé par Lin et al. en 2014, est dérivé de la méthode de Fisher et du test sigma-P [156], et intègre les variants de manière adaptative selon leur degré de significativité. Les variants rares sont testés individuellement et sont inclus dans l'analyse s'ils présentent une p-value inférieure à un seuil donné. Le seuil arbitraire de 0.05 étant peut-être trop « contraignant », celui-ci est alors déterminé de manière adaptative.

Soient $p_j, j\epsilon\{1, \dots, P\}$ les p-values obtenues en testant individuellement les variants rares (par exemple avec le test exact de Fisher). On pose T seuils de p-value possibles $\theta_1, \theta_2, \dots, \theta_t, \dots, \theta_T$ en dessous desquels on inclut les variants.

De plus, afin de distinguer le sens de l'effet des variants sur la susceptibilité de la maladie, deux statistiques S_t^+ et S_t^-, pour le seuil θ_t, sont calculées respectivement pour les variants enclin à être à risque et enclins à être protecteurs. La statistique de test finale est le maximum des deux statistiques S_t^+ et S_t^-.

La statistique pour les variants « potentiellement à risque » et pour le seuil t est

$$S_t^+ = -\sum_{j=1}^{P} \xi_j . I(p_j \leq \theta_t) . w_j . \log(p_j) \qquad \text{(I.2.22)}$$

avec w_j le poids accordé au variant j qui est celui décrit par Madsen and Browning (2009) pour le test wSum, $\xi_j = I(MAF_j^A \geq MAF_j^U)$ un indicateur du variant « potentiellement à risque ». De même la statistique pour les variants « potentiellement protecteurs » est $S_t^- = -\sum_{j=1}^{P}(1-\xi_j) . I(p_j \leq \theta_t) . w_j . \log(p_j)$. La statistique de test pour le seuil t est $S_t = \max(S_t^+, S_t^-)$.

On choisit le seuil de significativité de manière adaptative. Cela veut dire que le seuil de troncature t gardé est celui qui minimise la p-value p_t pour la statistique S_t. Le système d'évaluation de la significativité est complexe car se déroule en plusieurs étapes :

(i) La significativité de la statistique S_t, pour le seuil t, est évaluée par méthode de permutations des phénotypes. La p-value est alors $p_t = \frac{\left[\sum_{b=1}^{B} I\left(S_t^{(b)} \geq S_t\right)\right]+1}{B+1}$ avec les permutations $b \in \{1, \dots, B\}$.

(ii) La p-value minimale pour les T seuils de troncature est alors $MinP = \min_{t\in\{1,\dots,T\}} \frac{\left[\sum_{b=1}^{B} I\left(S_t^{(b)} \geq S_t\right)\right]+1}{B+1}$.

(iii) Cette p-value minimale $MinP$ ne peut être gardée telle quelle car cela entraînerait un grand taux d'erreurs. La distribution de la statistique de test $MinP$ sous H_0 est déterminée en calculant pour chaque permutation b, la p-value minimale $MinP^{(b)} = \min_{t\in\{1,\dots,T\}} \frac{\sum_{b\neq b'} I(S_t^{(b')} \geq S_t^{(b)})}{B}$. Il s'agit de comparer la p-value minimale obtenue avec la permutation b en utilisant les autres permutations $b' \neq b$.

(iv) La p-value finale est alors $\frac{\left[\sum_{b=1}^{B} I\left(MinP^{(b)} \geq MinP\right)\right]+1}{B+1}$.

L'introduction de covariables dans le test peut être effectuée lors du test simple marqueur avec l'utilisation du modèle de régression logistique. Toutefois la méthode implémentée par l'auteur sous R est le test exact de Fisher.

KBAC

Le test « *Kernel-Based Adaptive Clustering*» (KBAC) [109], proposé par Liu et Leal en 2010, est une méthode développée afin de mieux détecter les signaux d'associations en présence de nombreux variants rares neutres (non-causaux).

La grande différence avec les autres méthodes est la considération du génotype pour l'ensemble des variants et est appelé par la suite « **génotype multilocus** » (Figure 28). Soit $X_{KBAC_l}, l \in \{0, \dots, L\}$, les différents génotypes multilocus présentes au sein de la population étudiée, X_{KBAC_0} étant le génotype d'un individu ne présentant aucune mutation pour l'ensemble des P variants.

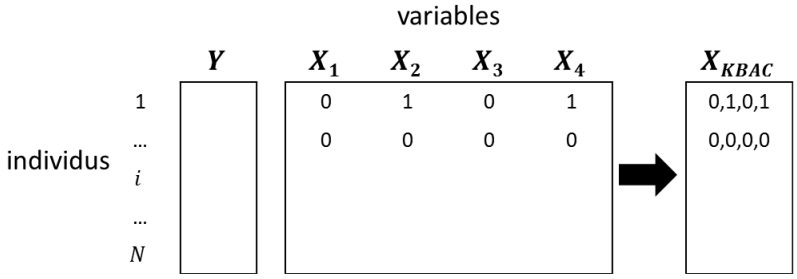


Figure 28. Construction de la variable X_{KBAC}.

X_{KBAC} est le génotype pour l'ensemble des variants 1 à 4 dans cet exemple.

KBAC compare les fréquences des génotypes multilocus entre les cas et les témoins. Sous l'hypothèse nulle, celles-ci sont identiques. L'hypothèse alternative, dans le cas bilatéral, est qu'il existe au moins un génotype multilocus pour lequel la fréquence est différente chez les cas et chez les témoins.

La statistique dans le cadre du test bilatéral est :

$$KBAC = \left(\sum_{l=1}^{L} \mathrm{w}_l \left(\frac{N_l^A}{N^A} - \frac{N_l^U}{N^U} \right) \right)^2 \qquad \text{(I.2.23)}$$

où w_l est un poids attribué au génotype multilocus X_{KBAC_l}.Il permet d'attribuer plus d'importance aux génotypes multilocus qui sont enrichis chez les cas, c'est-à-dire avec une probabilité plus grande d'être présent chez les cas. Ceci permet de mieux différencier les génotypes causaux des génotypes non-causaux.

Le risque de tirer un génotype l chez les cas, est estimé dans les données de la façon suivante :

$\widehat{R_l} = \frac{N_l^A}{N_l}$.

La densité de ce risque R_l sous l'hypothèse H_0 est déterminée par ce qu'on appelle une fonction noyau notée k_l^0. Dans le cadre de faibles effectifs, le noyau hypergéométrique est préféré. Le nombre de cas présentant le génotype l, N_l^A,sous H_0 suit la loi hypergéométrique $\mathcal{H}\left(N^A, \frac{N_l}{N}, N\right)$ et la fonction de densité de la variable aléatoire R_l est :

$$k_l^0(r_l) = P(R_l = r_l) = \frac{\binom{N_l}{N_l r_l}\binom{N - N_l}{N^A - N_l r_l}}{\binom{N}{N^A}}$$

Avec un noyau binomial, le nombre de cas présentant le génotype l, N_l^A,sous H_0suit une loi binomiale $\mathcal{B}\left(N_l, \frac{N^A}{N}\right)$. Ainsi

$$k_l^0(r_l) = P(R_l = r_l) = \binom{N_l}{N_l r_l}\left(\frac{N^A}{N}\right)^{N_l r_l}\left(1 - \frac{N^A}{N}\right)^{N_l(1-r_l)}$$

Le poids w_l est :

$$\mathrm{w}_l = \sum_{r_l \in \left\{\frac{0}{N_l}, \ldots, \widehat{R_l}\right\}} k_l^0(r_l) dr_l \qquad \text{(I.2.24)}$$

La p-value est calculée à partir de la distribution de la statistique sous H_0 obtenue par procédure de permutations des phénotypes.

Tests incorporant les positions

BOMP

Le test « *Burden Or Mutation Position test* » (BOMP) [37], proposé par Chen et al. en 2013, est la combinaison de deux tests : un test de type *burden* ($\mathrm{BOMP}_{\mathrm{burden}}$) qui compare le nombre d'individus portant au moins T mutations, T étant un seuil variable, et un test qui compare la distribution des positions des mutations sur le gène entre les cas et les témoins ($\mathrm{BOMP}_{\mathrm{position}}$). Les statistiques de ces deux tests sont basées sur les rapports de vraisemblances entre les modèles sous H_1 et sous H_0. Le test BOMP utilise la stratégie de calculer un score *burden* par fenêtre glissante sur le gène afin de tenir compte des positions des variants.

Test « burden » : $\mathrm{BOMP}_{\mathrm{burden}}$

Le test reprend la stratégie des tests « *burden* » avec le calcul d'un score génétique $\mathrm{S_i}$ par individu :

$$S_i = \sum_{j=1}^{P} X_{ij} \qquad \text{(I.2.25)}$$

Une variable binaire,$C_{BOMP}(T)$, est ensuite créée afin d'indiquer si ce score dépasse un seuil T (analogie : $T = 1$ pour CAST) :

$$C_{BOMP_i}(T) = \begin{cases} 1 & si\ S_i \geq T \\ 0 & si\ S_i < T \end{cases} \quad \text{(I.2.26)}$$

Pour le seuil T, on calcule la statistique $\Lambda_{burden}(T)$ qui correspond au logarithme du rapport de vraisemblance entre H_0 et H_1. On fait l'hypothèse que le nombre d'individus présentant au moins T mutations suit une loi de Bernoulli de paramètre p_T^A, p_T^U ou p_T pour les populations cas, témoin ou totale. Sous H_0, les probabilités qu'un individu présente au moins T mutations sont identiques chez les cas et les témoins, c'est-à-dire $p_T^A = p_T^U = p_T$. La statistique s'écrit :

$$\Lambda_{burden}(T) = \log\left(\frac{(\hat{p}_T^A)^{T_T^A}(1-\hat{p}_T^A)^{N^A - T_T^A}(\hat{p}_T^U)^{T_T^U}(1-\hat{p}_T^U)^{N^U - T_T^U}}{(\hat{p}_T)^{T_T}(1-\hat{p}_T)^{N-T_T}}\right)$$

avec T_T^A, T_T^U et T_T, les nombres d'individus présentant au moins T mutations chez les cas, les témoins et au total; $\hat{p}_T^A$, $\hat{p}_T^U$ et $\hat{p}_T$, les estimations des proportions.

Le seuil T est choisi tel qu'il maximise la statistique $\Lambda_{burden}(T)$ et la statistique de test est $\Lambda_{burden} = \max_T \Lambda_{burden}(T)$.

Test de la répartition des mutations : $BOMP_{position}$

Pour comparer la répartition des mutations sur le gène entre les cas et les témoins, un découpage du gène en M fenêtres est considéré.

Un score génétique est calculé par individu i et par fenêtre g_m , $m \in \{1, \dots, M\}$:

$$S_{i,m} = \sum_{j \in g_m} X_{ij} \quad \text{(I.2.27)}$$

Soit $\Lambda_{position}$ le logarithme du rapport de vraisemblance entre H_0 et H_1. On fait l'hypothèse que les nombres de mutations dans chaque fenêtre (Figure 29) suivent une loi multinomiale de paramètres $p_m^A, m \in \{1, \dots, M\}$ pour la population cas. Sous H_0, les nombres de mutations chez les cas et les témoins sont identiques quelle que soit la fenêtre g_m, c'est-à-dire $\forall\, m \in \{1, \dots, M\}\ p_m^A = p_m^U = p_m$.

La statistique s'écrit :

$$\Lambda_{position} = \log\left(\frac{\prod_{m=1}^{M}\left(\widehat{p_m^A}\right)^{T_m^A+1}\left(\widehat{p_m^U}\right)^{T_m^U+1}}{\prod_{m=1}^{M}\left(\widehat{p_m}\right)^{T_m+2}}\right)$$

avec T_m^A, T_m^U et T_m, les nombres de mutations présentes dans la fenêtre m chez les cas, chez les témoins et au total; $\hat{p}_m^A$, $\hat{p}_m^U$ et $\hat{p}_m$, les estimations des probabilités d'appartenir à une fenêtre m.

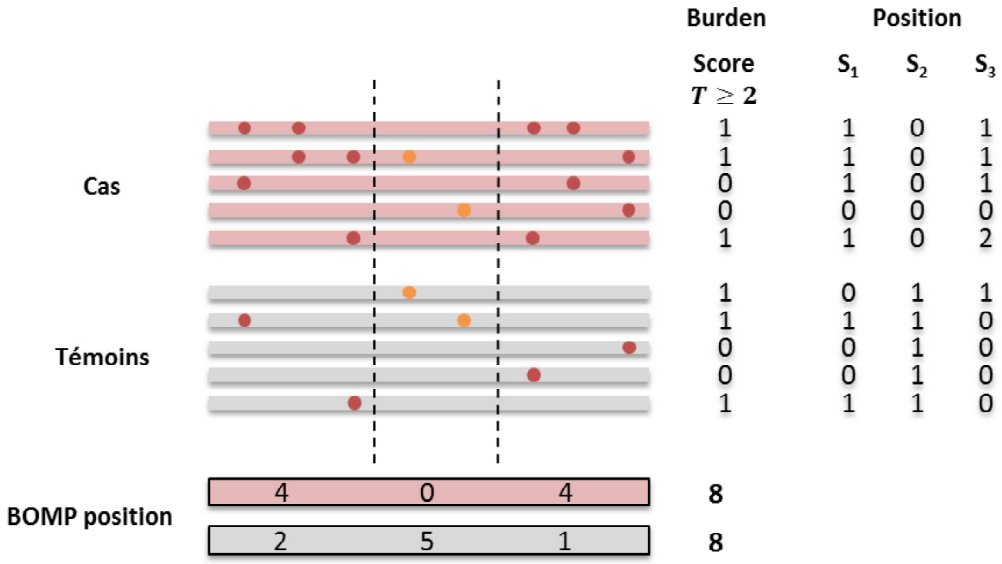


Figure 29. Calcul d'un score « *burden* » pour le gène et par région pour le test BOMP

Il s'agit d'un exemple de région génomique contenant des variants (ronds rouges et oranges) chez les cas et les témoins. On considère que cette région est importante pour la détermination du phénotype. Pour les méthodes BOMPposition, sont indiqués les nombres de mutations chez les cas (en rouge) et chez les témoins (en gris) pour chacune des trois fenêtres considérées. Les nombres de mutations chez les cas et les témoins sont identiques, la partie « *burden* » du test BOMP n'est pas en mesure de détecter une association. C'est la répartition des mutations qui est très importante dans ce cas-ci. Les probabilités de voir apparaître une mutation dans chacune des fenêtres sont clairement différentes entre les cas et les témoins.

Note : le découpage peut être déterminé par méthode de fenêtre glissante afin d'obtenir la statistique de test maximale.

Test BOMP total

Enfin, le test BOMP total combine les deux tests vus précédemment. La statistique de test Λ_{total} s'écrit alors $\Lambda_{total} = \Lambda_{burden} + \Lambda_{position}$

La distribution de Λ_{total} sous H_0 est obtenue par méthode de permutations des phénotypes. Il en est de même pour les statistiques de test Λ_{burden} et $\Lambda_{position}$.

DBM

Le test « *distance-based measure* » développé par Fier et al. (2012) [36], compare les distributions de distances inter-variants entre cas et témoins par le test d'Ansari-Bradley [172].

Soient $\boldsymbol{S^A}$ et $\boldsymbol{S^U}$ les séquences de positions pour les cas et les témoins. Plus précisément, dans ces vecteurs sont répétées les positions des variants $\left[m_j^A w_j\ \right]$ fois pour les cas et $\left[m_j^U w_j\ \right]$ fois pour les témoins (arrondis à l'entier). Les distances sont évaluées dans les vecteurs $\boldsymbol{D^A}$ et $\boldsymbol{D^U}$, en soustrayant deux éléments consécutifs des vecteurs $\boldsymbol{S^A}$ et $\boldsymbol{S^U}$. Un test d'Ansari-Bradley est effectué pour comparer les distributions des deux vecteurs de distances pondérées $\boldsymbol{D^A}$ et $\boldsymbol{D^U}$.

Pour souligner l'importance de la proximité spatiale entre les variants et pour contrôler une distribution irrégulière des fréquences alléliques; des poids dépendant des groupes (cas ou témoin) et des fréquences alléliques sont définis. Deux systèmes de poids sont utilisés en fonction des cas et des témoins. En se basant sur les fréquences alléliques estimées chez les cas, le poids est :

$$w_j^A = 1 + \frac{\left(\frac{m^A + 1}{m_j^A + 1}\right)}{\log\left(d_{j,min} + 1\right)}$$

avec $d_{j,min}$ la distance entre le variant j et son plus proche voisin. De même en se basant sur les fréquences estimées chez les témoins, le poids est :

$$w_j^U = 1 + \frac{\left(\frac{m^U + 1}{m_j^U + 1}\right)}{\log\left(d_{j,min} + 1\right)}$$

La statistique de test d'Ansari-Bradley est calculée pour les deux schémas de pondération décrits ci-dessus. La statistique de test finale est le maximum des deux statistiques. La significativité du test est évaluée par méthode de permutations des phénotypes.

KERNEL

Le test proposé par Schaid et al. (2013) [38] est inspiré de la méthode de regroupement spatial Tango [173]. Soit $\boldsymbol{\delta_{Kernel}}$ le vecteur de taille P des différences de fréquence allélique entres les cas et les témoins. La valeur pour un élément j de ce vecteur est :

$$\delta_{Kernel_j} = \frac{m_j^A}{\sum_{j=1}^{P} m_j^A} - \frac{m_j^U}{\sum_{j=1}^{P} m_j^U}$$

avec m_j^A et m_j^U les nombres d'allèles rares pour le variant j chez les cas et les témoins. L'information sur les positions des variants est contenue dans une matrice **A** de proximité entre les variants. La statistique de test a une forme quadratique et est :

$$\mathrm{Q} = \boldsymbol{\delta}'_{\boldsymbol{Kernel}}\mathbf{A}\boldsymbol{\delta}_{\boldsymbol{Kernel}}$$

La matrice **A** est aussi appelée matrice noyau car différentes mesures de proximité peuvent être calculées. La mesure utilisée par Schaid et al. (2013) est

$$A_{jj'}(c) = K\left(d'_{jj'}(c)\right) = \left(1 - d'_{jj'}(c)^2\right)^3$$

avec $d'_{jj'}(c) = \frac{d_{jj'}}{c \times maxd}$ avec $d_{jj'} = \left|l_{j'} - l_j\right|$ est la distance entre les variants j et j' en nombre de paires de bases ; $maxd$ est la distance maximale possible (renseignée par l'utilisateur : taille du gène considéré) ; c une constante prenant les valeurs allant de 0.1 à 1 avec un pas de 1. Cette transformation est nécessaire afin que $\left|d'_{jj'}\right| \leq 1$. La statistique de test finale est $Q_{max} = \max_c Q(c)$. La significativité du test est calculée avec une méthode de permutation des phénotypes.

CLUSTER

Le test CLUSTER est proposé par Lin (2014) [39], comme une extension du test ADA [123] (p65), et reprend la stratégie de Schaid et al. (2013) [38] (p72). Comme pour le test KERNEL, la statistique de test est de la forme :

$$Q = \boldsymbol{\delta}'_{\boldsymbol{CLUSTER}}\mathbf{A}\boldsymbol{\delta}_{\boldsymbol{CLUSTER}}$$

avec $\mathbf{A}$ la même matrice à noyau $P \times P$ de proximité entre les variants qui est utilisée pour le test KERNEL ; et $\boldsymbol{\delta_{CLUSTER}}$ un vecteur de taille P indiquant pour chaque variant génétique j une fonction de la p-value p_j du test simple-marqueur. Les éléments du vecteur reprennent les termes de la combinaison de p-values du test ADA. La valeur pour un variant j est :

$$\delta_{CLUSTER_j}(\mathrm{t}) = -\sqrt{\xi_j . I(p_j \leq \theta_t) . w_j . \log(p_j)}$$

avec w_j le poids accordé au variant j (voir le test wSum), $\xi_j = I\left(MAF_j^A \geq MAF_j^U\right)$ un indicateur du variant « potentiellement à risque ». Comme pour le test ADA, deux statistiques sont calculées pour les variants « potentiellement à risque » et les variants « potentiellement protecteurs ». La statistique de test est le maximum des deux statistiques. Il en est de même pour le seuil de p-value θ_t qui est déterminé de manière adaptative. La significativité est déterminée par un système de permutations (voir la description du test ADA).

PODKAT

Le test « *position-dependent kernel association test* » (PODKAT) est un test implémenté par Bodenhofer dans un package R. Ce test est inspiré des tests SKAT (p59) de Wu et al. (2011) [95] et utilise la même stratégie que Schaid et al. (2013) [38] et Lin (2014) [39] pour prendre en compte les positions des variants génétiques. La statistique de test s'écrit sous la forme :

$$\mathrm{Q} = (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}})' \mathbf{K} (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{N}})$$

avec le noyau dépendant des positions des variants

$$\mathbf{K} = \mathbf{XWAA'W'X'}$$

La matrice noyau $\mathbf{K}$ incorpore la matrice de poids de SKAT $\mathbf{W}$ et la matrice de proximité des variants $\mathbf{A}$. La mesure de proximité est :

$$A_{jj'} = \max\left(1 - \frac{1}{w} d_{jj'}, 0\right)$$

Le paramètre w est appelé "maximal radius of tolerance", et est par défaut de 1000 bp.

En comparaison avec le test KERNEL (p72) le paramètres w est équivalent au dénominateur $c \times maxd$. Cependant il faut noter quelques différences :

- pour le test KERNEL : $A_{jj'}(c) = \left(1 - \left(\frac{1}{\text{c×maxd}} d_{jj'}\right)^2\right)^3$
- pour le test PODKAT : $A_{jj'} = \max\left(1 - \frac{1}{w} d_{jj'}, 0\right)^3$

De plus si on écrit la statistique de test sous la forme :

$$Q = \boldsymbol{\delta}'_{PODKAT} \mathbf{A} \boldsymbol{\delta}_{PODKAT}$$

le vecteur $\boldsymbol{\delta}_{PODKAT} = (\boldsymbol{Y} - \hat{\boldsymbol{\mu}})' \mathbf{XW}$ informe pour chaque variant j la statistique du score pour le test simple marqueur et pondérée par le poids w_j. La valeur pour un variant j s'écrit :

$$\delta_{PODKAT\,j} = w_j u_j = w_j \sum_{i=1}^{N} X_{ij}(Y_i - \hat{\mu})$$

La significativité du test est évaluée par la méthode Davies [167] comme pour les tests SKAT, en supposant que la statistique de test suit un mélange de lois de chi-deux.

I.3- DOESTRARE : UN TEST DÉVELOPPÉ POUR DÉTECTER DES REGROUPEMENTS DE VARIANTS RARES DIFFÉRENTS CHEZ LES CAS

Principe de DoEstRare

DoEstRare, pour « *Density-oriented Estimation for Rare variant positions* » [41], est un test développé afin de détecter des regroupements de mutations rares à risque dans des régions localisées du gène. En effet on peut supposer que certains domaines protéiques seraient plus impliqués dans les mécanismes de développement d'une maladie. Pour cette raison, les personnes atteintes présenteraient plus de mutations rares dans ces régions génétiques très localisées. Pour répondre à cet objectif de détection de regroupements de mutations rares, le test compare les densités de probabilité des positions des mutations entre les cas et les témoins. Simultanément, DoEstRare permet de tester les fréquences alléliques globales entre les cas et les témoins, de la même manière que les *burden tests*.

Les hypothèses de ce test peuvent s'écrire de la façon suivante :

$$H_0 : f^A = f^U \ et \ p^A = p^U$$

$$H_1 : f^A \neq f^U \ ou \ p^A \neq p^U$$

avec f^A et f^U, les fonctions de densité des positions des variants rares; p^A et p^U, les moyennes pondérées des fréquences alléliques. Ces hypothèses s'interprètent de la manière suivante : un gène est considéré associé à la maladie si les mutations rares ne sont pas réparties de la même façon sur le gène ou si la fréquence globale de mutations est différente entre cas et témoins. Afin d'illustrer ceci, la Figure 30 présente trois situations différentes :

- le gène 1 ne présente ni de différence de répartition des mutations, ni de différence de fréquence allélique globale entre cas et témoins (situation sous H0)
- le gène 2 présente une différence de répartition des mutations mais pas de différence de fréquence allélique globale
- le gène 3 présente une différence de fréquence allélique globale, mais pas de différence de répartition des mutations.

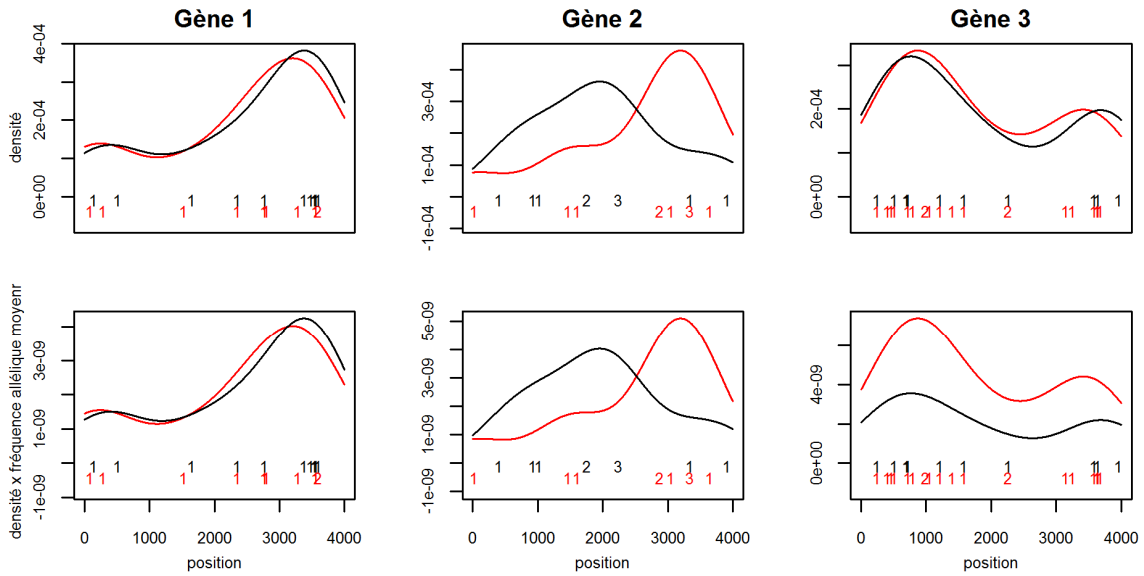


Figure 30. Illustration du principe de DoEstRare.

Les nombres de mutations rares sur le gène sont indiqués en rouge pour les témoins et en noir pour les cas. Les graphes du haut représentent la densité des mutations sur le gène, estimée par méthode à noyau. Les graphes du bas représentent cette densité multipliée par la fréquence allélique moyenne.

Structure de la statistique de test

Afin de tester l'hypothèse nulle, deux composantes entrent en jeu dans le calcul de la statistique de test: (i) les fonctions de densités des positions f^A et f^U ; (ii) les moyennes pondérées des fréquences alléliques p^Aet p^U. Afin de combiner ces deux composantes, la statistique de test est l'aire entre chaque courbe de densité multipliée par la moyenne des fréquences alléliques. Son expression est :

$$STAT = \int_1^{Lg} |\widehat{p^A} \times \hat{f}^A(pos) - \widehat{p^U}\hat{f}^U(pos)| dpos \qquad \text{(I.3.1)}$$

avec Lg la longueur du gène. Sans les estimations des composantes *burden*, p^A et p^U, cette statistique est similaire à la distance de la variation totale, utilisée pour calculer la distance entre deux fonctions de densité de probabilité f^A et f^U [174]. Voici deux cas particuliers de la statistique de test (I.3.1) :

Tableau 7. Écriture de la statistique de test selon l'hypothèse alternative

Condition	Statistique	Test
$f^A = f^U$	$STAT = \int_1^{Lg} \lvert\widehat{p^A} \times \hat{f}^A(pos) - \widehat{p^U}\hat{f}^A(pos)\rvert dpos$ $STAT = \int_1^{Lg} \lvert\widehat{p^A} \times -\widehat{p^U}\rvert \hat{f}^A(pos) dpos$ $STAT = \lvert\widehat{p^A} \times -\widehat{p^U}\rvert \int_1^{Lg} \hat{f}^A(pos) dpos$ $STAT = \lvert\widehat{p^A} - \widehat{p^U}\rvert$	*burden test*
$p^A = p^U$	$STAT = \int_1^{Lg} \lvert\widehat{p^A} \times \hat{f}^A(pos) - \widehat{p^A}\hat{f}^U(pos)\rvert dpos$ $STAT = \int_1^{Lg} \widehat{p^A} \lvert\hat{f}^A(pos) - \hat{f}^U(pos)\rvert dpos$ $STAT = \widehat{p^A} \int_1^{Lg} \lvert\hat{f}^A(pos) - \hat{f}^U(pos)\rvert dpos$	*position test*

Estimation des fonctions de densité

Les fonctions de densité f^A et f^U sont estimées en utilisant un noyau gaussien [175]. Les estimateurs à noyau sont :

$$\hat{f}^A(pos) = \frac{1}{bw}\sum_{j=1}^{P} w^A_{j,densité} \times K\left(\frac{pos - l_j}{bw}\right) \text{ et } \hat{f}^U(pos) = \frac{1}{bw}\sum_{j=1}^{P} w^U_{j,densité} \times K\left(\frac{pos - l_j}{bw}\right)$$

avec bw le paramètre de lissage et $K(.)$, the le noyau. Le noyau gaussien est $K(u) = \frac{1}{2\pi} e^{-\frac{u^2}{2}}$. Le paramètre de lissage bw (ou encore la taille de la fenêtre) utilisé est celui défini par Silverman et al. (1986) [175]. Il s'écrit $bw = 0.9 \times \min\left(\hat{\sigma}(l), \frac{R(l)}{1.34}\right) \times P^{-\frac{1}{5}}$ avec $\hat{\sigma}(l)$ et $R(l)$ l'écart-type et l'étendue de l'interquartile de la série de positions l.

Chaque position l_j intervient dans le calcul de l'estimation par noyau avec le poids $w^A_{j,densité}$ pour les cas et $w^U_{j,densité}$ chez les témoins. Ces poids correspondent au ratio du nombre de mutations pour la position l_j sur le nombre total de mutations pour l'ensemble des positions. Les poids s'écrivent de la façon suivante :

$$w^A_{j,densité} = \frac{m^A_j}{\sum_{j=1}^{P} m^A_j}$$

$$w^U_{j,densité} = \frac{m^U_j}{\sum_{j=1}^{P} m^U_j}$$

avec $m^A_j = \sum_{i=1}^{N^A} X_{ij}$ et $m^U_j = \sum_{i=1}^{N^U} X_{ij}$ les nombres observés de mutations rares pour le variant j chez les cas et les témoins respectivement.

Calcul des composantes *burden*

DoEstRare combine un test sur les positions et un test de type *burden*. Il compare la moyenne pondérée des fréquences alléliques, entre les cas et les témoins, pour l'ensemble des variants rares présents dans le gène. Les composantes *burden* sont :

$$\widehat{p^A} = \frac{1}{P} \sum_{j=1}^{P} \frac{w_j}{\sum_{j=1}^{P} w_j} \frac{m^A_j}{2N^A} \qquad \widehat{p^U} = \frac{1}{P} \sum_{j=1}^{P} \frac{w_j}{\sum_{j=1}^{P} w_j} \frac{m^U_j}{2N^U}$$

avec w_j le poids accordé au variant j. L'intégration des poids dans le calcul de la moyenne des fréquences alléliques permet de mieux discriminer les variants neutres des variants à risque. On souhaite accorder plus de poids aux variants les plus probablement impliqués dans la maladie. Pour le choix des poids, on suppose que les variants causaux sont tous délétères et qu'il n'y a pas de variant protecteur. Un poids plus important est alors accordé aux variants

enrichis chez les cas. Nous nous sommes inspirés des poids utilisés dans la statistique du test KBAC. On suppose que le nombre de mutations rares chez les cas suit sous H_0 une loi binomiale de paramètres $B\left(2N^A, \widehat{q_j^U}\right)$. La probabilité $\widehat{q_j^U}$ est l'estimation de la fréquence allélique chez les témoins:

$$\widehat{q_j^U} = \frac{m_j^U + 1}{2N^U + 2}$$

Le poids w_j est défini comme la probabilité de présenter chez les cas moins de mutations que la valeur observée m_j^A.

$$w_j = P\left(M_j^A \leq m_j^A\right) = \sum_{k=0}^{m_j^A} \binom{2N^A}{k} \left(q_j^U\right)^k \left(1 - q_j^U\right)^{2N^A - k}$$

Évaluation de la significativité-procédure de permutations des phénotypes

La significativité du test est évaluée par la procédure classique de permutations des phénotypes. Pour chaque permutation $b \in \{1, \ldots, B\}$, les statuts atteint/non-atteint sont aléatoirement permutés et la statistique $STAT^{(b)}$ est calculée. Etant donné que la statistique de test est une aire entre deux courbes, ce qui veut dire un nombre réel positif, la p-value est $\frac{\sum_{b=1}^{B} I\left(STAT^{(b)} \geq STAT\right)+1}{B+1}$ [176], avec B le nombre total de permutations.

Une procédure de permutation adaptative peut aussi être employée pour réduire les temps de calcul, dans le contexte de grandes tailles de données [177]. Par adaptatif, on considère que le nombre de permutations effectuées est variable selon le gène. Il est en effet logique de faire de nombreuses permutations dans le cas d'un gène très significatif pour parvenir à obtenir faible de l'ordre de 10^{-6}. Pour l'analyse d'exome, le nombre de gènes à tester est environ de 20 000, le seuil ajusté pour les tests multiples est $\alpha = \frac{0.05}{20000} = 2{,}5.10^{-6}$ par la méthode de Bonferroni. Ces permutations sont très coûteuses en temps de calcul, et doivent donc être réduites dans le cadre de gènes non significatifs.

Soit un succès $I\left(STAT^{(b)} \geq STAT\right)$, i.e. pour une permutation b la statistique obtenue par permutations est supérieure à la statistique observée. Lors de la procédure de permutation

adaptative, une des conditions d'arrêt est d'atteindre un nombre de succès R. Si ce nombre de succès n'est jamais atteint, l'autre condition d'arrêt est d'atteindre le nombre de permutations maximal B. Ces paramètres R et B sont déterminés par le seuil de significativité α voulu et la précision de l'estimation c.

Les algorithmes pour les deux procédures de permutations sont détaillés dans le Tableau 8.

Tableau 8. Algorithmes des procédures de permutations standard et adaptative.

Procédure standard	Procédure adaptative
Paramètre : $B \leftarrow$ nombre de permutations $STAT \leftarrow$ calcul de la statistique de statistique de test **for** $b = 1$ à $b = B$ $STAT^{(b)} \leftarrow$ calcul de la statistique pour la permutation b **if** $STAT^{(b)} \geq STAT$ $R_b \leftarrow R_b + 1$ **end** **end** $p - value \leftarrow \frac{R_b + 1}{B + 1}$	Paramètres : $\alpha \leftarrow$ seuil de significativité souhaité $c \leftarrow$ précision de la p-value souhaitée $B \leftarrow$ estimation du nombre de permutations maximal en fonction de α et c $R \leftarrow$ estimation du nombre de succès à atteindre pour l'arrêt des permutations en fonction de α et c. $STAT \leftarrow$ calcul de la statistique de statistique de test $R_b \leftarrow 0$ $b \leftarrow 0$ **while** $R_b < R$ et $b < B$ $STAT^{(b)} \leftarrow$ calcul de la statistique pour la permutation b **if** $STAT^{(b)} \geq STAT$ $R_b \leftarrow R_b + 1$ **end** $b \leftarrow b + 1$ **end** **if** $b < B$ $p - value \leftarrow \frac{R}{b}$ **else if** $b = B$ $p - value \leftarrow \frac{R_b+1}{B+1}$ **end**

II- COMPARAISON DE DIFFÉRENTES STRATÉGIES

II.1- ÉTUDE DE LA PERFORMANCE DES TESTS SUR LA BASE DE SIMULATIONS

Beaucoup de tests statistiques ont été développés ou adaptés pour les variants génétiques rares. Nous avons alors évalué les performances des principales stratégies à l'aide de simulations de données génétiques, en considérant comme indicateurs l'erreur de type I et la puissance.

L'erreur de type I d'un test statistique est la probabilité de rejeter à tort H_0 bien que H_0 soit vraie. Dans ce contexte, il s'agit de la probabilité de déclarer un gène significativement associé à la maladie bien qu'il n'y ait pas association. Une erreur de type I élevée se traduit, dans les résultats d'une étude, par un plus grand nombre de faux positifs. La puissance d'un test est la capacité de rejeter H_0 lorsque H_1 est vraie. Il s'agit alors de la capacité d'un test à détecter des signaux d'association. Ces deux critères de comparaison sont importants à indiquer ensemble, une puissance élevée peut être liée à une erreur de type I élevée. Le test idéal, qui ressortirait des comparaisons, serait un test avec une erreur de type I faible et une puissance élevée.

Pour la simulation d'un gène associé à la maladie, il existe de nombreuses possibilités de modèles dans le cadre d'un groupe de variants rares. Dans un premier temps nous avons comparé les principales stratégies des méthodes pour variants rares à l'aide de simulations basées sur les travaux de Basu et Pan (2011) [164]. Dans un second temps, afin d'évaluer la performance de tests incorporant des positions, dont DoEstRare, nous avons simulé, en utilisant un modèle génétique de coalescence, des scénarios avec des regroupements de variants rares à risque sur le gène. Ces deux schémas de simulation permettent d'une part de situer la puissance et l'erreur de type I de DoEstRare dans le cadre des comparaisons générales de tests pour variants rares, et d'autre part de mettre en avant l'intérêt d'intégrer les positions dans la statistique de test.

Simulations basées sur le travail de Basu et Pan (2011)

Objectifs

Afin de connaître la performance des tests, une première série de simulation, basée sur les travaux de Basu et al. (2011) [164], a été effectuée. Ceux-ci ont permis de comparer différents

tests pour de nombreux scénarios. Nous avons de même simulé de nombreux scénarios génétiques dans un premier temps. Nous présenterons dans ce manuscrit de thèse uniquement les résultats les plus intéressants, c'est-à-dire pour les scénarios faisant varier le nombre de variants rares neutres et l'effet délétère ou protecteur des variants. D'autres scénarios génétiques ont été envisagés, faisant varier la structure de corrélation entre les variants et la présence ou non de variants plus fréquents, mais ne seront pas présentés ici.

Nous nous intéressons particulièrement à la performance des tests lors de l'ajout de variants non causaux dans le groupe de variants à tester. Ces variants neutres, n'ayant aucun effet sur la susceptibilité d'être malade, sont une source de bruit supplémentaire pour les analyses et le vrai signal devient alors beaucoup plus difficile à détecter. Pour mieux discriminer les variants, certains des tests comme wSum, KBAC, SKAT et SKAT-O utilisent un système de pondération continu basé sur la fréquence allélique afin de mieux prendre en compte les variants les plus rares, pouvant avoir des effets plus forts. Les *variance-component tests* tels que C-alpha, SKAT et SKAT-O sont basés sur l'hypothèse de variance d'effet au sein du gène. Le test ADA, combinant les p-values des tests simple marqueur, utilise un poids binaire adaptatif pour classer le variant dans les catégories à risque ou neutre. Pour le test DoEstRare, nous avons opté pour un système de pondération proche de celui employé par le test KBAC, pour mieux distinguer les variants à risque des variants neutres.

La présence de variants protecteurs dans le groupe de variants rares à tester est aussi à considérer lors des analyses. Il est connu que des variants génétiques peuvent diminuer le risque d'être malade en agissant sur la régulation des voies de signalisation. Bien que la présence de variants rares protecteurs soit encore méconnue, il est important de connaître la performance des tests dans ce type de scénario génétique. Le test aSum emploie un système de pondération binaire afin de classer les variants supposés à risque et les variants supposés protecteurs en fonction du signe de la différence de fréquence allélique entre les cas et les témoins. Les *variance- component tests* ont été développé pour détecter la présence de variants protecteurs. Le système de pondération de la statistique de notre test DoEstRare, suppose que les variants sont à risque. Par conséquent DoEstRare peut donc ne pas détecter les gènes avec des variants protecteurs.

Méthodes

Simulation des données génétiques

La simulation de données génétiques est basée sur le méthode développée par Wang et Elston (2007) [178] et reprise dans la publication de Basu et Pan (2011) [164]. Les étapes de la simulation sont schématisées dans la Figure 31.

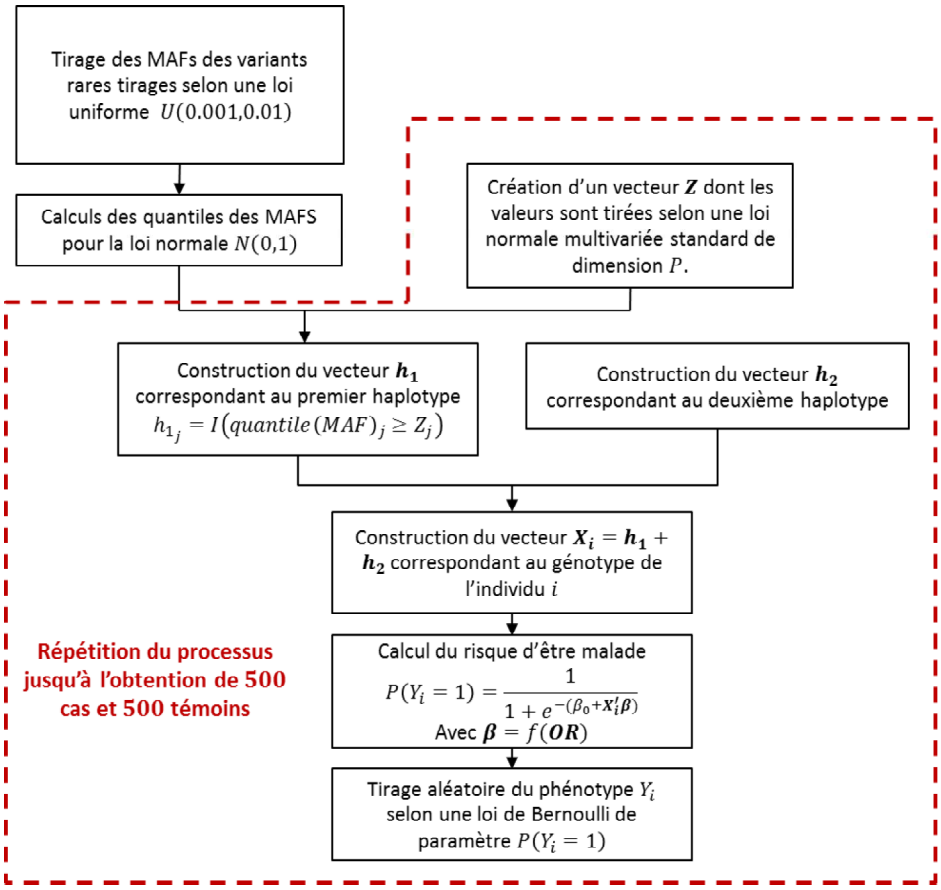


Figure 31. Schéma de la méthode de simulation de Basu et Pan (2011).

Les données génétiques simulées sont les génotypes de 500 cas et 500 témoins pour un groupe de variants rares. Ces variants rares ont une MAF dans la population tirée aléatoirement entre 0.001 et 0.01 selon une loi uniforme. Les variants rares sont considérés indépendants. Le nombre de variants rares analysés et les effets de ceux-ci sur la susceptibilité de la maladie varient en fonction des différents scénarios génétiques considérés.

La performance des tests est évaluée en termes d'erreur de type I et de puissance (formules précisées plus loin p86). Pour calculer les erreurs de type I, le phénotype des individus est

simulé sous l'hypothèse nulle. Et pour la puissance, le phénotype des individus est simulé sous l'hypothèse alternative. Le Tableau 9 permet de résumer les scénarios génétiques simulés. Tous les scénarios génétiques, sous l'hypothèse alternative, font intervenir 8 variants génétiques rares causaux, auxquels sont ajoutés 0, 4, 8, 16 ou 32 variants non causaux. Deux scénarios principaux sont envisagés selon la présence ou non de variants protecteurs :

- scénario 1 : les 8 variants rares causaux sont tous à risque et présentent un OR égal à 2. $\boldsymbol{OR_{causaux}} = (2, 2, 2, 2, 2, 2, 2, 2)$;
- scénario 2 : les 8 variants rares causaux sont à risque ou protecteurs. $\boldsymbol{OR_{causaux}} = \left(3, 3, 2, 2, 2, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$;

Les variants non causaux (ou neutres) ajoutés dans les analyses n'ont aucun effet sur la susceptibilité de la maladie et présentent des OR égaux à 1. Par conséquent, sous l'hypothèse nulle, tous les variants génétiques présentent des OR égaux à 1.

Tableau 9. Scénarios génétiques simulés avec le schéma de simulation de Basu et Pan (2011)

		Nombre de variants non causaux ajoutés				
		0	4	8	16	32
Hypothèse nulle	OR=1					
Hypothèse alternative $\boldsymbol{OR_{causaux}} =$	$(2, 2, 2, 2, 2, 2, 2, 2)$					
	$\left(3, 3, 2, 2, 2, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$					

Pour la simulation du phénotype, le modèle de susceptibilité de la maladie est le suivant :

$$logit\big(P(Y_i = 1|\boldsymbol{X_i})\big) = \alpha_0 + \boldsymbol{X_i'\beta}$$

avec $\boldsymbol{X_i'}$, le vecteur des génotypes de l'individu i échantillonné de la population générale ; α_0 , le risque de base ; et $\boldsymbol{\beta}$, le vecteur des coefficients de régression pour les effets génétiques. On pose le risque de base $\alpha_0 = \log\left(\frac{0,05}{1-0,05}\right)$ de façon à ce que 5% des personnes ne présentant pas de mutation rare soient atteintes (bien que cela ne soit pas très important pour la suite étant donné que les nombres de cas et de témoins sont fixes). Dans ce contexte de variants

rares, cette valeur est proche de la prévalence de la maladie. Le vecteur $\boldsymbol{\beta}$ correspondant aux logarithmes des OR posés pour chaque scénario génétique. Pour chaque individu i le phénotype 0 ou 1 est tiré aléatoirement selon une loi de Bernoulli de probabilité $P(\widehat{Y_i = 1}|\boldsymbol{X_i}) = \frac{1}{1+\exp\left(-(\alpha_0+\boldsymbol{X_i'\beta})\right)}$. Les individus sont intégrés dans le jeu de données jusqu'à l'obtention des 500 cas et 500 témoins.

Pour chaque scénario de simulation, nous effectuons 1000 réplicats, sur lesquels se base l'estimation des erreurs de type I et de puissance. Afin de réaliser ces simulations, nous avons utilisé le code de Wei Pan, qui est disponible depuis sa page personnelle (http://www.biostat.umn.edu/~weip/prog/BasuPanGE11/simRareSNP.R).

Analyses d'association pour les variants rares

Pour cette première comparaison de tests statistiques, nous avons appliqué un grand nombre de stratégies aux données ainsi simulées. Nous avons aussi inclus les tests incorporant les positions des variants dont le test DoEstRare que nous avons développé. Les différentes stratégies sont :

- les *burden tests* : CAST, Sum, wSum, aSum et VT
- le test avec le génotype multi-locus : KBAC
- les *variance-component tests* : C-alpha, SKAT et SKAT-O
- le test combinant les p-values : ADA
- les *position tests* : DBM, CLUSTER, KERNEL, PODKAT, BOMP et DoEstRare

Les tests appliqués ainsi que les détails de l'implémentation sont résumés dans le Tableau 10. Les tests CAST, SKAT, SKATO et PODKAT se basent sur une loi approchée pour le calcul de la p-value. La significativité pour les tests Sum, wSum, aSum et VT est soit basée sur une procédure de permutations standard, ou approchée pour des paramètres estimés à partir des permutations. Pour les tests se basant sur les permutations, nous avons effectué 500 permutations.

Note : Pour les *position tests*, les positions des variants sont tirées aléatoirement entre 1 et 10 000.

Tableau 10. Tests d'association comparés avec le modèle de simulation de Wei Pan

Test	Permutation	Implémentation type	Implémentation Lien	Implémentation Arguments	Dans la comparaison de Basu et Pan (2011)
CAST	Oui	Code R		NOTE : Ces tests *burden* sont implémentés avec le test du score pour un modèle de régression logistique.	OUI
Sum	Oui	Code R			OUI
wSum	Oui	Code R			OUI
aSum	Oui	Code R			OUI
VT	Oui	Code R			NON
KBAC	Oui	Package R KBAC	http://tigerwang.org/software/kbac	alpha = 999 procedure de permutation standard	OUI
C-alpha	Oui	Code R basé sur celui de Wei Pan	http://www.biostat.umn.edu/~weip/prog/BasuPanGE11/CalphaP.R		OUI
SKAT	Non	Package R SKAT (CRAN)	https://cran.r-project.org/src/contrib/SKAT_1.1.2.tar.gz		OUI (appelé SSUw)
SKAT-O	Non	Package R SKAT (CRAN)	https://cran.r-project.org/src/contrib/SKAT_1.1.2.tar.gz	method="optimal"	NON
MiST	Non	Package R MiST (CRAN)	https://cran.r-project.org/src/contrib/MiST_1.0.tar.gz	maf=maf maf : le vecteur des MAF calculées à partir des cas et témoins Z=rep(1,ncol(genotypes))	NON
ADA	Oui	Code R : page web de Wan-Yu Lin	http://homepage.ntu.edu.tw/~linwy/ADA.html	mafThr = 0.5	NON
DBM	Oui	Code R envoyé par l'auteur			NON
CLUSTER	Oui	Code R : page web de Wan-Yu Lin	http://homepage.ntu.edu.tw/~linwy/CLUSTER.html	mafThr = 0.5 max_d=$maxd$	NON
KERNEL	Oui	Code R			NON
BOMP	Oui	Software java	http://karchinlab.org/apps/appBomp.html		NON
PODKAT	Non	Code R de Bioconductor	https://www.bioconductor.org/packages/release/bioc/src/contrib/podkat_1.2.0.tar.gz	kernel="linear.podkat"	NON
DoEstRare	Oui	Code R			NON

Note : Afin d'harmoniser l'ensemble des résultats présentés dans ce manuscrit, les *burden tests* ont été implémentés avec un test de score permettant l'introduction de covariables.

Calculs de puissance et d'erreur de type I et analyse exploratoire

Le calcul de la puissance s'effectue sur $B = 1000$ réplicats et celui de l'erreur de type I sur $B = 10\,000$. Ces deux indicateurs de la performance des tests sont calculés de la façon suivante :

$$\left.\begin{matrix} Si\ H_0 & erreur\ de\ type\ I\ (\alpha = 5\%) \\ Si\ H_1 & puissance\ (\alpha = 5\%) \end{matrix}\right\} = \frac{\sum_{b=1}^{B} I(pvalue\ (b) \leq 0.05)}{B}$$

Afin d'explorer les ressemblances de profils de puissance des tests pour l'ensemble des différents scénarios génétiques, nous avons appliqué la méthode d'analyse multivariée. Dans les scénarios, nous avons fait varier deux paramètres : le nombre de variants non causaux et les effets des variants génétiques. Nous avons appliqué une Analyse Factorielle Multiple (AFM) [179] à la structure de données présentée dans la Figure 32 avec 17 individus (tests) et 2 tableaux de 5 variables (scénarios). Les groupes correspondent aux principaux scénarios 1 et 2, faisant varier les effets des variants.

L'AFM a été réalisée en employant le package R FactoMineR [180].

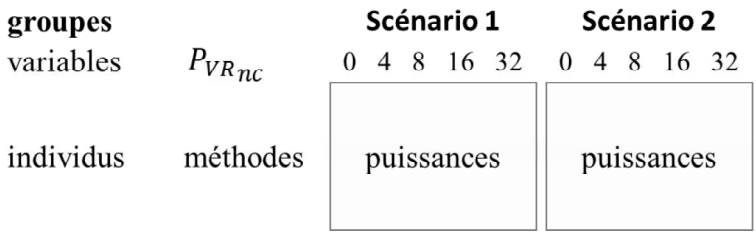


Figure 32. Structure des données pour l'AFM des puissances des tests pour différents scénarios de simulation.

Résultats

Erreurs de type I

Avant de comparer les puissances des tests, il est nécessaire de vérifier si les erreurs de type I sont proches du seuil $\alpha = 5\%$. La Figure 33 nous montre que les erreurs de type I sont

proches de 5%, à l'exception du test CAST qui semble conservateur[7]. Les valeurs sont en Annexe III dans le Tableau S 1.

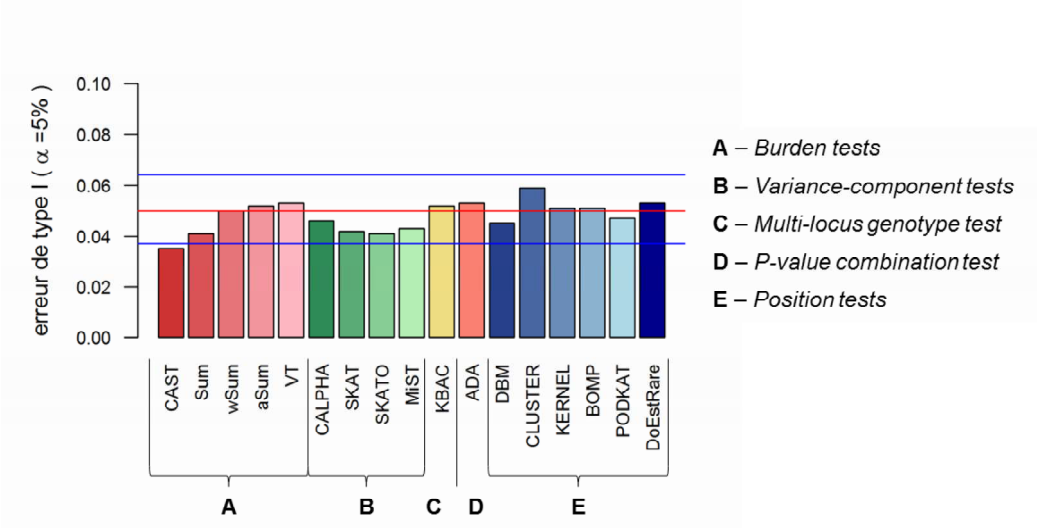


Figure 33. Erreurs de type I des tests statistiques pour le seuil α=5%, dans le cadre des simulations basées sur le travail de Basu et Pan (2011).

Puissances

Afin de comparer les puissances des tests, nous avons tracé les histogrammes de puissance dans la Figure 34 et les résultats de l'analyse factorielle multiple dans la Figure 35 pour mieux comparer les tests en fonction des scénarios considérés. Une interprétation plus approfondie de l'AFM est en Annexe IV.

Des tests se démarquent en termes de puissance

Certains tests se démarquent en termes de puissance pour l'ensemble des scénarios génétiques simulés. Les tests KBAC, SKAT-O, MiST et DoEstRare présentent de très bonnes puissances pour l'ensemble des scénarios génétiques considérés. Ils se situent dans les meilleurs tests en observant les histogrammes et sont repérables sur le graphe des individus de l'AFM (voir Figure 35). Le premier axe de l'AFM permet en effet de distinguer une source de variabilité commune aux deux groupes de variables (scénarios génétiques OR2 et ORs). Il oppose les

[7] Un test conservateur est un test dont le niveau de signification réel est inférieur au niveau de signification nominal. A l'inverse le test est dit anti-conservateur.

tests globalement puissants à ceux présentant des puissances peu élevées pour l'ensemble des scénarios.

KBAC, SKAT-O et MiST ont déjà été décrits dans la littérature, par Moutsianas et al. (2015) [181], pour montrer une bonne puissance dans de nombreux scénarios génétiques. Le test SKAT-O est beaucoup plus connu et utilisé dans les études d'association de maladies complexes. En effet il requiert moins de temps de calcul que KBAC car il se base sur une distribution nulle approchée pour l'évaluation de la significativité plutôt qu'une procédure de permutations.

Avec ces simulations, on peut constater que DoEstRare présente aussi une très bonne puissance dans ces scénarios très généraux envisagés par Basu et Pan (2011) pour la comparaison des tests. Pour rappel, dans ces scénarios les positions des variants sont tirées aléatoirement et aucun regroupement n'est simulé. Ces résultats sont très encourageants pour l'étude de DoEstRare dans le cas de regroupements de variants à risque, qui est plus détaillée dans la partie « Simulations de regroupements localisés de variants rares ». Les autres *position tests,* DBM, KERNEL, CLUSTER, PODKAT et BOMP, ne présentent pas de très bonnes puissances en comparaison avec les tests classiques.

Impact de l'ajout de variants neutres sur les puissances

Les histogrammes de la Figure 35 montrent que l'ajout de variants neutres aux variants causaux rend plus difficile la détection de l'association, avec une diminution globale de la puissance. Cependant on peut noter que cette diminution est plus ou moins conséquente selon les tests.

Les *burden tests* semblent plus sensibles à la présence de variants non causaux et voient leur puissance beaucoup diminuer en comparaison avec les autres tests. Les tests CAST, Sum et aSum ne sont pas en effet construits pour ces cas. Le test wSum accorde plus de poids aux variants les moins fréquents, et aurait peut-être eu une meilleure puissance si dans le modèle l'OR était une fonction décroissante de la MAF. Il en est de même pour le test VT qui sélectionne les variants selon plusieurs seuils de la MAF. Les *variance component tests* ont été conçus pour des gènes avec des variants présentant des effets variables, et ont une puissance qui diminue beaucoup moins fortement que celle des *burden tests*.

KBAC a aussi été développé pour mieux distinguer les génotypes multilocus à risque des autres et conserve une bonne puissance. Le test ADA, a aussi développé pour sélectionner de manière adaptative des variants plus susceptibles d'être en lien avec la maladie selon leur p-value au test simple-marqueur, et se comporte très différemment avec une augmentation puis une diminution de la puissance.

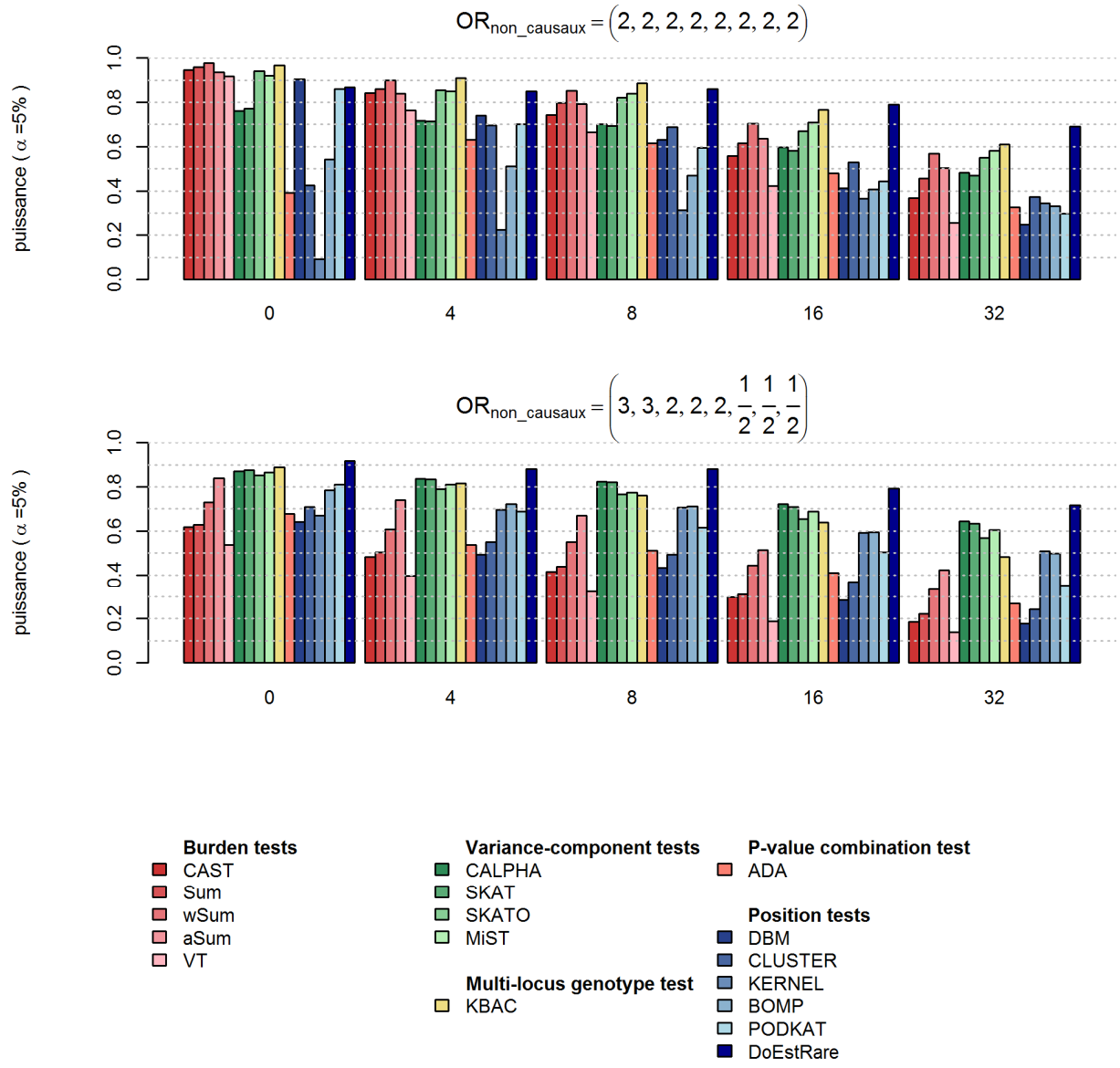


Figure 34. Puissances des tests statistiques pour le seuil α=5% dans le cadre des simulations basées sur le travail de Basu et Pan (2011).

Les tests intégrant les positions des variants visent à détecter des regroupements de variants à risque et présentent aussi des profils de puissance très différents des autres tests. Dans les simulations, les positions des variants ont été choisies aléatoirement entre 1 et 10000, et ne font pas intervenir de regroupements, ce qui pourrait expliquer ces profils différents. Des tests comme CLUSTER et KERNEL montrent des puissances pouvant augmenter avec l'introduction de variants non causaux.

PODKAT est un variance-component test dérivé du test SKAT, intégrant une matrice de proximité entre les variants dans la statistique. La puissance de ce test diminue très fortement, malgré qu'il soit très similaire à SKAT.

Enfin le test DoEstRare que nous avons développé conserve une très bonne puissance malgré l'introduction de variants non causaux. Celui-ci prend en compte les positions des variants génétiques pour détecter des différences de positions de mutations entre les cas et les témoins. Il permet aussi de comparer des nombres pondérés de mutations, avec des poids similaires à ceux du test KBAC pour mieux distinguer les variants à risque des autres.

Afin de détecter des groupes de variants rares impliqués dans des maladies complexes, il faut pouvoir détecter l'association malgré la présence de variants neutres dans les données. Beaucoup de stratégies ont été mises en place pour faire face à cet enjeu, comme le test de la variance des effets génétiques, la pondération ou la sélection des variants.

Dans la pratique, afin de mieux détecter des signaux d'association entre les variants génétiques et le phénotype, les variants sont sélectionnés à partir d'annotations fonctionnelles. Par exemple, les variants synonymes peuvent être exclus des analyses car ils ne modifient pas la constitution de la protéine. De nombreux scores de fonctionnalité permettent de décrire l'importance de l'impact de la mutation sur la fonction ou la régulation de la protéine codée.

Impact de l'ajout de variants protecteurs sur les puissances

Les profils de puissance sont très différents selon l'orientation des effets des variants, c'est-à-dire selon si tous les variants sont à risque (« OR2 ») ou selon un mélange de variants à risque et protecteurs (« ORs »). L'interprétation de l'AFM présentée en Annexe IV montre qu'en effet les profils de puissance des tests sont très différents selon les deux groupes. Avec le

graphe des variables présenté en Figure 35, on peut voir une nette séparation des deux groupes de variables.

Parmi les tests considérés comme globalement puissants dans l'ensemble des scénarios, les tests SKAT-O et MIST sont les combinaisons de *burden tests* et de *variance-component test*. La partie *variance-component test*, a été développée pour détecter la présence d'effets génétiques très différents (à risque, neutre, protecteur). C'est pourquoi on peut observer, à partir de la Figure 34 et de la Figure 35, que les tests C-alpha et SKAT présentent une très bonne puissance lorsqu'il y a un mélange de variants à risque et de variants protecteurs.

Les tests DoEstRare et KBAC ont aussi une puissance très élevée dans cette situation, malgré le fait que dans le calcul des poids, les variants sont considérés à risque ou neutres. En effet le test KBAC utilisé, se base sur une hypothèse unilatérale, bien qu'une version bilatérale peut aussi être utilisée. Les variants protecteurs dans la statistique, pour ces deux tests, ont des poids très faibles, ce qui permet de distinguer une association avec les variants à risque. Cependant, dans le cadre où tous les variants sont protecteurs, ces deux tests ne montreraient pas une très bonne puissance.

Contrairement aux *variance-component tests,* les *burden tests* VT, CAST et Sum présentent une puissance très diminuée avec ces variants aux effets opposés. En effet l'addition de mutations génétiques avec des effets opposés ne permet pas de distinguer de nombres de mutations différents chez les cas et les témoins. Le test aSum conserve une bonne puissance par rapport aux autres *burden tests* car le système de pondération adapté permet de différencier les effets à risque des effets protecteurs.

Résumé

Pour résumer l'ensemble de ces informations, des annotations ont été ajoutés sur le graphe des individus de l'AFM dans la Figure 35. Les tests statistiques SKAT-O, KBAC, MiST et DoEstRare se démarquent en termes de puissance pour de nombreux scénarios. Ils permettent de détecter le signal d'association de variants à risque malgré des groupes très hétérogènes avec des variants neutres ou des variants protecteurs.

Ces premiers résultats sont très encourageants pour l'utilisation de DoEstRare pour détecter des signaux d'association lors d'études de maladies complexes. Pour les scénarios très

généraux qui ont été considérés, la puissance de détection des signaux au seuil α=5% est entre 70% et 90%. Il faut toutefois être vigilent avec ces mesures, car elles dépendent fortement des nombres de cas et de témoins, et de la fréquence allélique. En effet, l'analyse de variants avec des fréquences très basses s'avère plus difficile pour n'importe quel test statistique employé, du fait du peu d'observations.

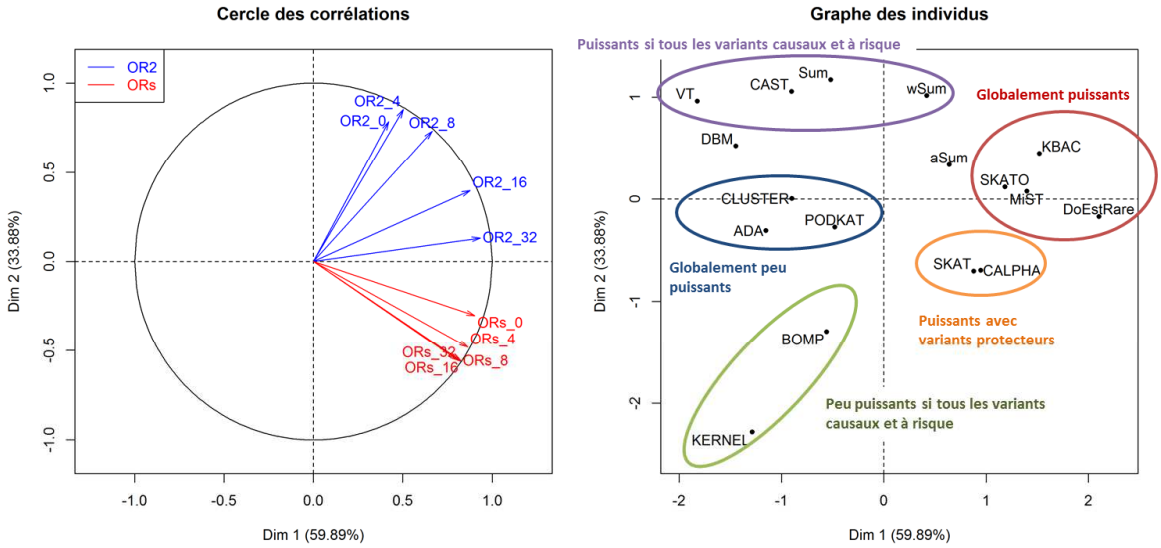


Figure 35. Résultats de l'AFM pour les axes 1 et 2 sur les profils de puissance des tests en fonction des scénarios.

L'AFM a été effectuée sur 17 invidus (tests) pour 2 tableaux de 5 variables (scénarios). Sur le graphe des individus, est aussi ajoutée une interprétation des profils de puissance des tests.

Simulations de regroupements localisés de variants rares

Afin d'évaluer les performances du test DoEstRare construit pour identifier des regroupements de variants génétiques rares dans les gènes, nous avons conçu des simulations de scénarios faisant intervenir les positions des variants à risque. En comparaison avec les simulations basées sur le travail de Basu et Pan (2011), ces simulations se rapprochent du scénario 1 (p82) avec les variants causaux tous à risque et une grande proportion de variants non causaux (plus proches des situations avec 16 ou 32 variants non causaux). Afin d'obtenir des données génétiques plus réalistes, avec une distribution des fréquences alléliques plus proche de ce qui est observé dans la population européenne, nous avons effectué ces

simulations avec le programme *cosi* [182] basé sur des modèles génétiques de coalescence. Ce travail a donné lieu à une publication dans la revue Plos One (Persyn et al. (2017) [41]) incluse en Annexe I.

Méthodes

Simulation des données génétiques

La construction du test DoEstRare a pour objectif de détecter des regroupements de variants à risque dans des régions spécifiques du gène. Pour plus de simplicité, on parlera par la suite de « *cluster* ». C'est pourquoi nous avons simulé principalement 3 grands scénarios (Figure 36) : (i) pas de *cluster*, (ii) un *cluster* et (iii) deux *clusters* de variants à risque.

Nous avons mené les simulations sur la base d'un modèle de coalescence implémenté dans le programme *cosi* [182]. Le principe est de simuler, à l'aide de ce programme, les haplotypes d'une population dite européenne, avec l'information sur les positions. A partir de ce « *pool* » d'haplotypes, nous avons échantillonné des individus et avons simulé leur phénotype selon un modèle de régression logistique.

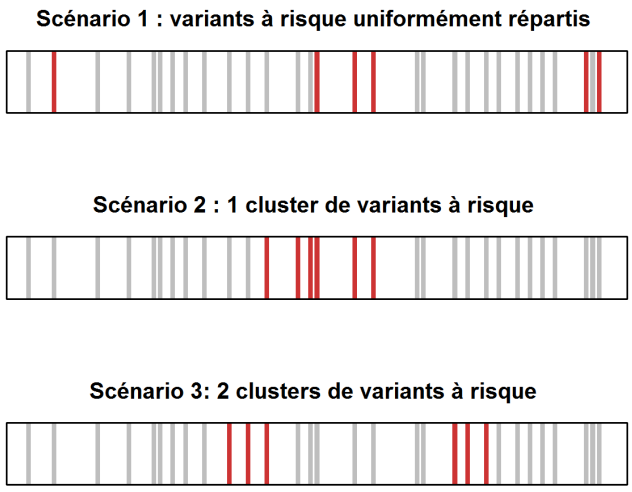


Figure 36. Scénarios génétiques simulés en fonction du regroupement des variants à risque

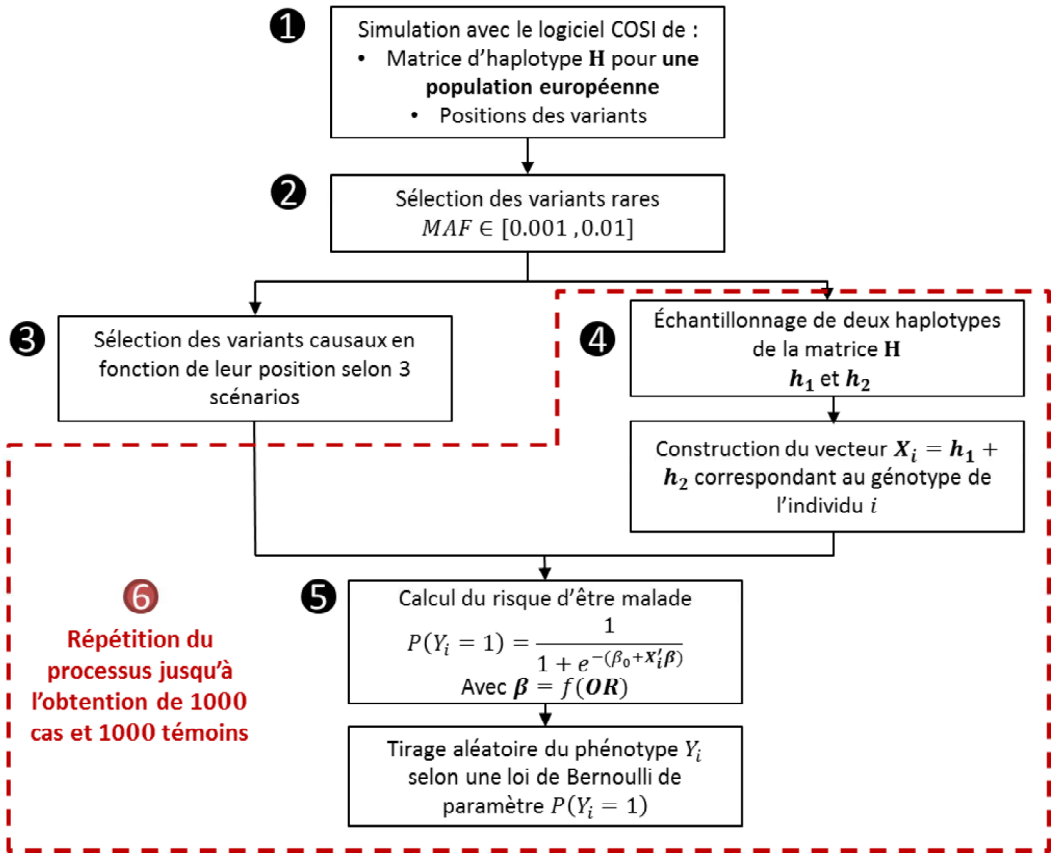


Figure 37. Schéma de simulation pour l'étude de la performance de DoEstRare.

Les étapes de simulation sont schématisées dans la Figure 37.

Étape 1

On génère une population de 10 000 haplotypes pour une région génétique de 10kb en utilisant le programme *cosi*. La population échantillonnée est dite européenne dans le modèle. Les paramètres correspondent au modèle appelé « *bestfit* » par les auteurs du programme Schaffner et al. (2005), qui correspond le mieux aux données de différentes populations mondiales après calibration. Ces paramètres ont été obtenus par calibration à partir de données de génétique de population. Un schéma de ce modèle démographique est présenté dans la Figure 38.

Étape 2

Les variants sont filtrés selon la MAF. Nous sélectionnons les variants présentant une MAF entre 0.001 et 0.01 (dans la population européenne des 10 000 haplotypes). Les variants trop

rares n'entrent pas dans le modèle de simulation afin d'éviter trop de variants à risque non observables dans les données.

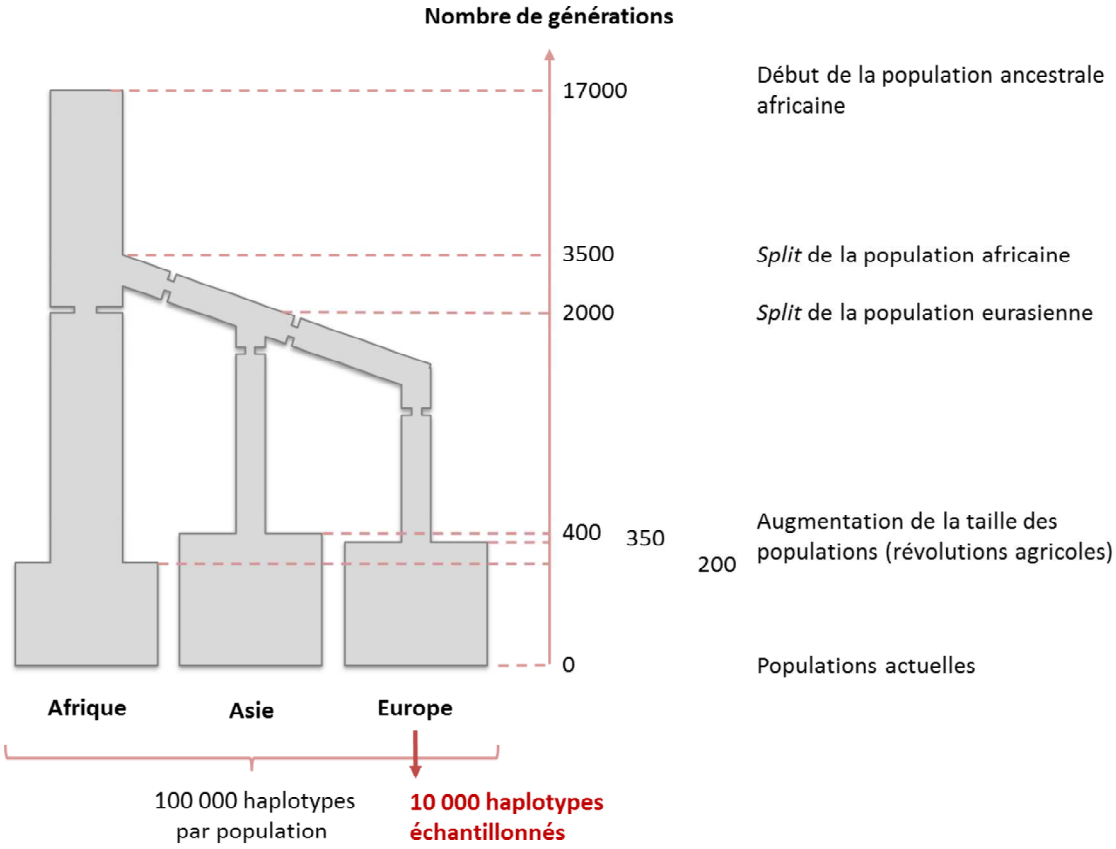


Figure 38. Modèle démographique selon Schaffner et al. (2005).
Les tailles effectives des populations sont schématisées par les différentes épaisseurs. Les *bottlenecks* ou goulots d'étranglement (la diminution de diversité génétique due à la séparation de populations) sont représentés par des constrictions temporaires.

Étape 3

Les variants causaux sont déterminés selon leur position sur le gène. Trois scénarios principaux sont considérés et sont détaillés plus loin.

Étape 4

Deux haplotypes de la population « européenne » sont échantillonnés afin de constituer le génotype d'un individu i.

Étape 5

Le phénotype de l'individu i est simulé selon le modèle de régression logistique suivant :

$$logit\big(P(Y_i = 1|\, \boldsymbol{X_i})\big) = \alpha_0 + \boldsymbol{X_i}^{\mathrm{T}}\boldsymbol{\beta}$$

avec $\alpha_0 = \log\left(\frac{0,05}{1-0,05}\right)$ le risque de base et $\boldsymbol{\beta}$ le vecteur des coefficients de régression pour les effets génétiques. Pour les coefficients de régression, $\beta_j = \log(\mathrm{OR}_j)$ avec OR_j =3 si le variant j est déterminé causal à l'étape 3, sinon $\beta_j = 0$. Le statut de maladie de l'individu i est échantillon selon une loi de Bernoulli de probabilité $P(Y_i = 1|\, X_i)$.

Étape 6

On répète les étapes 4 et 5 afin d'obtenir 1000 cas et 1000 témoins.

Les trois scénarios de simulation (Figure 36) font intervenir les positions des variants. Pour le premier scénario, pour lequel il n'y a pas de regroupement, les variants causaux sont tirées aléatoirement sur le gène. Pour les seconds et troisièmes scénarios, les variants causaux sont voisins dans une ou deux régions du gène. La position initiale du cluster correspond à la médiane des positions lorsqu'il y a un cluster, et aux quantiles 1/3 et 2/3 lorsqu'il y a deux clusters.

Chacun de ces trois scénarios est divisé en sous-scénario faisant intervenir la proportion de variants causaux. Le nombre de variants à risque choisis à l'étape 3 est fixé de façon à avoir 5%, 10%, 15% ou 20% de variants causaux au total dans le gène. Dans ce contexte, cette proportion est fortement liée à la taille du cluster dans le gène.

Analyses d'association pour les variants rares

Les tests statistiques utilisés sont les mêmes que ceux employés lors de la comparaison des tests pour les simulations basées sur la méthode de Basu et Pan (2011) (Tableau 10) (voir la partie **Analyses d'association pour les variants rares**, **p84**). La significativité des tests est estimée par procédure de permutation standard avec un nombre de permutations de 1000.

Seuls quelques tests, SKAT, SKAT-O, MiST et PODKAT, font exception en se basant sur une distribution approchée de la statistique de test sous H_0.

Note : Les tests CAST, wSum, aSum et VT sont implémentés différemment de ceux utilisés dans la publication. Pour une harmonisation des résultats dans la thèse, les *burden tests* se basent sur le calcul d'une statistique pour le test du score.

Résultats

Description des simulations

La distribution des fréquences simulées avec le logiciel *cosi* correspond environ à la distribution des fréquences alléliques dans la population européenne appelée NFE (*Non-Finish European*) de la base de données ExAC [8] (Figure 39).

Après extraction des variants rares avec une MAF comprise entre 0.001 et 0.01 dans la population simulée, le nombre médian de variants par gène de longueur 10kb varie vaut 30 (quantiles 2.5% et 97.5% = [18; 46]). Contrairement aux simulations effectuées précédemment, le nombre de variants rares par gène n'est pas fixe mais aléatoire.

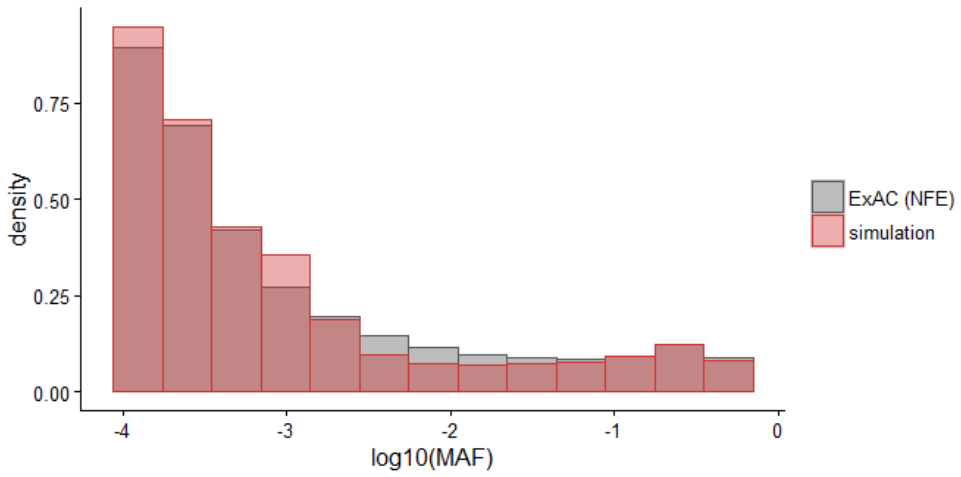


Figure 39.Comparaison de la distribution de la MAF entre les simulations et la base de données ExAC.

Erreur de type I

Les erreurs de type I, calculées à partir de 10 000 réplicats, sont correctes pour l'ensemble des tests (Figure 40) (tableaux de valeurs en Annexe V). Comme il a été énoncé dans la partie **Simulations basées sur le travail de Basu et Pan (2011) (p86)**, le test CAST semble conservateur avec une erreur de type I inférieure au seuil α=5%.

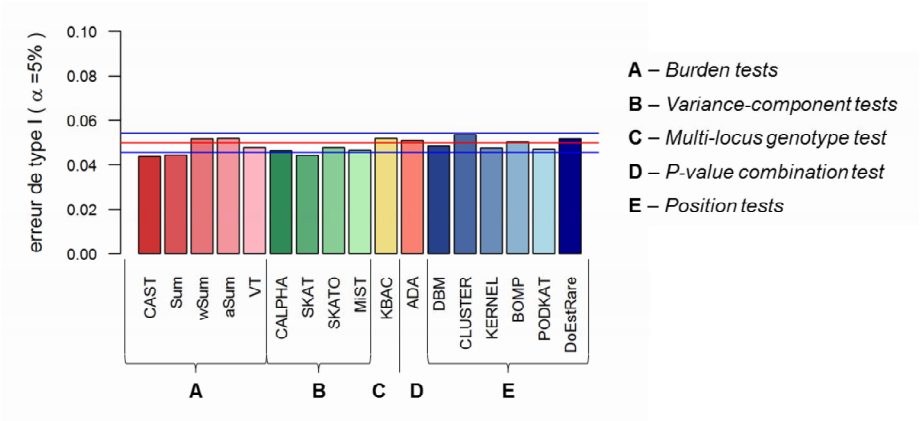


Figure 40. Erreurs de type I des tests statistiques pour le seuil α=5%, dans le cadre des simulations avec le logiciel *cosi*.

Puissances

Les résultats des puissances pour le seuil α=5% sont présentés dans la Figure 41 (tableaux de valeurs en Annexe V). Pour rappel, nous étudions la puissance des tests avec la simulation d'un gène associé avec la maladie sous trois scénarios principaux, en considérant la distribution des positions des variants à risque. Dans le scénario 1, les variants à risque sont distribués aléatoirement sur le gène. Dans les scénarios 2 et 3, les variants à risque sont localisés respectivement dans une et deux régions spécifiques du gène. Nous avons aussi fait varier la proportion de variants à risque entre 5%, 10%, 15% et 20%. Cette proportion est intimement liée avec la taille des regroupements.

Puissances des tests sans regroupement de variants à risque

En se focalisant sur les résultats du scenario 1, où les variants sont distribués aléatoirement sur le gène, les *variance-component tests* et KBAC sont plus puissants que les *burden tests*, du fait du grand nombre de variants neutres. Ceci est consistant avec les observations effectuées précédemment, dans la partie **Simulations basées sur le travail de Basu et Pan**

(2011). Les *position tests* BOMP et DoEstRare se comportent très bien malgré qu'il n'y ait pas de regroupements de variants à risque. Avec les simulations précédentes, DoEstRare présentait en effet une bonne puissance globale, tandis que BOMP présentait une bonne puissance face aux autres tests en présence de beaucoup de variants neutres.

Puissances des tests avec regroupement de variants à risque

Avec les scénarios 2 et 3, les variants à risque sont regroupés dans un une et deux régions génétiques. Si on compare les résultats entre les différents scénarios avec la Figure 42, on observe que les tests n'incorporant pas les positions dans la statistique de test ne montrent pas de gain de puissance. Ceci nous permet bien de comparer les puissances des *position tests* selon les différents scénarios.

DoEstRare obtient une meilleure puissance en présence d'un *cluster* pour les proportions P10, P15 et P20. On peut supposer que lorsque la proportion est faible avec 5% de variants causaux (P5), la région génétique concernée est très petite et ne comporte pas suffisamment de variants pour être mieux détectée avec DoEstRare. Les tests CLUSTER, KERNEL et PODKAT présentent aussi une puissance supérieure ; ils semblent donc mieux adaptés à l'identification de régions à risque. Malgré cette augmentation de puissance, seul le test PODKAT montre une bonne puissance par rapport aux autres tests avec le scénario 2, pour des proportions allant de 15% à 20% de variants causaux (P15 et P20). Ainsi seuls les *position tests* BOMP, DoEstRare et PODKAT montrent de bonnes puissances avec la présence de *clusters*.

On peut aussi noter que les puissances obtenues dans le cas du scénario 3 (avec 2 *clusters*) sont dans de nombreux cas plus faibles par rapport à celles obtenues dans le scénario 2 (présence d'un seul *cluster*). Cette différence pourrait s'expliquer par la façon dont on a simulé les différents scénarios. À proportion de variants causaux égale, le *cluster* du scénario 2 est divisé en deux *clusters* de taille plus réduite pour le scénario 3.

En conclusion, ces simulations mettent en évidence les bonnes performances de DoEstRare par rapport aux autres tests, que ce soit avec ou sans regroupement de variants à risque.

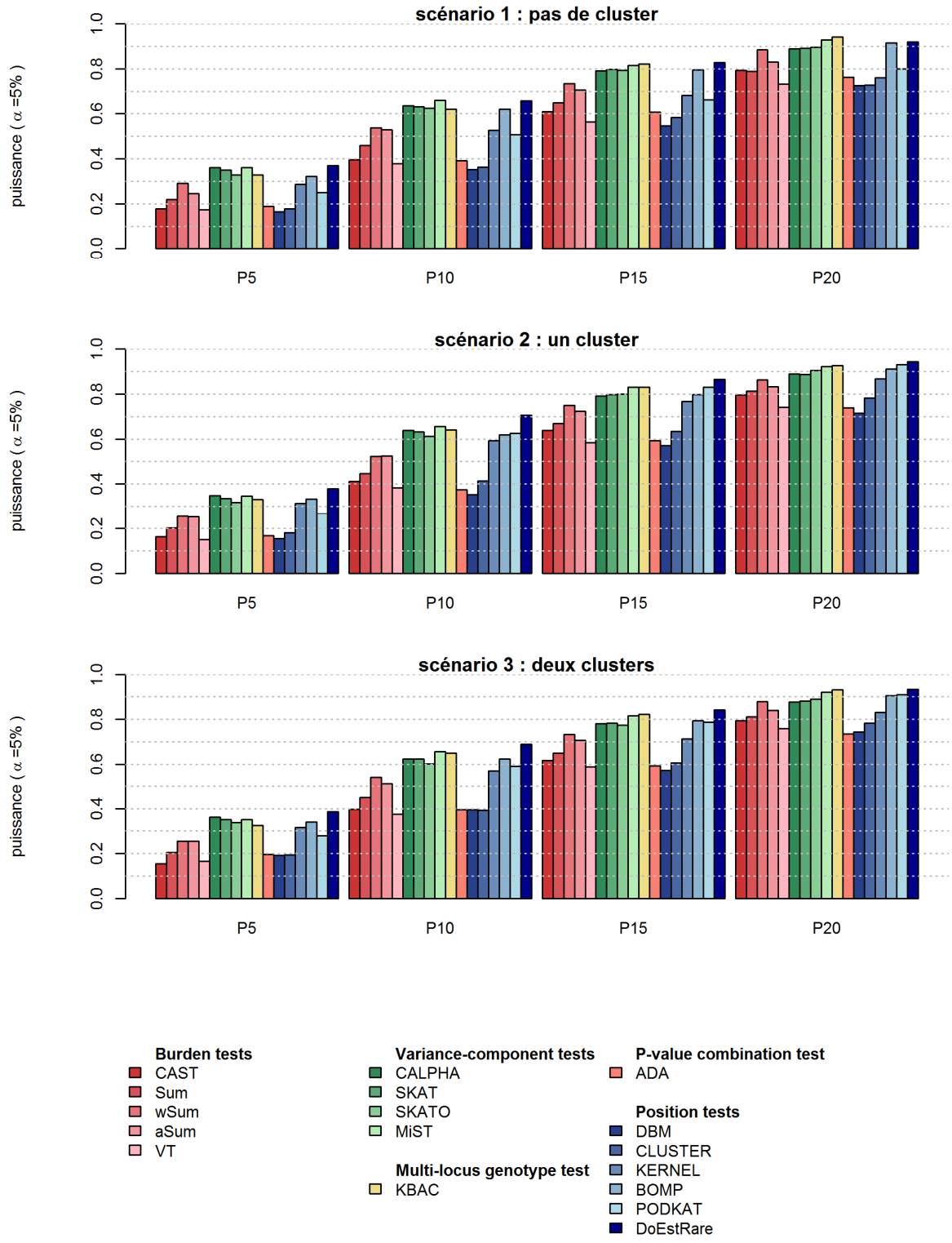


Figure 41. Puissances des tests au seuil α=5% dans le cadre de localisations différentes des mutations à risque.

P5, P10, P15 et P20 correspondent aux scénarios avec des proportions respectives de 5% , 10%, 15%, 20% de variants causaux dans le modèle de simulation.

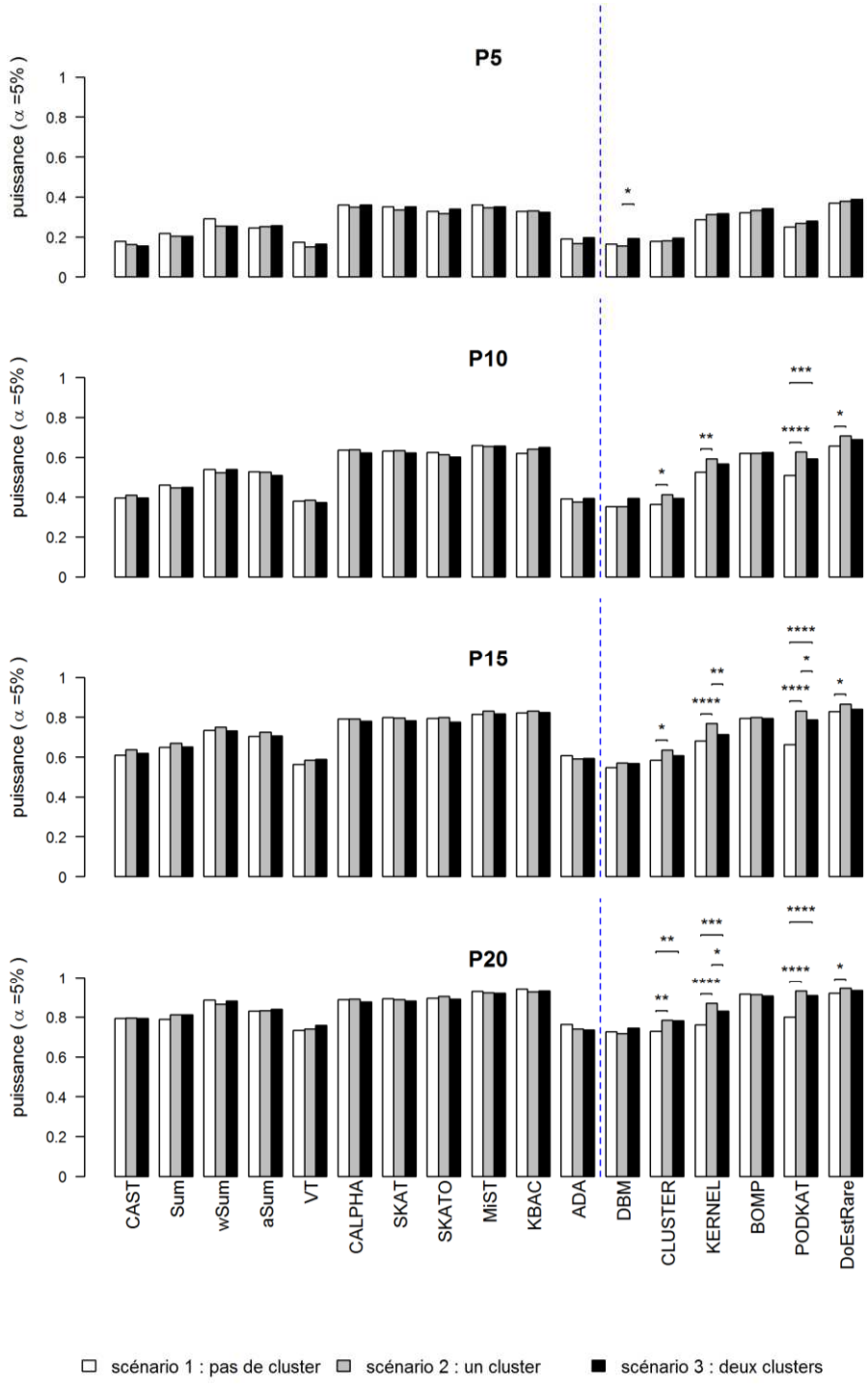


Figure 42. Comparaison des puissances entre les différents scénarios.

Un test exact de Fisher a été effectué pour tester l'indépendance du taux de gènes significatifs pour α=5% et les scénarios comparés. Nous avons comparé les scénarios 1 et 2, 2 et 3, et 1 et 3. .ns: p>0.05, *: p≤0.05, **: p≤0.01, ***: p≤0.001, ****: p≤0.0001. La ligne bleue en tirets sépare les *position tests* (DBM, CLUSTER, KERNEL, PODKAT, BOMP et DoEstRare) du reste des tests.

II.2- APPLICATIONS DES TESTS À DES DONNÉES GÉNÉTIQUES RÉELLES

Les simulations ont permis de comparer la performance des tests statistiques en fonction de différents scénarios génétiques. La grande variété de stratégies est due à la complexité de tester des groupes de variants rares. Selon le modèle biologique supposé, certains tests sont plus performants que d'autres. Or, dans la réalité nous ne connaissons pas la réelle architecture génétique derrière les maladies, et le choix du test à utiliser reste un enjeu. L'application de ces tests à de vraies données génétiques nous permet de comparer les résultats de significativités des tests. Les tests reposant sur différentes hypothèses, on s'interroge à la similarité des résultats obtenus. Pour répondre à cette question nous avons appliqué différents tests statistiques à des données de séquençage pour deux maladies : le syndrome de Brugada (BrS, *Brugada Syndrome*) et la maladie d'Alzheimer d'apparition précoce (EOAD, *early-onset Alzheimer disease*).

Le syndrome de Brugada est une arythmie cardiaque rare (prévalence estimée autour de 1/1000) étudiée au sein de l'institut du thorax. Une étude d'association a été réalisée sur les variants génétiques rares pour cette maladie et a été publiée par Le Scouarnec et al. (2015) [183]. La maladie d'Alzheimer d'apparition précoce est définie pour les personnes âgées de moins de 65 ans. Les données pour la maladie d'Alzheimer d'apparition précoce ont été fournies par le CHU de Rouen d'après la publication de Nicolas et al. (2015) [184]. Les différentes étapes de prétraitement et d'analyse statistique des données, pour chaque maladie étudiée, sont résumées dans le Tableau 11 et sont détaillées dans la section **Traitement et analyse des données (p104)**.

Tableau 11. Tableau récapitulatif de l'analyse des données pour le BrS et l'EOAD

		BrS	EOAD
Recrutement		**Recrutement des cas et des témoins**	
	Nombre de cas	167 patients atteints de BrS	486 patients atteints d'EOAD
	Nombre de témoins	167 patients atteints de RAC	592 participants au projet FREX
Séquençage		**Séquençage ciblé**	
	Capture	HaloPlex™, kit à façon (Agilent Technologies)	SureSelect Human All Exon kits (Agilent Technologies)
	Séquençage	Illumina HiSeq	Illumina HiSeq
	Régions génétiques séquencées	Exons de 163 gènes candidats +/- 10bp	Exome
Prétraitement des données par les plateformes bioinformatiques		**Alignement**	
	Logiciel	BWA-MEM	BWA
		Calling / Détection des variants	
	Algorithmes	GATK Samtools	GATK
		Contrôle qualité	
	Contrôle-qualité individus	• profondeur de lecture moyenne >=100x • européen	• sexe concordant • exclus si le taux d'hétérozygosité est significativement plus élevé • indicateur de parenté pi_hat<18.5% • exclus si contanmination
	Contrôle qualité séquences	• Identifiés avec les 2 algorithmes Samtools et GATK • *quality score*>25 "QUAL"	• *genotype read depth*>10 "DP" • *genotype quality*>50 "GQ" • *Variant Quality Score* >-2 "VQSLOD"
		Annotation des variants	
		VEP	VEP
Prétraitement des données supplémentaire		**Délimitation des régions à analyser**	
	Définition des régions génétiques analysées	Pour chaque gène, la région génétique analysée est l'ensemble des séquences CDS +/- 10 bp, tout transcrit confondu.	Pour chaque gène autosomal, est sélectionné le transcrit codant pour une protéine avec la taille CDS la plus longue. La délimitation correspond à la concaténation des CDS.
		Filtre des individus	
			Séquencés avec les technologies : • Agilent SureSelect Human All Exon V5 • Agilent SureSelect Human All Exon V5UTR
	Nombre de cas	167 patients atteints de BrS	431 patients atteints d'EOAD
	Nombre de témoins	167 patients atteints de RAC	555 participants au projet FREX
		Filtre des variants	
		MAF ExAC (NFE) ≤ 0.01 MAF cas ≤ 0.05 MAF témoins ≤ 0.05	
		pvF<0.01	pvF<1e-04
		Filtre des gènes	
		Nombre de SNV>1 Nombre de variants rares total cas >0 Nombre de variants rares total témoins >0	
	Nombres analysés	58 gènes (1 462 variants)	17,409 gènes (273 390 variants)

pvF : p-value pour le test exact de Fisher effectué sur les variants individuellement.

Traitement et analyse des données

Données pour le syndrome de Brugada

Les données génétiques pour l'étude du syndrome de Brugada, décrites par Le Scouarnec et al. (2015) [183], ont été générées par le séquençage de 163 gènes candidats pour diverses pathologies cardiaques, pour 167 cas et 167 témoins. Dans la liste des gènes candidats, 21 gènes avaient étés repérés, pour le syndrome de Brugada, à partir de la littérature et des études familiales. Ces 21 gènes sont : *ABCC9*, *CACNA1C*, *CACNA2D1*, *CACNB2*, *GPD1L*, *FGF12*, *HCN4*, *HEY2*, *KCND3*, *KCNE3*, *KCNE1L*, *KCNH2*, *KCNJ8*, *PKP2*, *RANGRF*, *SCN1B*, *SCN2B*, *SCN3B*, *SCN5A*, *SCN10A* et *TRPM4*.

Les patients atteints du syndrome de Brugada ont été recrutés au centre de référence des troubles du rythme cardiaque d'origine génétique se situant au CHU de Nantes. Cette structure s'assure de l'expertise et de la prise en charge des patients atteints du syndrome de Brugada au niveau national. Les témoins sont des patients atteints de rétrécissement aortique calcifié (RAC). Cette pathologie est considérée suffisamment éloignée du syndrome de Brugada, pour être utilisée de référence. Au moment de l'étude qui a été publiée, les données du projet FREX, de génétique de population, n'étaient pas disponibles. L'avantage est que les patients atteints de RAC sont plus suivis quant aux maladies cardiovasculaires et ne présentent pas de syndrome de Brugada.

Le contrôle qualité des données a été effectué par la plateforme bio-informatique de l'institut du thorax. Les individus et les variants génétiques dont les données sont de mauvaise qualité ont été retirés selon les seuils décrits dans Le Scouarnec et al. (2015) et résumés dans le Tableau 11.

Les groupes de variants rares correspondent à l'ensemble des variants présents dans les séquences ciblées pour l'étude d'un gène candidat. Les séquences ciblées sont les régions codantes +/- 10 pb. Dans ce cadre ce qu'on considère un « gène » dans le Tableau 11, est la juxtaposition de toutes régions codantes +/10 pb. Pour les tests intégrant l'information des positions dans la statistique, les positions des variants ont été calculées en fonction des débuts et des fins de chaque séquence.

À partir des données, nous avons effectué un prétraitement supplémentaire en fonction de la MAF des variants génétiques dans les bases de données externes et calculées à partir des

données. Comme Le Scouarnec et al. (2015), nous nous sommes basés sur la MAF dans la population générale pour déterminer les variants génétiques rares. Nous avons choisi de fixer le seuil de MAF<1% dans la base de données ExAC pour la population NFE (*Non-Finnish Europeans*). Nous avons aussi retenu les variants présentant une MAF<5% chez les cas ou les témoins car sont susceptibles d'être dus à des erreurs techniques. Les variants ont été testés individuellement avec le test exact de Fisher ; ceux présentant une p-value p<0.01 ont aussi été retirés. Ce filtre permet d'éviter d'analyser des variants trop significatifs pouvant trop influencer le résultat du test pour le gène. Aussi ces variants ont été étudiés séparément et sont des faux-positifs.

Les variants n'ont pas été filtrés selon leur conséquence fonctionnelle sur la protéine codée, comme c'est le cas dans l'analyse effectuée par Le Scouarnec et al. (2015). En effet nous désirions être beaucoup moins sélectifs sur les variants pour ne pas trop réduire les données.

Données pour la maladie d'Alzheimer d'apparition précoce

Les données génétiques pour la maladie d'Alzheimer d'apparition précoce, décrites par Nicolas et al. (2015) [184], ont été collectées pour 486 patients atteints. Un séquençage d'exome a été effectué pour chacun de ces patients avec les kits de capture d'*Agilent SureSelect Human All Exon*.

Pour le choix de la population témoin, les données du projet FREX ont été utilisées. Au moment de l'étude de Nicolas et al. (2015), 500 échantillons d'ADN avaient été séquencés. Pour l'analyse dans le cadre de ce doctorat, 96 échantillons d'ADN supplémentaires collectés par le centre de Dijon ont été séquencés et ajoutés aux données.

Le contrôle qualité a été effectué par les plateformes bioinformatiques du centre de Rouen et sont décrits dans la publication et résumés dans le Tableau 11. Suite à ce contrôle qualité, 484 patients EOAD et 583 témoins FREX sont gardés.

Différentes versions de kit de capture ont été utilisés au fil de l'évolution du projet, allant de *Agilent SureSelect Human All Exon V1* à *Agilent SureSelect Human All Exon V5.* Un filtre a été effectué sur les individus en fonction de design de capture pour limiter les biais techniques dus à l'utilisation de différentes versions de kit de capture. La majorité étant *Agilent SureSelect*

Human All Exon V5 (et *V5 UTR*), nous avons exclu ceux avec un autre design. Cependant nous n'avons pas filtré les individus en fonction de leur histoire familiale.

L'annotation des variants génétiques a été effectuée par les plateformes informatiques avec l'outil *Variant effect Predictor* (VEP) disponible sur le site de la base de données *ensembl* [85] .Pour l'analyse de variants génétiques rares, il est nécessaire de définir les groupes de variants rares en fonction de ces annotations. Un groupe de variants rares, dans cette analyse, inclut les variants appartenant à l'ensemble des séquences codantes pour un transcrit génétique. Un gène correspond très fréquemment à plusieurs transcrits. Nous avons choisi d'analyser un transcrit codant pour une protéine par gène, celui avec la taille CDS (*coding DNA sequence*) la plus longue. Cette information sur la taille CDS des transcrits a été récupérée à partir du site *ensembl biomart* pour le même génome de référence ayant servi aux annotations avec l'outil VEP. Si des transcrits dans les annotations présentaient la même longueur, seulement un seul a été choisi aléatoirement. Les positions des variants correspondent aux positions CDS fournies par VEP.

Nous avons filtré les variants selon s'ils appartenaient aux régions CDS. Nous avons procédé au même filtre des variants que pour l'analyse du syndrome de Brugada. Nous avons sélectionné les variants génétiques avec une MAF<1% dans la population NFE de ExAC, avec une MAF<5% chez les cas et les témoins, et avec une p-value p>1e-4 pour le test exact de Fisher.

Analyse statistique

Pour l'analyse d'association, après les différents filtres, sont testés les gènes avec au moins deux SNV et présentant au moins un variant rare chez les cas et chez les témoins. En effet certaines méthodes ne peuvent être appliquées lorsque ces conditions ne sont pas satisfaites.

Nous avons analysé dans un premier temps les données BrS avec les différents tests ayant servi aux comparaisons de puissance. Pour chaque test, à l'exception de SKAT, SKAT-O, MiST et PODKAT, la p-value est estimée à partir d'une procédure standard de permutations des phénotypes. Le nombre de permutations est de 1000 pour chaque gène, sauf pour le gène majeur SCN5A connu pour la maladie, où le nombre de permutations est passé à 200 000.

Pour l'analyse des données EOAD, la procédure de permutations adaptative est nécessaire pour gagner en temps de calcul. Cela permet aussi d'éviter de placer différents nombres de permutations pour les gènes majeurs. Nous avons choisi comme paramètres le seuil de significativité $\alpha=5.10^{-7}$ avec une précision c=0.2 (voir Tableau 8 p79). Les tests ADA, CLUSTER et BOMP ne sont pas utilisés car difficilement modifiables pour cette procédure de permutations. Nous n'avons pas non plus utilisé le test VT car trop long, et le test DBM à cause du trop grand nombre de messages d'erreur en raison de gènes non analysables (divisions par zéro et etc).

Les différents tests utilisés ainsi que les procédures de permutations sont résumés dans le Tableau 12 suivant.

Note : À la différence de l'article publié pour DoEstRare, nous avons implémenté les *burden tests* en se basant sur un modèle de régression logistique avec un test du score. Ceci ne change pas grandement les résultats mais permet d'harmoniser l'emploi des méthodes dans la thèse.

Tableau 12. Utilisation des tests et évaluation de la significativité

	BrS	**EOAD**
CAST	Permutation standard	Permutation adaptative
Sum	Permutation standard	Permutation adaptative
wSum	Permutation standard	Permutation adaptative
aSum	Permutation standard	Permutation adaptative
VT	Permutation standard	**Trop long**
KBAC	Permutation standard	Permutation adaptative
C-alpha	Permutation standard	Permutation adaptative
SKAT	Loi approchée	Loi approchée
SKAT-O	Loi approchée	Loi approchée
MiST	Loi approchée	Loi approchée
ADA	Permutation standard	**Non adaptable**
DBM	Permutation standard	**Trop d'erreurs**
CLUSTER	Permutation standard	**Non adaptable**
KERNEL	Permutation standard	Permutation adaptative
BOMP	Permutation standard	**Non adaptable**
PODKAT	Loi approchée	Loi approchée
DoEstRare	Permutation standard	Permutation adaptative

Analyse exploratoire des résultats de significativité

Afin de résumer les différences de résultats pour les différents tests d'association, nous avons appliqué une ACP normée sur les -log10(p-value) avec les gènes en individus et les différents tests en variables. Avec cette ACP, le critère de maximisation est la variabilité des résultats de significativité des différents gènes. Les composantes permettent ainsi de distinguer les gènes les plus significatifs des gènes les moins significatifs pour un ou plusieurs tests. Nous nous intéressons surtout à la liaison entre les variables. Nous voulons savoir si les différences de significativité entre gènes sont similaires d'un test à l'autre. Cela permettrait aussi de faire des groupes de tests fournissant des résultats de significativité similaires.

Pour l'ACP, le test DBM a été retiré n'était pas applicable pour un grand nombre de gènes. Pour chacune des deux ACP, les tests en actif sont les 12 tests appliqués aux données BrS et aux données EOAD. Pour l'ACP menée sur les résultats d'association du BrS, nous avons ajouté en variables illustratives les 4 tests VT, ADA, CLUSTER et BOMP qui n'ont pas pu être appliqués aux données EOAD.

L'ACP a été effectuée avec le package R FactoMineR [180].

Résultats

Étude de la similarité des résultats de significativité

Nous avons effectué une analyse d'association pour deux maladies, qui sont le syndrome de Brugada (BrS) et la maladie d'Alzheimer d'apparition précoce (EOAD). Les données BrS incluent 163 gènes candidats séquencés pour 167 cas et 167 témoins. Les données EOAD sont les données d'exome pour 431 cas et 555 témoins. Suite aux différents filtres, 58 gènes candidats ont été testés avec 16 méthodes pour les données BrS et 17 409 gènes autosomaux ont été testés avec 12 méthodes pour les données EOAD.

Pour l'ACP, nous nous concentrons sur les trois premières composantes principales pour expliquer la variabilité des résultats de significativité entre gènes. Elles expliquent environ 90% de la variabilité totale pour les données BrS et 86% pour les données EOAD. Les valeurs propres, les graphes des individus, ainsi que les valeurs de corrélation entre les variables et les composantes principales sont présentés en Annexe VII et Annexe VIII.

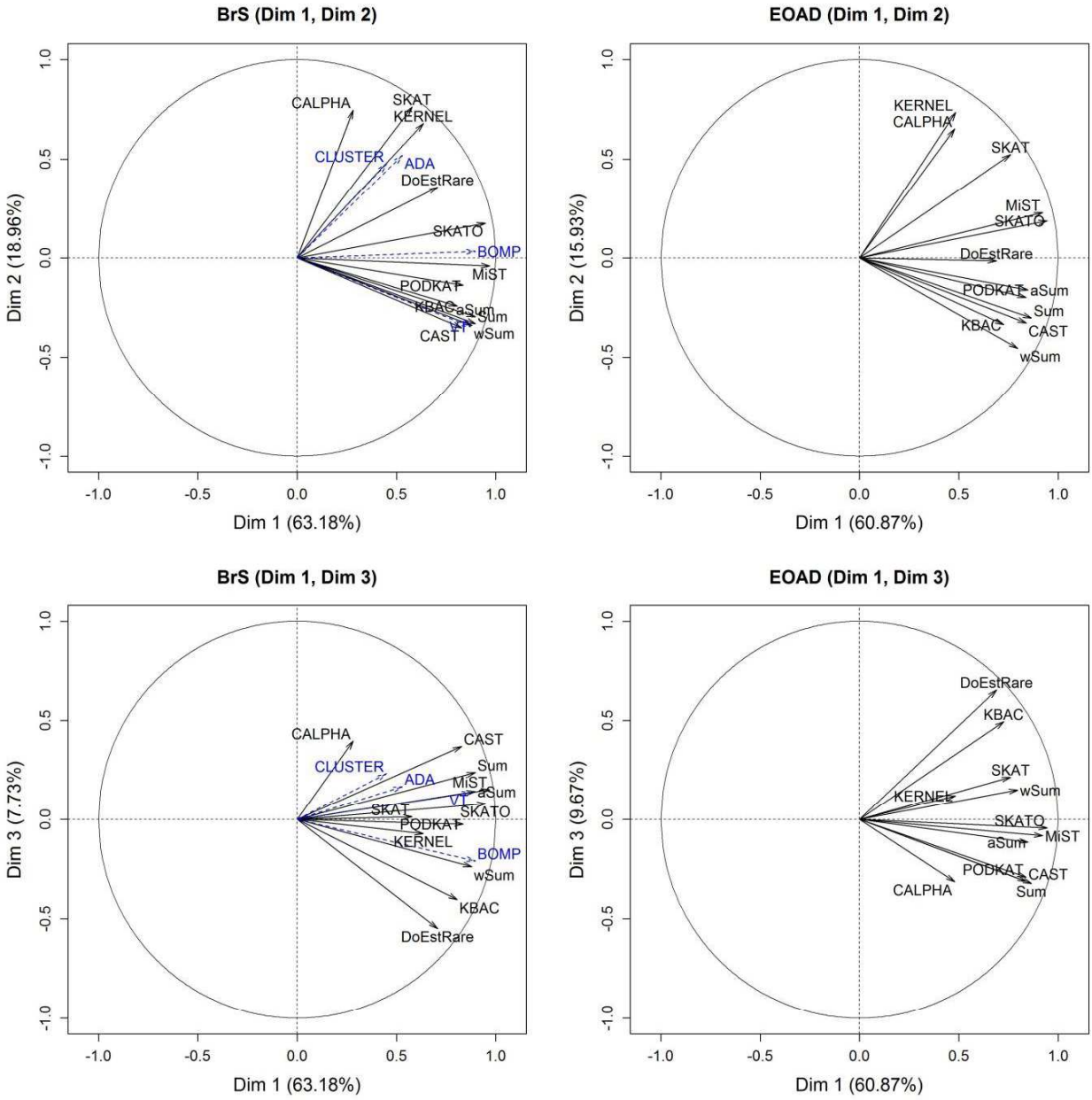


Figure 43. Cercle des corrélations de l'ACP pour les données BrS (gauche) et pour les données EOAD (droite).

L'ACP est appliquée aux $-\log_{10}$(p-value) obtenus à partir de 12 différents tests. Les données BrS incluent 58 gènes et les données EOAD incluent 17 409 gènes autosomaux. Les variables illustratives, ne servant pas à la construction des axes de l'ACP, sont indiquées par des flèches en pointillés bleus.

Les cercles de corrélation pour les deux ACP, en Figure 43, montrent une certaine cohérence générale entre les résultats des différents tests. En effet la première composante principale, pour chaque ACP, permet d'opposer les gènes les moins significatifs aux gènes les plus significatifs pour la quasi-totalité des méthodes. On peut toutefois observer des différences entre les tests. C-alpha, KERNEL et SKAT fournissent les résultats les plus différents des autres et contribuent en partie à la deuxième composante de chaque ACP. Enfin, la troisième composante permet d'opposer les tests CAST et Sum aux tests DoEstRare et KBAC.

Dans l'analyse des données BrS, le gène détecté par le plus grand nombre de tests est *SCN5A*, gène majeur impliqué dans le BrS (Tableau 13). En considérant le seuil de significativité corrigé (α=0.05/58=0.00086) par la méthode Bonferroni pour les tests multiples, *SCN5A* est le seul gène significatif.

Tableau 13. P-values pour le gène *SCN5A*

CAST	Sum	wSum	aSum	MiST	VT	CALPHA	SKAT
0,00252	0,00097	0,00014	0,00101	0,00100	0,00054	0,18729	0,02146

SKATO	KBAC	DoEstRare	ADA	CLUSTER	KERNEL	PODKAT	BOMP
0,00025	0,00048	0,00361	0,07707	0,14769	0,00975	0,00049	0,00007

Les cellules en rouge correspondent aux valeurs significatives en considérant le seuil α=0.05/58=0.00086.

En Annexe IX, sont présentés les résultats pour les gènes candidats liés au BrS (10 sur les 21 du design ont été testés après filtre) pour les 16 tests. Nous pouvons observer les positions de ces gènes sur le graphe des individus de l'ACP dans la figure suivante (Figure 44). En termes de significativité, ils sont très éloignés du gène *SCN5A*.

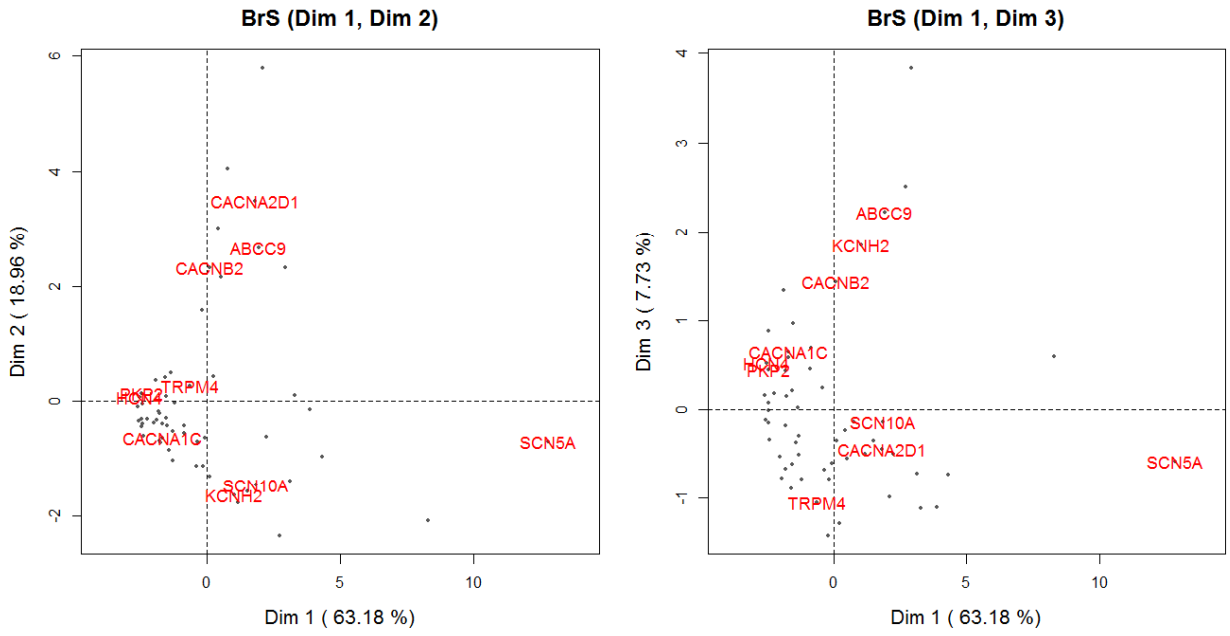


Figure 44. Graphe des individus de l'ACP pour les données BrS.

Les gènes en rouge sont les gènes candidats liés au syndrome de Brugada.

Quant à l'analyse des données EOAD, en considérant le seuil de significativité après correction de Bonferroni, $\alpha=0.05/17409=2.9.10^{-6}$, nous n'avons pas pu identifier le gène *SORL1* décrit dans la publication de Nicolas et al. (2015). Ceci est dû aux étapes de filtrage que nous avons modifiées, quant au choix des individus, pour être au plus proche des filtres pour l'analyse du BrS. Les résultats obtenus avec le test DoEstRare sont représentés à l'aide du Manhattan plot en Figure 45. Le gène le plus significatif, *KRTAP5-5*, est avec certitude un faux-positif, avec les variants rares présents dans une zone de quelques paires de bases et chez les mêmes individus (voir Annexe X). Le second gène le plus significatif avec le test DoEstRare, CELA3B, est lui aussi un faux-positif, pour la même raison.

Des analyses supplémentaires sont requises pour valider les résultats de l'analyse. Les Q-Q plots suggèrent une inflation des p-values pour certains tests (voir Annexe X). Ces tests sont KBAC(λ_{50}=1.13), SKAT(λ_{50}=1.15), SKAT-O(λ_{50}=1.14), MiST(λ_{50}=1.36) et DoEstRare (λ_{50}=1.13) (définition du *genomic inflation factor* λ_{50} dans la partie p31). Pour les tests SKAT et SKAT-O, nous soupçonnons que l'ajustement des p-values pour les petits échantillons entraîne une inflation. Sans ajustement, les tests sont trop conservateurs et ne permettraient pas de détecter les associations. Une des solutions serait d'effectuer des permutations pour

mieux estimer la p-value. Mais nous n'avons pu le mettre en pratique, car les temps de calcul sont beaucoup trop longs. L'origine des biais statistiques pour les autres tests reste à explorer au niveau de la structure de la population et des erreurs de génotype, bien qu'aucune structure de population n'ait été signalée par Nicolas et al. (2015). On peut aussi noter que le test C-alpha se comporte anormalement avec beaucoup de p-values estimées à 1, d'où l'estimation de l'inflation λ_{50}=0.22. Pour les tests CAST, Sum, wSum, aSum, KERNEL et PODKAT, qui présentent une inflation correcte avec $\lambda_{50} \leq 1.05$, nous ne pouvons identifier aucun gène significatif au seuil $\alpha=2.9.10^{-6}$ (0.05/17409).

Note : Ayant effectué une ACP, l'inflation ne devrait avoir aucun impact sur l'interprétation. En effet les résultats de significativité sont centrés et réduits pour cette analyse exploratoire.

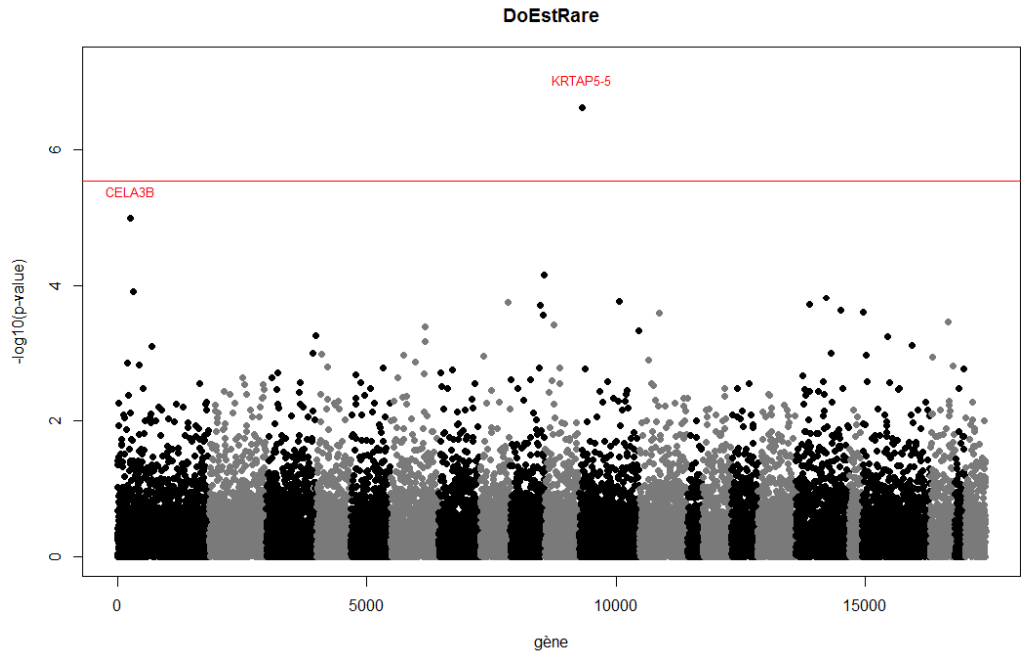


Figure 45. Manhattan plot de l'analyse d'association de l'EOAD avec le test DoEstRare.
Les gènes sont triés selon la position du début du gène. Les chromosomes autosomaux sont représentés par des couleurs alternées. La ligne rouge correspond au seuil de significativité $\alpha=2.9.10^{-9}$.

Variabilité des résultats liés aux filtres.

Avec l'analyse d'association sur les données EOAD, nous n'avons pas retrouvé le gène *SORL1* qui a été identifié par l'équipe de Rouen [184]. Nous avons alors essayé différents

filtres et appliqué les tests les plus rapides. En plus de CAST, qui a été utilisé dans l'analyse effectuée par Nicolas et al. (2015), nous avons aussi appliqué les tests Sum, wSum, SKAT et SKAT-O.

Dans notre comparaison, nous avons étudié l'impact de ne prendre en compte que les individus présentant une histoire familiale de la maladie. L'étude de Nicolas et al. (2015) s'est en effet concentré sur ces personnes, avec l'hypothèse d'une composante génétique plus élevée. Nous avons étudié l'impact du choix des variants en n'incorporant que les variants les plus fonctionnels avec l'annotation de *Sequence Ontology*[86]. Les termes *Sequence Ontology* retenus sont ceux avec impact prédit élevé ou modéré. Cela concerne les termes : SO:0001893 *transcript_ablation*; SO:0001574 *splice_acceptor_variant*; SO:0001575 *splice_donor_variant*; SO:0001587 *stop_gained*; SO:0001589 *frameshift_variant*; SO:0001578 *stop_lost*; SO:0002012 *start_lost*; SO:0001889 *transcript_amplification*, SO:0001821 *inframe_insertion*; SO:0001822 *inframe_deletion*; SO:0001583 *missense_variant*; SO:0001818 *protein_altering_variant*. Certains termes ne sont retrouvés, tels que les variants dans les sites d'épissage génétique, car nous nous sommes focalisés sur les régions CDS pour l'étude des positions. Les nombres de cas et les nombres de gènes analysés, pour les différents filtres, sont indiqués dans le Tableau 14. Pour l'ensemble des analyses, 15 996 gènes ont été testés.

Tableau 14. Filtres effectués sur les données EOAD.

	Tous les individus	**Les individus avec une histoire familiale de la maladie**
Tous les variants rares	« alllind_all_var » 431 cas et 555 témoins 17 898 gènes analysés	« famind_allvar » 185 cas et 555 témoins 17 502 gènes analysés
Les variants rares dits fonctionnels	« allind_functvar » 431 cas et 555 témoins 16 783 gènes analysés	« famind_functvar » 185 cas et 555 témoins 15 996 gènes analysés

Afin de pouvoir résumer cette variabilité de résultats selon les différents filtres, nous avons appliqué une AFM sur le tableau des -log10(p-value). Les individus statistiques sont les gènes, les groupes sont les différents filtres, et les variables correspondent aux tests. Nous avons considéré dans l'analyse, seulement les gènes les plus significatifs, avec une p-

value<=10^{-3} pour au moins un test. Ceci permet de mieux observer la variabilité parmi ces 165 gènes les plus significatifs.

Pour interpréter l'AFM, nous nous concentrons sur les trois premiers axes qui expliquent respectivement 36.47%, 23.89% et 13.68%, soit un total de 74.04% de l'inertie. Nous pouvons observer de grandes différences de résultats selon le choix des cas à incorporer dans l'analyse d'association, ce qui est visible avec le deuxième axe de l'AFM. Les résultats diffèrent aussi, dans une moindre mesure, selon le choix des variants à incorporer dans l'analyse.

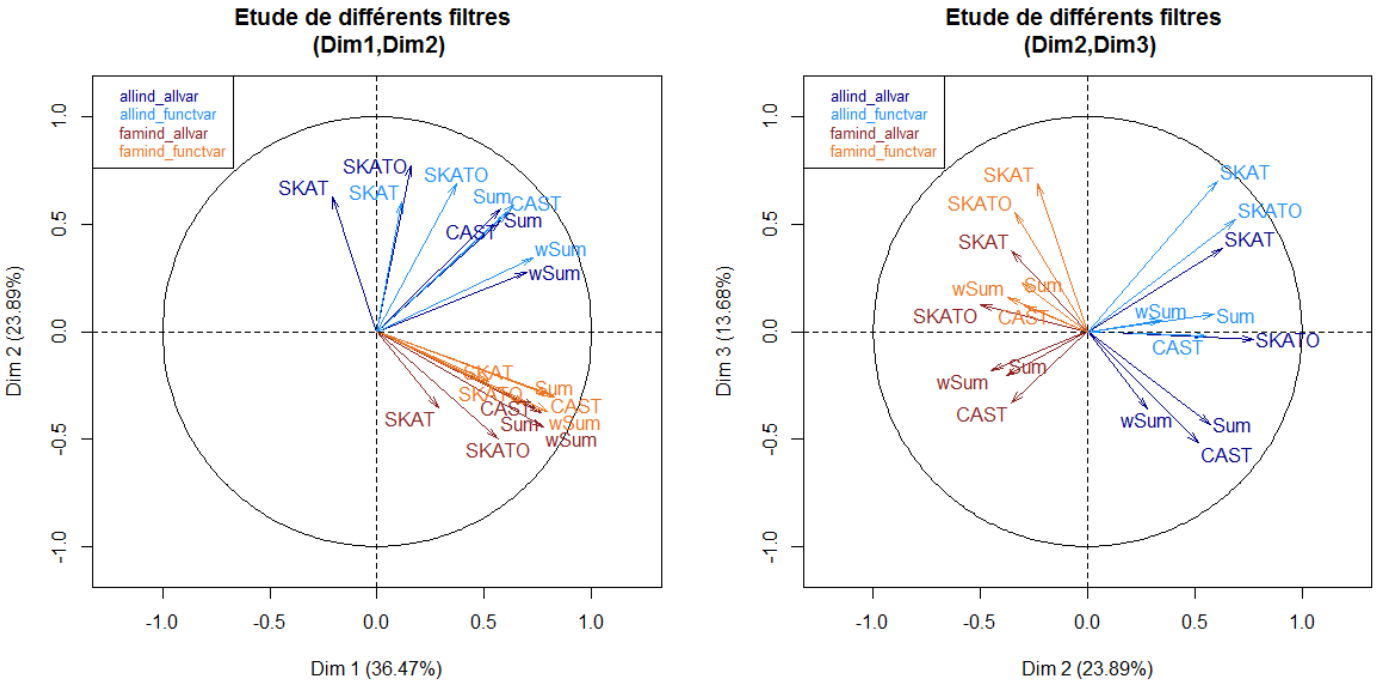


Figure 46. Cercles des corrélations de l'AFM pour l'étude de l'impact du choix des individus et des variants.

L'AFM est effectuée sur les -log10(p-value). Les individus statistiques sont 165 gènes présentant une p-value<=10^{-3} pour au moins un des tests. Les groupes sont les 4 filtres essayés. Les variables sont les différents tests appliqués.

Le gène *SORL1* n'est significatif pour aucun des filtres choisis (Tableau 15), mais est proche du seuil de significativité α=3.13e-06 (15996 gènes pour le filtre « famind_functvar »). Il reste tout de même loin du niveau de significativité décrit par Nicolas et al. (2015) avec une p-value de 3.82e-7. Après discussion, le filtre réalisé dans la publication a été beaucoup

stringent sur les variants fonctionnels avec les annotations des logiciels Polyphen2 HumDiv, Mutation Taster et SIFT [185]. Nous avons de plus retiré certains cas qui avaient été séquencé avec un design de capture plus ancien, pour éviter les biais statistiques. Cependant ces individus, ayant été les premiers inclus, étaient importants pour les résultats de l'analyse.

Nous n'avons pas trouvé d'autres gènes significatifs avec ces différents filtres. Il aurait été intéressant d'appliquer les tests KBAC et DoEstRare, en plus des 5 tests présentés dans cette partie.

Tableau 15. Résultats de significativité pour le gène *SORL1*.

filtre	CAST	SKAT	SKATO	Sum	wSum
allind_allvar	0.011	0.468	0.007	0.005	0.001
allind_functvar	0.014	0.487	0.006	0.006	4.24e-04
famind_allvar	9.59e-04	0.076	1.64e-04	2.57e-04	1.69e-05
famind_functvar	5.69e-04	0.057	9.06e-05	1.30e-04	4.48e-06

III- DISCUSSION

Quel test d'association utiliser pour l'analyse des variants génétiques rares ?

Depuis une dizaine d'années, de nombreuses méthodes statistiques ont été développées pour tester l'association entre un groupe de variants rares et une maladie. En pratique les tests les plus utilisés sont les tests CAST (ou appelé CMC), wSum, SKAT et SKAT-O, car simples d'utilisation et rapides grâce à l'absence de procédures de permutations. De nouvelles méthodes paraissent chaque année avec des améliorations et des enjeux différents. Il est dans ce cadre très difficile de savoir quel test choisir. C'est pourquoi nous avons mené une étude pour comparer les principales stratégies de test au moyen de simulations et de l'application à des données de séquençage.

Nous avons mené des simulations de données génétiques sous différents scénarios pour comparer la puissance et l'erreur de type I des tests. Certains tests se démarquent des autres en termes de **puissance** dont KBAC, SKAT-O, MiST et notre test DoEstRare. Ces tests sont notamment adaptés à la présence de variants neutres dans les données, qui ajoutent « du bruit » diminuant la capacité de détecter les signaux d'association.

L'application des tests à de vraies données génétiques s'est révélée très informative. Les tests fournissent des **résultats de significativité plus ou moins similaires**. On peut observer un certain consensus dans les résultats obtenus avec les *burden tests*. Il est aussi intéressant de constater que les gènes les plus significatifs peuvent différer d'un test à l'autre. Par exemple le test SKAT fournit des résultats très différents de ceux obtenus avec les *burden tests*. C'est pourquoi il est souvent recommandé dans la littérature d'**utiliser plusieurs tests statistiques** reposant sur des hypothèses différentes pour identifier le maximum de gènes. En effet les mécanismes biologiques sous-jacents sont complexes et peuvent varier d'un gène à un autre.

Bien sûr de nombreux autres tests sont parus dans la littérature, mais nous avons fait le choix d'étudier les principales stratégies. Il est important de retenir que **chaque stratégie présente des avantages et des inconvénients**, comme il a déjà été résumé pour les *variance-component tests* et les *burden tests* dans les revues de Lee et al. (2014) [30] et de Bomba et al. (2017) [32]. Nous avons de plus exploré l'intérêt de prendre en compte les positions des variants pour identifier d'éventuels regroupements de variants rares à risque.

Des approches permettant de **combiner les différents tests statistiques** sont proposées dans la littérature. Parmi les tests présentés dans cette thèse, SKAT-O est lui-même un test permettant de combinant un *burden test* et le *variance-component test* SKAT. DoEstRare, est aussi une combinaison d'une stratégie *burden* et d'un test comparant les distributions des positions des variants. Pour aller plus loin, Greco et al. (2016) [186] propose de combiner les p-value obtenues de différents tests avec une méthode de Fisher.

D'autres facteurs importants jouent sur le choix d'un test statistique plutôt qu'un autre comme **la facilité d'utilisation** et le **temps de calcul**. Les tests les plus utilisés CAST, SKAT et SKAT-O sont très rapides et faciles d'utilisation car ne reposant pas sur des procédures de permutations des phénotypes pouvant être très lourdes en temps de calcul. Le développement d'outils bioinformatiques est très important pour la facilité d'utilisation des tests. Par exemple des outils bioinformatiques comme RVTESTS (2016) [187] ou SEQSpark (2017) [188], permettent d'effectuer différentes étapes d'analyse des variants rares à partir de données de séquençage, avec l'implémentation de quelques tests qui sont les plus utilisés (CAST, Sum, wSum, VT, SKAT et SKAT-O).

DoEstRare

DoEstRare a été développé afin de pouvoir identifier des gènes où les variants à risque sont concentrés dans une région spécifique. L'équipe de génétique de l'institut du thorax, étudiant le prolapsus valvulaire mitral, a remarqué que des variants rares sont localisés dans une région du gène *FLNA* pour les patients atteints de cette maladie [33]. Très peu de méthodes ont été développées pour ce type de scénario génétique et c'est pourquoi nous avons développé DoEstRare.

L'idée de ce test est de combiner une stratégie *burden* agrégeant les fréquences alléliques à une comparaison des distributions des positions des variants. Les deux études de simulation ont montré que DoEstRare conserve une **très bonne puissance dans les nombreux scénarios génétiques étudiés**. De plus sa puissance augment légèrement avec la présence d'un *cluster* de variants à risque dans le gène. D'autres scénarios de simulation avec des variants protecteurs pourraient être considérés comme ceux présentés dans la Figure 47, bien que leur pertinence reste discutable d'un point de vue biologique. La présence de variants à risque et de variants protecteurs au sein d'une même région spécifique paraît peu vraisemblable et

DoEstRare n'y est pas adapté de par sa construction. Ce scénario est tout de même envisagé lors de simulations par Wang et al. (2017) [189], avec une méthode permettant de localiser une fenêtre génétique idéale pour appliquer un test d'association tel que SKAT et SKAT-O. Cette méthode consiste à appliquer le test sur différentes partitions successives du gène, de plus en plus fines.

Avec l'application à des données de séquençage pour deux maladies, le syndrome de Brugada et la maladie d'Alzheimer d'apparition précoce, nous avons remarqué que DoEstRare fournit **des résultats de significativité des gènes différents des autres tests**. Cette stratégie repose en effet sur une autre hypothèse, permettant d'explorer de nouvelles pistes génétiques pour expliquer les maladies. Ceci est nettement plus flagrant pour la maladie d'Alzheimer avec les gènes les plus significatifs étant très différents de ceux observés avec les autres tests. Cependant les gènes que nous avons identifiés sont des faux-positifs et présentent des séquences très courtes mal séquencées.

Plusieurs **pistes d'amélioration statistiques** pour DoEstRare sont à considérer :

- Utilisation de poids autres que ceux définis dans la description ;
- Adaptation de DoEstRare à des traits quantitatifs ;
- Utilisation d'autres noyaux pour l'estimation des fonctions de densité des positions dans le but de réduire le temps de calcul.

Le **système de pondération** implémenté pour le test DoEstRare, inspiré de KBAC, repose sur l'hypothèse que les variants causaux sont à risque et non protecteurs. Nous avons envisagé une construction de la statistique de test afin de prendre en compte les variants protecteurs, sans grand succès. Cependant il serait envisageable de rendre le système de pondération plus flexible dans le package R, en laissant la possibilité à l'utilisation d'autres poids. En effet il serait intéressant de pouvoir incorporer des informations fonctionnelles disponibles dans les bases de données. Kircher et al. (2014) [190] décrivent un score appelé « *Combined Annotation Dependent Depletion* » (CADD) permettant de combinant différentes annotations des bases de données. Avec le même objectif, Ionita-Laza et al. (2016) [91] ont mis en place une mesure appelée « Eigen ». Ces scores de fonctionnalité sont utilisés afin d'évaluer la pathogénicité des variants. Ces mesures peuvent être utilisées comme poids dans la statistique de test DoEstRare.

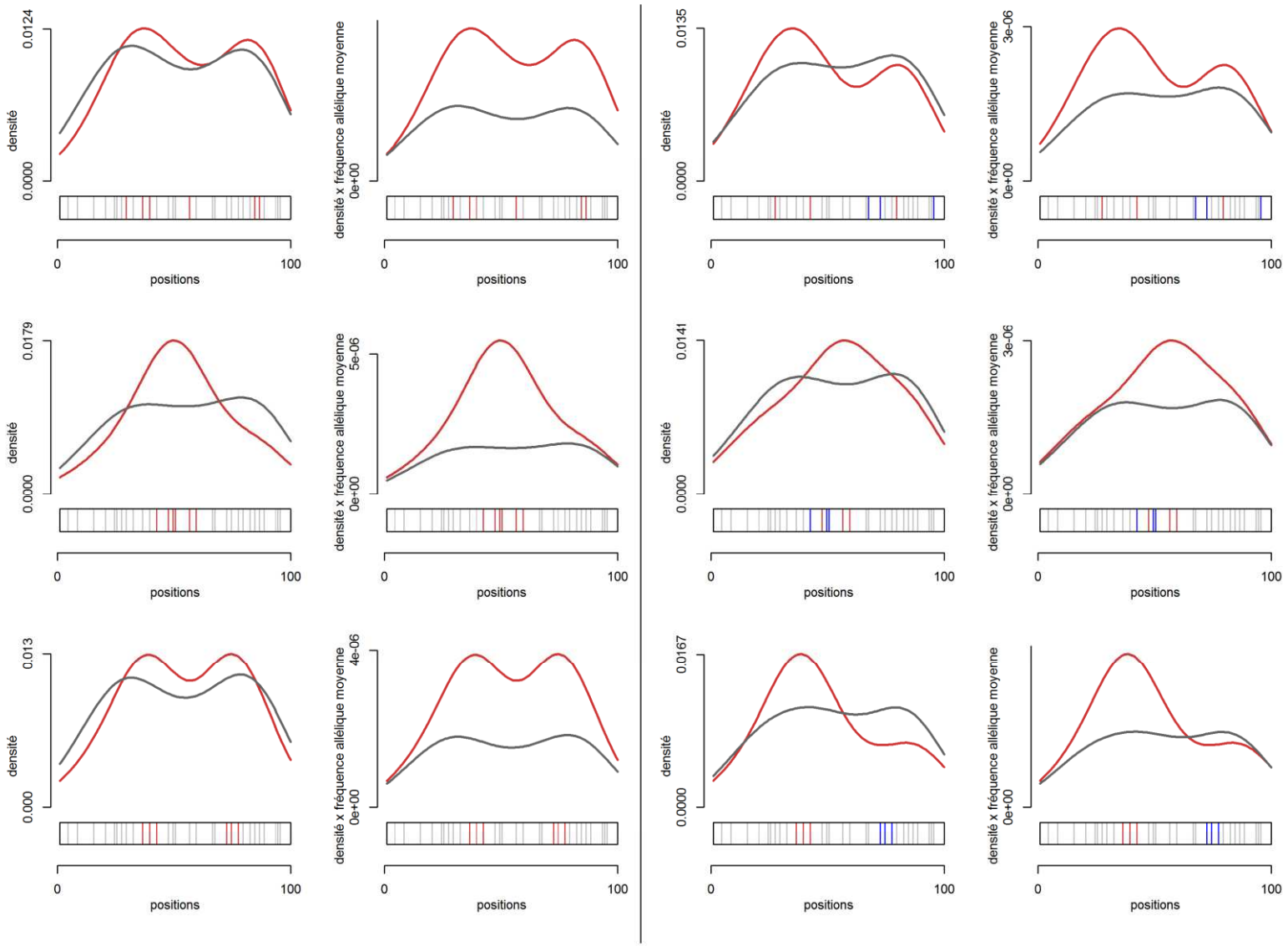


Figure 47. Différentes possibilités de scénarios en fonction de la localisation des variants à risque ou protecteurs.

Les graphes représentent les densités des positions des mutations et les densités multipliées par la fréquence allélique moyenne. Les courbes en rouge son pour les cas et les courbes grises pour les témoins. Différents scénarios génétiques sont considérés en fonction de la répartition des variants à risque (rouge) et des variants protecteurs (bleu). Les variants neutres sont grisés.

DoEstRare compare les densités de position des mutations entre les cas et les témoins et est donc adapté à des traits binaires. Pour l'adapter à un **trait quantitatif**, une idée serait de classifier les individus en deux groupes. Li et al. (2011)[191], Barnett et al. (2013)[192] et Zhou et al. (2016)[193], discutent de l'avantage de l'échantillonnage d'individus avec des phénotypes extrêmes pour mieux identifier les facteurs génétiques lors des analyses d'association de variants génétiques rares. Ce type d'échantillonnage permettrait d'enrichir le nombre de variants rares chez les individus avec un phénotype extrême. Par exemple, des traits quantitatifs pour l'hypertension et l'obésité peuvent être étudiés de cette manière. Cependant il est moins évident d'analyser un trait quantitatif basé sur un échantillonnage aléatoire des individus dans la population.

Enfin, une autre voie d'amélioration de DoEstRare serait la recherche d'un noyau pour l'estimation de la densité afin de réduire les temps de calcul. En effet avec le noyau gaussien qui est employé, la densité est estimée pour chaque position du gène. Selon la taille du gène, les temps de calcul peuvent être longs. Une adaptation de l'estimation de la densité en fonction des régions riches ou pauvres en mutations rares serait bénéfique.

Le test DoEstRare est implémenté dans un **package R** déposé sur le CRAN (https://CRAN.R-project.org/package=DoEstRare), afin de pouvoir être utilisé facilement. De plus, la statistique de test a été implémentée dans le langage C afin de plus de rapidité au niveau des calculs des fonctions de densité. Cependant, en comparaison avec les autres tests statistiques DoEstRare est plutôt long en temps de calcul. Les deux procédures de permutation des phénotypes classique et adaptative ont été implémentées. Nous recommandons très fortement la procédure adaptative pour réduire les temps de calcul. Ce package est pour le moment très minimaliste avec la seule fonction permettant d'appliquer le test. D'autres fonctions doivent y être intégrées pour, par exemple, visualiser la répartition des variants rares dans le gène, ce qui est important pour l'interprétation des éléments significatifs.

Enjeux perçus lors de l'analyse de données réelles

Lors de l'application des tests statistiques à des données de séquençage pour le syndrome de Brugada (BrS) et la maladie d'Alzheimer d'apparition précoce (EOAD), des enjeux d'analyse ont été mis en évidence tels que la variabilité des résultats avec le choix des filtres et les gènes faux-positifs dus les erreurs de séquençage.

L'observation la plus frappante est la **variabilité des résultats** obtenus lors des analyses en fonction des filtres choisis pour l'analyse des données EOAD. Ayant appliqué des filtres différents de ceux de Nicolas et al. (2015) [184] (pour être cohérent avec l'analyse effectuée sur les données BrS), nous n'avons pu identifier le gène *SORL1* décrit. Selon les individus et les variants incorporés dans les analyses, les résultats sont très différents. En effet les gènes les plus significatifs ne sont pas les mêmes, selon le choix de prendre tous les individus ou ceux ayant une histoire familiale de la maladie. Ceci est aussi le cas pour le filtre sur les conséquences fonctionnelles des variants. Ces étapes sont tout aussi décisives que le choix du test statistique pour l'interprétation des résultats.

Cette variabilité des résultats peut s'expliquer du fait de la **faible fréquence des variants analysés** et du **manque de puissance** entrainé. Selon l'échantillonnage réalisé dans la population les résultats des études d'association peuvent être très différents, variant entre non significatif et significatif. Pour illustrer ce propos, la Figure 48 présente les résultats d'une simulation d'un groupe de 8 variants rares avec une MAF donnée et un OR de 2. Une grande population de cas et de témoins est simulée, le test SKAT-O (parce qu'il globalement de bons résultats) est alors appliqué à un échantillonnage de 1000 cas et 1000 témoins. Pour observer la diversité des résultats, l'échantillonnage est répété 1000 fois. La puissance de détecter des associations, i.e. la proportion de gènes significatifs dans la simulation, est plus importante avec des variants plus fréquents, avec l'identification du gène quel que soit l'échantillonnage. A l'inverse, pour des variants très rares, la puissance est très faible, les résultats sont globalement non significatifs. Enfin, pour les variants rares ou peu fréquents, la puissance est suffisamment élevée pour détecter l'association, mais les résultats peuvent être significatifs ou non significatifs selon l'échantillonnage réalisé dans la population. Par exemple pour une MAF de 0.01, la p-value peut varier entre 0.089 et $1{,}98.10^{-14}$. Le schéma de simulation est bien sûr très simple mais permet d'illustrer le manque de robustesse des résultats des tests selon le choix des individus et des variants incorporés dans l'analyse. Une des solutions pour remédier à ce manque de robustesse est d'augmenter la puissance avec une réplication de l'analyse avec un grand nombre d'individus.

Il est difficile de recommander une taille d'échantillon pour obtenir une puissance de 80%, étant donné l'architecture génétique dépendant du gène analysé et du test employé. Auer et al. (2016) [194] discutent des tailles d'échantillon requises pour atteindre une puissance de 80% avec les tests SKAT, CAST et Sum au seuil $\alpha=2.5.10^{-6}$. Dans leur analyse, une maladie avec

une prévalence de 1% est simulée à partir de données d'exome avec tous les variants rares (MAF<0.01) ayant un OR de 1.5. Selon leur simulation, 30% des gènes nécessiteraient 100 000 individus dans l'étude, et seulement 1.25% des gènes seraient détectables avec 10 000 individus. L'analyse de variants rares nécessite des tailles d'échantillon de grande échelle pour détecter des associations, même avec l'utilisation de tests développés à cet effet. Bien sûr l'OR envisagé par Auer et al. (2016) est faible et des variants avec des effets plus forts devraient être considérés, même si une telle configuration favorable serait certainement contrebalancée, dans des données réelles, par le nombre de variants neutres.

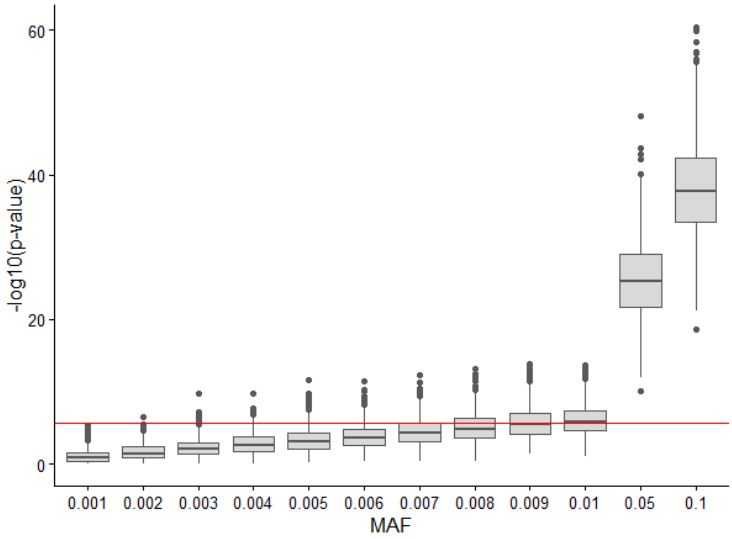


Figure 48. Variabilité de la significativité des tests pour différents échantillonnages de la population selon la MAF.

Ces simulations sont réalisées à partir du schéma de Basu et Pan (2011) pour le scénario avec 8 variants causaux avec un OR égal à 2, et sans variant non causal. La MAF des variants est fixée pour chaque simulation, et non tirée aléatoirement, afin de voir l'évolution de la significativité des tests. Un grand jeu de données avec 2 000 000 d'individus cas et témoins est généré et est ensuite partagé en 1000 parties pour les tests. Le test réalisé est SKAT-O, car rapide d'exécution. La ligne rouge correspond au seuil de significativité corrigé $\alpha=2.5.10^{-6}$ (20 000 gènes).

Le **contrôle qualité** est aussi une étape très importante pour l'interprétation, car des faux-positifs peuvent survenir avec les erreurs de séquençage. Cook et al. (2014) [195] ont évalué l'impact d'erreurs de génotypes pour les tests d'association pour variants rares. Ces erreurs peuvent diminuer la puissance et/ou augmenter les erreurs de type I des tests. Les auteurs font la distinction entre ce qu'ils appellent « les erreurs non-différentielles » et les « erreurs

différentielles ». Les erreurs non-différentielles sont dues un processus d'erreur indépendant du phénotype, i.e. la probabilité d'erreur de génotype est la même chez les cas et les témoins. Les erreurs non-différentielles mènent à une diminution de puissance de détection des associations, avec un maintien des erreurs de type I dans la plupart des cas. Les erreurs différentielles sont dues à un processus d'erreur associé au phénotype, i.e. les probabilités d'erreur sont différentes chez les cas et les témoins. Ces erreurs entraînent une augmentation des erreurs de type I, ce qui implique une augmentation du nombre de faux-positifs en pratique. Ces erreurs différentielles sont un facteur de confusion au même titre que les la stratification de population. Elles peuvent survenir lorsque le séquençage a été effectué séparément pour les cas et les témoins. Par exemple, Cook et al. (2014) discutent de l'utilisation des bases de données publiques pour le choix des témoins. Avec ces données, il est difficile de savoir si le processus d'erreur des génotypes est le même que chez les cas.

Les tests pour variants rares peuvent être plus ou moins sensibles aux erreurs de séquençage selon leur répartition. Par exemple, si les erreurs de séquençage surviennent surtout chez l'un des deux groupes, cas ou témoins, les *burden tests* détectent une différence globale de variants rares. Lors de l'analyse des données EOAD, DoEstRare a permis de mettre en évidence des gènes significatifs à cause de petites régions mal séquencées chez certains individus. Pour pallier ces erreurs de séquençage, on pourrait s'inspirer d'autres tests qui ont été développés pour prendre en compte la qualité de séquençage dans la statistique de test [144,150].

En conclusion, tester des groupes de variants rares est une tâche complexe lors de laquelle de nombreux facteurs influent sur les résultats. Contrairement à l'analyse des variants fréquents, les protocoles à suivre, ne sont pas encore bien déterminés. Des recommandations ont toutefois été émises pour détecter le plus d'associations possibles, telles que l'utilisation de tests reposant sur des hypothèses différentes et l'étude des données selon plusieurs filtres.

Dans cette partie sur les tests d'association adaptés à l'étude des variants rares, nous n'avons pas envisagé la présence de structures de population. Or des cas et des témoins provenant d'origines géographiques différentes peuvent être un facteur de confusion pour l'interprétation des résultats des tests. Ceci est bien connu pour les variants fréquents, comme nous l'avons rappelé dans la partie **Préalables de génétique**, mais la question se pose aussi dans le cadre d'analyse de variants rares.

PARTIE II : ÉTUDE DE LA STRUCTURE DE POPULATION DANS LE CADRE DES TESTS D'ASSOCIATION POUR VARIANTS RARES

I- ÉTUDE DE L'IMPACT DE LA STRUCTURE DE POPULATION SUR LES TESTS D'ASSOCIATION POUR VARIANTS RARES

I.1- BIBLIOGRAPHIE

L'impact de la stratification de population est bien connu dans le cadre des GWAS menées sur les variants fréquents, avec une augmentation du nombre de faux-positifs. Lorsque les cas et les témoins proviennent de différentes populations géographiques, il est difficile de savoir si les signaux d'association sont identifiés à cause de différence de statut atteint/non atteint ou à cause de différences de fréquences alléliques entre les populations.

Différentes études ont montré que la stratification de population a un impact sur les résultats des analyses d'association de variants rares. Tintle et al. (2011) [42] et Mathieson et McVean (2012) [43] ont montré, avec l'analyse d'un trait quantitatif, qu'une structure de population **à l'échelle mondiale** entraine une augmentation des erreurs de type I. Tintle et al. (2011) résument les résultats des analyses d'association pour variants rares menées entre des données de séquençage d'exons et des phénotypes simulés, dans le cadre du Genetic Analysis Workshop 17. Les données de séquençage d'exons proviennent du projet 1000 Genomes et incluent 697 personnes non apparentées provenant de 7 populations mondiales différentes [196]. Cette étude met en évidence une grande inflation des erreurs de type I (jusqu'à 50%), des probabilités de faux-positifs élevées (jusqu'à 90%) et une puissance faible de détecter les variants causaux. Dans une autre étude, Mathieson et McVean (2012) ont montré, au moyen de simulations de structures spatiales de populations, des niveaux d'inflation différents selon la fréquence des variants, rare ou fréquent, et la distribution spatiale de la maladie. Lorsque la probabilité de maladie, i.e. d'échantillonner un cas, est très restreinte géographiquement, l'inflation est nettement plus importante pour les variants rares que pour les variants fréquents. Les **variants rares,** étant plus récents, sont en effet **plus localisés géographiquement.** L'échelle spatiale simulée par Mathieson et McVean (2012) se

rapproche d'une échelle mondiale étant donné la valeur élevée de F_{ST} qui est d'environ 0.01 (pour une définition du F_{ST}, voir la partie L'indice de fixation F_{ST} entre populations, p18).

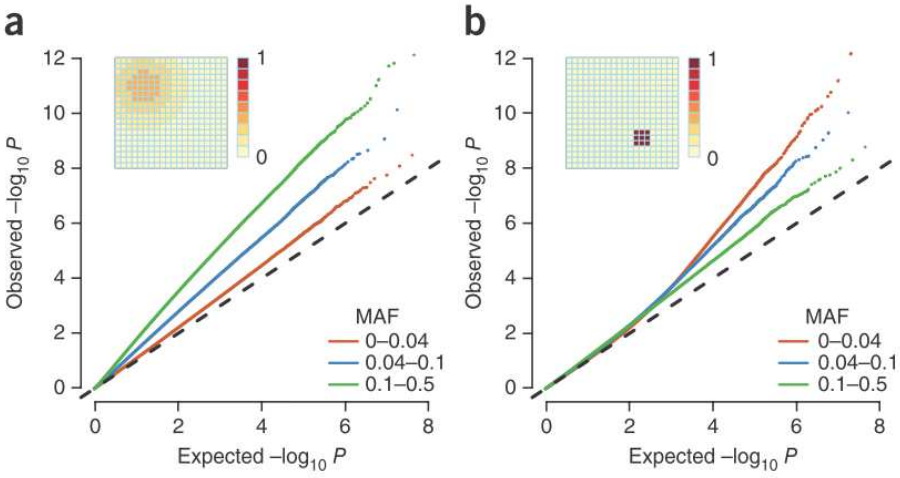


Figure 49. Différentes inflations selon l'analyse de variants rares ou de variants fréquents selon Mathieson et McVean (2012).

Ces Q-Q plots sont réalisés à partir des p-values obtenues à partir des analyses d'association pour un facteur de risque non-génétique avec une distribution spatiale étendue (a) ou restreinte (b). La grille en haut à gauche représente la répartition spatiale des personnes malades.

La stratification de population à l'échelle mondiale entraîne des inflations p-values extrêmement élevées lors de l'analyse de variants rares. O'Connor et al. (2013) [45] et Zawistowski et al. (2014) [47] ont montré l'impact de cette structure de population à une échelle plus fine, telle que **l'échelle européenne**.

O'Connor et al. (2013) ont utilisé le logiciel *cosi* [182] pour simuler 5 populations européennes avec la calibration des paramètres effectuée à partir de données de séquençage d'exome de 316 euro-américains. Afin de préciser le degré de variabilité génétique entre les populations, les F_{ST} indiqués dans l'article sont compris entre 0.005 et 0.01. Différents *burden tests*, tels que CAST, Sum, wSum, VT, et d'autres tests que nous n'avons pas décrits RareCover [107], StepUp et Stepdown [114] présentent des inflations d'erreurs de type I en présence d'une structure de population Européenne.

Zawistowski et al. (2014) ont étudié l'impact d'une structure encore plus fine au moyen de données simulées à partir de données de séquençages pour diverses populations européennes. Les valeurs de F_{ST} varient entre 0.000626 et 0.000866, montrant une structure de population très fine. Ils ont considéré les deux principales catégories de test, avec CAST, CMAT [112], GRANVIL [197] et wSum, pour la catégorie des *burden tests*, et SKAT, C-alpha pour la catégorie des *variance-component tests*. Leur résultat est la présence d'une inflation plus importante pour les *variance component tests* dans le cadre d'une structure très fine de population.

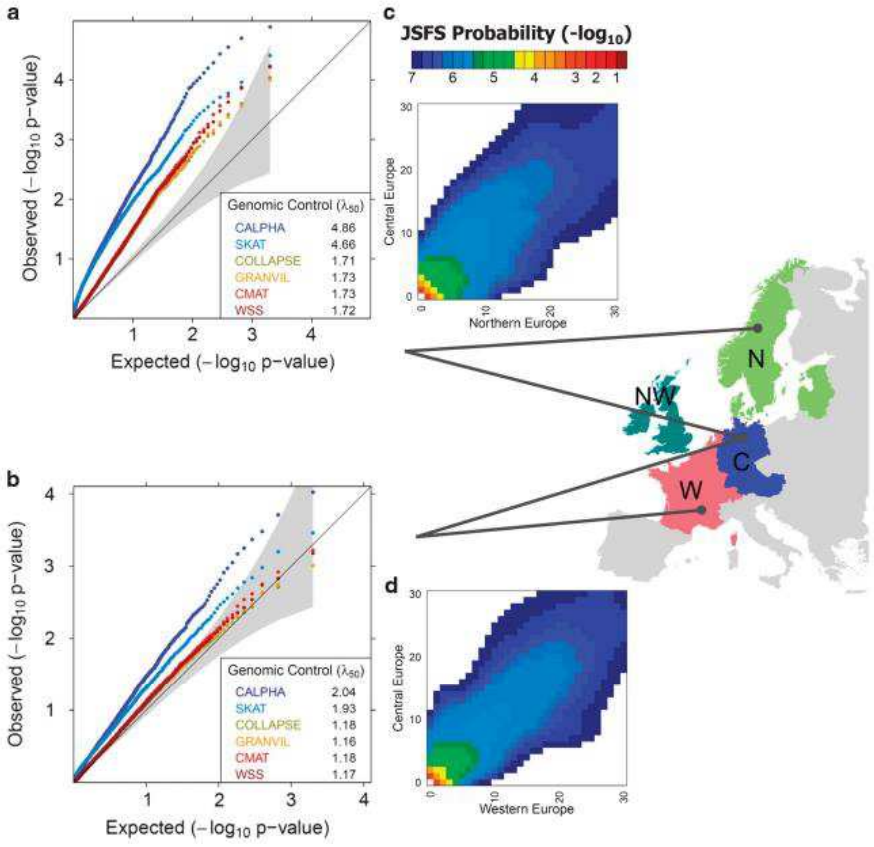


Figure 50. Inflation des p-values lors de la comparaison de populations européennes d'après Zawistowski et al. (2014).

Les données sont simulées à partir des JSFS (*joint site frequency spectrum*) obtenus de l'étude des données de séquençage de deux populations européennes. Le JSFS est la distribution des nombres d'allèles pour deux populations. Les tests sont appliqués avec les cas provenant tous d'une même première population et les témoins d'une deuxième population.

Pour aller plus loin, Zawistowski et al. (2014) montrent que selon la distribution des allèles dans les deux populations, les deux **catégories de tests se comportent de façon différente** en termes d'inflation. Ils considèrent alors deux paramètres appelés « *allele sharing* » et « *weighted symmetry* ». L'*allele sharing* (AS) quantifie les différences de fréquences alléliques entre les deux populations. Il s'agit de la probabilité que deux individus présentent le même allèle mais proviennent de populations différentes. AS=1 signifie qu'il n'y a pas de différences de fréquences alléliques. La *weighted symmetry* (WS) quantifie les différences d'abondance globale de variants rares entre les populations. WS=1 signifie que les fréquences globales de variants rares sont les mêmes dans les deux populations. Lorsque les nombres moyens de variants rares sont différents dans les deux populations, les *burden tests* montrent des inflations plus élevées. Lorsque l'AS s'éloigne de 1, ce sont les *variance-component tests* qui présentent le plus d'inflation. Ces deux paramètres dépendent de nombreux paramètres démographiques. Par exemple, une croissance démographique rapide entraine une augmentation du nombre de variants rares. Ainsi des populations avec des taux de croissance différents sont susceptibles de présenter des différences globales de nombres de variants.

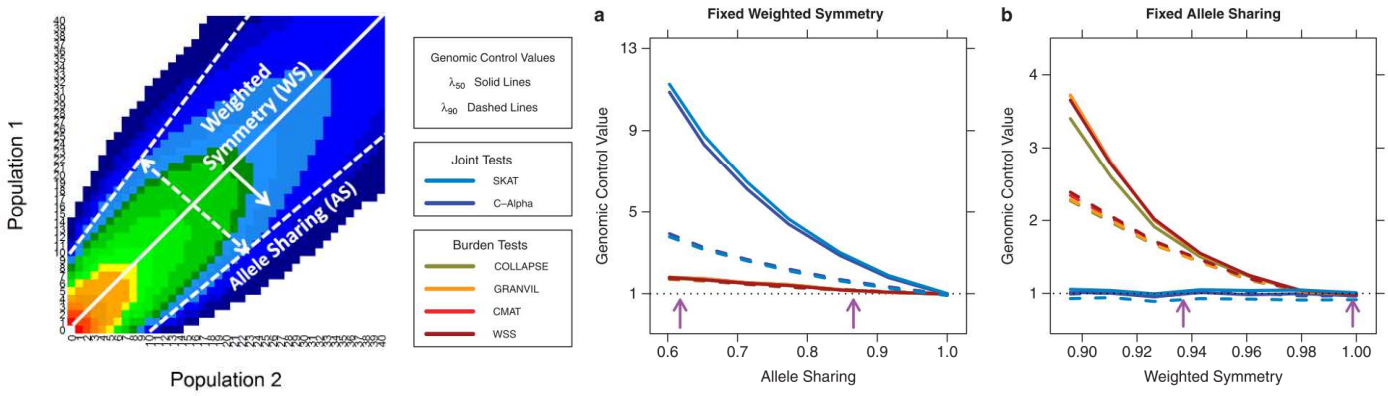


Figure 51. Niveaux d'inflation pour les *variance-component tests* et les *burden tests* selon la structure de population d'après Zawistowski et al. (2014).

Le graphique de gauche permet d'illustrer les notions d'*allele sharing* (AS) et de *weighted symmetry* (WS) à partir du graphe de JSFS. Les graphes de droite présentent le niveau d'inflation en fonction de ces paramètres AS et WS. La valeur « *Genomic control value* » ou aussi appelé « *Genomic inflation factor* » est un indicateur du niveau d'inflation des p-values.

I.2- OBJECTIFS

Précédemment nous avons vu que l'impact d'une structure de population sur les résultats d'association de variants rares a été évalué à différentes échelles géographiques relativement larges. La Figure 52 permet de résumer les échelles étudiées, avec l'indication des valeurs de F_{ST} pour deux études. Les valeurs pour différentes populations européennes proviennent de l'étude de Nelis et al. (2009) [57]. Les valeurs pour différentes villes françaises proviennent de l'étude des données d'exome pour le projet FREX et données par Génin et al. (2016) [198]. Les valeurs de F_{ST} pour les données FREX sont présentes en Annexe XI.

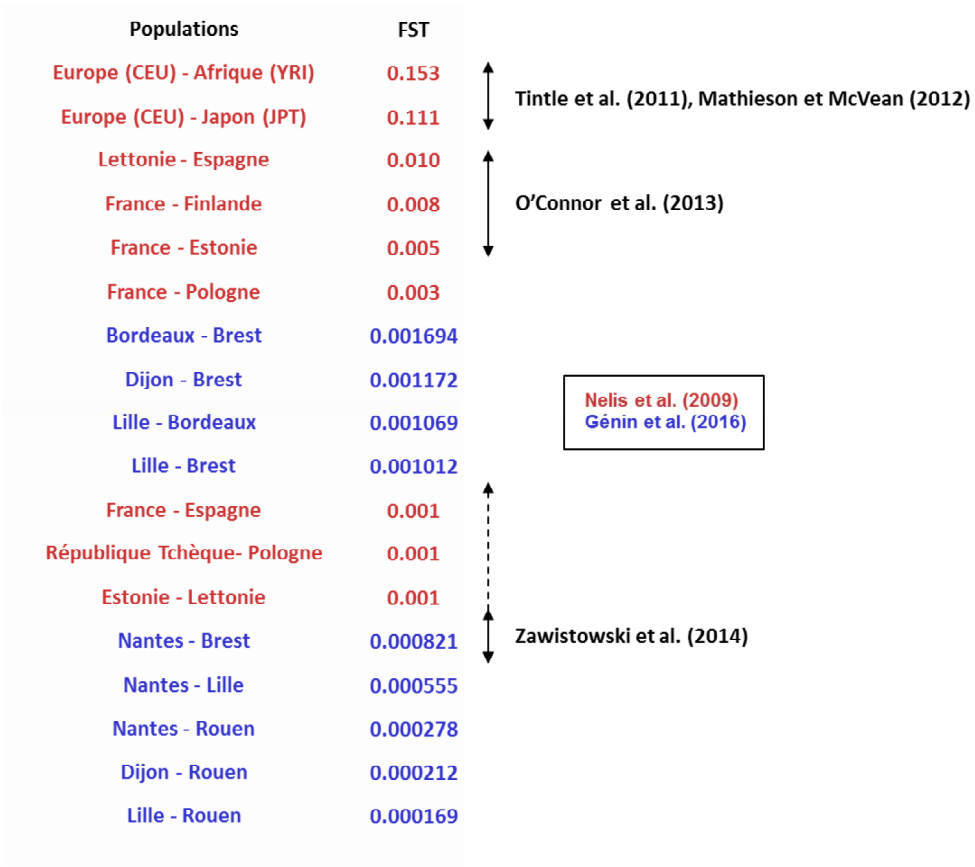


Figure 52. Valeurs de F_{ST} pour différentes populations.

Les valeurs de F_{ST} sont de différentes couleurs selon l'étude dans lesquelles elles ont été reportées. Les publications indiquées sur le côté droite correspondent aux différentes études de l'impact de la structure de population sur les tests d'association pour variants rares. Étant donné que les valeurs de F_{ST} sont arrondies à 0.001 près pour l'étude de Nelis et al. (2009), il est difficile de savoir si les valeurs correspondent à celles de Zawistowski et al. (2014) (flèche en pointillés).

Zawistowski et al. (2014) [47] ont présenté les résultats pour une structure géographique fine avec des F_{ST} allant de 0.0006 et 0.0009. Si l'on se fie aux valeurs de F_{ST} indiquées par Nelis et al. (2009) et Génin et al. (2016), ceci correspond à une structure très fine, i.e. des pays européens voisins ou deux régions françaises assez proches.

Dans cette partie, nous avons pour objectif (i) d'étudier l'impact de **structures de population très fines** sur les résultats d'association pour variants rares, (ii) d'étudier l'impact sur les **différentes catégories de test**, (iii) de discuter de la pertinence des **systèmes de pondération** couramment employés basés sur les fréquences alléliques.

I.3- MÉTHODOLOGIE

Simulation de populations

Le schéma de simulation des données est présenté dans la Figure 53. Comme pour les **Simulations de regroupements localisés de variants rares**, les données génétiques ont été simulées au moyen du logiciel *cosi* [182]. Le modèle démographique appelé « *bestfit* », de Schaffner et al. (2005), a été repris en ajoutant deux sous-populations A et B s'étant séparées de la population européenne il y a 80 générations et dont la taille est de 10 000 haplotypes. Afin d'étudier différents niveaux de stratification de population nous avons fait varier le taux de migration entre ces deux populations, entre 0, 0.001, 0.01, 0.025, 0.05 et 0.1. Ce taux de migration est lié à la distance géographique des populations, avec le taux de migration décroissante avec la distance.

Il est important de garder à l'esprit que dans ce travail, notre intérêt principal concerne l'étude de l'impact de la présence d'une structure de population sur le taux d'erreurs de type I. C'est pourquoi les phénotypes/génotypes sont simulés sous l'hypothèse nulle, i.e. pas d'association entre le phénotype et le gène. Nous considérons dans les simulations que les 1000 cas proviennent de la même population A, et les 1000 témoins proviennent des deux populations A et B avec une proportion variant entre 0%, 25%, 50%, 75% et 100% des individus de la population B. Ces scenarios correspondent au cas où on peut séquencer tous les patients de notre étude et où on a la possibilité d'utiliser des témoins extérieurs (déjà financés) ainsi qu'un nombre, plus ou moins important, de témoins internes mieux appariés.

Les variants rares sont filtrés selon la MAF dans la population A étant la population d'origine des cas de l'étude. Seuls les variants avec une MAF comprise entre 0.001 et 0.01 sont gardés dans l'étude.

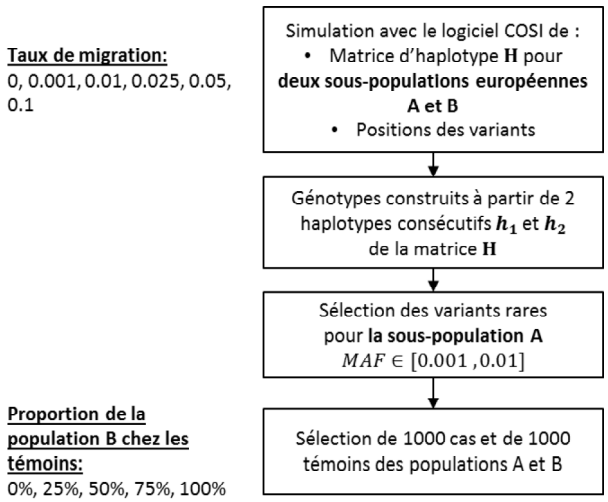


Figure 53. Schéma de simulation pour l'étude l'impact de la stratification de population à échelle fine sur les tests d'association pour variants rares.

Nous avons simulé 10 000 gènes de 10kb selon ce modèle, afin d'avoir des estimations précises des erreurs de type I (calcul des erreurs de type I détaillé à la p86).

Note : Afin de tester la méthode de simulation, nous avions fait des essais préalables avec 16 populations (Annexe XII). Ceci nous a permis de voir que la structure génétique permet de bien refléter la structure géographique simulée. Cependant nous avons préféré une approche plus simple avec 2 populations pour pouvoir interpréter facilement les résultats d'erreurs de type I.

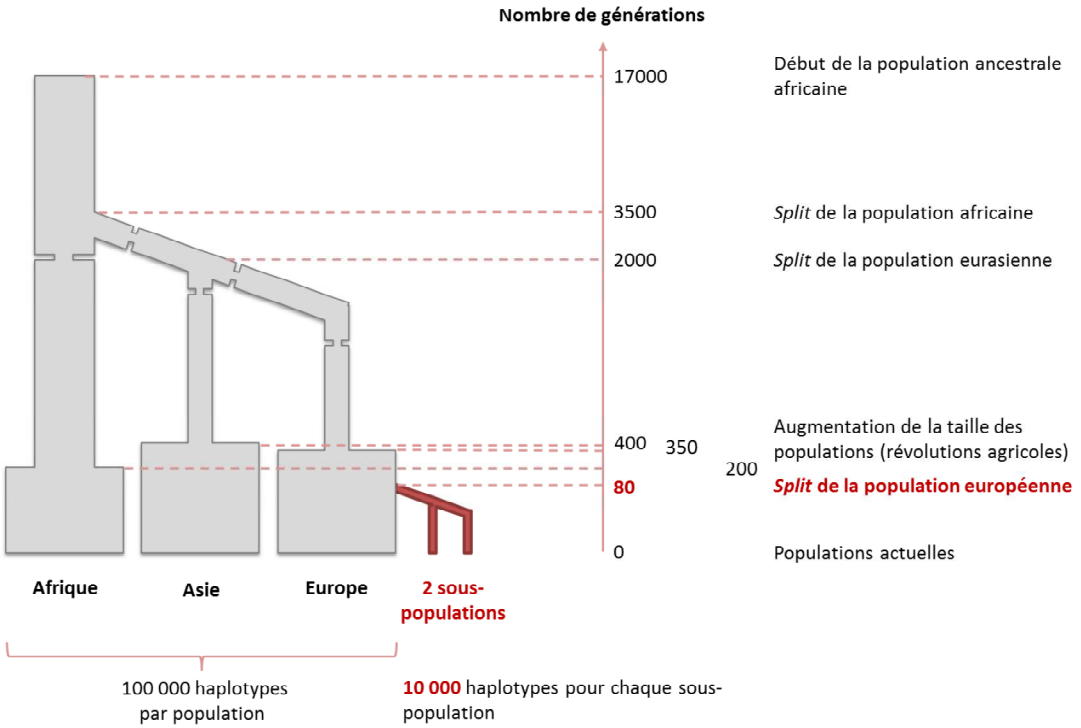


Figure 54. Modèle démographique basé sur celui de Schaffner et al. (2005) pour l'étude de l'impact de stratification de population sur les résultats des tests d'association.

Les tailles effectives des populations sont schématisées par les différentes épaisseurs. Les *bottlenecks* ou goulots d'étranglement (la diminution de diversité génétique due à la séparation de populations) sont représentés par des constrictions temporaires.

Analyse exploratoire des données de simulation

L'échelle géographique est plus ou moins fine selon le paramètre de migration choisi. Afin de pouvoir donner une indication d'échelle, nous avons calculé l'indicateur F_{ST} (p18) entre les deux populations pour chacun des paramètres de migration. Le F_{ST} entre les populations A et B est calculé selon la méthode d'estimation de Weir et Cockerham (1984) [56]. Nous avons utilisé la fonction « calc_wcFst_spop_pairs » venant du répertoire github https://github.com/ekfchan/evachan.org-Rscripts. Cette estimation est effectuée à partir de l'ensemble des variants fréquents non-corrélés (MAF≥5% et $r^2 \leq 0.2$) des 10 000 gènes simulés, avec un échantillonnage de 1000 personnes pour chacune des populations A et B.

Analyse d'association

Nous avons étudié l'impact d'une structure de population sur trois catégories de test dont les deux principales, i.e. *burden tests* et les *variance-component tests*, et les *position tests*. Pour les *burden tests*, nous avons choisi d'appliquer les tests CAST, Sum, wSum, aSum ; et pour les *variance-component tests* nous avons choisi SKAT et SKAT-O. Le test KBAC a aussi été utilisé car il montrait une bonne puissance lors des simulations précédentes. Et enfin pour les *position tests*, les tests PODKAT et DoEstRare ont été appliqués.

De nombreuses méthodes utilisent un système de pondération basé sur le calcul des MAF à partir des données. Parmi les tests utilisés, les tests utilisant un système de pondération pour prioriser les variants sont les tests wSum, SKAT, SKAT-O, KBAC, PODKAT et DoEstRare. Pour étudier l'impact du choix de la pondération, nous avons appliqué différentes versions pondérées des tests Sum, SKAT et SKAT-O : (i) sans pondération ; (ii) avec pondération de Madsen et Browning (2009) en se basant sur la MAF chez l'ensemble des cas et des témoins [98] ; et (iii) la pondération de Wu et al. (2011) [95]. Pour le test Sum, nous avons aussi effectué le test wSum qui utilise le système de pondération de Madsen et Browning (2009) en se basant sur la MAF chez les témoins. Afin d'observer la différence entre les différents systèmes de pondération, la Figure 55 représente les poids en fonction de la MAF.

Pour rappel, Madsen et Browning (2009) proposent pour le test wSum un poids basé sur le calcul de la fréquence allélique chez les témoins. Les variants avec une faible fréquence chez les témoins ont un poids plus élevé. Soient w_j le poids pour le variant j, et $\widehat{MAF_j^U}$ l'estimation de la MAF du variant j chez les témoins. Le poids w_j est alors $\frac{1}{\sqrt{N.\widehat{MAF_j^U}.\left(1-\widehat{MAF_j^U}\right)}}$. Pour pouvoir comparer avec le système de pondération de Wu et al. (2011), nous avons utilisé l'estimation $\widehat{MAF_j}$ de la MAF pour l'ensemble des cas et des témoins.

Wu et al. (2011), proposent un poids similaire à celui de Madsen et Browning pour SKAT. Cependant le poids w_j est basé sur l'estimation de la MAF du variant pour l'ensemble des cas et des témoins et $w_j = dbeta\left(\widehat{MAF_j}, 1{,}25\right)$. La loi beta offre plus de flexibilité quant à la distribution des poids. SKAT étant un test développé pour l'analyse de variants fréquents et de variants rares, les paramètres 1 et 25 sont choisis afin de ne pas donner un poids quasi-nul aux variants peu fréquents.

Tableau 16. Résumé des notations pour les différents systèmes de pondération.

	Notations		
$w_j = 1.$	Sum	SKAT	SKATO
$w_j = dbeta(\widehat{MAF_j}, 1{,}25)$	wSum_betaMAFtot	wSKAT_betaMAFtot	wSKATO_betaMAFtot
$w_j = \frac{1}{\sqrt{N.\widehat{MAF_j}.(1-\widehat{MAF_j})}}.$	wSum_MAFtot	wSKAT_MAFtot	wSKATO_MAFtot
$w_j = \frac{1}{\sqrt{N.\widehat{MAF_j^U}.(1-\widehat{MAF_j^U})}}.$	wSum_MAFctrl		

Les différents systèmes de notation sont résumés dans le Tableau 16. Nous ne pouvons utiliser les tests SKAT et SKAT-O avec des poids basés sur l'estimation de la MAF chez les témoins. En effet, sans procédure de permutations appropriée, on aurait une augmentation des erreurs de type I même en absence de stratification de population. À chaque permutation, le statut cas/témoin est modifié et une nouvelle estimation de la MAF chez les témoins est nécessaire. Cette procédure n'est pas considérée dans le package R SKAT.

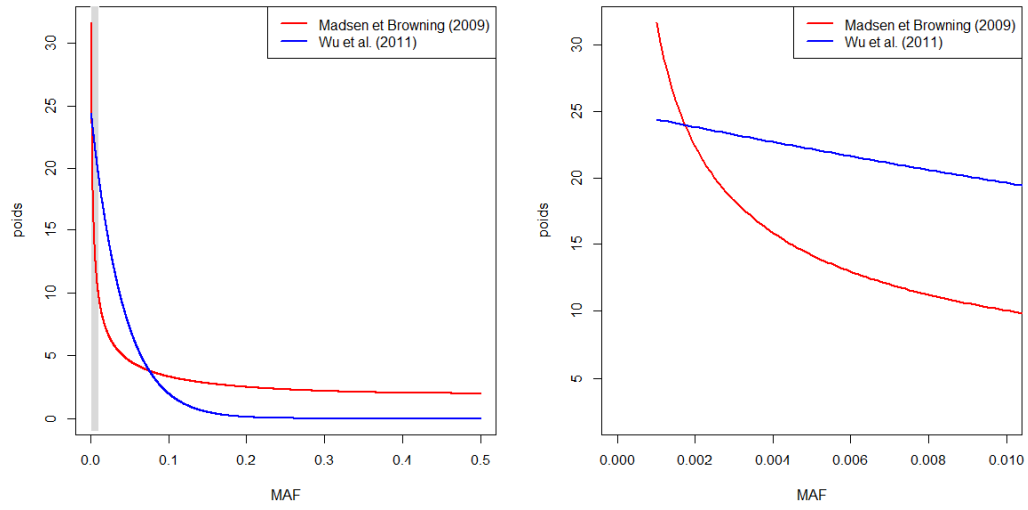


Figure 55. Systèmes de pondération de Madsen et Browning (2009) et Wu et al. (2011).
La figure de gauche représente le poids pour les variants avec une MAF variant entre 0.001 et 0.5. La figure de droite permet de se concentrer sur l'intervalle de MAF entre 0.001 et 0.01.

Pour l'évaluation de la significativité des tests, à l'exception de SKAT, SKAT-O et PODKAT, une procédure de permutations adaptative est effectuée avec pour paramètres α=0.01 et c=0.2.

Note : Comme pour toutes les analyses effectuées précédemment, les *burden tests* sont employés avec un test du score via un modèle de régression logistique. Ceci permettra par la suite d'intégrer facilement des covariables traduisant les structures de population pour corriger les tests.

AFM sur les résultats d'erreurs de type I

Afin d'étudier la variabilité des erreurs de type I entre les tests statistiques, nous avons appliqué une AFM [179] pour résumer l'information issue des 16 scénarios avec une structure de population fine (choix des scénarios avec structure fine discuté dans la partie Résultats). La structure des données analysées avec l'AFM est présentée dans la Figure 56. Les individus sont les 14 tests statistiques appliqués. Les 4 tableaux analysés conjointement correspondent aux 4 taux de migration et sont constitués de 4 variables mesurant la proportion de témoins appartenant à la population B. Il est à noter que nous aurions pu aussi choisir comme groupe les 4 proportions de témoins de la population B.

En réalisant l'AFM nous perdons l'information d'inflation par rapport au seuil α mais nous pouvons distinguer les tests ayant erreurs de type I élevées par rapport à ceux ayant des erreurs de type I faibles pour certains ou tous les scénarios.

groupes	**migration**	**0.01**	**0.025**	**0.05**	**0.1**
variables	% pop B	25 50 75 100	25 50 75 100	25 50 75 100	25 50 75 100
individus	tests	Erreurs de type I	Erreurs de type I	Erreurs de type I	Erreurs de type I

Figure 56. Structure des données pour l'AFM sur les erreurs de type I des tests en présence d'une stratification de population.

I.4- RÉSULTATS

Échelles géographiques des populations simulées

Différentes échelles géographiques ont été simulées en faisant varier le taux de migration. Plus le taux de migration est élevé, plus les populations sont proches géographiquement. La Figure 57 resitue les scénarios simulés par rapport aux observations de F_{ST} effectuées pour les populations européennes d'après Nelis et al. (2009) [57] et pour les villes françaises pour le projet FREX d'après Génin et al. (2016) [198]. Les valeurs de F_{ST} ont été calculées à partir des variants fréquents non-corrélés de l'ensemble des 10 000 gènes de 10kb simulés, entre les deux populations A et B de 1000 individus chacune. Les nombres de variant pour le calcul du F_{ST} sont indiqués dans le Tableau 17.

Les populations pour les taux de migration de 0 et 0.001correspondent à des populations européennes éloignées (voire intercontinentales pour le taux de migration 0). Le taux de migration de 0.01se rapproche d'une situation avec des populations de pays voisins ou de villes françaises éloignées. Enfin les structures que nous considérons très fines sont celles pour les taux de migration de 0.025, 0.05 et 0.1, avec des valeurs de F_{ST} proches de celles pour des villes françaises assez proches géographiquement.

Pour la suite de la présentation des résultats, nous présentons les scénarios avec des taux de migration entre 0.01 et 0.1. Les structures simulées avec les taux de migration plus faibles correspondent à des échelles géographiques trop larges se rapprochant de populations européennes éloignées et déjà étudiées dans la littérature.

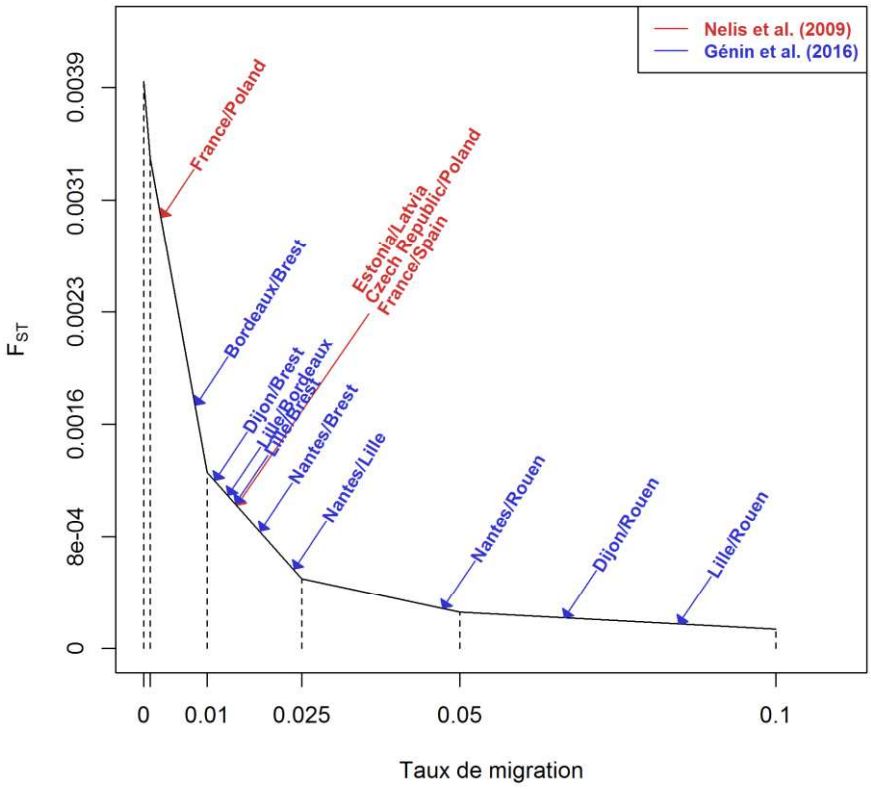


Figure 57. F_{ST} en fonction du paramètre taux de migration des simulations.

Tableau 17. Valeurs de F_{ST} entre les deux populations simulées en fonction du taux de migration.

Taux de migration	Nombre de SNP	F_{ST}
0	37 449	0.003940
0.001	37 532	0.003410
0.01	37 583	0.001226
0.025	37 294	0.000487
0.05	37 718	0.000251
0.1	37 561	0.000132

Erreurs de types I des tests pour les différentes structures simulées

Les erreurs de type I en absence d'une structure géographique très fine

En absence de stratification de population (Figure 58), i.e. les 1000 cas et 1000 témoins proviennent tous de la population A, les erreurs de type I pour le seuil α=5% sont toutes autour de 5%. Les tests CAST et Sum semblent conservateurs, avec des erreurs de type I inférieures. Ces résultats sont similaires aux observations effectuées précédemment (Figure 33, Figure 40).

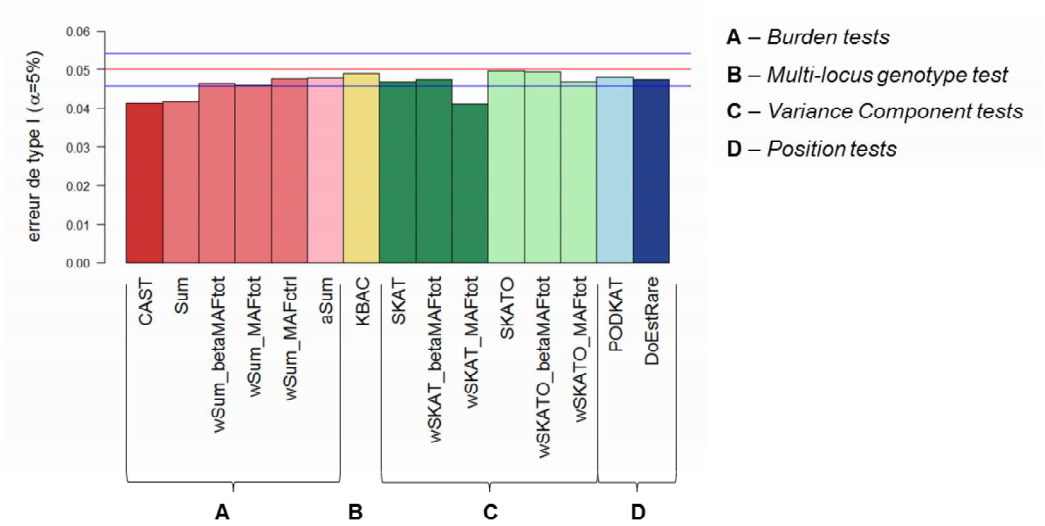


Figure 58. Erreurs de type I pour le seuil α=5% sans stratification de population.

La ligne rouge correspond au seuil α=5%, et les lignes bleues correspondent à un intervalle de confiance de 95% pour affirmer que l'erreur de type I est différente de 5%. Pour cet intervalle de confiance on considère que le nombre d'erreurs suit une loi binomiale de probabilité 5% et de taille maximale 10 000 (nombre de réplicats).

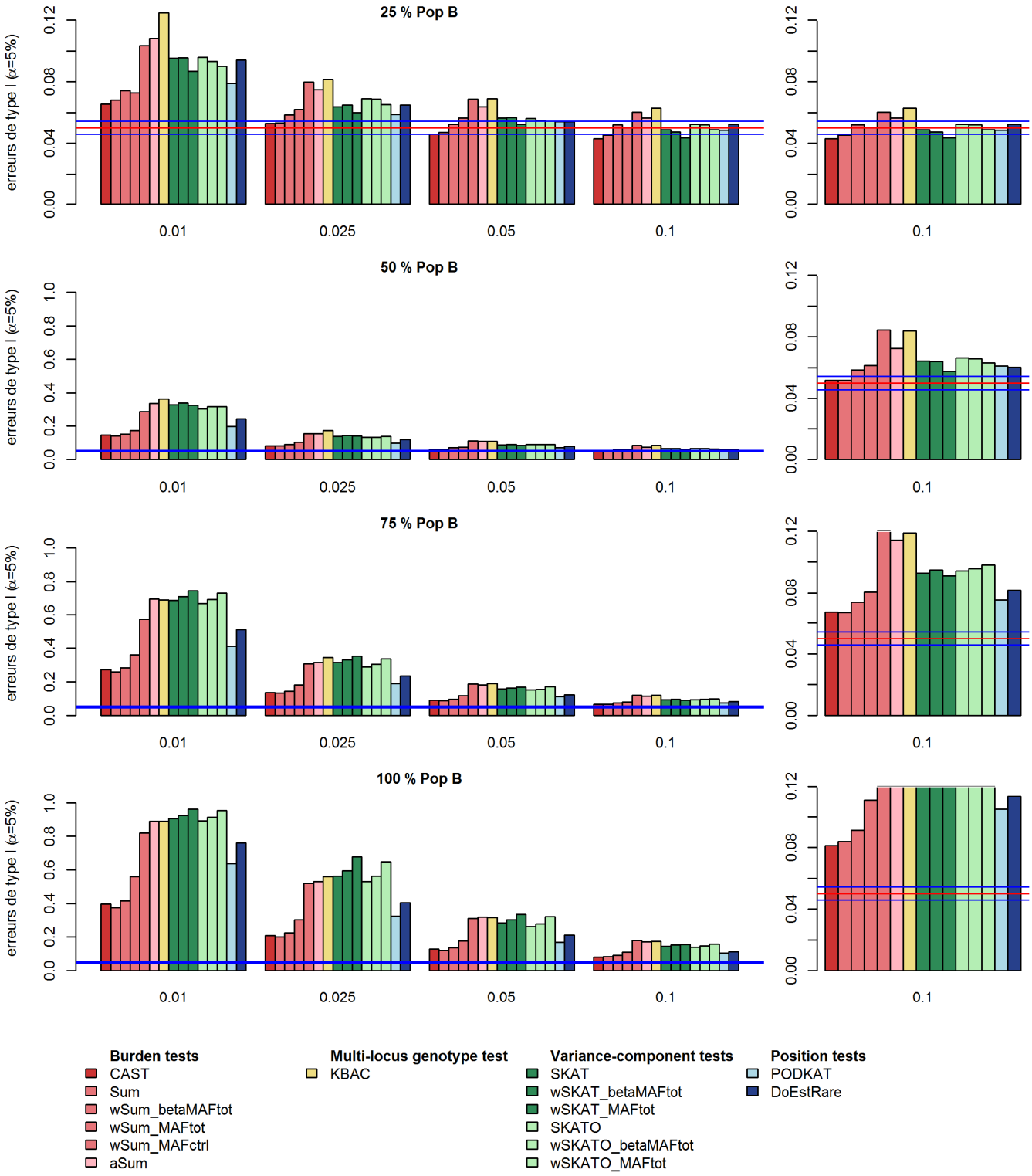


Figure 59. Erreurs de type I au seuil α=5% avec une structure de population fine.

Le graphe est découpé en plusieurs parties selon l'échelle utilisée. Pour le paramètre de migration 0.1 et la proportion 25% de témoins venant de la population B, l'échelle utilisée est de 0 à 0.12. Sinon l'échelle utilisée est de 0 à 1. La ligne rouge correspond au seuil α=5%, et les lignes bleues correspondent à un intervalle de confiance de 95% pour affirmer que l'erreur de type I est différente de 5%. Pour cet intervalle de confiance on considère que le nombre d'erreurs suit une loi binomiale de probabilité 5% et de taille maximale 10 000 (nombre de réplicats).

Les erreurs de type I en présence d'une structure géographique très fine

Avec une structure de population fine, c'est-à-dire avec des taux de migration variant entre 0.01 et 0.1, on observe une inflation des erreurs de type I pour la majorité des scénarios (Figure 59). Le scénario avec un taux de migration de 0.1 et 25% des témoins provenant de la population B présente des erreurs de type I correctes pour la majorité des tests, à l'exception de KBAC et wSum_MAFctrl.

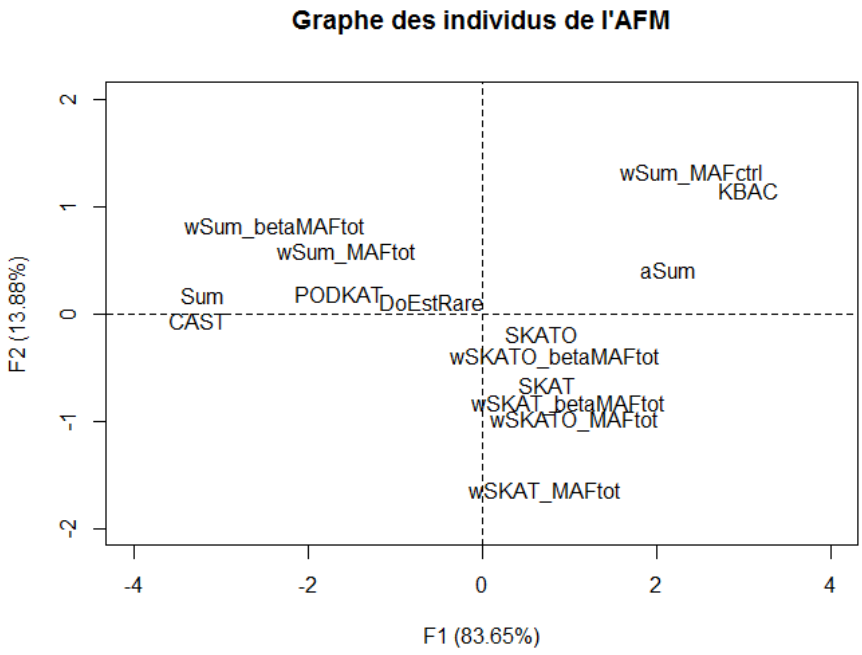


Figure 60. Graphe des individus de l'AFM sur les profils d'erreur de type I au seuil α=5%.
AFM réalisée conjointement sur les 4 tableaux (taux de migration) d'erreurs de type I croisant 15 individus (tests) et 4 variables (proportion de témoins de la population B).

Une sensibilité à la structure de population différente selon le test

Des différences d'inflation sont observables entre les différents tests d'association pour variants rares. Afin de résumer l'information pour les 16 scénarios, nous avons effectué une AFM sur les 4 tableaux d'erreurs de type I correspondant aux taux de migration (Figure 60). Nous avions aussi effectué une analyse ComDim (pour « *Common Dimensions* ») [199] avec la même stratégie d'analyse conjointe de différents tableaux, mais les résultats étant sensiblement les mêmes, nous ne présentons que les résultats issus de l'AFM. Le premier axe

de l'AFM explique 84% de la variabilité totale. Il oppose les *variance-component tests*, KBAC et wSum_MAFctrl et aSum avec une erreur de type I élevée pour de nombreux scénarios, aux tests CAST et Sum, wSum_betaMAFtot qui ont une erreur de type I plus faible. Les tests PODKAT et DoEstRare ont des erreurs de type I souvent intermédiaires, bien que DoEstRare ait une erreur de type I plutôt élevée dans certains scénarios.

D'après ces résultats, les tests construits pour détecter la présence d'effets variables dans le gène, sont les plus impactés par la présence d'une structure de population. Les *variance-component tests* sont adaptés à la présence de variants avec des effets différents tels que les mélanges de variants protecteurs, à risque et neutres. Le test aSum permet de distinguer les variants à risque des variants protecteurs. Enfin le test KBAC a surtout été développé avec l'idée de distinguer les variants à risque des variants neutres.

Nous pouvons mettre ces résultats en relation avec les observations effectuées par Zawistowski et al. (2014) [47], qui montrent une plus grande inflation avec les *variance-component tests*. Selon la distribution jointe des fréquences alléliques pour les deux populations, les inflations varient entre les tests. Lors de déséquilibres du nombre global de variants entre les populations, les *burden tests* seraient plus impactés que dans les scénarios que nous avons simulés ici.

L'influence du système de pondération choisi

Le système de pondération permettrait aussi d'expliquer les différences d'erreurs de type I. Nous avons fait varier les poids pour les tests Sum, SKAT et SKAT-O. Il est clair que les erreurs de type I augmentent, dans beaucoup de scénarios, avec l'utilisation de poids basés sur l'estimation de la MAF dans les données. Pour rappel, dans ces simulations, les cas proviennent tous d'une même population tandis que les témoins proviennent des deux populations A et B. Si une grande proportion de témoins provient de la population B, la MAF estimée pour l'ensemble des cas et des témoins peut ne pas être représentative de la MAF chez les cas.

On peut aussi noter une grande inflation pour le test wSum_MAFctrl avec un système de pondération reposant sur l'estimation de la MAF chez les témoins. En effet, cette estimation n'est pas représentative de la MAF de la population des cas.

Selon le système de pondération « betaMAFtot » ($w_j = dbeta(\widehat{MAF_j}, 1{,}25)$) ou « MAFtot ($w_j = \frac{1}{\sqrt{N.\widehat{MAF_j}.(1-\widehat{MAF_j})}}$.), il peut y avoir des différences d'inflation. Par exemple, dans le cadre des scénarios avec 75% et 100% des témoins de la population B, les tests avec le système de pondération « MAFtot » présentent une plus grande inflation. Les poids décroissent plus rapidement en fonction de la MAF, et permettent de mieux discriminer les variants.

Nous n'avons pas fait varier les poids des tests KBAC, PODKAT et DoEstRare. KBAC présente une très grande erreur de type I en comparaison avec les autres tests. Ce test considère les combinaisons des allèles sur le gène testé, et compare des fréquences de génotypes multi-locus. Il est aussi construit afin de mieux différencier les génotypes multilocus à risque des neutres. La pondération adoptée est basée sur l'estimation de la fréquence du génotype multilocus chez les témoins. Cette fréquence estimée peut ne pas être semblable à celle des cas, à cause de la stratification de population, ce qui permettrait d'expliquer en partie une telle erreur de type I. Enfin notre test DoEstRare utilise des poids similaires à la stratégie de KBAC, en se basant sur la MAF des variants chez les témoins, et présente une erreur de type I assez élevée dans l'ensemble des scénarios.

II- CORRECTION DES TESTS D'ASSOCIATION POUR LA STRUCTURE DE POPULATION

II.1- BIBLIOGRAPHIE

Méthodes de prise en compte de la stratification de population

Afin de remédier à l'inflation des résultats, la stratification de population doit être prise en compte avant, pendant et après l'analyse. Avant l'analyse, les témoins avec un profil génétique similaire à celui des cas doivent être sélectionnés. Avec une structure de population évidente, i.e. avec des populations totalement distinguables, l'analyse doit être effectuée séparément sur chaque population. Si les populations sont proches géographiquement et difficilement distinguables, une correction doit être réalisée pendant l'analyse avec l'incorporation de l'information géographique dans la statistique de test. Enfin si une inflation est tout de même visible dans les résultats d'analyse, même après correction, un ajustement doit être réalisé sur les p-values.

Dans le cadre de l'analyse d'association de variants fréquents, la méthode de correction la plus utilisée est l'ajout de covariables dans le modèle de régression logistique simple-marqueur (voir p27). Les covariables sont en général les premières composantes de l'ACP effectuée sur les données génétiques de l'ensemble des individus[63]. L'objectif de l'ACP est de capturer l'information sur un nombre d'axes réduits, d'une structure génétique au sein des individus, reflétant la structure géographique (voir section **Analyse exploratoire de la structure de population, p23**).

En appliquant cette même stratégie aux *burden tests*, le modèle de régression logistique est

$$logit\big(P(Y_i = 1|S_i, \boldsymbol{Z_i})\big) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \beta S_i \qquad \text{(II.1.1)}$$

avec S_i le *burden score*, $\boldsymbol{Z_i'}$ le vecteur des coordonnées de l'individu i pour les premiers axes de l'ACP. Il en est de même pour les *variance-component tests* SKAT et SKAT-O où le modèle s'écrit

$$logit(P(Y_i = 1|\boldsymbol{X_i}, \boldsymbol{Z_i})) = \alpha_0 + \boldsymbol{Z_i'\alpha} + \sum_{j=1}^{P} \beta_j w_j X_{ij} \qquad \text{(II.1.2)}$$

avec $\beta_j \in \{1, \dots, P\}$ les effets aléatoires des variants du gène. Cette méthode de correction n'est pas applicable à tous les tests d'association pour variants rares.

L'efficacité de la correction par la méthode ACP a déjà étudiée dans le cadre des variants rares, notamment par Mathieson et McVean (2012) [43], Zhang et al. (2013) [200] et Jiang et al. (2013) [46] **pour une population mondiale**. Mathieson et McVean (2012) ont effectué une ACP avec soit des variants avec une MAF≥0.04, soit avec des variants rares définis avec une MAF<0.04. Selon les scénarios envisagés, la correction par ACP est soit très efficace, soit ne permet pas de corriger l'inflation. De plus si l'ACP est effectuée sur les variants rares, elle n'apporte pas d'amélioration en comparaison de l'ACP sur les variants fréquents. Zhang et al. (2013) ont distingué trois catégories de variants : les variants rares (MAF<1%), les variants peu fréquents (MAF>1% et MAF<5%) et les variants fréquents (MAF>5%). Les ACP, effectuées sur les variants fréquents ou peu fréquents, permettent de bien distinguer les deux populations africaine et européenne. La correction par l'intégration des premières composantes est efficace pour enlever l'inflation. Cependant la correction par l'ACP sur les variants peu fréquents, entraine un sur-ajustement avec une perte de puissance. Zhang et al. (2013) montrent aussi qu'une ACP menée sur les variants rares ne permet pas de bien séparer les populations. Une difficulté supplémentaire est le choix du nombre de composantes à incorporer dans le modèle, qui est en général arbitraire. Zhang et al. (2013) ont essayé 1, 10 et 20 premières composantes, dans le modèle. Ce nombre doit être choisi selon l'information retenue par les premières composantes. Enfin, à une échelle plus fine, avec une **population euro-américaine**, O'Connor et al. (2013) [201] montrent une très bonne correction des *burden tests* en intégrant les premières composantes principales de l'ACP.

Une autre stratégie couramment employée pour prendre en compte la structure de population, pour l'étude des variants fréquents, est l'utilisation d'un modèle de régression logistique mixte avec une composante polygénique aléatoire [66,202–204]. Dans le modèle est ajoutée une composante polygénique aléatoire u dont la structure de corrélation est informée par une matrice de similarité génétique entre individus. Le modèle de régression logistique dans le cadre d'un test simple-marqueur est :

$$logit\left(P\left(Y_i = 1 \middle| \mathrm{X_{ij}}\right)\right) = \alpha_0 + \beta X_{ij} + u_i \qquad \text{(II.1.3)}$$

avec u_i la composante aléatoire polygénique pour l'individu i. Les effets aléatoires u_i sont supposés suivre une distribution $\mathcal{N}\left(0, \sigma_G^2 \mathbf{K}\right)$. σ_G^2 est la variance de la composante génétique. La matrice **K** est une matrice de similarité génétique entre individus. Cette méthode est très peu étudiée dans le cadre des études d'association pour les variants rares. Mathieson et McVean (2012) l'ont étudiée en plus de la correction par l'ACP, pour le trait quantitatif simulé, bien que les résultats ne soient pas très différents. Listgarten et al. (2013) [205], ont effectué des analyses supplémentaires à partir des travaux de Mathieson et McVean (2012) et montrent une bonne correction de leur algorithme pour le modèle logistique mixte.

D'autres approches pour prendre en compte la structure de population ont été mises au point dans le cadre des variants rares [206–210]. Nous en présentons les grandes lignes, nous renvoyons le lecteur intéressé à la publication y faisant référence dans le texte. Les approches de Mao et al. (2013)[207] et de Wang et al. (2015)[208] sont basées sur l'existence et la mise en évidence d'un mélange de sous-populations auxquelles appartiennent les individus de l'échantillon. L'assignation à chaque population est effectuée à partir d'algorithmes comme ADMIXTURE [211] qui permettent d'inférer l'origine de ces individus. L'inconvénient de ces méthodes est qu'elles ne permettent pas de corriger les autres tests existant. Epstein et al. (2012)[206] et Lee et al. (2015)[209] se sont tournés vers une meilleure estimation de la distribution de la statistique sous l'hypothèse nulle est estimée, avec une procédure de permutation ou de ré-échantillonnage tenant compte des covariables pour la stratification de la population. L'avantage de la méthode d'Epstein et al. (2012) est qu'elle peut être adaptée à n'importe quel test statistique, tandis que Lee et al. (2015) ont préféré améliorer l'efficacité des calculs dans le cadre de tests du score.

Objectifs

L'efficacité des méthodes de correction dépend de la capture de l'information géographique avec les données génétiques. Dans les différentes études présentées précédemment, la méthode classique de correction, par l'intégration de composantes principales dans le modèle

de régression logistique, semble être efficace dans la plupart des cas. Elle ne permet pas de corriger les biais statistiques, par exemple, lorsque la composante environnementale de la maladie est très localisée géographiquement, avec la concentration de personnes malades dans certaines régions.

Hormis ce cas particulier, les études se sont principalement concentrées sur des populations très distinctes, à une grande échelle géographique, qui seraient en pratique analysées séparément pour une méta-analyse. Or les variants rares étant plus localisés géographiquement, la détection de structures fines de population est très importante pour la correction des tests. Nous avons poursuivi l'analyse des données simulées précédemment (cf partie **Simulation de populations, p130**) avec l'incorporation des composantes principales pour les tests basés sur des modèles de régression logistique.

II.2- MÉTHODES

Analyse exploratoire ACP

Pour l'analyse exploratoire, nous avons considéré l'ensemble des variants fréquents non-corrélés des 10 000 réplicats pour effectuer l'ACP. Comme pour le choix des variants rares, nous nous sommes basés sur la fréquence allélique dans la population A pour définir les variants fréquents avec une MAF≥0.05. Sont gardés ensuite les variants présentant une corrélation $r^2 < 0.2$. Cette analyse exploratoire est effectuée pour chaque scénario en fonction du taux de migration et de la proportion de témoins de la population B.

La méthode ACP que nous avons utilisée est implémentée dans le programme *smartpca* [63,212]. Par rapport à l'ACP classique, *smartpca* présente des légères différences.

La principale différence réside dans la première étape de l'ACP qui est la normalisation de la matrice des génotypes **X** de dimension $N \times P$ (dans ce contexte, ce sont l'ensemble des variants fréquents du génome). La valeur normalisée $X_{ij,n} = \frac{X_{ij} - \mu_j}{\sqrt{p_j(1-p_j)}}$ avec un centrage par la valeur moyenne μ_j et une réduction par une valeur proportionnelle à l'écart-type pour une variable aléatoire suivant une loi binomiale de probabilité p_j.

La matrice diagonalisée s'apparente à une matrice de produits scalaires entre individus $\mathbf{X}_n\mathbf{X}'_n$ de dimension $N \times N$ au lieu de la matrice de corrélation entre variants génétiques dans le cadre classique de l'ACP. En effet étant donné que le nombre de variants est largement supérieur au nombre d'individus, la matrice $P \times P$ aurait été très longue à diagonaliser. La diagonalisation d'une matrice $N \times N$ au lieu d'une matrice $P \times P$ ne change pas les résultats les premières composantes principales restant identiques à un facteur près.

Analyse d'association

Nous avons analysé les mêmes données que pour l'étude de l'impact d'une structure de population fine (p130). Les tests appliqués sont ceux reposant sur un modèle de régression logistique permettant l'intégration de covariables :

- Les *burden tests* : CAST, Sum, wSum, aSum
- Les *variance-component tests* : SKAT et SKAT-O
- Le *position test* : PODKAT

D'après la description de KBAC par Liu et Leal (2010) [109], nous aurions pu corriger la statistique de test avec des covariables. Cependant cette correction du test n'est pas implémentée dans le package R KBAC.

Comme pour les analyses effectuées précédemment, une procédure de permutation adaptative est réalisée pour les tests CAST, Sum, wSum et aSum, avec les paramètres de seuil à atteindre et de précision $\alpha = 0.01$ et $c=0.2$.

Nous nous sommes focalisés sur les scénarios avec des structures géographiques fines, avec un taux de migration entre 0.01 et 0.1. De plus nous n'avons pas analysé le scénario avec 100% des témoins venant de la population B, car nous craignons que dans certaines analyses la structure de population discrimine parfaitement le phénotype, ce qui rend impossible le test de l'effet génétique à l'aide d'un modèle de régression logistique.

II.3- RÉSULTATS

ACP pour la recherche de structures de populations

Les analyses ACP sont effectuées sur l'ensemble des 1000 cas et 1000 témoins, pour environ 37000 variants fréquents non corrélés. Avec la visualisation des graphes des individus, il est net que les populations se distinguent quand les taux de migrations sont bas, c'est-à-dire avec les structures de population les plus larges, similaires à des pays voisins ou à des régions françaises éloignées (Figure 61). Nous n'avons pas analysé le scénario avec 100% des témoins de la population B, car les populations sont totalement confondues avec le statut cas-témoin, lorsque les taux de migrations sont de 0.01 et 0.025. Il n'aurait pas été possible de calculer la statistique de test pour l'effet génétique, à cause des covariables expliquant parfaitement le phénotype.

On peut aussi noter que les variances expliquées par les premiers axes sont très faibles. Ceci est dû au très grand nombre de composantes (1999 composantes = 2000-1 individus) et de la structure de population fine. Lorsque les populations sont bien distinguables sur le graphe de l'ACP, c'est principalement la première composante principale qui se démarque. Mais lorsque les populations sont confondues, il est difficile de choisir le nombre de composantes car de nombreuses composantes principales présentent une valeur propre supérieure à 1 (Annexe XIV). Les graphes des individus dans la Figure 61 sont pour les scénarios avec 25% et 100% des témoins de la population B ; les autres scénarios sont présentés en Annexe XIV.

Efficacité de la correction par la méthode ACP

Dans de nombreux scénarios, l'intégration des deux premières composantes principales comme covariables dans le modèle permet de corriger parfaitement les inflations qui avaient été observées, avec des erreurs de type I proches du seuil α=5% (Figure 62, tableaux de valeurs en Annexe XV). Cependant pour le scénario avec un taux de migration de 0.1, et lorsque 50% ou 75% des témoins proviennent de la population B, l'inflation n'est pas corrigée. Pour ces scénarios, aucune structure de population n'était en effet visible sur le plan 1-2 de l'ACP. Nous avons tenté de corriger ces inflations en intégrant plus de covariables dans le modèle, mais les résultats restent sensiblement les mêmes. Pour ajuster les tests, il aurait été nécessaire de détecter la structure de population, ce qui n'est pas le cas.

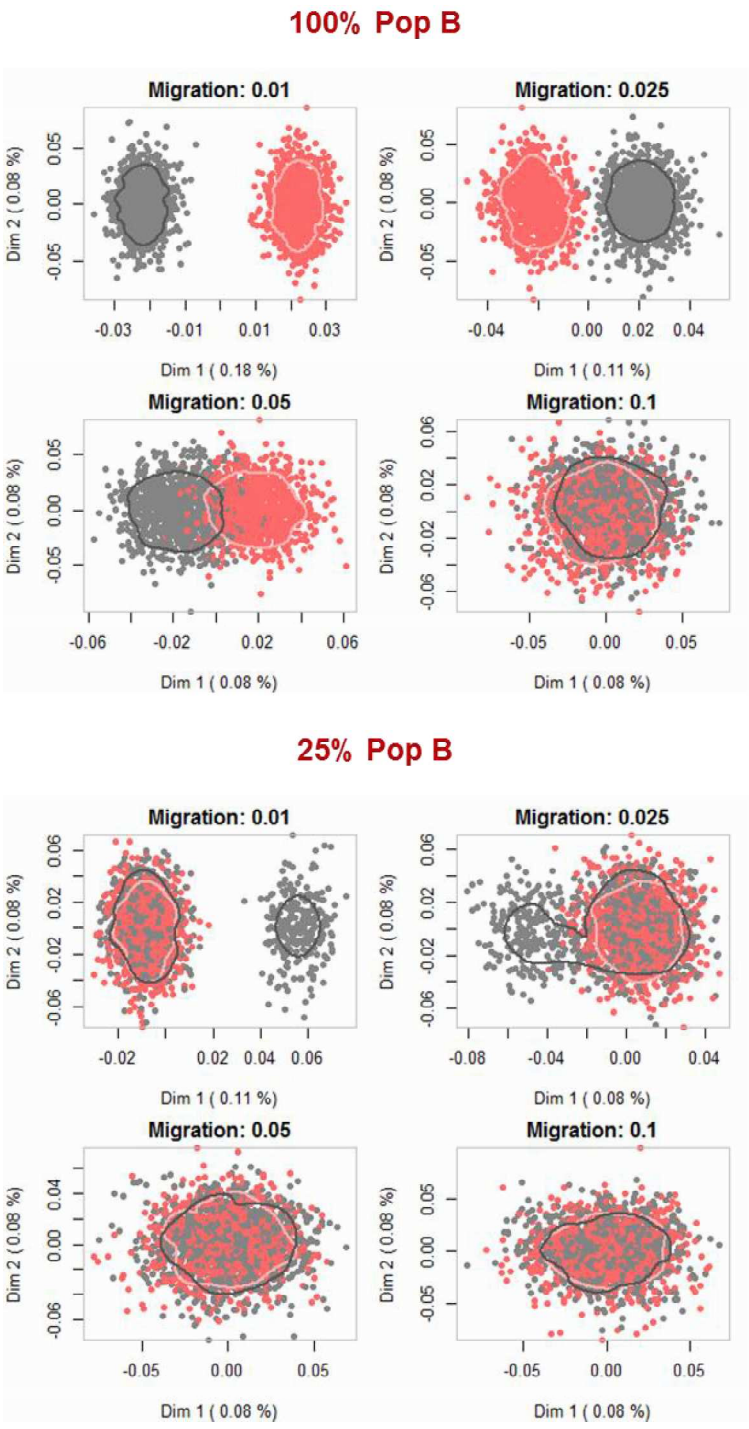


Figure 61. Graphe des individus pour les 2 premières dimensions de l'ACP et pour les scénarios avec 100% et 25% des témoins de la population B.

Les cas sont représentés en rouge et les témoins en gris. Les « patatoïdes » sont tracées avec à l'aide de la fonction `locfit` du package R `locfit`. Elles contiennent 75% des individus.

Ainsi la correction des tests d'association pour variants rares pour diminuer les inflations d'erreurs de type I n'est pas toujours efficace pour des structures de populations très fines, semblables à des régions françaises très proches géographiquement.

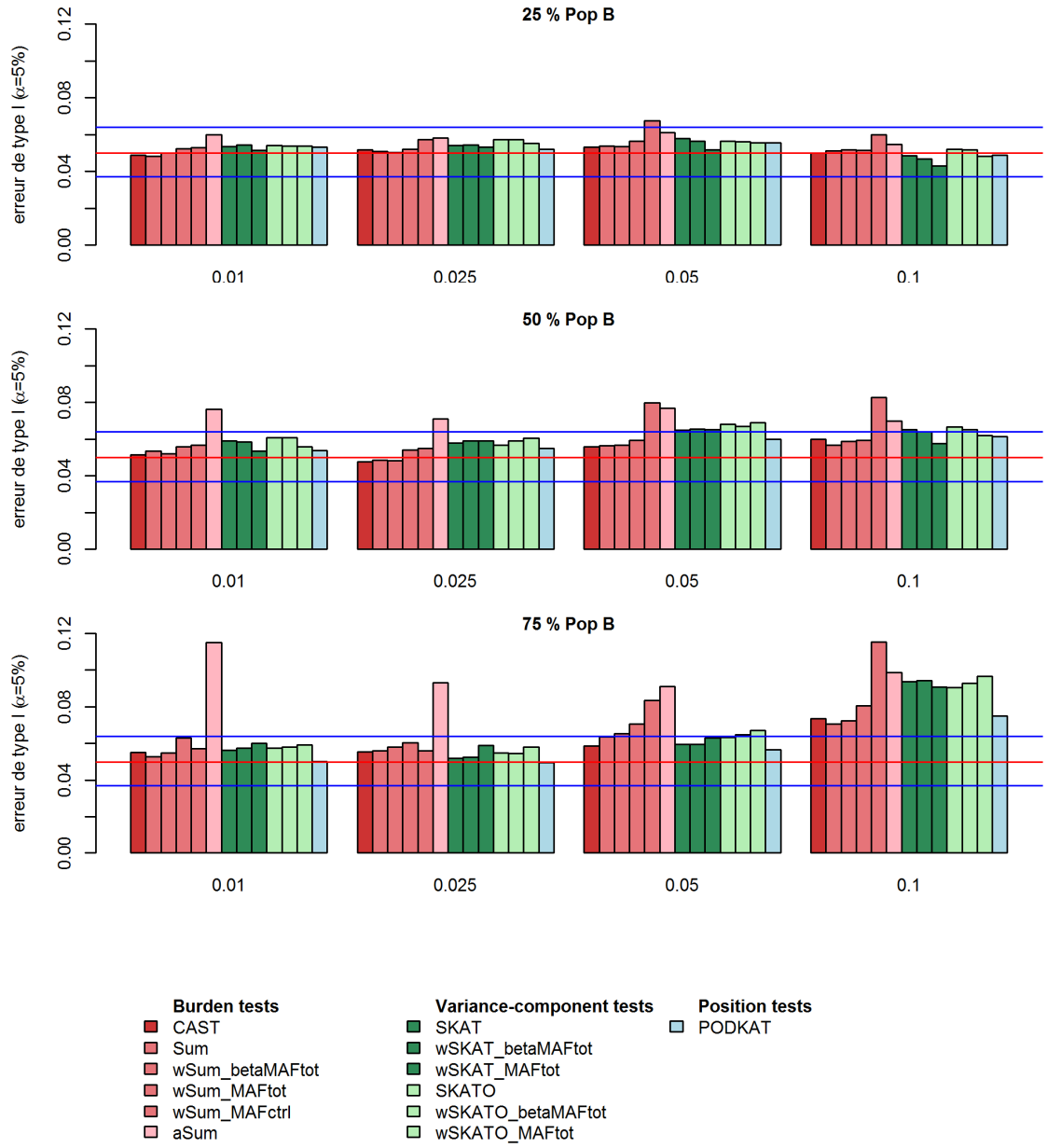


Figure 62. Erreurs de type I au seuil α=5% suite à la correction des tests pour la stratification de population.

L'ajustement pour la structure de population a été effectué en incorporant les 2 premières composantes principales.

Note : On peut noter un comportement du test aSum très différent de celui des autres tests. Dans cette méthode, les variants sont d'abord testés individuellement pour différencier les variants protecteurs des variants à risque, puis enfin les tester ensemble. Un ajustement est réalisé à chaque étape. Il est difficile de comprendre pourquoi ce test se comporte ainsi.

III- DISCUSSION

Afin de minimiser les coûts d'analyse, les études d'association génétique utilisent de plus en plus des bases de données extérieures pour le choix des témoins. Le choix des témoins est très important dans le cadre de l'analyse de variants rares car, étant plus récents, ils sont susceptibles d'être plus localisés géographiquement. L'impact sur les résultats des tests, avec une inflation des p-values et donc une augmentation du nombre de faux positifs, a déjà été prouvé à différentes échelles [42,43,46,47,201]. Nous avons voulu évaluer cet impact à des échelles encore plus fines. On montre en effet, à partir de simulations, des augmentations d'erreurs de type I même en présence d'une structure géographique très fine pouvant se comparer à des régions françaises proches en termes de F_{ST}.

Dans notre étude les tests les plus impactés sont les tests KBAC et les variance-component tests, tandis que les *burden tests* tels que CAST et Sum présentent les erreurs de type I les plus basses. En tenant compte des observations effectuées par Zawistowski et al. (2014) [47], les *burden tests* présenteraient une inflation plus importante dans le cas d'un déséquilibre global important du nombre de variants entre les deux populations. Le nombre de variants rares dans une population dépend de facteurs démographiques tels que l'augmentation de la taille de la population. Lorsqu'il y a une grande croissance de la population, le nombre de variants rares augmente au sein de la population. Nous aurions pu envisager dans notre modèle démographique de simulation, des croissances exponentielles de population différentes. Cependant le réalisme de ce modèle reste à discuter pour des populations très proches géographiquement.

Notre étude présente en effet certaines limites, par le nombre de populations choisi et la composition des cas et des témoins. Nous avons fait le choix de considérer seulement deux populations pour la facilité d'interprétation des résultats. De plus les cas proviennent tous de la même population, tandis que les témoins proviennent de deux populations différentes. Afin que ce soit plus réaliste les cas et les témoins devraient provenir de différentes populations avec des proportions différentes. Bien sûr il est difficile de connaître les modèles démographiques de vraies populations, telle que la population française. Ceux-ci sont le cadre de recherche des études de génétique des populations.

Nous avons ensuite étudié l'efficacité de l'incorporation de composantes principales dans les modèles de régression logistique de certaines méthodes d'association. Dans la plupart des scénarios, où la structure de population est visible sur le graphe des individus de l'ACP, cette approche est efficace avec des erreurs de type I correctes. Il faut toutefois noter, qu'en pratique les individus avec des profils trop différents sont écartés de l'analyse, pour éviter tout biais statistique, mais nous les avons tout de même incorporés comme c'est le cas pour les autres études présentées dans la littérature. La méthode de correction reste tout de même efficace avec des erreurs de type I ramenées à 5%. Il aurait aussi été judicieux d'étudier les puissances des tests après la correction afin de voir si elles sont maintenues. En effet il est important de conserver une bonne puissance de test tout en diminuant l'erreur de type I.

Dans d'autres scénarios, les populations ne sont pas distinguables à partir de l'ACP. Pour ces scénarios avec des structures très fines de population et avec de nombreux témoins à l'origine démographique différente, nous n'arrivons pas à corriger les inflations observées en termes d'erreurs de type I. Avec cette possibilité d'inflation même en présence d'une structure de population fine et non identifiable, il est important en pratique de vérifier la présence de biais dans les résultats à partir des Q-Q plots et de l'indice d'inflation génomique (*genomic inflation factor*).

Pour pouvoir remédier à cette inflation lors de l'analyse d'association, il faudrait pouvoir identifier la structure de population. Nous avons employé l'ACP car il s'agit de la méthode de référence lors des études d'association classiques pour les variants fréquents. Comme nous l'avons introduit précédemment, différentes études ont envisagé l'utilisation de l'ACP sur des variants fréquents, peu fréquents ou rares [43,200]. Dans cette étude, nous avons appliqué l'ACP sur les variants fréquents pour refléter la structure géographique. O'Connor et al. (2015) [213] ont observé que l'analyse exploratoire basée sur les variants rares permet d'observer des structures génétiques à plus petite échelle. Il serait donc intéressant de voir si l'analyse effectuée sur l'ensemble des variants rares permettrait de mieux corriger l'inflation pour les structures les plus fines.

Des programmes bioinformatiques comme fineSTRUCTURE [214] et ADMIXTURE [211] sont très utilisés par la communauté scientifique s'intéressant à la génétique des populations. Ils permettraient d'identifier des structures plus fines que l'ACP. Cependant il faut pouvoir utiliser l'information ressortie de ces programmes pour l'incorporer dans la statistique de test.

Avec fineSTRUCTURE, l'une des sorties est une matrice de coancestralité, semblable à une matrice de similarité. Une analyse MDS à partir de cette matrice de similarité pourrait être envisageable, pour ensuite utiliser les premières composantes dans les modèles. Avec le programme ADMIXTURE, une méthode bayésienne permet de calculer la probabilité d'un individu d'appartenir à différentes populations, dont le nombre est choisi par validation croisée. On peut imaginer d'utiliser ces vecteurs de probabilité comme covariables dans les modèles. De plus fineSTRUCTURE se base sur les régions IBD pour calculer la proximité entre deux individus. Or le schéma de simulation que nous avons employé ne permet pas de bien identifier les régions IBD car nous avons considéré des petites régions génétiques de 10kb. Il faudrait ainsi pouvoir simuler de grandes régions génétiques, pour tester l'emploi de fineSTRUCTURE.

Au lieu d'incorporer des covariables dans le modèle de régression logistique avec les effets génétiques, nous aurions pu aussi employer d'autres méthodes décrites précédemment, telles que la régression logistique avec une composante polygénique aléatoire, et des méthodes de permutations tenant compte de covariables. Le modèle logistique mixte aurait permis de contourner la question du nombre de composantes principales de l'ACP à introduire. En effet la distribution de la composante aléatoire du modèle mixte est basée sur une matrice de similarité génétique entre individus. La méthode de permutations d'Epstein et al. (2012) [206] présente l'avantage d'être applicable à tous les tests utilisant des procédures de permutation pour évaluer la significativité de la statistique. Cette méthode effectue d'abord une régression logistique avec d'expliquer le phénotype en fonction de covariables, pour ensuite permuter les individus tout en maintenant cette structure. Nous aimerions étudier de plus près cette méthode car elle serait applicable au test DoEstRare que nous avons développé.

Enfin le choix des témoins lors d'une analyse d'association est très important pour éviter les biais. En pratique ces études considèrent le pays d'origine et l'ethnie pour le choix des témoins. Cependant ces informations ne sont pas toujours renseignées et des méthodes d'analyse exploratoire comme l'ACP sont effectuées pour distinguer des structures génétiques au sein des données. Nous avons considéré des méthodes de correction d'analyse, le choix au préalable des témoins est tout aussi important. Epstein et al. (2012) [215] ont développé une approche pour le choix optimal de témoins, avec l'appariement basé sur un score de stratification. Des méthodes pour le choix des témoins seront nécessaires, dans l'avenir, pour

la sélection des témoins parmi la grande quantité de données, dans le cadre de l'analyse de variants rares.

CONCLUSION GÉNÉRALE

I- PRINCIPAUX RÉSULTATS

De nombreuses méthodes statistiques ont été développées ou adaptées pour les études d'association entre des variants génétiques rares et des maladies. En raison de la faible fréquence des variants, ces méthodes présentent toutes la même stratégie de tester l'information pour un groupe de variants, en général situés dans un même gène lors des études de séquençages d'exomes. La grande diversité de tests statistiques s'explique du fait de la difficulté de tester un groupe de variants présentant des fréquences alléliques et des effets différents. Beaucoup de questions se posent quant au choix du test statistique à employer. Les tests les plus couramment utilisés sont CAST, Sum, wSum, SKAT et SKAT-O. Afin de mieux comprendre les avantages et les inconvénients des tests, nous avons comparé un certain nombre de stratégies au moyen de simulations de données selon différents scénarios génétiques et de l'application à des données de séquençage.

À partir des simulations selon les modèles très généraux envisagés par Basu et Pan (2011) [164] pour leur comparaison de tests, nous montrons que KBAC, SKAT-O et MiST se démarquent en termes de puissance, et semblent adaptés à un grand nombre de scénarios. Des *burden tests*, comme CAST, Sum et wSum sont adaptés pour détecter des groupes de variants avec le même effet sur la maladie, impliquant très peu de variants neutres. A l'inverse, les *variance-component tests* SKAT et C-alpha sont construits de façon à détecter des variants avec des effets très différents, avec un mélange de variants à risque, neutres et protecteurs. Selon la structure génétique sous-jacente de la maladie, ces tests seront plus ou moins puissants pour détecter le gène. Avec l'application des tests aux données de séquençage pour le syndrome de Brugada et la maladie d'Alzheimer d'apparition précoce, nous avons noté des différences de significativité entre les gènes les plus significatifs selon les tests employés. Il est alors recommander en pratique d'utiliser des tests, reposant sur des hypothèses très différentes, afin de couvrir le maximum de structures possibles. Les mécanismes biologiques impliqués dans les maladies sont en effet complexes, et certainement variables selon les gènes.

Peu de tests prennent en compte l'information sur les positions des variants dans la région testée, nous avons alors développé le test DoEstRare afin d'identifier des regroupements de

variants à risque, en comparant des distributions de variants. Les simulations ont démontré une très bonne puissance du test DoEstRare dans de nombreux scénarios au même rang que SKAT-O et KBAC. Nous avons aussi observé que DoEstRare fournit des résultats de significativité différents de ceux obtenus avec les autres tests. Ceci encourage l'utilisation de DoEstRare afin d'explorer de nouvelles pistes génétiques pour expliquer les maladies, notamment dans le cadre de répartitions des variants rares différentes entre cas et témoins.

Par ailleurs, dans le cadre du développement d'études basées sur le séquençage génome entier d'un grand nombre de patients pour une population de référence, comme prévu dans le plan France Médecine Génomique 2025 [15], il sera de plus en plus important de choisir les témoins de manière appropriée pour les études génétiques. Nous avons étudié l'impact d'une structure de population sur les résultats des tests d'association pour variants rares, avec la volonté de représenter des échelles géographiques à petite échelle. Nous montrons une légère inflation des erreurs de type I pour certains tests même en présence d'une proportion importante de témoins venant d'une population très proche de celle des cas. En effet, certains tests sont plus sensibles que d'autres à la présence de différentes populations. De plus, la méthode de correction classique qui est l'incorporation de composantes principales de l'ACP pour les tests basés sur des modèles de régression logistique, ne permet de corriger totalement le biais statistique dû à la présence de différentes populations géographiques très proches. L'inflation des tests après correction est relativement peu élevée mais reste à considérer pour l'interprétation des résultats. Ceci rejoint l'enjeu des études de génétique de population avec la recherche de méthodes d'analyse exploratoire permettant de mieux capturer l'information de modèles démographiques via l'analyse des profils génétiques.

II- VERS L'ANALYSE DE GÉNOMES ENTIERS

Les études d'association pour les variants rares sont en plein essor avec l'analyse d'énormes données de séquençage. Les données qui ont été analysées dans ce projet sont issues du séquençage de gènes candidats pour l'étude du syndrome de Brugada et de l'exome pour l'étude de la maladie d'Alzheimer d'apparition précoce. Cependant avec le développement rapide des technologies de séquençage et la diminution des coûts, les études se tourneront progressivement vers le séquençage de génome entier (*whole-genome sequencing*). L'exome ne représente qu'environ 2% du génome entier, ce qui implique que beaucoup de variations

génétiques présentes dans les séquences non-codantes ne sont pas étudiées. Les données de génome entier, permettent de mener des études d'association classiques sur les variants fréquents ainsi que des tests sur les variants rares. En continuation des projets VACARME et FREX, le projet FranceGenRef en cours permettra d'obtenir le séquençage de génomes pour une base de données de référence de la population française.

Avec l'analyse d'association de génomes, d'autres enjeux statistiques et informatiques se présentent. Pour les deux études d'association qui ont été présentées, les groupes de variants sont facilement définissables car ils correspondent aux gènes, ou plutôt à l'ensemble des séquences codantes du gène. Avec l'étude du génome-entier, les groupes de variants rares sont plus difficiles à définir. Morrison et al. (2017) [216] discutent de pratiques à adopter pour les études d'association basées sur le séquençage génome-entier. Dans cette étude sont présentées et mises en application les étapes méthodologiques pour l'analyse d'association à l'échelle du génome. L'étude d'association est effectuée pour 10 traits quantitatifs facteurs de risque cardiovasculaires pour 1860 afro-américains. Pour étudier le génome complet, les groupes de variants rares (MAF≤5%) ont été définis par (1) fenêtre génomique glissante, (2) domaine régulateur, (3) premier intron des gènes. L'approche des fenêtres glissantes considère le génome entier sans connaissance de structures biologiques et sans hypothèse de fonctionnalité. Les paramètres de taille de fenêtre et de glissement sont choisis de manière arbitraire et doivent être choisis minutieusement. Des tailles de fenêtre trop petites sont difficilement analysables car ne contenant pas assez de variants rares. La détection d'association significative peut être difficile avec des tailles de fenêtre trop grandes, car risquant de présenter trop de « bruit » diminuant la puissance des tests. Morrison et al. (2017) ont choisi les paramètres de 4kb pour la taille de fenêtre et de 2kb pour le glissement. Afin de se concentrer plus particulièrement sur les séquences non-codantes, dans leur étude, les motifs de régulation regroupent les différentes structures comme le promoteur, les régions UTR (*untranslated regions*) 3' et 5', le premier intron des gènes, et les *enhancers* (voir Figure 63). Les variants fréquents (MAF>5%) sont analysés individuellement avec les tests simple-marqueur.

L'identification de fenêtres optimales pour les tests d'association fait le sujet d'études [189]. L'utilisation de la stratégie de DoEstRare, comparant des fonctions de densité entre cas et témoins, permettrait d'éviter de spécifier des tailles de fenêtres arbitraires. Une idée serait aussi d'appliquer DoEstRare sur des fenêtres glissantes.

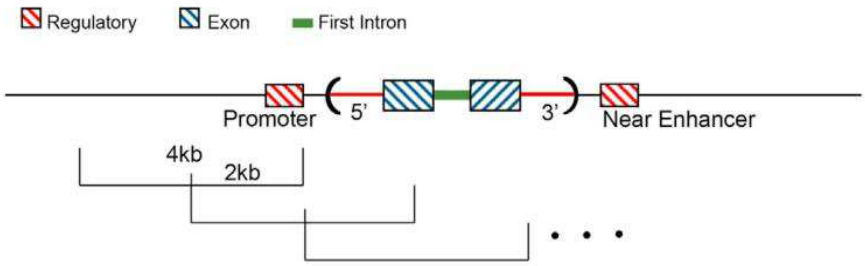


Figure 63. Motifs fonctionnels annotés pour l'étude de Morrison et al. (2017).

Ces analyses statistiques requièrent l'utilisation d'outils bioinformatiques performants pour gérer les grands volumes de données à analyser. Dernièrement, un outil appelé SEQSpark [188], a été publié par Zhang et al. (2017) pour les études d'association menées sur les variants génétiques rares à l'échelle de l'exome ou du génome. Les groupes de variants rares peuvent être définis par les gènes, les fenêtres glissantes ou d'autres régions de régulation. Seulement quelques tests sont implémentés tels que CAST (CMC), Sum, wSum, VT, SKAT et SKAT-O. Tous ces tests sont implémentés considérant un modèle de régression logistique pouvant intégrer des covariables.

BIBLIOGRAPHIE

1. Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004 Oct 21;431(7011):931–45.

2. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.

3. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The International HapMap Project. Nature. 2003 décembre;426(6968):789–96.

4. The International HapMap Consortium. A haplotype map of the human genome. Nature. 2005 Oct 27;437(7063):1299–320.

5. Karakachoff M, Duforet-Frebourg N, Simonet F, Le Scouarnec S, Pellen N, Lecointe S, et al. Fine-scale human genetic structure in Western France. Eur J Hum Genet EJHG. 2015 Jun;23(6):831–6.

6. Projet VaCaRMe. Site de VaCaRMe ! [Internet]. 2014 [cited 2014 Aug 28]. Available from: http://www.vacarme-project.org/

7. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLOS Biol. 2015 Jul 7;13(7):e1002195.

8. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18;536(7616):285–91.

9. Lejeune J, Gautier M, Turpin R. Etude des chromosomes somatiques de neuf enfants mongoliens. Comptes Rendus Hebd Seances Acad Sci. 1959 Mar 16;248(11):1721–2.

10. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980 May;32(3):314–31.

11. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, et al. A polymorphic DNA marker genetically linked to Huntington's disease. Nature. 1983 Nov 17;306(5940):234–8.

12. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005 Apr 15;308(5720):385–9.

13. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet. 2005 février;6(2):95–108.

14. Marquet P, Longeray P-H, Barlesi F, Ameye V, Augé P, Cazeneuve B, et al. Recherche translationnelle : médecine personnalisée, médecine de précision, thérapies ciblées : marketing ou science ? Thérapie. 2015 Jan 1;70(1):1–10.

15. Aviesan. France Médecine Génomique 2025 [Internet]. 2016 Jun. Available from: http://www.gouvernement.fr/sites/default/files/liseuse/7433/master/index.htm

16. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009 Jun 9;106(23):9362–7.

17. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan 4;45(D1):D896–901.

18. Maher B. Personal genomes: The case of the missing heritability. Nat News. 2008 Nov 5;456(7218):18–21.

19. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8;461(7265):747–53.

20. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010;11(6):446–50.

21. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001 Jul;69(1):124–37.

22. Iyengar SK, Elston RC. The genetic basis of complex traits: rare variants or "common gene, common disease"? Methods Mol Biol Clifton NJ. 2007;376:71–84.

23. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. Am J Hum Genet. 2008 Jan;82(1):100–12.

24. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr Opin Genet Dev. 2009 Jun;19(3):212–9.

25. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet. 2010 Jun;11(6):415–25.

26. Gibson G. Rare and common variants: twenty arguments. Nat Rev Genet. 2011 Feb;13(2):135–45.

27. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nat Genet. 2008 Jun;40(6):695–701.

28. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010 Nov 1;11(11):773–85.

29. Wagner MJ. Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. Pharmacogenomics. 2013 Mar;14(4):413–24.

30. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014 juillet;95(1):5–23.

31. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. Genome Med. 2015;7(1):16.

32. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome Biol [Internet]. 2017 Dec [cited 2017 Jul 31];18(1). Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1212-4

33. Duval D, Lardeux A, Le Tourneau T, Norris RA, Markwald RR, Sauzeau V, et al. Valvular dystrophy associated filamin A mutations reveal a new role of its first repeats in small-GTPase regulation. Biochim Biophys Acta BBA - Mol Cell Res. 2014 Feb;1843(2):234–44.

34. Robertson SP, Twigg SRF, Sutherland-Smith AJ, Biancalana V, Gorlin RJ, Horn D, et al. Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. Nat Genet. 2003 avril;33(4):487–91.

35. Ionita-Laza I, Makarov V, ARRA Autism Sequencing Consortium, Buxbaum JD. Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. Am J Hum Genet. 2012 Jun 8;90(6):1002–13.

36. Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, et al. "Location, Location, Location": a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. Bioinforma Oxf Engl. 2012 Dec 1;28(23):3027–33.

37. Chen Y-C, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, et al. A hybrid likelihood model for sequence-based disease association studies. PLoS Genet. 2013;9(1):e1003224.

38. Schaid DJ, Sinnwell JP, McDonnell SK, Thibodeau SN. Detecting genomic clustering of risk variants from sequence data: cases versus controls. Hum Genet. 2013 Nov;132(11):1301–9.

39. Lin W-Y. Association testing of clustered rare causal variants in case-control studies. PloS One. 2014;9(4):e94337.

40. Bodenhofer U. PODKAT: An R Package for Association Testing Involving Rare and Private Variants. R package version 1.0.3; 2015.

41. Persyn E, Karakachoff M, Le Scouarnec S, Le Clézio C, Campion D, Consortium FE, et al. DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease. Wang K, editor. PLOS ONE. 2017 Jul 24;12(7):e0179364.

42. Tintle N, Aschard H, Hu I, Nock N, Wang H, Pugh E. Inflated Type I Error Rates When Using Aggregation Methods to Analyze Rare Variants in the 1000 Genomes Project Exon Sequencing Data in Unrelated Individuals: Summary Results from Group 7 at Genetic Analysis Workshop 17. Genet Epidemiol. 2011;35(Suppl 1):S56–60.

43. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012 Mar;44(3):243–6.

44. Babron M-C, de Tayrac M, Rutledge DN, Zeggini E, Génin E. Rare and low frequency variant stratification in the UK population: description and impact on association tests. PloS One. 2012;7(10):e46519.

45. O'Connor TD, Kiezun A, Bamshad M, Rich SS, Smith JD, Turner E, et al. Fine-Scale Patterns of Population Stratification Confound Rare Variant Association Tests. PLoS ONE [Internet]. 2013 Jul 4 [cited 2015 Jun 8];8(7). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3701690/

46. Jiang Y, Epstein MP, Conneely KN. Assessing the impact of population stratification on association studies of rare variation. Hum Hered. 2013;76(1):28–35.

47. Zawistowski M, Reppell M, Wegmann D, St Jean PL, Ehm MG, Nelson MR, et al. Analysis of rare variant population structure in Europeans explains differential stratification of gene-based tests. Eur J Hum Genet EJHG. 2014 Sep;22(9):1137–44.

48. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine-scale genetic structure of the British population. Nature. 2015 Mar 19;519(7543):309–14.

49. Liu Q, Nicolae DL, Chen LS. Marbled inflation from population structure in gene-based association studies with rare variants. Genet Epidemiol. 2013 Apr;37(3):286–92.

50. Felsenfeld G, Groudine M. Controlling the double helix. Nature. 2003 Jan 23;421(6921):448–53.

51. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006 Feb 1;7(2):85–97.

52. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nat Rev Genet. 2008 Jun;9(6):477–85.

53. Wright S. The genetical structure of populations. Ann Eugen. 1951 Mar;15(4):323–54.

54. Wright S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. Evolution. 1965;19(3):395–420.

55. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nat Rev Genet. 2009 Sep;10(9):639–50.

56. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38(6):1358–70.

57. Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, et al. Genetic Structure of Europeans: A View from the North–East. PLoS ONE [Internet]. 2009 May 8 [cited 2016 Oct 6];4(5). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675054/

58. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet Lond Engl. 2003 Feb 15;361(9357):598–604.

59. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

60. Itan Y, Jones BL, Ingram CJ, Swallow DM, Thomas MG. A worldwide correlation of lactase persistence phenotype and genotypes. BMC Evol Biol. 2010;10(1):36.

61. Bezzina CR, Barc J, Mizusawa Y, Remme CA, Gourraud J-B, Simonet F, et al. Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. Nat Genet. 2013 Sep;45(9):1044–9.

62. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. 2006 Oct 1;7(10):781–91.

63. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006 Aug;38(8):904–9.

64. Hosmer DW, Lemeshow S. Applied logistic regression. 2. ed. New York, NY: Wiley; 2000. 373 p. (Wiley series in probability and statistics).

65. Y Benjamini YH. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57:289–300.

66. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010 juillet;11(7):459–63.

67. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999 Dec;55(4):997–1004.

68. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010 Sep 15;26(18):2336–7.

69. Ioannidis JPA, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet. 2009 May;10(5):318–29.

70. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan 1;42(D1):D1001–6.

71. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, et al. Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet. 2008 May;40(5):575–83.

72. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, et al. Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet. 2008 mai;40(5):584–91.

73. Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, et al. Many sequence variants affecting diversity of adult human height. Nat Genet. 2008 mai;40(5):609–15.

74. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008 May;9(5):356–69.

75. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of heritability for human height. Nat Genet. 2010 Jul;42(7):565–9.

76. Almal SH, Padh H. Implications of gene copy-number variation in health and diseases. J Hum Genet. 2012 Jan;57(1):6–13.

77. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat Rev Genet. 2017 May 30;18(7):441–51.

78. Carlborg Ö, Haley CS. Opinion: Epistasis: too often neglected in complex trait studies? Nat Rev Genet. 2004 Aug;5(8):618–25.

79. Phillips PC. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008 Nov;9(11):855–67.

80. Khoury MJ, Wacholder S. Invited Commentary: From Genome-Wide Association Studies to Gene-Environment-Wide Interaction Studies--Challenges and Opportunities. Am J Epidemiol. 2008 Nov 25;169(2):227–30.

81. Thomas D. Gene–environment-wide association studies: emerging approaches. Nat Rev Genet. 2010 Mar 9;11(4):259–72.

82. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma Oxf Engl. 2009 Jul 15;25(14):1754–60.

83. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009 Aug 15;25(16):2078–9.

84. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303.

85. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016 Dec [cited 2017 Jul 11];17(1). Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4

86. Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, et al. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005;6(5):R44.

87. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

88. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–9.

89. Li MJ, Wang J. Current trend of annotating single nucleotide variation in humans – A case study on SNVrap. Methods. 2015 Jun;79–80:32–40.

90. Butkiewicz M, Bush WS. In Silico Functional Annotation of Genomic Variation. Curr Protoc Hum Genet. 2016 Jan 1;88:Unit 6.15.

91. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet. 2016 Jan 4;48(2):214–20.

92. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008 Sep;83(3):311–21.

93. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007 Feb 3;615(1–2):28–56.

94. Morris AP, Zeggini E. An Evaluation of Statistical Approaches to Rare Variant Analysis in Genetic Association Studies. Genet Epidemiol. 2010 Feb;34(2):188–93.

95. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011 Jul 15;89(1):82–93.

96. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. Am J Hum Genet. 2012 Aug 10;91(2):224–37.

97. Chen LS, Hsu L, Gamazon ER, Cox NJ, Nicolae DL. An exponential combination procedure for set-based association tests in sequencing studies. Am J Hum Genet. 2012 Dec 7;91(6):977–86.

98. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009 Feb;5(2):e1000384.

99. Wang J, Zhao Z, Cao Z, Yang A, Zhang J. A probabilistic method for identifying rare variants underlying complex traits. BMC Genomics. 2013;14 Suppl 1:S11.

100. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. Genet Epidemiol. 2013 May;37(4):334–44.

101. Li Q, Zhang H, Yu K. Approaches for evaluating rare polymorphisms in genetic association studies. Hum Hered. 2010;69(4):219–28.

102. Navon O, Sul JH, Han B, Conde L, Bracci PM, Riby J, et al. Rare variant association testing under low-coverage sequencing. Genetics. 2013 Jul;194(3):769–79.

103. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010;70(1):42–54.

104. Tyrer JP, Guo Q, Easton DF, Pharoah PDP. The admixture maximum likelihood test to test for association between rare variants and disease phenotypes. BMC Bioinformatics. 2013 Jun 6;14:177.

105. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010 Jun 11;86(6):832–8.

106. Fang H, Hou B, Wang Q, Yang Y. Rare variants analysis by risk-based variable-threshold method. Comput Biol Chem. 2013 Oct;46:32–8.

107. Bhatia G, Bansal V, Harismendy O, Schork NJ, Topol EJ, Frazer K, et al. A covering method for detecting genetic associations between rare variants and common phenotypes. PLoS Comput Biol. 2010;6(10):e1000954.

108. Ayers KL, Cordell HJ. Identification of grouped rare and common variants via penalized logistic regression. Genet Epidemiol. 2013 Sep;37(6):592–602.

109. Liu DJ, Leal SM. 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 Genet. 2010 Oct 14;6(10):e1001156.

110. Li Y, Byrnes AE, Li M. To identify associations with rare variants, just WHaIT: Weighted haplotype and imputation-based tests. Am J Hum Genet. 2010 Nov 12;87(5):728–35.

111. Liang F, Xiong M. Bayesian detection of causal rare variants under posterior consistency. PloS One. 2013;8(7):e69633.

112. Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet. 2010 Nov 12;87(5):604–17.

113. Clarke GM, Rivas MA, Morris AP. A flexible approach for the analysis of rare variants allowing for a mixture of effects on binary or quantitative traits. PLoS Genet. 2013;9(8):e1003694.

114. Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PloS One. 2010;5(11):e13584.

115. Cheng KF, Lee JY, Zheng W, Li C. A powerful association test of multiple genetic variants using a random-effects model. Stat Med. 2014 May 20;33(11):1816–27.

116. King CR, Rathouz PJ, Nicolae DL. An evolutionary framework for association testing in resequencing studies. PLoS Genet. 2010 Nov 11;6(11):e1001202.

117. Larson NB, Schaid DJ. Regularized rare variant enrichment analysis for case-control exome sequencing data. Genet Epidemiol. 2014 Feb;38(2):104–13.

118. Zhang L, Pei Y-F, Li J, Papasian CJ, Deng H-W. Efficient utilization of rare variants for detection of disease-related genomic regions. PloS One. 2010;5(12):e14288.

119. Won S, Kim Y, Lange C. On rare-variant analysis in population-based designs: decomposing the likelihood to two informative components. Hum Hered. 2013;76(2):76–85.

120. Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. Genet Epidemiol. 2011 Jan;35(1):57–69.

121. Kuk AYC, Nott DJ, Yang Y. A stepwise likelihood ratio test procedure for rare variant selection in case-control studies. J Hum Genet. 2014 Apr;59(4):198–205.

122. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C. A new testing strategy to identify rare variants with either risk or protective effect on disease. PLoS Genet. 2011;7(2):e1001289.

123. Lin W-Y, Lou X-Y, Gao G, Liu N. Rare variant association testing by adaptive combination of P-values. PloS One. 2014;9(1):e85728.

124. Sul JH, Han B, He D, Eskin E. An optimal weighted aggregated association test for identification of rare variants involved in common diseases. Genetics. 2011 May;188(1):181–8.

125. Logsdon BA, Dai JY, Auer PL, Johnsen JM, Ganesh SK, Smith NL, et al. A variational Bayes discrete mixture test for rare variant association. Genet Epidemiol. 2014 Jan;38(1):21–30.

126. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an Unusual Distribution of Rare Variants. PLoS Genet. 2011 Mar 3;7(3):e1001322.

127. Pan W, Shen X. Adaptive Tests for Association Analysis of Rare Variants. Genet Epidemiol. 2011 Jul;35(5):381–8.

128. Sun H, Wang S. A power set-based statistical selection procedure to locate susceptible rare variants associated with complex traits with sequencing data. Bioinforma Oxf Engl. 2014 Aug 15;30(16):2317–23.

129. Gordon D, Finch SJ, De La Vega FM, De La Vega F. A new expectation-maximization statistical test for case-control association studies considering rare variants obtained by high-throughput sequencing. Hum Hered. 2011;71(2):113–25.

130. Sha Q, Zhang S. A rare variant association test based on combinations of single-variant tests. Genet Epidemiol. 2014 Sep;38(6):494–501.

131. He L, Pitkäniemi J, Sarin A-P, Salomaa V, Sillanpää MJ, Ripatti S. Hierarchical Bayesian model for rare variant association analysis integrating genotype uncertainty in human sequence data. Genet Epidemiol. 2015 Feb;39(2):89–100.

132. Zhang Q, Irvin MR, Arnett DK, Province MA, Borecki I. A data-driven method for identifying rare variants with heterogeneous trait effects. Genet Epidemiol. 2011 Nov;35(7):679–85.

133. Chen Z, Yang W, Liu Q, Yang JY, Li J, Yang M. A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. BMC Bioinformatics. 2014;15 Suppl 17:S3.

134. Lin D-Y, Tang Z-Z. A General Framework for Detecting Disease Associations with Rare Variants in Sequencing Studies. Am J Hum Genet. 2011 Sep 9;89(3):354–67.

135. Turkmen AS, Yan Z, Hu Y-Q, Lin S. Kullback-Leibler distance methods for detecting disease association with rare variants from sequencing data. Ann Hum Genet. 2015 May;79(3):199–208.

136. Quintana MA, Berstein JL, Thomas DC, Conti DV. Incorporating model uncertainty in detecting rare variants: the Bayesian risk index. Genet Epidemiol. 2011 Nov;35(7):638–49.

137. Fouladi R, Bessonov K, Van Lishout F, Van Steen K. Model-Based Multifactor Dimensionality Reduction for Rare Variant Association Analysis. Hum Hered. 2015;79(3–4):157–67.

138. Dai Y, Jiang R, Dong J. Weighted selective collapsing strategy for detecting rare and common variants in genetic association study. BMC Genet. 2012 Feb 6;13:7.

139. Wu B, Pankow JS, Guan W. Sequence Kernel Association Analysis of Rare Variant Set Based on the Marginal Regression Model for Binary Traits. Genet Epidemiol. 2015 Sep;39(6):399–405.

140. Yuan A, Chen G, Zhou Y, Bentley A, Rotimi C. A novel approach for the simultaneous analysis of common and rare variants in complex traits. Bioinforma Biol Insights. 2012;6:1–9.

141. Austin E, Shen X, Pan W. A Novel Statistic for Global Association Testing Based on Penalized Regression. Genet Epidemiol. 2015 Sep;39(6):415–26.

142. Liu Y, Huang CH, Hu I, Lo S-H, Zheng T. Association screening for genes with multiple potentially rare variants: an inverse-probability weighted clustering approach. BMC Proc. 2011 Nov 29;5 Suppl 9:S106.

143. Lee W, Lee D, Pawitan Y. Likelihood ratio and score burden tests for detecting disease-associated rare variants. Stat Appl Genet Mol Biol. 2015 Nov;14(5):481–95.

144. Asimit JL, Day-Williams AG, Morris AP, Zeggini E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. Hum Hered. 2012;73(2):84–94.

145. Zhou Y, Wang Y. Detecting association of rare and common variants by adaptive combination of P-values. Genet Res. 2015 Oct 6;97:e20.

146. Wang Y, Chen Y-H, Yang Q. Joint rare variant association test of the average and individual effects for sequencing studies. PloS One. 2012;7(3):e32485.

147. Greco B, Hainline A, Arbet J, Grinde K, Benitez A, Tintle N. A general approach for combining diverse rare variant association tests provides improved robustness across a wider range of genetic architectures. Eur J Hum Genet EJHG. 2015 Oct 28;

148. Urrutia E, Lee S, Maity A, Zhao N, Shen J, Li Y, et al. Rare variant testing across methods and thresholds using the multi-kernel sequence kernel association test (MK-SKAT). Stat Interface. 2015;8(4):495–505.

149. Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genet Epidemiol. 2012 Sep;36(6):561–71.

150. Hu Y-J, Liao P, Johnston HR, Allen AS, Satten GA. Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls. PLoS Genet. 2016 May;12(5):e1006040.

151. Wang K, Fingert JH. Statistical tests for detecting rare variants using variance-stabilising transformations. Ann Hum Genet. 2012 Sep;76(5):402–9.

152. Fang H, Zhang H, Yang Y. Poisson Approximation-Based Score Test for Detecting Association of Rare Variants. Ann Hum Genet. 2016 Jul;80(4):221–34.

153. Sha Q, Wang S, Zhang S. Adaptive clustering and adaptive weighting methods to detect disease associated rare variants. Eur J Hum Genet EJHG. 2013 Mar;21(3):332–7.

154. Sun R, Weng H, Hu I, Guo J, Wu WKK, Zee BC-Y, et al. A W-test collapsing method for rare-variant association testing in exome sequencing data. Genet Epidemiol. 2016;40(7):591–6.

155. Li H, Chen J. Efficient unified rare variant association test by modeling the population genetic distribution in case-control studies. Genet Epidemiol. 2016;40(7):579–90.

156. Cheung YH, Wang G, Leal SM, Wang S. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. Genet Epidemiol. 2012 Nov;36(7):675–85.

157. Hasegawa T, Kojima K, Kawai Y, Misawa K, Mimori T, Nagasaki M. AP-SKAT: highly-efficient genome-wide rare variant association test. BMC Genomics. 2016 Sep 21;17(1):745.

158. Xu C, Ladouceur M, Dastani Z, Richards JB, Ciampi A, Greenwood CMT. Multiple regression methods show great potential for rare variant association tests. PloS One. 2012;7(8):e41694.

159. Li Y-M, Xu C, Xiang Y, Peng C, Deng H-W. An adaptive strategy for association analysis of common or rare variants using entropy theory. J Hum Genet. 2017 Apr 6;

160. Zhan H, Xu S. Adaptive ridge regression for rare variant detection. PloS One. 2012;7(8):e44173.

161. Sofer T. BinomiRare: A robust test of the association of a rare variant with a disease for pooled analysis and meta-analysis, with application to the HCHS/SOL. Genet Epidemiol. 2017 Jul;41(5):388–95.

162. Brisbin A, Jenkins GD, Ellsworth KA, Wang L, Fridley BL. Localization of association signal from risk and protective variants in sequencing studies. Front Genet. 2012;3:173.

163. Sugasawa S, Noma H, Otani T, Nishino J, Matsui S. An efficient and flexible test for rare variant effects. Eur J Hum Genet EJHG. 2017 Jun;25(6):752–7.

164. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genet Epidemiol. 2011 Nov;35(7):606–19.

165. Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting Disease Associations due to Linkage Disequilibrium Using Haplotype Tags: A Class of Tests and the Determinants of Statistical Power. Hum Hered. 2003 Nov 14;56(1–3):18–31.

166. Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol. 2004 Dec;27(4):415–28.

167. Davies RB. Algorithm AS 155: The Distribution of a Linear Combination of $\chi 2$ Random Variables. J R Stat Soc Ser C Appl Stat. 1980 Jan 1;29(3):323–33.

168. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. Genet Epidemiol. 2009 Sep;33(6):497–507.

169. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostat Oxf Engl. 2012 Sep;13(4):762–75.

170. Liu H, Tang Y, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. Comput Stat Data Anal. 2009 Feb;53(4):853–6.

171. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd; 1925.

172. Ansari AR, Bradley RA. Rank-Sum Tests for Dispersions. Ann Math Stat. 1960 Dec;31(4):1174–89.

173. Tango T. Statistical Methods for Disease Clustering [Internet]. 2010 [cited 2015 Oct 23]. Available from: https://books.google.fr/books?id=4rT2sv151b0C&pg=PR5&lpg=PR5&dq=Tango+T+(2010)+Statistical+methods+for+disease+clustering.+Springer,+New+York&source=bl&ots=1jGUCxWMDP&sig=tCgGZwNExL5dxkGo-mPyfGzn3e0&hl=fr&sa=X&ved=0CEoQ6AEwBmoVChMIkbXy-uTYyAIVwbIeCh2Elg0w#v=onepage&q=Tango%20T%20(2010)%20Statistical%20methods%20for%20disease%20clustering.%20Springer%2C%20New%20York&f=false

174. Feller W. An Introduction to Probability Theory and Its Applications. Wiley; 1971. 656 p.

175. Silverman BW. Density Estimation for Statistics and Data Analysis. CRC Press; 1986. 190 p.

176. Ernst MD. Permutation Methods: A Basis for Exact Inference. Stat Sci. 2004 Nov;19(4):676–85.

177. Che R, Jack JR, Motsinger-Reif AA, Brown CC. An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use. BioData Min. 2014;7:9.

178. Wang T, Elston RC. Improved Power by Use of a Weighted Score Test for Linkage Disequilibrium Mapping. Am J Hum Genet. 2007 Feb;80(2):353–60.

179. Pagès J. Analyse factorielle multiple avec R. Les Ulis: EDP Sciences; 2013. 253 p.

180. Lê S, Josse J, Husson F, others. FactoMineR: An R package for multivariate analysis. J Stat Softw. 2008;25(1):1–18.

181. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. 2015 Apr;11(4):e1005165.

182. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005 Nov 1;15(11):1576–83.

183. Le Scouarnec S, Karakachoff M, Gourraud J-B, Lindenbaum P, Bonnaud S, Portero V, et al. Testing the burden of rare variation in arrhythmia-susceptibility genes provides new insights into molecular diagnosis for Brugada syndrome. Hum Mol Genet. 2015 May 15;24(10):2757–63.

184. Nicolas G, Charbonnier C, Wallon D, Quenez O, Bellenguez C, Grenier-Boley B, et al. SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. Mol Psychiatry. 2015 Aug 25;

185. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep 1;38(16):e164–e164.

186. Greco B, Hainline A, Arbet J, Grinde K, Benitez A, Tintle N. A general approach for combining diverse rare variant association tests provides improved robustness across a wider range of genetic architectures. Eur J Hum Genet EJHG. 2016 May;24(5):767–73.

187. Zhan X, Hu Y, Li B, Abecasis GR, Liu DJ. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data: Table 1. Bioinformatics. 2016 May 1;32(9):1423–6.

188. Zhang D, Zhao L, Li B, He Z, Wang GT, Liu DJ, et al. SEQSpark: A Complete Analysis Tool for Large-Scale Rare Variant Association Studies Using Whole-Genome and Exome Sequence Data. Am J Hum Genet. 2017 Jul;101(1):115–22.

189. Wang MH, Weng H, Sun R, Lee J, Wu WKK, Chong KC, et al. A Zoom-Focus algorithm (ZFA) to locate the optimal testing region for rare variant association tests. Bioinformatics [Internet]. 2017 Mar 11 [cited 2017 Aug 17]; Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btx130

190. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Feb 2;46(3):310–5.

191. Li D, Lewinger JP, Gauderman WJ, Murcray CE, Conti D. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. Genet Epidemiol. 2011 Dec;35(8):790–9.

192. Barnett IJ, Lee S, Lin X. Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. Genet Epidemiol. 2013 Feb;37(2):142–51.

193. Zhou Y-J, Wang Y, Chen L-L. Detecting the Common and Individual Effects of Rare Variants on Quantitative Traits by Using Extreme Phenotype Sampling. Genes. 2016 Jan 14;7(10):2.

194. Auer PL, Reiner AP, Wang G, Kang HM, Abecasis GR, Altshuler D, et al. Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. Am J Hum Genet. 2016 Oct;99(4):791–801.

195. Cook K, Benitez A, Fu C, Tintle N. Evaluating the impact of genotype errors on rare variant tests of association. Front Genet. 2014;5:62.

196. Almasy L, Dyer TD, Peralta J, Kent JW, Charlesworth JC, Curran JE, et al. Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc. 2011;5(Suppl 9):S2.

197. Mägi R, Kumar A, Morris AP. Assessing the impact of missing genotype data in rare variant association analysis. BMC Proc. 2011;5(Suppl 9):S107.

198. Génin E, Dina C, Ludwig T, Quenez O, Letort S, Lindenbaum P, et al. Are population-specific panels of exomes useful to identify disease variants: Insights from the French Exome Project. Vancouver: Presented at the 69th Annual Meeting of The American Society of Human Genetics; 2016.

199. Hanafi M. Nouvelles propriétés de l'analyse en composantes communes et poids spécifiques. J Société Fr Stat. 2008;149(2):75–97.

200. Zhang Y, Guan W, Pan W. Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol. 2013 Jan;37(1):99–109.

201. O'Connor TD, Kiezun A, Bamshad M, Rich SS, Smith JD, Turner E, et al. Fine-scale patterns of population stratification confound rare variant association tests. PloS One. 2013;8(7):e65834.

202. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010 Apr;42(4):355–60.

203. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010 Apr;42(4):348–54.

204. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. Am J Hum Genet. 2016 Apr;98(4):653–66.

205. Listgarten J, Lippert C, Heckerman D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. Nat Genet. 2013 May;45(5):470–1.

206. Epstein MP, Duncan R, Jiang Y, Conneely KN, Allen AS, Satten GA. A Permutation Procedure to Correct for Confounders in Case-Control Studies, Including Tests of Rare Variation. Am J Hum Genet. 2012 Aug;91(2):215–23.

207. Mao X, Li Y, Liu Y, Lange L, Li M. Testing genetic association with rare variants in admixed populations. Genet Epidemiol. 2013 Jan;37(1):38–47.

208. Wang X, Zhang S, Li Y, Li M, Sha Q. A Powerful Approach to Test an Optimally Weighted Combination of Rare Variants in Admixed Populations. Genet Epidemiol. 2015 Mar 10;

209. Lee S, Fuchsberger C, Kim S, Scott L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case–control studies. Biostatistics. 2015 Sep 11;kxv033.

210. Sha Q, Zhang K, Zhang S. A Nonparametric Regression Approach to Control for Population Stratification in Rare Variant Association Studies. Sci Rep. 2016 Nov 18;6:37444.

211. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep 1;19(9):1655–64.

212. Patterson N, Price AL, Reich D. Population Structure and Eigenanalysis. PLoS Genet. 2006;2(12):e190.

213. O'Connor TD, Fu W, NHLBI GO Exome Sequencing Project, ESP Population Genetics and Statistical Analysis Working Group, Emily Turner, Mychaleckyj JC, Logsdon B, et al. Rare variation facilitates inferences of fine-scale population structure in humans. Mol Biol Evol. 2015 Mar;32(3):653–60.

214. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of Population Structure using Dense Haplotype Data. Copenhaver GP, editor. PLoS Genet. 2012 Jan 26;8(1):e1002453.

215. Epstein MP, Duncan R, Broadaway KA, He M, Allen AS, Satten GA. Stratification Score Matching Improves Correction for Confounding by Population Stratification in Case-Control Association Studies. Genet Epidemiol. 2012 Apr;36(3):195–205.

216. Morrison AC, Huang Z, Yu B, Metcalf G, Liu X, Ballantyne C, et al. Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. Am J Hum Genet. 2017 Feb 2;100(2):205–15.

Annexe I PUBLICATIONS

PLOS ONE

RESEARCH ARTICLE

DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease

Elodie Persyn[1], Matilde Karakachoff[1,2], Solena Le Scouarnec[1], Camille Le Clézio[3], Dominique Campion[3], French Exome Consortium[¶], Jean-Jacques Schott[1,2], Richard Redon[1,2☯], Lise Bellanger[4☯]*, Christian Dina[1,2☯]*

1 INSERM, CNRS, UNIV Nantes, l'institut du thorax, Nantes, France, **2** CHU Nantes, l'institut du thorax, Nantes, France, **3** Inserm U1079, Rouen University, Normandy Center for Genomic Medicine and Personalized Medicine, Normandy University, Rouen, France, **4** Laboratoire de Mathématiques Jean Leray, UMR CNRS 6629, Nantes, France

☯ These authors contributed equally to this work.
¶ Membership of the French Exome Consortium is provided in the Acknowledgments.
* lise.bellanger@univ-nantes.fr (LB); christian.dina@univ-nantes.fr (CD)

Citation: Persyn E, Karakachoff M, Le Scouarnec S, Le Clézio C, Campion D, Consortium FE, et al. (2017) DoEstRare: A statistical test to identify local enrichments in rare genomic variants associated with disease. PLoS ONE 12(7): e0179364. https://doi.org/10.1371/journal.pone.0179364

Editor: Kai Wang, Columbia University Medical Center, UNITED STATES

Received: January 26, 2017

Accepted: May 29, 2017

Published: July 24, 2017

Data Availability Statement: The generation of simulated data is fully explained within the paper and can be reproduced with the publicly available software cosi. Due to ethical restriction imposed by Paris Necker Ethic Committee, EOAD case-control data are available upon request. Interested researchers may inquire about access to these data by contacting Dr. Alexis Brice, the Director of the Institut du Cerveau et de la Moelle épinière – ICM, at alexis.brice@icm-institute.org. These data were initially published in Nicolas et al. 2015 (Mol. Psych.). The BrS data initially published in Le

Abstract

Next-generation sequencing technologies made it possible to assay the effect of rare variants on complex diseases. As an extension of the "common disease-common variant" paradigm, rare variant studies are necessary to get a more complete insight into the genetic architecture of human traits. Association studies of these rare variations show new challenges in terms of statistical analysis. Due to their low frequency, rare variants must be tested by groups. This approach is then hindered by the fact that an unknown proportion of the variants could be neutral. The risk level of a rare variation may be determined by its impact but also by its position in the protein sequence. More generally, the molecular mechanisms underlying the disease architecture may involve specific protein domains or intergenic regulatory regions. While a large variety of methods are optimizing functionality weights for each single marker, few evaluate variant position differences between cases and controls. Here, we propose a test called DoEstRare, which aims to simultaneously detect clusters of disease risk variants and global allele frequency differences in genomic regions. This test estimates, for cases and controls, variant position densities in the genetic region by a kernel method, weighted by a function of allele frequencies. We compared DoEstRare with previously published strategies through simulation studies as well as re-analysis of real datasets. Based on simulation under various scenarios, DoEstRare was the sole to consistently show highest performance, in terms of type I error and power both when variants were clustered or not. DoEstRare was also applied to Brugada syndrome and early-onset Alzheimer's disease data and provided complementary results to other existing tests. DoEstRare, by integrating variant position information, gives new opportunities to explain disease susceptibility. DoEstRare is implemented in a user-friendly R package.

Scouarnec et al. 2015 (HMG) referred to in the present paper are available from figshare: https://doi.org/10.6084/m9.figshare.4903814.

Funding: This work was supported by a grant from The French Regional Council of Pays de la Loire (RFI VaCaRMe: Recherche, Formation et Innovation, Vaincre les maladies Cardiovaculaires, Respiratoires et Métaboliques). The website for this project is available at the following URL address http://www.vacarme-project.org/.

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

Introduction

Genome-wide association studies (GWASs) have identified numerous common haplotypes associated with a wide variety of complex diseases [1]. However these common variants often present low effects on disease susceptibility and do not explain the whole heritability [2]. Rare variants with stronger effects may explain, among other factors, the missing heritability [3]. These variants are defined with a minor allele frequency (MAF) often arbitrarily set between 0.1% and 1%, depending on the disease prevalence.

The huge advances in genome sequencing are now enabling association studies on rare variants. However single-marker tests (consisting in testing each variant individually) are not suitable in the context of low-frequency alleles, for which immoderately large sample sizes would be required to obtain sufficient power to detect association signals. Many specific statistical methods have thus been developed to test the association between complex diseases and groups of rare variants [4–14]. For efficiency reasons, groups of rare variants often correspond to gene coding sequences, which are biological units easy of interpretation.

One challenge is to deal with the heterogeneous nature of genetic variants. Indeed groups of rare variants are likely to include a non-negligible proportion of neutral variants and causal variants with various effect sizes. Different strategies have been adopted to detect an association in the presence of neutral variants, such as keeping only putative deleterious variants with an adaptive method [7,14] or using continuous weights based on functional potential [6,8,11,12]. It is also recognized that all positions are not equal, corresponding to various protein domains (within a gene) or putative functionality (genome). For instance, Robertson et al. (2003) found that pathogenic mutations are localized in various domains of the *FLNA* gene, causing diverse congenital malformations [15]. Only a few tests take also into account genetic positions, in order to detect clusters of disease-risk variants residing within specific domains of given proteins [16–21].

We developed a new statistical test, named DoEstRare for "Density-oriented Estimation for Rare variant positions", to detect both global enrichment in rare alleles and localized clusters of disease-risk rare variants (DRVs), by integrating position information. The DoEstRare statistic consists in comparing simultaneously the mutation position densities, estimated by kernel method, and the overall average allele frequencies between cases and controls. To better discriminate neutral from causal variants (deleterious or protective), we incorporated a weight system in the computation of average allele frequencies. A similar approach was used in the Kernel-Based Adaptive Cluster (KBAC) test [8], on multi-locus genotypes.

To assess the performance of DoEstRare, we compared its power and type I error to other existing tests by analyzing simulated data. We conducted simulations, based on the backward coalescent model implemented in the COSI program [22], under different scenarios varying the position distribution of DRVs. We considered three main scenarios: in the first scenario DRVs are uniformly distributed on the gene, in the second and third scenarios DRVs are respectively clustered in one and two areas. We also varied the proportion of causal variants which is linked to the window sizes of clustered areas. From these simulations we show that our test is among the most powerful statistical tests and perform even better with the presence of one cluster of DRVs.

We also applied association tests based on rare alleles on two real datasets to assess the consistency between significance results and evaluate the properties of our test in real settings. The studied pathologies were Brugada syndrome (BrS), with data from Le Scouarnec et al. (2015) [23] and early-onset Alzheimer's disease (EOAD), from Nicolas et al. (2015) [24]. Interestingly, we show that DoEstRare provides slightly different significance results from other tests. DoEstRare is indeed based on a different hypothesis and could be used to explore new research insights, involving variant positions.

Results

Overview of DoEstRare

DoEstRare test aims to compare mutation position probability distributions on the region of interest (e.g. a gene) between cases and controls. If the distributions of rare variant positions are different in cases and controls, the gene is considered associated with the disease. This pattern could be expected when a specific domain of the protein is involved in the pathogenicity and causal mutations cluster in specific areas of the gene.

Simultaneously, we test a global burden hypothesis in DoEstRare to consider aggregated counts of mutations across the gene. Cases and controls may present equal mutation position distributions but different probabilities to present a rare mutation. This burden hypothesis consists in comparing average allele frequencies across the gene.

The hypotheses of our test can be formulated as followed:

$$H_0 : f^A = f^U \text{ and } p^A = p^U$$

$$H_1 : f^A \neq f^U \text{ or } p^A \neq p^U$$

with f^A and f^U, the mutation position density functions in affected (A) and unaffected (U) individuals; p^A and p^U, the average allele frequencies. To illustrate these hypotheses, Gene 1 from Fig 1 is not associated with the disease, as there is no difference in terms of position distribution and total mutation count. DoEstRare aims to identify situations like Gene 2 and Gene 3, where the mutation position distribution or the mutation number differs between cases and controls.

Further details about the construction of DoEstRare are described in the Methods section.

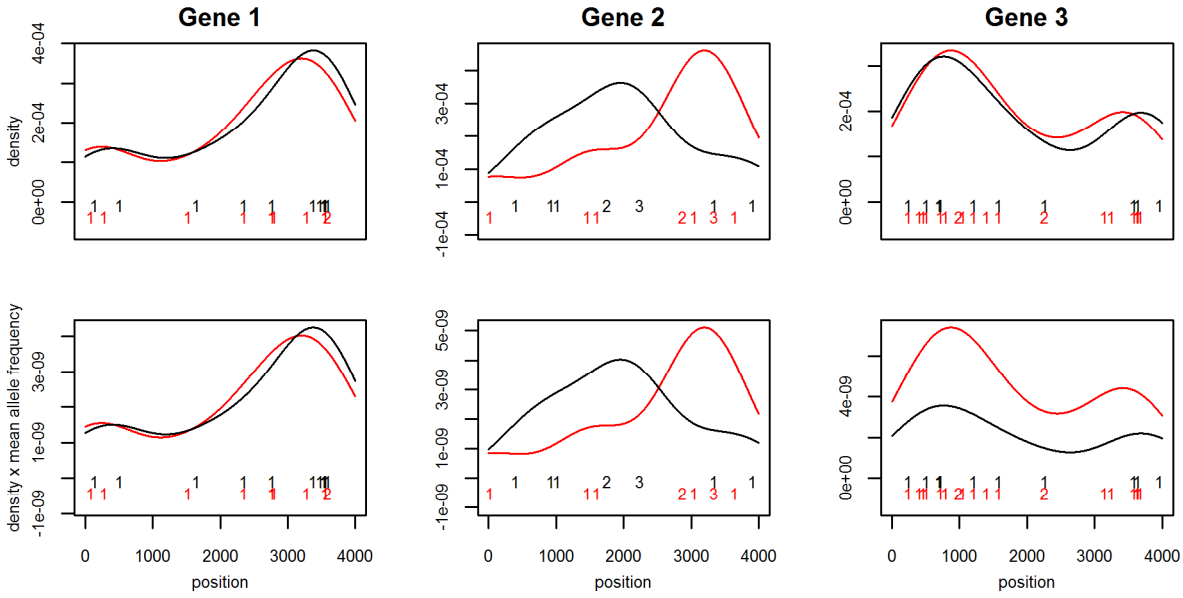


Fig 1. DoEstRare method illustration. The rare allele counts are represented on the gene in cases (red) and controls (black). From these counts are computed the density of mutation positions on the gene (top), and this density multiplied by the mean allele frequency (bottom), which is used by DoEstRare. A non-parametric method, the kernel density estimation using a Gaussian kernel, was used to estimate the mutation position density. Three genes have been simulated. Gene 1 presents the same mutation number (10) and the same mutation position distribution in cases and controls (no association with the disease). Gene 2 presents the same mutation number (10) but different mutation position distributions (association). Finally Gene 3 presents the same mutation position distribution but different mutation numbers (10 in controls and 20 in cases) (association).

https://doi.org/10.1371/journal.pone.0179364.g001

Performance of DoEstRare

We conducted simulations to assess the type I error and the power of DoEstRare and 14 other association tests on rare variants, covering a large spectrum of existing strategies. The tests we compared are summarized in the Table 1 with a categorization inspired by the review from Lee et al. (2014) [25].

We simulated a 10kb gene with the backward coalescent model available in the COSI program [22], for a population of 10,000 haplotypes. We repeated the simulations to obtain 10,000 replicates of this haplotype set. The median number of simulated variants is 202 (2.5% and 97.5% quantiles: [172; 235]). The median number of variants with a MAF between 0.001 and 0.01 is 30 (2.5% and 97.5% quantiles: [18; 46]). From the population, we simulated the phenotype in order to obtain 1,000 cases and 1,000 controls.

Type I error analysis. We assessed type I errors for each test under comparison by simulating under the null hypothesis of no association between the gene and the disease status. Type I error results are based on 10,000 replicates and the Fig 2 shows that all tests have correct type I error rates in terms of inflation of p-values [see S1A Table for type I error values]. Moreover some tests such as CAST and SKAT seem to be conservative. It has already been observed by Wu et al. (2011) that SKAT could be conservative with small sample sizes at low α levels [11].

Power analysis. For the power analysis, we simulated a disease-associated gene under three main scenarios, varying the causal variant positions (Fig 3). In a first scenario, the causal variants are uniformly distributed on the gene. In the second and third scenarios, they are clustered in one or two specific areas of the gene. We also varied the proportion of causal variants between 5%, 10%, 15% and 20%.

Focusing on scenario 1 results (Fig 4), in which DRVs are not clustered, an obvious observation is a power decrease with higher proportions of neutral variants [see S1B Table]. Nevertheless, in the context of no cluster of DRVs, some statistical tests seem to be more sensitive to the neutral variant inclusion than others. As it was observed with the comparison made by

Table 1. Rare variant association tests under comparison.

Positions	Category	Description of the strategy	Methods
No	Burden tests	Computation of a genetic score per individual corresponding to a binary variable.	CAST[4]
		Computation of a genetic score per individual corresponding to a weighted sum of genotypes.	WSS[6], VT[7], aSum[9]
	Variance-component tests	Test the variance of genetic effects.	C-alpha[10], SKAT[11], SKAT-O[12]
	P-value combination tests	Combination of p-values from single-marker tests.	ADA[14]
	Multi-genotype pattern	Analysis of multi-locus genotypes.	KBAC[8]
Yes	Sliding-window tests	A statistic is computed by genetic sliding window.	BOMP[18]
	Kernel matrix tests	A kernel matrix is used in the statistic to take into account physical distance between variants.	CLUSTER[20], KERNEL[19], PODKAT[21]
	Test on inter-marker distances	Physical distances between rare variants are computed. Weighted distance distribution functions are compared between cases and controls.	DBM[16]
	Rare variant density test	Comparison of rare variant position distributions and average allele frequencies on the gene, between cases and controls.	DoEstRare

Abbreviations: ADA, adaptive combination of P-values for rare variant association testing; aSum, data-adaptive sum test; BOMP, burden or mutation position; CAST, cohort allelic sum test; CLUSTER, test from Lin (2014); DBM, distance-based measure; KBAC, kernel-based adaptive cluster; KERNEL, test from Schaid et al. (2013); PODKAT, position-dependent kernel association test; SKAT, sequence kernel association test; VT, variable threshold; WSS, weighted sum statistic.

https://doi.org/10.1371/journal.pone.0179364.t001

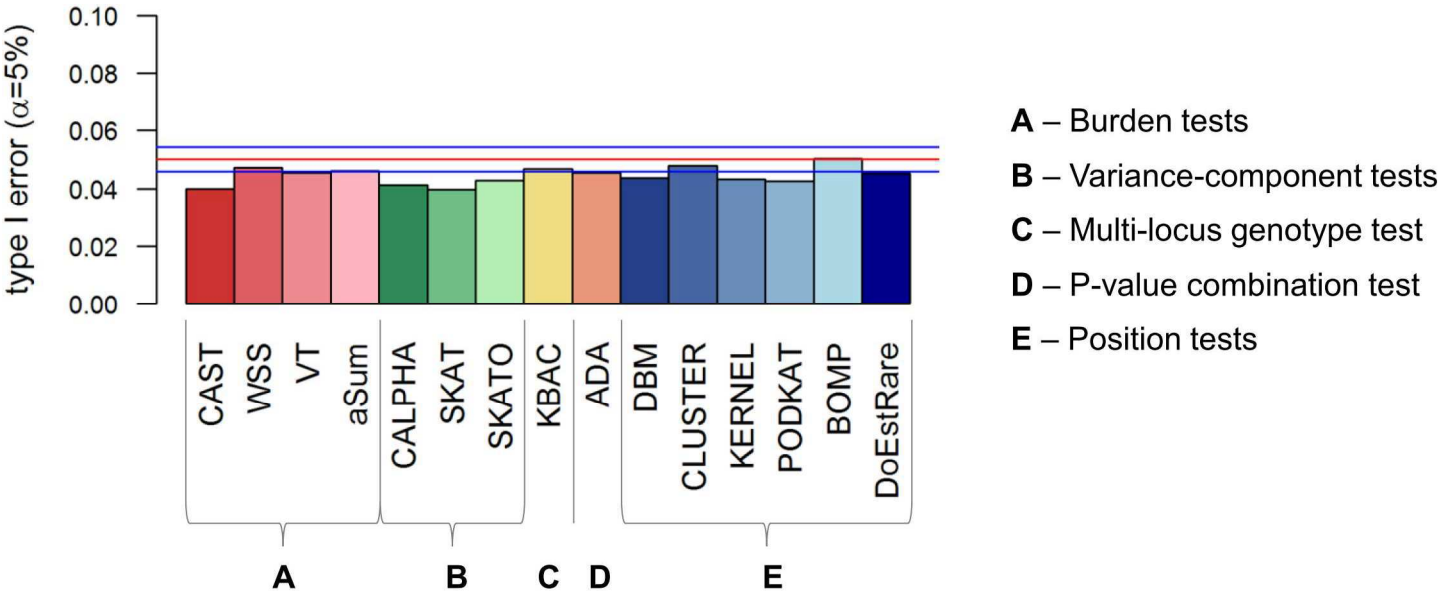


Fig 2. Type I error results at nominal level α = 5% based on 10,000 replicates. The red line corresponds to α = 5% and blue lines correspond to 95% confidence interval. Confidence interval is computed assuming that the number of false positives follows a binomial distribution with parameters 10,000 and 0.05.

https://doi.org/10.1371/journal.pone.0179364.g002

Basu and Pan (2011) [26], burden tests suffer from an important loss of power in a context of many non-causal variants, compared to variance-component tests. KBAC is also less noise-sensitive than burden tests, confirming that its statistic succeeds in better dissociating causal signals from noise. However ADA, which selects adaptively variants, is far less powerful than other tests in this context.

Regarding tests incorporating position information ("position tests"), DBM and CLUSTER tests perform badly with any proportion of DRVs. KERNEL and PODKAT tests are a bit more powerful than DBM and CLUSTER in this context of randomly distributed DRVs.

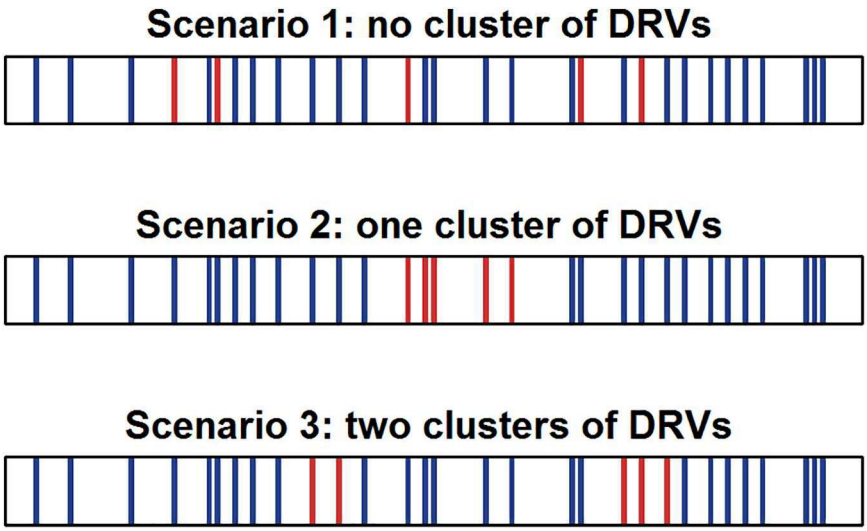


Fig 3. Simulation scenarios varying the DRV distribution. Each box represents a SNV on the gene. Blue boxes: non-causal variants. Red boxes: DRVs.

https://doi.org/10.1371/journal.pone.0179364.g003

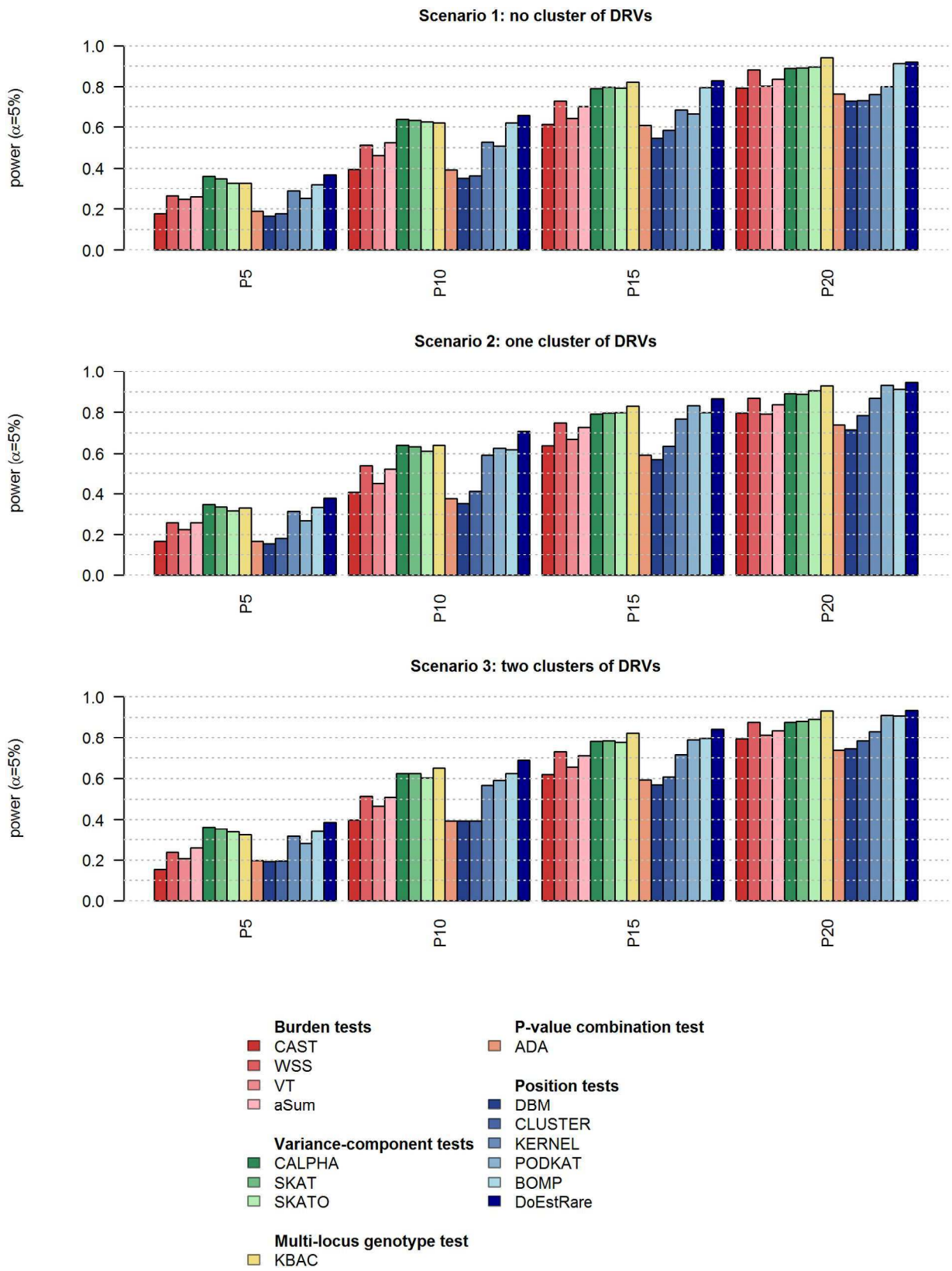


Fig 4. Power results at nominal level α = 5% based on 1,000 replicates. P5, P10, P15 and P20 correspond to 5%, 10%, 15% and 20% of DRVs in the gene. DRVs: disease-risk variants.

https://doi.org/10.1371/journal.pone.0179364.g004

Finally BOMP and DoEstRare test perform well with any proportion of DRVs, compared to the other tests, with DoEstRare slightly better than BOMP, especially with low proportions of DRVs. DoEstRare is among the most powerful tests with variance-component tests and KBAC.

Focusing on scenarios 2 and 3, in which DRVs are clustered respectively in one and two areas, the tests that do not incorporate gene positions display obviously the same power as in scenario 1 [see S1C and S1D Table]. Tests incorporating position information such as CLUSTER, KERNEL and PODKAT and DoEstRare show a power increase in scenario 2, i.e. with one cluster of DRVs, for some proportions of DRVs [see S1A Fig for power comparison between scenarios]. However we can observe a power decrease with two small clusters compared to one large cluster. In these scenarios with clustered DRVs, variance-component tests and KBAC are still more powerful than a majority of position tests. Finally DoEstRare is the most powerful test in every simulated scenario from main scenarios 2 and 3.

Application of DoEstRare to real data

We also applied our newly developed test and other association tests on rare variants to real data in order to study the significance similarities across the analyzed genes. We performed association tests on two pathologies, BrS and EOAD, in order to evaluate the stability of the comparison. Due to the number of tests under comparison, we used a multidimensional approach as Jeanmougin et al. (2010) [27]. We analyzed the significance similarities using a Principal Component Analysis (PCA) [28] on $-\log_{10}$(p-values) data.

BrS data. We applied 15 statistical tests (DoEstRare and 14 other tests) to BrS data. Sequence data for 163 candidate genes were available for 167 cases and 167 controls. Rare variants are defined as showing an MAF below 1%, and residing in coding DNA sequences (CDS) regions +/- 10 bp. We excluded from the PCA, all genes with missing p-values in at least one test. The reason of missing p-values is often due to the low number of variants in the gene. For some statistical tests such as DoEstRare and KERNEL, missing p-values are also due to the absence of rare mutations in cases or controls. We also excluded the DBM test, which returned more missing p-values than other tests. The PCA was performed on the remaining 58 genes (36% of the 163 genes in the targeted sequencing design) for the remaining 14 rare variant association tests. These genes represent a total of 1,462 rare variants, with a median of 15 rare variants per gene (min: 5; max: 441). In order to exhaustively compare BrS PCA results with EOAD PCA results, we set extra statistical tests (ADA, aSum, BOMP, CLUSTER, VT), which were performed only on BrS, as illustrative variables.

The cumulative inertia explained by the first three axes of the PCA is about 88.31% of the total inertia. From the BrS correlation circle (**Fig 5**), all statistical tests are positively correlated to the first PC (56.53% of inertia). The second PC (22.72%) opposes C-alpha, SKAT and KERNEL tests (ADA and CLUSTER illustrative tests) on one side, and WSS and KBAC tests (aSum and VT illustrative tests) on the other side. A third PC which explains only 9.05% of the inertia opposes DoEstRare to CAST.

EOAD data. We applied DoEstRare and 8 rare variant association tests to the EOAD dataset, which contain whole exome sequences from 431 cases and 555 controls. Rare variants are still defined with an MAF inferior to 1% and a location in CDS regions. As for the previous analysis on BrS, we removed from the PCA, genes presenting a missing p-value for at least one test. We analyzed 17,409 autosomal protein-coding genes with 9 statistical tests. These genes represent a total of 273,390 rare variants, corresponding to a median number of variants per gene of 11 (min: 2; max: 901).

The first three axes explain 86.72% of the total inertia (**Fig 5**). As for BrS data, all statistical tests are positively correlated with the first PC (54.89% of inertia). The second PC (19.96%)

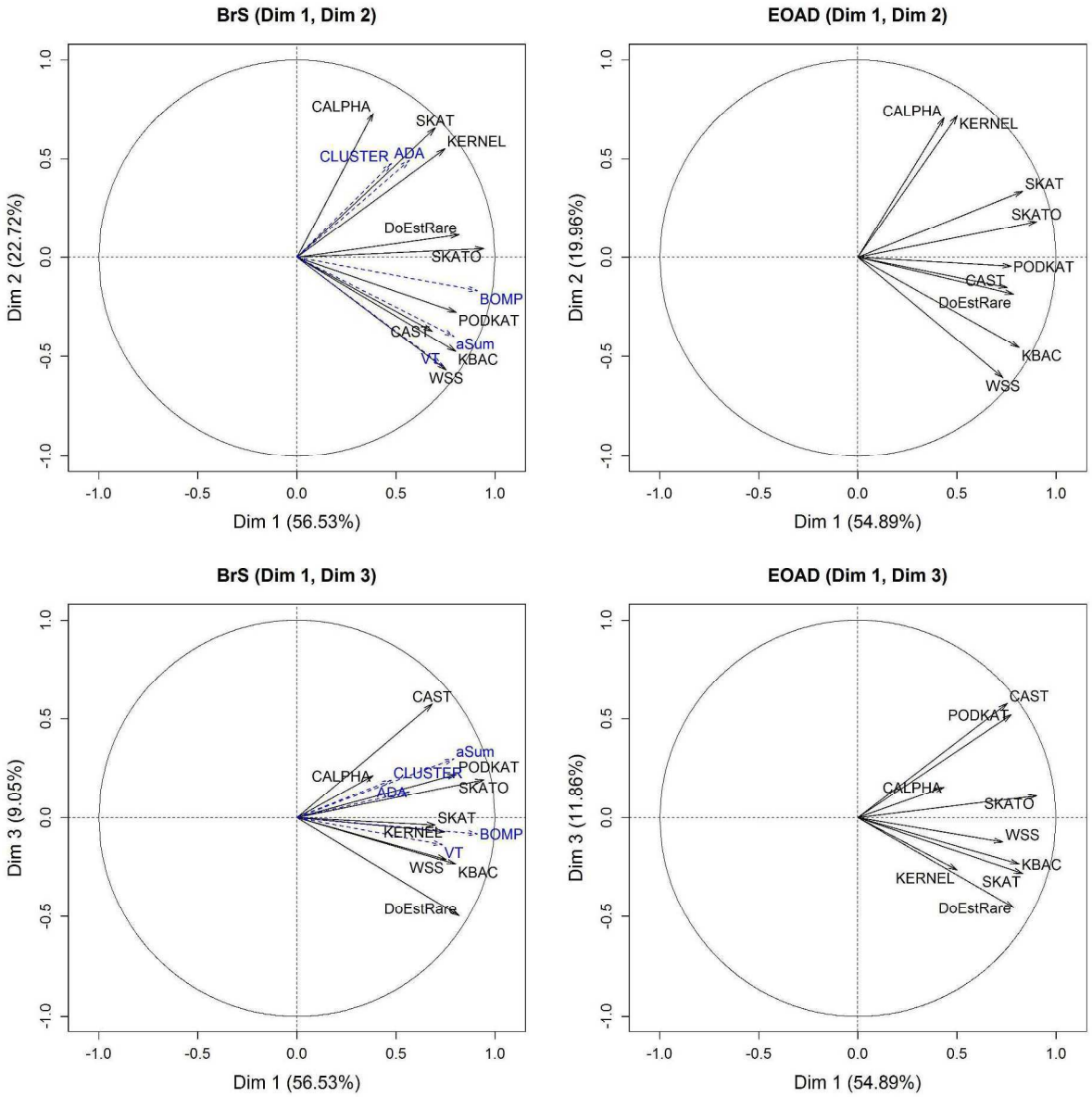


Fig 5. PCA correlation circle for BrS data (left) and EOAD data (right). PCA on -$\log_{10}$(p-value) is generated from the application of 9 association tests. BrS data includes 58 candidate genes. EOAD data includes 17409 autosomal protein coding genes. Illustrative variables are represented with blue dashed arrows.

https://doi.org/10.1371/journal.pone.0179364.g005

opposes mainly C-alpha and KERNEL tests to WSS and KBAC tests. Finally the third PC (11.86%) opposes DoEstRare to CAST and PODKAT tests.

These axes summarize significance differences between the association tests. Focusing on DoEstRare, two association signals stand out by their significance: *KRTAP5-5* (transcript: ENST00000399676, p = 3.8e-07) and *CELA-3B* (transcript: ENST00000337107, p = 8e-06). *KRTAP5-5* is in the list of the 10 most significant association signals with only the tests SKAT (p = 3.8e-05) and SKAT-O (p = 6.4e-05). This gene is certainly a false positive gene as rare mutations in cases are clustered in a 3 bp region within a repetitive element and are carried by a few individuals. The second most associated gene, *CELA3B*, is far less significant for all the other tests, the minimum p-value obtained with SKAT (p = 2.6e-04). In this gene, rare variants

present a different position distribution between cases and controls, and tend to cluster in a small genetic region in cases. On the contrary, the gene *NIPAL4* (transcript: ENST00000311946) with a high significance with most of the tests (CAST, WSS, SKAT-O, KBAC, PODKAT), is not well identified by DoEstRare (p = 1.13e-03). We can observe in this gene, a clear difference in the number of rare mutations between cases and controls, but not a clear difference of position distributions.

The significance levels of these three genes are shown in Manhattan plots [S2A–S2I Fig]. The localization of rare variants in cases and controls are represented in S3A–S3C Fig.

Computation times

Computation times of the different statistical tests that are used in this paper are indicated in the **Table 2**. These times are based on the analysis a 10kb gene, simulated under the null hypothesis, and including 30 variants for 1000 cases and 1000 controls. For the standard permutation and bootstrap procedures, we did respectively 500 permutations or resampling.

We note big differences in computation times between statistical tests. Of course these values depend highly on the implementation we used [see S1 Text for implementation details]. The most used statistical tests are CAST, SKAT and SKAT-O, as they are fast-running without a permutation or bootstrap procedure. DoEstRare computation time is quite high with a standard permutation procedure. This time can be greatly reduced with the adaptive permutation procedure we implemented and should be used in practice. The density estimation can be furthermore optimized in terms of computation time.

Discussion

Here we propose a new association test for rare variants, called DoEstRare, to identify clusters of DRVs in genes. The DoEstRare strategy combines a "position test" and a burden test, so that

Table 2. User CPU times for the different methods.

Test	Permutations/Bootstrap	Average time per gene (sec)	Total time (1000 genes)
CAST	No	0.013	0h 0min 13sec
WSS	Yes	26.221	7h 17min 1sec
VT	Yes	111.678	31h 1min 18sec
aSum	Yes	10.582	2h 56min 22sec
CALPHA	Yes	3.598	0h 59min 58sec
SKAT	No	0.091	0h 1min 31sec
SKAT	Yes (bootstrap)	1.326	0h 22min 6sec
SKAT-O	No	1.051	0h 17min 31sec
SKATO	Yes (bootstrap)	160.124	44h 28min 44sec
KBAC	Yes	0.187	0h 3min 7sec
ADA	Yes	25.318	7h 1min 58sec
DBM	Yes	7.933	2h 12min 13sec
CLUSTER	Yes	27.095	7h 31min 35sec
KERNEL	Yes	6.450	1h 47min 30sec
PODKAT	No	0.108	0h 1min 48sec
PODKAT	Yes (bootstrap)	1.459	0h 24min 19sec
BOMP	Yes	4.757	1h 19min 17sec
DoEstRare	Yes (standard)	22.617	6h 16min 57sec
DoEstRare	Yes (adaptive)	12.916	3h 35min 16sec

https://doi.org/10.1371/journal.pone.0179364.t002

it is still adapted to cases with randomly distributed DRVs. We also used, in the burden component, a weighting system to better discriminate risk variants from neutral variants.

First, we compared type I errors and powers, by conducting simulations under several genetic scenarios. These simulations scenarios were designed to assess statistical power in the context of high proportions of neutral variants. Simulations showed that DoEstRare is systematically the most powerful, alone or in conjunction with others for every scenario we simulated. DoEstRare performs well with or without clusters of DRVs. We also noticed by simulating scenarios varying proportion of DRVs, that DoEstRare is less noise-sensitive than burden tests.

In our power analysis, we also compared different strategies for testing rare variant association with a disease. We confirm that variance-component tests (C-alpha, SKAT, SKAT-O) and KBAC are better adapted to noise than burden tests and are powerful in every simulated scenario. There is no common benchmark in the literature for simulation designs to compare efficiently our results. However, Moutsianas et al. (2015) [29] compared different rare variant association tests and found that SKAT-O and KBAC have the highest mean power, across simulated scenarios varying sample sizes, effect locus sizes and neutral variant proportions, which is in accordance with our results. In our scenarios, SKAT-O, SKAT and C-alpha tests behave the same way because we simulated scenarios with high proportions of neutral variants. Compared to DoEstRare, these tests are still as powerful in most scenarios.

Most of the existing tests incorporating position information, except PODKAT and BOMP, are less powerful than variance-component tests and KBAC, in every scenario. KERNEL, DBM and PODKAT are more powerful in the presence of a cluster of DRVs, confirming the use of position information in their statistic. We also noted some power differences between scenarios with one and two clusters of DRVS. Indeed some tests are significantly less powerful in the scenario with two clusters. This observation may be related to cluster sizes being smaller in this scenario, for the same proportion of DRVs.

Simulation results depend on many factors due to the complexity of a group of rare variants. In our simulated scenarios we varied the proportion of causal variants in the gene and also the number of clusters of DRVs. Currently, there is still poor knowledge about the underlying biological mechanisms, likely to differ greatly between diseases. That is why it is hard to assess the realism of these two parameters. We used constant ORs for disease risk variant effects while it is unlikely that all causal variants in a gene have the same effect. Some authors suppose very rare variants to present stronger effects and set the regression coefficients of the logistic model as a decreasing function of the MAF [11,12]. Moreover, every simulated gene is a 10kb region while the real genetic structure is more complicated with very heterogeneous gene sizes and variant numbers.

We could assess the power of our statistical test DoEstRare and usual tests, with the analysis of simulated data. However, to test their behavior on biological datasets, we have applied up to 14 association tests, in addition to DoEstRare, on rare variants to BrS and EOAD data and investigated the significance similarities between the results from the different tests. This comparison may help in choosing the best test combination when designing a rare variants project and in interpreting differential test results. Both BrS and EOAD PCA representations underlined the same tendencies. All statistical tests are correlated with the first principal component, which means that significance results provided by different tests show globally the same tendency. The second source of inertia in significance results is due to some statistical tests giving partially different results. This may be related to the underlying genetic structure behind the disease. The main significance differences, according to the second principal component, are between the group including C-alpha and KERNEL, and the group including WSS and KBAC. The third component, which explains a low part of the inertia, opposes DoEstRare to CAST.

Finally SKAT-O seems to be a good compromise as it provides similar results than other tests. We could observe some differences in the correlation structure when analyzing BrS data and EOAD data. For instance, the correlation of PODKAT and SKAT significance results with the other tests differs between BrS and EOAD datasets. One explanation is that correlation structure between statistical results may depend on the studied pathology, as statistical tests assume different biological models. Interestingly, although we observe some correlations between association tests based on similar hypotheses, we do not observe a clear categorization. This issue should be further investigated with the study of other diseases.

With the application of rare variant association tests to real data, we highlighted several practical issues. In this article, we considered analyzing all rare variants without prior functional information. Criteria to incorporate only the most potential disease-risk rare variants in the analysis are not well defined in the literature, as the genetic architecture is specific of the studied disease. It is possible to take into account only those variants that have been previously functionally annotated [25,30,31], for example with sequence ontology terms describing their impact on the coded protein, in the context of gene analysis. Regarding the tests that incorporate position information, the delineation of the analyzed region is even more important as positions are relative to the bounds. It should be added that using tests to detect clusters of DRVs is irrelevant when the analyzed gene contains very few variants. The observed number of rare variants per gene depends greatly on the sample population size of the study. Indeed by screening more individuals in a study, it is more probable to identify very rare mutations. The sampling design is also an important factor in association analyses, as Nicolas et al. (2015) [24] identified an association signal when applying rare variant association test to EOAD patients with a positive family history. In this article we chose to analyze all 431 EOAD patients, whose 185 patients present a family history.

When applying statistical tests to real data, the use of an adaptive permutation procedure [32] is needed to reduce computational times and we implemented it for most of tests. Even using this strategy, encountering a high association signal may be time-consuming compared to association tests using an approximate statistic distribution under the null hypothesis. Moreover, taking into account position information is very useful to discover clusters of DRVs, but the delineation of the analyzed region is a supplementary step in the analysis workflow requiring reasoning. For the analysis of BrS data, we chose to use the definition of captured coding sequences. For the EOAD data, as capture designs were differing among patients, we used CDS annotations. Of course, by analyzing CDS regions, we may have missed few important variants situated in splice regions. The presence of clustered rare variants is, in some cases, due to technical artefacts as some regions are difficult to sequence. In EOAD results, a gene presents a very high significance with DoEstRare but is a false positive due to a cluster of rare mutations in a very small region for a very few cases. DoEstRare is also interesting to explore these short genomic regions with a low sequencing quality.

The relevance of a test depends highly on the underlying biological structure when assessing the association between a group of rare genetic variants and a disease. Each statistical test for rare variants is based on relative complex assumptions aiming to translate mathematically the disease mechanisms. We recommend, in order to identify the maximum association signals, to perform different statistical tests covering various strategies.

DoEstRare enables to incorporate position information and is powerful to detect clusters of disease risk variants. DoEstRare is a good alternative test to use in addition to classical strategies in order to explore genetic architectures involving functional domains. This test is advantageous in the context of analyzing long transcripts which are susceptible to contain important sub-regions. However, the density estimation of rare variant positions may be limited by testing a small group of variants. As discussed previously, DoEstRare is sensitive to short genetic

areas that are not well sequenced, which may result in the presence of false positives. The quality control is an important step in the analysis of rare variants, and DoEstRare can also be used to identify problematic sequences.

Our newly developed test DoEstRare is so far designed for deleterious rare alleles in case/control studies. In this article we did not take into account genetic population structure in rare variant association tests. However it has already been shown that this could impact significance results [33,34]. It can be developed in four directions: (i) adapting weights in the DoEstRare statistic in order to consider a mixture of protective and deleterious variants, or incorporating functional information; (ii) applying DoEstRare to quantitative traits by using a latent binary status, whose distribution probability depends on the phenotype distribution; (iii) exploring the choice of the kernel used in the density function estimations to reduce computational times; (iv) incorporating population stratification components in the computation of mean allele frequencies.

Methods

Notations

Let $\mathbf{X}$ be the matrix of genotypes with X_{ij} the count of rare alleles for the i-th individual and j-th rare variant, varying between 0, 1 or 2 rare alleles. Let $\boldsymbol{Y}$ be the vector of phenotypes with $Y_i = 1$ if the i-th individual is a case, $Y_i = 0$ if else. Let l_j be the position of the j-th rare variant. The number of affected (A) and unaffected (U) individuals are respectively N^A and N^U, with N the total number of individuals. The number of rare variants in the gene is P.

Tests under comparison

A first category of rare variant association tests is called burden tests [4,6,7,9], and consists in summarizing (or collapsing) the genetic information across the variants or across the individuals into a single value. A simple approach consists in computing for each individual a genetic score corresponding to a weighted sum of minor allele counts

$$S_i = \sum_{j=1}^{P} w_j X_{ij}$$

with w_j the weight for variant j. For example, the WSS test accords a more important weight to rare variants, as they may be more likely to have an effect on disease susceptibility. Weights differ between approaches according to biological assumptions. Finally the association between the genetic score and the disease status is tested.

Another category are the variance-component tests [10,11]. It has been developed to deal with the presence of opposite effects (deleterious and protective variants) and difference of effect magnitude (moderate effect to no effect). They test unusual variance of genetic effects on disease susceptibility in a group of variants. SKATs enable also more complex disease susceptibility models than the linear logistic regression model.

Because the proportion of deleterious and protective rare alleles is not known, some statistical tests combine both a burden test and a variance-component test such as SKAT-O [12] in order to preserve highest power. Indeed it has been observed by Basu et al. (2011) [26] that burden tests perform best when the gene includes a large amount of causal variants with the same effect whereas variant-component tests perform best in situations with protective or a lot a non-causal variants. Combined tests are developed to take advantage of the both strategies. To simplify our classification, we put SKAT-O into the variance-component test category.

Another strategy in association tests is to combine p-values obtained by single-marker tests that are commonly used in GWASs such as the ADA test [14].

The KBAC test [8] is a test considering multi-locus genotypes, i.e. the combinations of alleles across the genetic region of interest. This strategy aims to account for potential within-gene or between-gene interactions. It also uses a weighting system to better account for potential risk variants.

Finally several tests incorporate physical positions of rare variants [16–21]. A simple approach to detect a cluster of DRVs in a gene is to use a sliding-window approach. The genetic region of interest is divided into windows and a rare variant association test is performed for each window. Because the size and the location of the cluster are usually unknown, sliding windows of different sizes are commonly considered. This strategy is used in the BOMP test, proposed by Chen et al. (2013), and the test developed by Ionita-Laza et al. (2012). We didn't use the scan test from Ionita-Laza et al. (2012) as it is more suitable for large regions. Other tests such as KERNEL, CLUSTER and PODKAT, incorporate position information in a kernel matrix that measures distances between pairs of rare variants. In this article we propose a rare variant association test, DoEstRare, which is a combination of a burden test and a test comparing mutation position distributions.

Tests are described with more details in supplementary methods [see S1 Text].

DoEstRare test

Computation of the test statistic. DoEstRare aims to compare both the rare variant position distributions and allele frequencies between cases and controls. To test these two aspects, the statistic computes the area between the two mutation position density curves, each multiplied by the corresponding mean allele frequency, computed across all rare variants:

$$STAT = \int_1^{Lg} |\widehat{p^A} \times \widehat{f}^A(pos) - \widehat{p^U}\widehat{f}^U(pos)|dpos$$

with Lg denotes the length of the gene in bp. $\widehat{p^A}$, $\widehat{p^U}$, $\widehat{f}^A$ and $\widehat{f}^U$ are estimators for respectively mean allele frequencies and position density functions whose computation will be explained in next sections. Without the burden components $\widehat{p^A}$ and $\widehat{p^U}$, the statistic is similar to the total variation distance, used to compute a distance between the two probability density functions $\widehat{f}^A$ and $\widehat{f}^U$ [35].

Estimation of density functions. Position density functions f^A and f^U are estimated using a non-parametric way with the Gaussian kernel density estimation [36].

$$\widehat{f}^A(pos) = \frac{1}{bw}\sum_{j=1}^{P} w^A_{j,density} \times K\left(\frac{pos - l_j}{bw}\right) \text{ and } \widehat{f}^U(pos) = \frac{1}{bw}\sum_{j=1}^{P} w^U_{j,density} \times K\left(\frac{pos - l_j}{bw}\right)$$

with bw the bandwidth (smoothing parameter) and $K(.)$, the Gaussian kernel.

$w^A_{j,density}$ and $w^U_{j,density}$ are ratios of mutations at the lj-th position in cases and controls.

$$w^A_{j,density} = \frac{m^A_j}{\sum_{j=1}^{P} m^A_j}$$

$$w^U_{j,density} = \frac{m^U_j}{\sum_{j=1}^{P} m^U_j}$$

with $m^A_j = \sum_{i=1}^{N^A} X_{ij}$ and $m^U_j = \sum_{i=1}^{N^U} X_{ij}$ the observed counts of mutations for the j-th variant in cases and controls.

Burden components. To test the burden hypothesis, we estimate a weighted allele frequency average in cases and controls. The weight system enables to better discriminate high potential causal variants from neutral variants. The burden component expressions are:

$$\widehat{p^A} = \frac{1}{P}\sum_{j=1}^{P} \frac{w_j}{\sum_{j=1}^{P} w_j} \frac{m_j^A}{2N^A} \qquad \widehat{p^U} = \frac{1}{P}\sum_{j=1}^{P} \frac{w_j}{\sum_{j=1}^{P} w_j} \frac{m_j^U}{2N^U}$$

with w_j the weight for the j-th variant.

Under the assumption that all causal variants are deleterious, (i.e. variants that are enriched in cases present a more important weight), we assume that the count M_j^A of rare mutations in cases for a variant j follows, under the null hypothesis, a binomial distribution $B(2N^A, \widehat{q_j^U})$ with $\widehat{q_j^U}$ the estimate of the minor allele frequency in controls:

$$\widehat{q_j^U} = \frac{m_j^U + 1}{2N^U + 2}$$

The weight w_j is defined as the probability to present less than the observed count m_j^A.

$$w_j = P(M_j^A \leq m_j^A) = \sum_{k=0}^{m_j^A} \binom{2N^A}{k} (q_j^U)^k (1 - q_j^U)^{2N^A - k}$$

Significance. The significance of the test is evaluated with a standard phenotype permutation procedure. For each permutation $b \in \{1,\ldots,B\}$, the phenotypes labels are randomly shuffled (permuted) and the statistic $STAT^{(b)}$ is calculated. As the statistic is an area, which means a positive real number, the p-value is defined, in the context of standard phenotype permutation procedure, by $\frac{\sum_{b=1}^{B} (STAT^{(b)} \geq STAT) + 1}{B+1}$ [37], with B the total number of permutations. An adaptive permutation procedure can also be used to reduce computational times, in the context of large data [32].

Simulation framework

We conducted genetic simulation studies to evaluate and compare the performance of DoEstRare in terms of power and type I error. Our simulation workflow for each replicate is described with the following steps. It briefly consists in generating a haplotype matrix from which are sampled cases and controls, the disease risk model being a logistic regression model (see **Fig 6**).

Step 1: We generate 10,000 haplotypes for a 10kb region using a backward coalescent model implemented in the COSI program [22]. Parameters correspond to what is called the "bestfit" model by Schaffner et al. (2005), which were obtained by calibration. Haplotypes are sampled from what corresponds to the European population in this model.

Step 2: In the haplotype matrix generated in step 1, we select rare variants with a MAF $\in$ [0.001;0.01]. We set a minimal MAF to avoid a lot of non-observed causal mutations in the simulated data, a framework leading to the null hypothesis model because of the very low frequency of a large proportion of variants.

Step 3: We determine rare causal variants for the logistic regression model. Causal variants are determined according different scenarios related to their positions on the gene. These scenarios are explained further in the simulation framework description.

Step 4: We sample two haplotypes from the haplotype matrix generated in step 1 to constitute the genotype data X_i for the i-th individual.

Step 5: The phenotype of the i-th individual is simulated with the following logistic regression model:

$$logit(P(Y_i = 1|X_i)) = \beta_0 + \boldsymbol{X_i}^T\boldsymbol{\beta}$$

with β_0 the intercept and $\boldsymbol{\beta}$ the vector of regression coefficients for the genetic effects. We set $\beta_0 = \log\left(\frac{0,05}{1-0,05}\right)$ so that 5% of individuals without any rare mutation are affected. In the context of rare variants, this value is close to the disease prevalence. For regression coefficients, we chose $\beta_j = \log(\mathrm{OR}_j)$ with $OR_j = 3$ if the j-th is a causal variant, else $\beta_j = 0$. The disease status of the individual i is sampled according to the Bernoulli distribution of probability $P(Y_i = 1|X_i)$.

Step 6: We repeat steps 4 and 5 until we obtain 1,000 cases and 1,000 controls.

Three main scenarios (Fig 3) are considered in relation to the positions of causal variants. In a first scenario, DRVs are not clustered in any specific area and are randomly sampled without replacement on the whole gene. In the second and third scenarios, DRVs cluster respectively in one and two areas. In these scenarios, initial causal variant positions correspond to the median of all variant positions for the second scenario, and to the quantiles 1/3 and 2/3 for the third scenario. Then DRVs are extended from initial causal positions to the neighbor variants until the specified number of DRVs is reached. We set the number of DRVs so that the proportions vary between 5%, 10%, 15% and 20% of the total variants within a gene (noted scenarios P5, P10, P15 and P20). In the context of clustered DRVs, these proportions are related to cluster window sizes. Indeed, cluster window sizes are larger with higher proportions of DRVs.

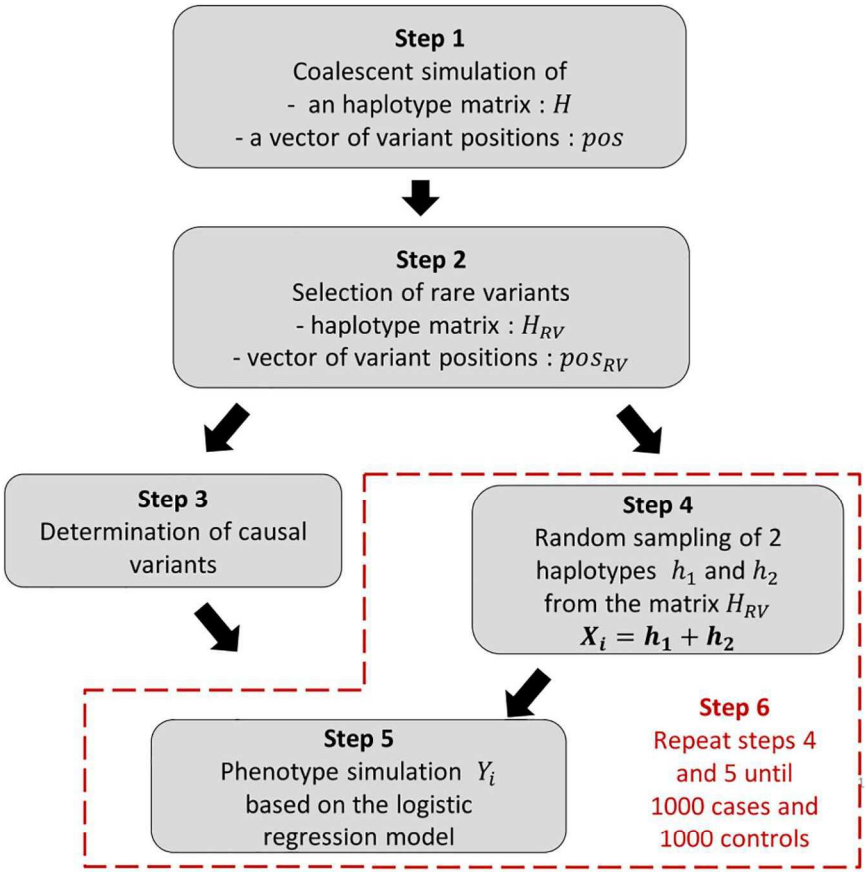


Fig 6. Simulation workflow. The different steps of the simulation workflow are further detailed in the article. The red dashed frame represents the step 6 which consists in the repetition of steps 4 and 5.

https://doi.org/10.1371/journal.pone.0179364.g006

The performances of the different tests (see section Tests under comparison) are compared in terms of power and type I error. We defined the type I error and the power as

$$\frac{\sum_{r=1}^{R} I(p - value_r \leq \alpha)}{R} = \begin{cases} type\ I\ error(\alpha) & \text{if } H_0 \text{ situation} \\ power(\alpha) & \text{if } H_1 \text{situation} \end{cases}$$

with $r \in \{1,\ldots,R\}$ the replicate index. We did 1,000 replicates for each scenario of the power analysis and 10,000 replicates for type I error analysis. Power and type I errors were computed for the test significance level at $\alpha = 5\%$.

We applied DoEstRare and 14 other rare variant association tests on simulated data. We set B = 500 permutations for each permutation-based test, i.e. all tests except CAST, SKAT, SKAT-O and PODKAT.

Real data

We performed DoEstRare and other rare variant association tests (see section Tests under comparison) on BrS and EOAD data. We analyzed the significance result similarities between the different tests by using a PCA [28]. The PCA data is the matrix containing $(-log_{10}(p\text{–}value))_{jt}$ for gene j in row and statistical test t in column. The PCA was performed with the R package FactoMineR [38].

Data from the BrS study. We applied DoEstRare and 14 rare variant association tests to BrS data published by Le Scouarnec et al. (2015) [23]. In this study, rare variant association tests were conducted to identify new genes of susceptibility for BrS. A significant enrichment of *SCN5A* rare variant carriers was observed in BrS patients.

In this study, cases include 167 patients diagnosed with BrS, and controls include 167 individuals aged over 65-year old and showing no history of cardiac arrhythmia. Both cases and controls are individuals of European origin.

This is a candidate gene study in which coding sequences of 163 candidate genes have been captured and sequenced. In this publication, burden test results for 45 genes are published. These genes have been previously shown to be related to cardiac arrhythmias or conduction defects and/or sudden cardiac death. The functional units tested are genes and more specifically the coding regions with a margin of 10 bp to take into account splicing sites.

In the present study, rare genetic variants were defined as variants with a MAF<1% in the Exome Aggregation Consortium (ExAC) database for the Non-Finnish European population (NFE) (release 0.3 downloaded at ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/ExAC.r0.3.sites.vep.vcf.gz) [39]. To avoid false positives, we excluded variants found in more than 5% of cases or controls but absent from ExAC database. Unlike the burden test framework described in the original publication, we analyzed all variants regardless their Sequence Ontology (SO) terms, estimated with Ensembl (http://www.ensembl.org), which describe the type of consequence of the mutation.

We evaluated significance of association tests using a standard permutation procedure (excluding CAST, SKAT, SKAT-O and PODKAT) with 1,000 phenotype permutations for all genes except *SCN5A*, which is a major gene implicated in BrS and where we performed 200,000 permutations.

Note: More details about population sampling, sequencing and variant calling can be found in the publication of Le Scouarnec et al. (2015) [23].

Data from the Alzheimer study. We performed rare variant association tests on EOAD data from Nicolas et al. (2015) [24]. In this study, a significant enrichment of SORL1 rare variants was detected in EOAD patients.

Published data, after quality control preprocessing, include 498 controls from 5 different French cities and 484 EOAD patients recruited by the French National CNR-MAJ consortium (205 patients with positive family history). We brought some modifications to the initial design in order to allow a robust comparison between tests. First, we decided to minimize technical biases by including only cases and controls that were sequenced by either Agilent SureSelect Human All Exons V5 or Agilent SureSelect Human All Exons V5UTR capture designs. Additional controls from another French city were also included, due to the French Exome project's progress. Our analysis finally compared 555 controls to 431 EOAD patients.

In this exome study, we annotated variants with Variant Effect Predictor (Ensembl). For each protein coding gene, we analyzed the "canonical" transcript: the transcript which presents (1) the longest CDS length, (2) if CDS lengths are equal, the longest transcript length with UTR regions. A total of 19,076 autosomal protein coding genes were annotated. For association tests incorporating position in the transcript, we used CDS regions and each variant was annotated with a CDS position.

Filters on genetic variant are the same as for the BrS data analysis. Rare genetic variants present a MAF<1% in the ExAC database for the NFE population and present a MAF in cases and in controls both less than 5%. To avoid false positives we excluded all rare variants which were very significant, as they could influence results from gene-based tests. These variants need to be checked apart from the analysis. Variants with a p-value less than 1e-04 by single-marker test (Fisher exact) were removed.

Due to the data size, we used the adaptive permutation procedure described by Che et al. (2014) to reduce computational times [32]. We set, as parameters, the adjusted nominal significance level to α = 5e-07, with a precision of c = 0.2. We chose to not apply all statistical tests and selected several tests per category. We applied CAST and WSS as burden tests, SKAT and SKAT-O as variance component tests, KBAC as a multi-locus genotype test, KERNEL, PODKAT and DoEstRare as position tests. Some statistical tests as ADA and CLUSTER were not easily adaptable for adaptive permutation and thus were not included in this round of analyses.

Supporting information

S1 Table. Power and type I error tables. Values of type I errors and powers assessed with the analysis of simulated data.
(PDF)

S1 Fig. Power comparison between simulated scenarios. The Fig A is another illustration of power results to better compare simulated scenarios, represented in Fig 4.
(PDF)

S2 Fig. Manhattan plots for EOAD results. From Fig A to Fig I, are represented significance results for the 17,409 autosomal genes that were analyzed. Only the names of the three genes, *KRTAP5-5*, *CELA3B* and *NIPAL4*, are indicated. The red line corresponds to a significance level of 2.5e-06 (5% adjusted with a Bonferroni correction for 20,000 genes).
(PDF)

S3 Fig. Mutation position density plots for EOAD results. From Fig A to Fig C, are represented allele counts and mutation position densities for the three genes, *KRTAP5-5*, *CELA3B* and *NIPAL4*, in cases and controls. Density functions for mutation positions were estimated with a Gaussian kernel.
(PDF)

S1 Text. Supplementary methods. Further details about the rare variant association tests we compared.
(PDF)

Acknowledgments

The authors would like to thank Genomics and Bioinformatics Core Facility of Nantes (GenoBiRD, Biogenouest). Computations were also performed on the "Centre de calcul intensif des Pays de la Loire" (CCIPL) computer "Erdre". We are also grateful to French clinical network against inherited cardiac arrhythmias. We gratefully acknowledge Pierre Lindenbaum and Floriane Simonet for technical support in bioinformatics, biostatistics and data management.

Consortia: The FREX Consortium's principal investigators are Emmanuelle Génin, Dominique Campion, Jean-François Dartigues, Jean-François Deleuze, Jean-Charles Lambert, and Richard Redon. Collaborators are as follows: bioinformatics group (Thomas Ludwig, Benjamin Grenier-Boley, Sébastien Letort, Pierre Lindenbaum, Vincent Meyer, Olivier Quenez), statistical genetics group (Christian Dina, Céline Bellenguez, Camille Charbonnier-Le Clézio, Joanna Giemza), data collection (Stéphanie Chatel, Claude Férec, Hervé Le Marec, Luc Letenneur, Gaël Nicolas, Karen Rouault), and sequencing (Delphine Bacq, Anne Boland, Doris Lechner).

Author Contributions

Conceptualization: EP RR LB CD.

Formal analysis: EP MK SLS.

Funding acquisition: JJS RR CD.

Investigation: EP.

Methodology: EP RR LB CD.

Project administration: RR LB CD.

Resources: JJS RR.

Software: EP.

Supervision: CLC DC The French Exome Consortium JJS RR LB CD.

Visualization: EP.

Writing – original draft: EP.

Writing – review & editing: MK SLS RR LB CD.

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan 1; 42(D1):D1001–6.
2. Maher B. Personal genomes: The case of the missing heritability. Nat News. 2008 Nov 5; 456 (7218):18–21.
3. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct 8; 461(7265):747–53. https://doi.org/10.1038/nature08494 PMID: 19812666

4. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat Res. 2007 Feb 3; 615(1–2):28–56. https://doi.org/10.1016/j.mrfmmm.2006.09.003 PMID: 17101154
5. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am J Hum Genet. 2008 Sep; 83(3):311–21. https://doi.org/10.1016/j.ajhg.2008.06.024 PMID: 18691683
6. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009 Feb; 5(2):e1000384. https://doi.org/10.1371/journal.pgen.1000384 PMID: 19214210
7. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei L-J, et al. Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet. 2010 Jun 11; 86(6):832–8. https://doi.org/10.1016/j.ajhg.2010.04.005 PMID: 20471002
8. Liu DJ, Leal SM. 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 Genet. 2010 Oct 14; 6(10):e1001156. https://doi.org/10.1371/journal.pgen.1001156 PMID: 20976247
9. Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Hum Hered. 2010; 70(1):42–54. https://doi.org/10.1159/000288704 PMID: 20413981
10. Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an Unusual Distribution of Rare Variants. PLoS Genet. 2011 Mar 3; 7(3):e1001322. https://doi.org/10.1371/journal.pgen.1001322 PMID: 21408211
11. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011 Jul 15; 89(1):82–93. https://doi.org/10.1016/j.ajhg.2011.05.029 PMID: 21737059
12. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. Am J Hum Genet. 2012 Aug 10; 91(2):224–37. https://doi.org/10.1016/j.ajhg.2012.06.007 PMID: 22863193
13. Cheung YH, Wang G, Leal SM, Wang S. A fast and noise-resilient approach to detect rare-variant associations with deep sequencing data for complex disorders. Genet Epidemiol. 2012 Nov; 36(7):675–85. https://doi.org/10.1002/gepi.21662 PMID: 22865616
14. Lin W-Y, Lou X-Y, Gao G, Liu N. Rare variant association testing by adaptive combination of P-values. PloS One. 2014; 9(1):e85728. https://doi.org/10.1371/journal.pone.0085728 PMID: 24454922
15. Robertson SP, Twigg SRF, Sutherland-Smith AJ, Biancalana V, Gorlin RJ, Horn D, et al. Localized mutations in the gene encoding the cytoskeletal protein filamin A cause diverse malformations in humans. Nat Genet. 2003 avril; 33(4):487–91. https://doi.org/10.1038/ng1119 PMID: 12612583
16. Fier H, Won S, Prokopenko D, AlChawa T, Ludwig KU, Fimmers R, et al. "Location, Location, Location": a spatial approach for rare variant analysis and an application to a study on non-syndromic cleft lip with or without cleft palate. Bioinforma Oxf Engl. 2012 Dec 1; 28(23):3027–33.
17. Ionita-Laza I, Makarov V, ARRA Autism Sequencing Consortium, Buxbaum JD. Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. Am J Hum Genet. 2012 Jun 8; 90(6):1002–13. https://doi.org/10.1016/j.ajhg.2012.04.010 PMID: 22578327
18. Chen Y-C, Carter H, Parla J, Kramer M, Goes FS, Pirooznia M, et al. A hybrid likelihood model for sequence-based disease association studies. PLoS Genet. 2013; 9(1):e1003224. https://doi.org/10.1371/journal.pgen.1003224 PMID: 23358228
19. Schaid DJ, Sinnwell JP, McDonnell SK, Thibodeau SN. Detecting genomic clustering of risk variants from sequence data: cases versus controls. Hum Genet. 2013 Nov; 132(11):1301–9. https://doi.org/10.1007/s00439-013-1335-y PMID: 23842950
20. Lin W-Y. Association testing of clustered rare causal variants in case-control studies. PloS One. 2014; 9(4):e94337. https://doi.org/10.1371/journal.pone.0094337 PMID: 24736372
21. Bodenhofer U. PODKAT: An R Package for Association Testing Involving Rare and Private Variants. R package version 1.0.3; 2015.
22. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005 Nov 1; 15(11):1576–83. https://doi.org/10.1101/gr.3709305 PMID: 16251467
23. Le Scouarnec S, Karakachoff M, Gourraud J-B, Lindenbaum P, Bonnaud S, Portero V, et al. Testing the burden of rare variation in arrhythmia-susceptibility genes provides new insights into molecular

diagnosis for Brugada syndrome. Hum Mol Genet. 2015 May 15; 24(10):2757–63. https://doi.org/10.1093/hmg/ddv036 PMID: 25650408

24. Nicolas G, Charbonnier C, Wallon D, Quenez O, Bellenguez C, Grenier-Boley B, et al. SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. Mol Psychiatry. 2015 Aug 25;
25. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. Am J Hum Genet. 2014 juillet; 95(1):5–23. https://doi.org/10.1016/j.ajhg.2014.06.009 PMID: 24995866
26. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genet Epidemiol. 2011 Nov; 35(7):606–19. https://doi.org/10.1002/gepi.20609 PMID: 21769936
27. Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, Guedj M. Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. PloS One. 2010; 5(9):e12336. https://doi.org/10.1371/journal.pone.0012336 PMID: 20838429
28. Bellanger L, Tomassone R. Exploration de données et méthodes statistiques: Data analysis & Data mining avec le logiciel R. Ellipses; 2014. 479 p.
29. Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS Genet. 2015 Apr; 11(4):e1005165. https://doi.org/10.1371/journal.pgen.1005165 PMID: 25906071
30. Li B, Liu DJ, Leal SM. Identifying rare variants associated with complex traits via sequencing. Curr Protoc Hum Genet Editor Board Jonathan Haines Al. 2013 Jul;Chapter 1:Unit 1.26.
31. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. Genome Med. 2015; 7(1):16. https://doi.org/10.1186/s13073-015-0138-2 PMID: 25709717
32. Che R, Jack JR, Motsinger-Reif AA, Brown CC. An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use. BioData Min. 2014; 7:9. https://doi.org/10.1186/1756-0381-7-9 PMID: 24976866
33. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012 Mar; 44(3):243–6. https://doi.org/10.1038/ng.1074 PMID: 22306651
34. O'Connor TD, Kiezun A, Bamshad M, Rich SS, Smith JD, Turner E, et al. Fine-scale patterns of population stratification confound rare variant association tests. PloS One. 2013; 8(7):e65834. https://doi.org/10.1371/journal.pone.0065834 PMID: 23861739
35. Feller W. An Introduction to Probability Theory and Its Applications. Wiley; 1971. 656 p.
36. Silverman BW. Density Estimation for Statistics and Data Analysis. CRC Press; 1986. 190 p.
37. Ernst MD. Permutation Methods: A Basis for Exact Inference. Stat Sci. 2004 Nov; 19(4):676–85.
38. Lê S, Josse J, Husson F, others. FactoMineR: An R package for multivariate analysis. J Stat Softw. 2008; 25(1):1–18.
39. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug 18; 536(7616):285–91. https://doi.org/10.1038/nature19057 PMID: 27535533

Human Molecular Genetics, 2015, 1–7

doi: 10.1093/hmg/ddv036
Original Article

ORIGINAL ARTICLE

Testing the burden of rare variation in arrhythmia-susceptibility genes provides new insights into molecular diagnosis for Brugada syndrome

Solena Le Scouarnec[1,2,3,†], Matilde Karakachoff[1,2,3,4,†], Jean-Baptiste Gourraud[1,2,3,5,†], Pierre Lindenbaum[1,2,3,5], Stéphanie Bonnaud[1,2,3,5], Vincent Portero[1,2,3], Laëtitia Duboscq-Bidot[1,2,3], Xavier Daumy[1,2,3], Floriane Simonet[1,2,3], Raluca Teusan[1,2,3], Estelle Baron[1,2,3], Jade Violleau[1,2,3,5], Elodie Persyn[1,2,3], Lise Bellanger[3,6], Julien Barc[7,8], Stéphanie Chatel[1,2,3,5], Raphaël Martins[9], Philippe Mabo[9], Frédéric Sacher[10], Michel Haïssaguerre[10], Florence Kyndt[1,2,3,5], Sébastien Schmitt[3,11], Stéphane Bézieau[3,11], Hervé Le Marec[1,2,3,5], Christian Dina[1,2,3,5], Jean-Jacques Schott[1,2,3,5], Vincent Probst[1,2,3,5] and Richard Redon[1,2,3,5,*]

[1]Inserm, UMR 1087, l'institut du thorax, Nantes, France, [2]CNRS, UMR 6291, Nantes, France, [3]Université de Nantes, Nantes, France, [4]Institute of Clinical Physiology, National Research Council, Pisa, Italy, [5]CHU Nantes, l'institut du thorax, Service de Cardiologie, Nantes, France, [6]Laboratoire de Mathématiques Jean Leray, UMR CNRS 6629, Nantes, France, [7]Department of Clinical and Experimental Cardiology, Academic Medical Center, Amsterdam, The Netherlands, [8]ICIN-Netherlands Heart Institute, Utrecht, The Netherlands, [9]CHU Rennes, Service de Cardiologie, France, [10]CHU Bordeaux, Service de Cardiologie, LYRIC Institute, France and [11]CHU Nantes, Service de Génétique Médicale, Nantes, France

*To whom correspondence should be addressed at: l'institut du thorax, Inserm UMR 1087, CNRS UMR 6291, IRS-UN, 8 Quai Moncousu, BP 70721, 44007 Nantes Cedex 1, France; Tel: +33 228080141; Fax: +33 228080130; Email: richard.redon@inserm.fr

Abstract

The Brugada syndrome (BrS) is a rare heritable cardiac arrhythmia disorder associated with ventricular fibrillation and sudden cardiac death. Mutations in the *SCN5A* gene have been causally related to BrS in 20–30% of cases. Twenty other genes have been described as involved in BrS, but their overall contribution to disease prevalence is still unclear. This study aims to estimate the burden of rare coding variation in arrhythmia-susceptibility genes among a large group of patients with BrS. We have developed a custom kit to capture and sequence the coding regions of 45 previously reported arrhythmia-susceptibility genes and applied this kit to 167 index cases presenting with a Brugada pattern on the electrocardiogram as well as 167 individuals aged over 65-year old and showing no history of cardiac arrhythmia. By applying burden tests, a significant enrichment in rare coding variation (with a minor allele frequency below 0.1%) was observed only for *SCN5A*, with rare coding variants carried by 20.4% of

[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Received: November 18, 2014. **Revised and Accepted:** January 31, 2015

cases with BrS versus 2.4% of control individuals ($P = 1.4 \times 10^{-7}$). No significant enrichment was observed for any other arrhythmia-susceptibility gene, including *SCN10A* and *CACNA1C*. These results indicate that, except for *SCN5A*, rare coding variation in previously reported arrhythmia-susceptibility genes do not contribute significantly to the occurrence of BrS in a population with European ancestry. Extreme caution should thus be taken when interpreting genetic variation in molecular diagnostic setting, since rare coding variants were observed in a similar extent among cases versus controls, for most previously reported BrS-susceptibility genes.

Introduction

The Brugada syndrome (BrS, MIM #601144) is a rare cardiac arrhythmia disorder associated with syncope and sudden cardiac death, which is characterized by a coved ST-segment and J-point elevation ≥0.2 mV, followed by a negative T wave in the right precordial leads of the electrocardiogram (ECG) (1).

Mutations in the *SCN5A* gene, which encodes the major cardiac sodium channel $Na_v1.5$, have been causally related to BrS in 20–30% of cases (2). Twenty other genes have been described as involved in BrS, mostly by candidate gene approaches. These genes encode subunits of cardiac sodium channels (*SCN1B*, *SCN2B*, *SCN3B*, *SCN10A*), L-type calcium channels (*CACNA1C*, *CACNB2*, *CACNA2D1*), potassium channels (*KCNH2*, *KCNE3*, *KCNE1L/KCNE5*, *KCND3*, *KCNJ8*, *ABCC9*), other channels (*TRPM4*, *HCN4*), channel-interacting proteins (*GPD1L*, *RANGRF/MOG1*, *SLMAP*), a desmosomal protein (*PKP2*) and a fibroblast growth factor (*FGF12*) (2–8). However, the prevalence of mutations in these other genes among BrS cases is still unclear. For example, the reported mutation prevalence of the *CACNA1C* gene ranges from 2 to 12% in the literature (2). Noteworthy, in a report describing the presence of putative pathogenic mutations among 12 of these genes in 20% of BrS cases, the authors showed that the mutations in all genes except *SCN5A* account for <5% of cases in total (9). Conversely, mutations in the *SCN10A* gene were recently reported as carried by 16.7% of patients with BrS (4).

The emergence of next-generation sequencing (NGS) technologies enables to screen multiple genes simultaneously and identify multiple rare genetic variants in patients. As a consequence, since rare variants are a common feature in the general population, there is an increasingly recognized risk of false-positive reports of causality (10). Studies are now required to determine which genes are significantly associated with BrS and interpret genetic data accurately in the context of molecular diagnosis. In this study, we have tested by targeted sequencing 45 genes previously involved in cardiac arrhythmias (including 21 BrS-susceptibility genes) in 167 index cases with BrS as well as 167 ethnically matched control individuals, in order to determine which genes carry a burden of rare genetic variants in cases versus controls and therefore contribute significantly to BrS.

Results

Prevalence of rare variants in BrS susceptibility genes

The 167 patients with BrS included in this study were sequenced on average to a mean coverage depth of 749× per sample—with 98% of the targeted regions covered at least 10 times—in order to ensure the detection of most heterozygous variants (Supplementary Material, Table S1). Variants were considered as relevant if rare—i.e. reported with a minor allele frequency (MAF) below 0.1% in the European population—and if predicted to alter the protein sequence (see Supplementary Material, Methods for details). All relevant variants following these criteria—hereafter called 'functional' variants—are listed in Supplementary Material, Table S2.

Rare functional variants in *SCN5A* were identified in 20.4% of our BrS population (34 out of 167 cases, 35 variants; see Fig. 1, Table 1 and Supplementary Material, Table S2), in accordance with previous reports (2,9,11). Among these 35 variants, one is non-sense (NM_198056.2:c.535C>T, R179Ter), two are frameshift deletions (NM_198056.2:c.2559delT, F853LfsTer16 and NM_198056.2:c.5314delC, L1772WfsTer15) and one affects a splice site (NM_198056.2:c.3963+2T>C). In accordance with previous reports (9,12–14), we found that rare functional variation in *SCN5A* is associated with longer PR interval (197 ± 21 versus 174 ± 25 ms; $P < 0.001$) and longer QRS duration (109 ± 13 versus 101 ± 12 ms; $P = 0.005$). In our study, spontaneous BrS ECG pattern (67 versus 36%; $P = 0.002$) was more prevalent in *SCN5A* rare variant carriers (Table 2).

Interestingly, three patients carry two rare missense variants in *SCN5A*: BrS_15 (NM_198056.2:c.1003T>A, C335S and c.3121G>A, D1041N), BrS_27 (NM_198056.2:c.1844G>A, G615E and c.4247C>A, A1416E), and BrS_28 (NM_198056.2:c.1333C>G, H445D and c.5339A>G, E1780G). The six variants are highly conserved between species, with every GERP (Genomic Evolutionary Rate Profiling) conservation score above 4. The three patients present with a spontaneous Brugada pattern, one case being symptomatic (unexplained syncope without any arrhythmia in the follow up). The mean PR interval in the presence of two *SCN5A* variants [213 (190–230) ms] tends to be longer than among all *SCN5A* variant carriers, while the mean QRS interval is similar [110 (109–111) ms].

Rare functional variants in *SCN10A*, *CACNA1C*, *PKP2*, *CACNB2*, *KCNH2* and *TRPM4* were found in 1–6% of BrS cases (Table 1). One missense variant in *SCN10A* (NM_006514.2:c.3674T>C, I1225T) was previously reported in a BrS case (4). Only one rare variant (1/167 patients, 0.6% of the BrS population) was identified in each of the *KCND3*, *CACNA2D1*, *HEY2*, *SCN2B* and *SCN3B* genes. Note that the *SCN3B* missense variant (NM_018400.3:c.29T>C, L10P) was previously reported in one unrelated patient with BrS of German, Swedish and Native American descent (15) and is absent from the 1000 Genomes and ESP (NHLBI GO Exome Sequencing Project) European populations. No rare functional variant was identified in *ABCC9*, *SCN1B*, *RANGRF*, *FGF12*, *GPD1L*, *HCN4*, *KCNE1L/KCNE5*, *KCNE3* and *KCNJ8*.

Testing significant enrichment in rare functional variants

A substantial amount of rare functional variation was observed in arrhythmia-susceptibility genes when analysing a group of 167 ethnicity-matched individuals with no history of cardiac arrhythmia on the same targeted sequencing system (Supplementary Material, Table S2). Therefore, a burden test was applied to detect genes with a significant enrichment in rare variation in cases versus controls. *SCN5A* was the only BrS-susceptibility gene showing a significant enrichment associated with BrS with 20.4% of cases, compared with 2.4% of controls, carrying at least one rare functional variant in this gene ($P = 1.4 \times 10^{-7}$, Table 1). To reproduce this observation with an independent control set and evaluate if a set of external data could be used for

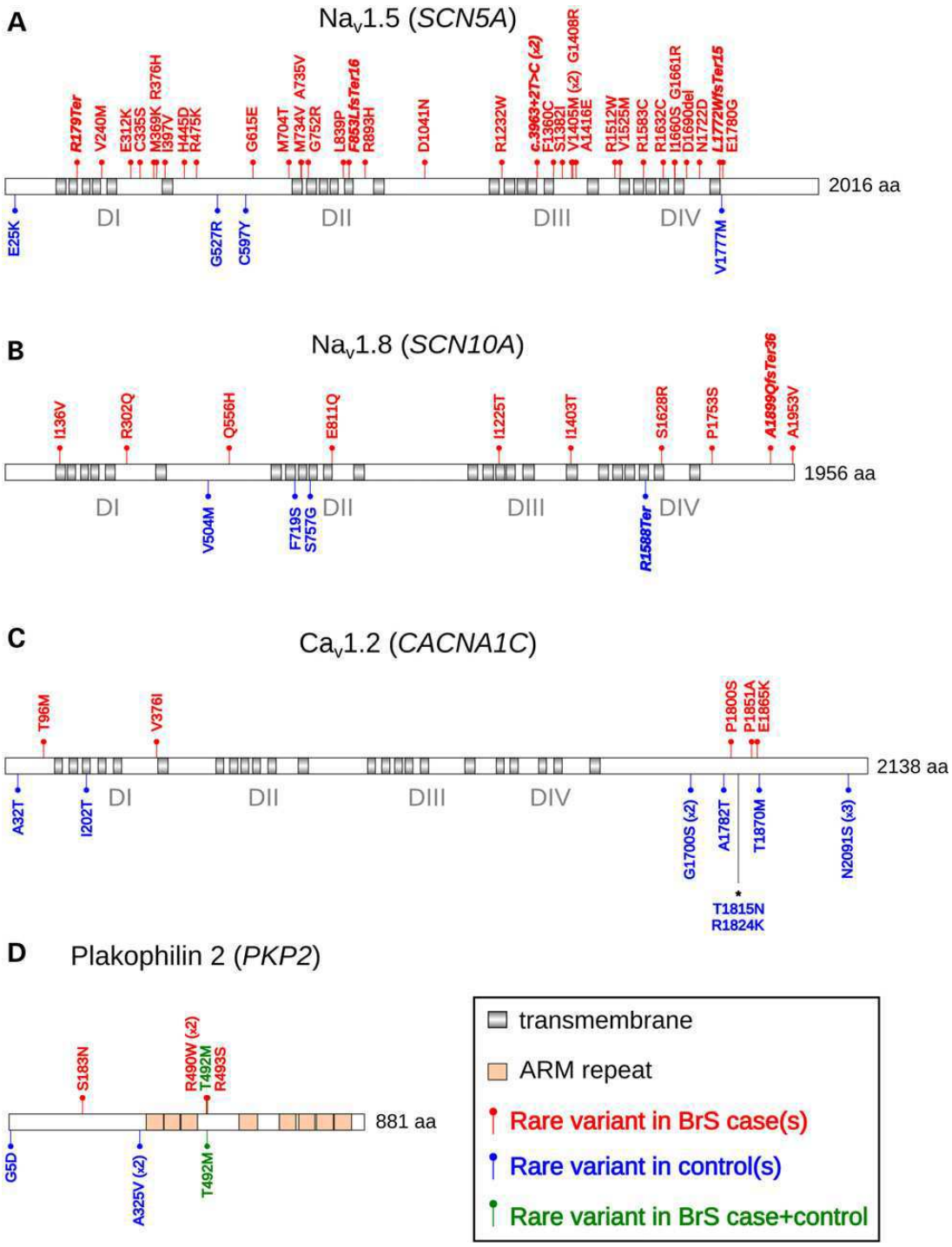


Figure 1. Location of the rare functional variants detected among cases and controls in four previously reported BrS-susceptibility genes: *SCN5A*, *SCN10A*, *CACNA1C* and *PKP2*. Rare functional variants (with an MAF <0.1%) detected in cases (red, above), controls (blue, below) or both cases and controls (green) are represented. If a rare variant was found in more than one case or control, the number of subjects is indicated in brackets. Domain annotations were obtained from UniProt. Non-sense, frameshift and splice site variants (four in *SCN5A* and two in *SCN10A*) are shown in bold/italics. (**A**) *SCN5A* (sodium channel, Na$_v$1.5), NCBI: NP_932173.1, UniProtKB: Q14524. (**B**) *SCN10A* (sodium channel, Na$_v$1.8), NCBI: NP_006505.2, UniProtKB: Q9Y5Y9. (**C**) *CACNA1C* (calcium channel, Ca$_v$1.2), NCBI: NP_000710.5, UniProtKB: Q13936–12. The asterisk points to two rare missense variants located in an alternative exon in isoform NP_001161097.1 (p.Thr1815Asn and p.Arg1824Lys). (**D**) *PKP2* (plakophilin-2), NCBI: NP_004563.2, UniProtKB: Q99959.

future studies, access was requested to the variant calls from 881 whole-exome sequences generated by the UK10K consortium. The significant association of *SCN5A* with BrS was confirmed with this independent control set: 20.4% of cases versus 2.4% of controls ($P = 1.7 \times 10^{-15}$, Table 1). *SCN5A* was again the only gene showing significant enrichment in rare functional variants in cases versus controls (Table 1).

Although not significant, a slight tendency to enrichment in rare functional variants in cases was also observed with both control sets for *SCN10A* (internal: 6 versus 2.4%, $P = 0.170$;

Table 1 Burden tests results for 45 genes linked to cardiac arrhythmias

Gene	BrS cases ($n = 167$)	Internal controls ($n = 167$)	P-value 1	UK10K controls ($n = 881$)	P-value 2
BrS-susceptibility genes					
SCN5A	20.4% (34)	2.4% (4)	1.4×10^{-7}[a]	2.4% (21)	1.7×10^{-15}[a]
SCN10A	6% (10)	2.4% (4)	0.170	3.5% (31)	0.131
CACNA1C	3% (5)	6.6% (11)	0.199	2% (18)	0.395
PKP2	3% (5)	2.4% (4)	1	1.7% (15)	0.348
CACNB2	1.8% (3)	1.2% (2)	1	0.9% (8)	0.396
KCNH2	1.2% (2)	3.6% (6)	0.283	1.6% (14)	1
TRPM4	1.2% (2)	3% (5)	0.448	1.9% (17)	0.754
KCND3	0.6% (1)	1.2% (2)	1	1.6% (14)	0.488
CACNA2D1	0.6% (1)	0.6% (1)	1	3.3% (29)	0.072
HEY2	0.6% (1)	0.6% (1)	1	0.1% (1)	0.293
SCN2B	0.6% (1)	0.6% (1)	1	0.5% (4)	0.581
SCN3B	0.6% (1)	0.6% (1)	1	0.5% (4)	0.581
ABCC9	—	3% (5)	0.061	1.1% (10)	0.379
SCN1B	—	1.8% (3)	0.248	0.3% (3)	1
RANGRF	—	0.6% (1)	1	0.2% (2)	1
FGF12	—	—	—	0.7% (6)	0.597
GPD1L	—	—	—	0.1% (1)	1
HCN4	—	—	—	1.6% (14)	0.144
KCNE1L	—	—	—	1% (9)	0.369
KCNE3	—	—	—	0.1% (1)	1
KCNJ8	—	—	—	0.5% (4)	1
Other susceptibility genes					
AKAP9	6% (10)	4.2% (7)	0.620	7.9% (70)	0.431
ANK2	4.2% (7)	3.6% (6)	1	5% (44)	0.844
RYR2	3.6% (6)	4.8% (8)	0.786	4.9% (43)	0.417[b]
TRDN	3.6% (6)	1.8% (3)	0.502	2.5% (22)	0.431
CASQ2	2.4% (4)	4.2% (7)	0.542	1.7% (15)	0.526
CACNA1D	2.4% (4)	1.8% (3)	1	5.1% (45)	0.161
NUP155	2.4% (4)	1.2% (2)	0.685	1.9% (17)	0.761
KCNQ1	1.8% (3)	—	0.248	1% (9)	0.691[b]
SNTA1	1.2% (2)	1.8% (3)	1	1.2% (11)	1
KCNJ5	1.2% (2)	—	0.498	0.6% (5)	0.310
DPP6	0.6% (1)	2.4% (4)	0.371	0.2% (2)	0.406
KCNA5	0.6% (1)	2.4% (4)	0.371	4.3% (38)	0.014
NKX2-5	0.6% (1)	1.8% (3)	0.623	0.5% (4)	0.581
GJA1	0.6% (1)	1.2% (2)	1	0.2% (2)	0.406
EMD	0.6% (1)	0.6% (1)	1	1.1% (10)	1
KCNE1	0.6% (1)	0.6% (1)	1	0.5% (4)	0.581
KCNJ2	0.6% (1)	0.6% (1)	1	0.6% (5)	1
CAV3	0.6% (1)	—	1	0.6% (5)	1
GATA4	0.6% (1)	—	1	0.2% (2)	0.406
NPPA	0.6% (1)	—	1	0.3% (3)	0.501
SCN4B	—	1.8% (3)	0.248	0.5% (4)	1
LMNA	—	0.6% (1)	1	0.8% (7)	0.605
GJA5	—	0.6% (1)	1	0.2% (2)	1
KCNE2	—	—	—	0.3% (3)	1

The proportion of BrS cases/controls (in %) carrying at least one rare 'functional' variant in each gene is shown. The corresponding number of cases/controls is indicated in brackets.

[a]Significant enrichment in rare 'functional' variants in BrS cases.

[b]One variant presenting two non-reference alleles was excluded for the CAST analysis.

UK10K: 6 versus 3.5%, $P = 0.131$), *PKP2* (internal: 3 versus 2.4%, $P = 1$; UK10K: 3 versus 1.7%, $P = 0.348$) and *CACNB2* (internal: 1.8 versus 1.2%, $P = 1$; UK10K: 1.8 versus 0.9%, $P = 0.396$). The same burden test was applied to 24 additional genes selected on the basis of their reported involvement in monogenic forms of cardiac arrhythmia and conduction defects: no significant enrichment was observed for any of these genes (Table 1). Rare variants were detected in large genes such as *ANK2*, *AKAP9* and *RYR2* in more than 3% of both case and control populations.

Distribution of rare variation across protein sequences

Four genes display rare functional variation in more than 2% of BrS cases: *SCN5A*, *SCN10A*, *CACNA1C* and *PKP2* (Fig. 1). Most variants identified in *SCN5A* among BrS cases alter the DI–DIV transmembrane domains of the $Na_v1.5$ protein. Conversely, three out of four rare variants identified among controls were located in the N-terminal domain or the DI–DII linker region known as harbouring variations with unlikely pathogenicity (16). The distribution of rare variants across *SCN10A* ($Na_v1.8$) seemed less predictive

Table 2. Genotype–phenotype correlations in relation to SCN5A status

	SCN5A positive	*SCN5A* negative	*P*-value
Number of BrS cases	34	133	—
Age	44 ± 14	49 ± 14	0.051
Symptoms	41% (14)	44% (59)	0.888
Male gender	65% (22)	82% (109)	0.051
Baseline BrS ECG pattern	67% (23)	36% (49)	0.002*
Heart rate (bpm)	70 ± 12	72 ± 12	0.260
PR interval (ms)	197 ± 21	174 ± 25	<0.001*
QRS duration (ms)	109 ± 13	101 ± 12	0.005*
QTc interval (ms)	410 ± 31	416 ± 32	0.304

*Significant at $P < 0.05$.

of their potential pathogenicity, three variants carried by control individuals being located in transmembrane domains.

Most variants affecting *CACNA1C* among cases (3/5) as well as controls (6/8) were located within the C-terminal tail of $Ca_v1.2$ (Fig. 1C), indicating that genetic variation may be extensive and benign in this region. Finally, one particular *PKP2* interval coding for four amino acids was the site of three rare variants detected among BrS patients: NM_004572.3:c.1468C>T (R490W, 2 cases), NM_004572.3:c.1475C>T (T492M, 1 case, 1 control) and NM_004572.3:c.1479G>C (R493S, 1 case). This interval, albeit exhibiting low evolutionary conservation, might be a preferential site for rare variants causally related to BrS (Supplementary Material, Table S2).

Discussion

Is *SCN5A* the only major BrS-susceptibility gene?

Three genes are currently considered as major susceptibility genes for BrS, possibly explaining 50% of cases: *SCN5A*, *SCN10A* and *CACNA1C* (17). However, given that the coding portions of these three genes are the largest among BrS-susceptibility genes, one could expect to find higher levels of variation affecting the corresponding transcripts. To take into account such biological parameters, burden tests were performed by comparing rare genetic variants detected in BrS cases with those identified in two independent control populations. In view of these burden tests, *SCN5A* appears to be the only gene—among the genes previously reported as involved in cardiac arrhythmia—to contribute significantly to the occurrence of BrS (~20% of BrS cases) in populations of European origin.

The prevalence of mutations/variants in BrS populations can vary greatly between studies in the absence of clear guidelines for the frequency definition of a 'rare' variant. In this study, we considered variants as 'rare' if the frequency observed in ethnically matched individuals from public databases was <0.1%. When considering only variants with an MAF below 0.1%, the proportion of *SCN10A* carriers among BrS cases reported by Hu *et al.* falls to 7.3% (versus 16.7% reported with a 0.5% threshold), a figure similar to the frequency of 6% reported in the present study. While 3.6% of BrS cases carried a rare functional variant in *CACNA1C*, rare variants were also identified in 2–6.6% of control individuals, and thus, no significant enrichment was observed among cases. In addition, rare genetic variation is extensive in the C-terminal region of $Ca_v1.2$, in both case and control groups. While this domain was previously reported as harbouring a high proportion of rare variants in BrS cases [50–66%, (18,19)], the observations reported here confirm that *CACNA1C* mutations in isolated BrS are unlikely and may be restricted to BrS patients also presenting with a short QT interval (9).

How to interpret rare genetic variation?

The burden test results demonstrate the usefulness of considering the extent of rare genetic variation in the general population when testing the contribution of rare alleles in the susceptibility to rare diseases. As discussed above, this strategy is particularly relevant for large genes or transcripts, such as *SCN5A*, *SCN10A*, *CACNA1C*, *PKP2*, *KCNH2* and *TRPM4*. Furthermore, several genes included in this study harbour rare variants in less than 1% of the control population (*HEY2*, *SCN2B*, *SCN3B*, *RANGRF*, *FGF12*, *GPD1L*, *KCNE3*, *KCNJ8*), indicating that although no statistical enrichment in rare variation could be observed for these genes, individual rare variants observed in patients with BrS might still be causally related to the cardiac electrical anomaly.

One common strategy to evaluate the potential contribution of any rare or private coding variant to disease is to interrogate functional prediction algorithms. However, a substantial proportion of the rare variants detected in controls are predicted to be deleterious or damaging with these tools (Supplementary Material, Table S2). Conversely, a subset of rare variants detected in BrS cases, although predicted as benign with the same tools, may impact the function of the encoded protein. For example, one variant in *SCN3B*, carried by one patient in our study, was previously reported in another unrelated case with BrS and is absent from European populations in public variation databases. Although this variant is predicted to be tolerated or benign by SIFT and PolyPhen-2 (20,21), it leads to a reduction in the sodium current density (15). This illustrates that prediction tools have limited power to predict the causality of rare missense variants and should be used with great caution, especially if used to guide clinical decision-making. In the absence of experimental data regarding the functional effect of one rare genetic variant, segregation analysis in relatives remains the strategy of choice to further estimate its causal relationship with disease.

Which insights into molecular diagnosis?

The current study indicates that *SCN5A* is the only major susceptibility gene for BrS identified so far, with rare coding variants in this gene accounting for ~20% of cases. Although additional genes such as *SCN10A*, *PKP2* and *CACNB2* may contribute to BrS susceptibility in small subsets of cases, the aetiology of this cardiac electrical anomaly remains largely unknown. Rare genetic variation identified in arrhythmia-susceptibility genes among patients with BrS should thus be interpreted with precaution, as the only criterion available so far to evaluate the arrhythmic risk is the presence of the Brugada ECG pattern. Genetic testing in BrS should thus be restricted to *SCN5A* in clinical diagnostic setting, as genetic counselling is unlikely to relieve patient distress under current knowledge on other reported susceptibility genes.

The authors of this study have recently reported that common genetic alleles at the *SCN5A-SCN10A* and *HEY2* loci are associated with BrS, with an unexpectedly high cumulative effect of the risk alleles on disease susceptibility (22). Further investigations addressing the combined effect of rare and common genetic variants on the BrS-specific ECG pattern are now required to shed light on the genetic aetiology of this complex arrhythmia disorder and better address risk stratification based on genetic information.

Limitations

NGS technologies allow high-throughput genetic screening in the context of molecular diagnosis, but some relevant variation may have been missed due to lower coverage depth for particular exon sequences. Controls have not undergone any pharmacological test to unmask a putative BrS-specific ECG pattern, and therefore, it is possible that the control sets are not devoid of such cases. We did not consider the possibility that rare alleles at risk and protective may occur in the same gene and therefore did not formally apply tests suited to this model.

Conclusions

This study demonstrates that, among the arrhythmia-susceptibility genes reported so far in the literature, only *SCN5A* accounts for a significant proportion of cases with BrS, with rare coding variants altering this gene in ~20% of cases versus <3% of individuals from a reference population. For every other gene, including *SCN10A* and *CACNA1C*, no enrichment of rare coding variation was observed. Extreme caution should thus be taken to avoid false-positive reports of causality in the context of genetic counselling, as rare coding variation with potential functional effect in arrhythmia-susceptibility genes is extensive among the general population.

Materials and Methods

This study was carried out in accordance with the ethical guidelines of the Declaration of Helsinki and with French guidelines for clinical and genetic research. Informed written consent was obtained from each individual who agreed to participate in the genetic study.

A full description of the methods is available as Supplementary Material.

Patient and control populations

A group of 167 index cases of European origin and diagnosed with BrS on the basis of the second consensus conference (1) was included in the study. Seventy-three cases (44%) were symptomatic and ECG examination revealed a baseline BrS pattern for 43% of the patients. The mean age at diagnosis was 48-year old (range: 38–58). The mean heart rate was 72 ± 12, with PR, QRS and QTc intervals of 179 ± 26, 104 ± 14 and 415 ± 32 ms, respectively. Two physicians reviewed the clinical and ECG data independently. The first control set ('internal') comprised 167 individuals of European origin aged over 65-year old and showing no history of cardiac arrhythmia. The second control set ('UK10K') included 881 whole-exome sequences released by the UK10K consortium (http://www.uk10k.org).

Targeted sequencing

We developed a custom design based on the HaloPlex™ technology (Agilent Technologies) to perform high-throughput sequencing of the coding regions of 45 genes previously linked to cardiac arrhythmias or conduction defects and/or sudden cardiac death, including 21 genes linked to BrS. Targeted coding regions (exons) ± 10 bp correspond to 141 kb of genomic sequence. Multiplex amplification and library preparation were performed following the manufacturer's instruction. Libraries were pooled to an equimolar concentration, then pools were diluted to a 4 pM final concentration before proceeding to 100 bp paired-end Illumina sequencing on MiSeq for the validation study (39 patients) and on HiSeq for the main study (167 cases/167 controls).

Variant calling

Raw sequence reads were aligned to the reference genome GRCh37 using BWA-MEM (version 0.7.5a) after removing adapter sequences with Cutadapt v1.2. GATK was used for indel realignment and base recalibration. Variants were called and considered for further analyses if found by both GATK UnifiedGenotyper (version 2.8) and Samtools mpileup (version 0.1.19), with a minimum quality score of 25. Variants were defined as rare if the MAF was <0.1% in the European population according to the 1000 genomes phase 1 data (379 individuals) and the NHLBI GO Exome Sequencing Project (ESP) data (4300 individuals). The potential pathogenicity of variants was determined using the Variant Effect Predictor (Ensembl) and filtering was performed using Knime4Bio (23).

Validation of the targeted sequencing system

Thirty-nine patients for whom genetic variation had previously been identified in arrhythmia-susceptibility genes were recruited to evaluate the performance of the targeted sequencing system. After DNA capture, sequencing on the Illumina MiSeq system resulted in a mean coverage depth of 141× per sample, with 94% of the targeted regions covered at least 10 times (Supplementary Material, Table S1). The variant-calling pipeline allowed the automatic detection of 67 out of the 68 (98.5%) genetic variants previously identified by capillary sequencing. The only undetected variant is a substitution located at a chromosomal position covered <10 times. In order to limit the risk of missing genetic variants, subsequent sequencing was performed on the Illumina HiSeq system, which enables higher coverage at reasonable cost, and only samples with a mean coverage depth above 100× were included in further analyses.

Burden tests

Burden tests were carried out to compare the proportion of cases and controls carrying at least one rare variant with potential functional consequence in a given gene. Two tests were performed for each gene: (i) 167 BrS cases versus 167 internal controls and (ii) 167 BrS cases versus 881 UK10K controls. First, we performed a Fisher's exact test for each variant, comparing allele counts in cases versus controls, to exclude variants with a *P*-value of <0.01, which would be potential false positives. We also excluded variants found in 5% or more of cases and/or controls but absent from European populations from the 1000 Genomes Project and/or ESP. For the second burden test (using UK10K data), only variants located in UK10K exome capture regions ± 100 bp were considered. Finally, the Cohort Allelic Sums Test (CAST) (24), based on a Fisher's exact test, was applied for each gene. Any *P*-value lower than 0.05 was considered as suggestive of association.

Supplementary material

Supplementary Material is available at *HMG* online.

Acknowledgements

We would like to thank the French clinical network against inherited cardiac arrhythmias, in particular the Hospitals of Tours, Brest, la Roche-sur-Yon, La Rochelle, Bayonne and la Réunion.

We are also grateful to the members of the Genomics and Bioinformatics Core Facility of Nantes (Biogenouest) for their technical expertise. This study makes use of data generated by the UK10K Consortium as controls. A full list of the investigators who contributed to the generation of these data is available from http://www.UK10K.org. The UK10K project was funded by the Wellcome Trust under award WT091310.

Conflict of Interest statement. None declared.

Funding

This work was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM, ATIP-Avenir program to R.R.); the French Regional Council of Pays-de-la-Loire; and the Fondation pour la Recherche Médicale (FRM grant DEQ20140329545).

References

1. Antzelevitch, C., Brugada, P., Borggrefe, M., Brugada, J., Brugada, R., Corrado, D., Gussak, I., LeMarec, H., Nademanee, K., Perez Riera, A.R. *et al.* (2005) Brugada syndrome: report of the second consensus conference: endorsed by the Heart Rhythm Society and the European Heart Rhythm Association. *Circulation*, **111**, 659–670.
2. Wilde, A.A.M. and Behr, E.R. (2013) Genetic testing for inherited cardiac disease. *Nat. Rev. Cardiol.*, **10**, 571–583.
3. Riuró, H., Beltran-Alvarez, P., Tarradas, A., Selga, E., Campuzano, O., Vergés, M., Pagans, S., Iglesias, A., Brugada, J., Brugada, P. *et al.* (2013) A missense mutation in the sodium channel β2 subunit reveals SCN2B as a new candidate gene for Brugada syndrome. *Hum. Mutat.*, **34**, 961–966.
4. Hu, D., Barajas-Martínez, H., Pfeiffer, R., Dezi, F., Pfeiffer, J., Buch, T., Betzenhauser, M.J., Belardinelli, L., Kahlig, K.M., Rajamani, S. *et al.* (2014) Mutations in SCN10A are responsible for a large fraction of cases of Brugada Syndrome. *J. Am. Coll. Cardiol.*, **64**, 66–79.
5. Wang, Q., Ohno, S., Ding, W.-G., Fukuyama, M., Miyamoto, A., Itoh, H., Makiyama, T., Wu, J., Bai, J., Hasegawa, K. *et al.* (2014) Gain-of-function KCNH2 mutations in patients with Brugada syndrome. *J. Cardiovasc. Electrophysiol.*, **25**, 522–530.
6. Hu, D., Barajas-Martínez, H., Terzic, A., Park, S., Pfeiffer, R., Burashnikov, E., Wu, Y., Borggrefe, M., Veltmann, C., Schimpf, R. *et al.* (2014) ABCC9 is a novel Brugada and early repolarization syndrome susceptibility gene. *Int. J. Cardiol.*, **171**, 431–442.
7. Cerrone, M., Lin, X., Zhang, M., Agullo-Pascual, E., Pfenniger, A., Chkourko Gusky, H., Novelli, V., Kim, C., Tirasawadichai, T., Judge, D.P. *et al.* (2014) Missense mutations in plakophilin-2 cause sodium current deficit and associate with a brugada syndrome phenotype. *Circulation*, **129**, 1092–1103.
8. Hennessey, J.A., Marcou, C.A., Wang, C., Wei, E.Q., Wang, C., Tester, D.J., Torchio, M., Dagradi, F., Crotti, L., Schwartz, P.J. *et al.* (2013) FGF12 is a candidate Brugada syndrome locus. *Heart Rhythm*, **10**, 1886–1894.
9. Crotti, L., Marcou, C.A., Tester, D.J., Castelletti, S., Giudicessi, J. R., Torchio, M., Medeiros-Domingo, A., Simone, S., Will, M.L., Dagradi, F. *et al.* (2012) Spectrum and prevalence of mutations involving BrS1- through BrS12-susceptibility genes in a cohort of unrelated patients referred for Brugada syndrome genetic testing: implications for genetic testing. *J. Am. Coll. Cardiol.*, **60**, 1410–1418.
10. MacArthur, D.G., Manolio, T.A., Dimmock, D.P., Rehm, H.L., Shendure, J., Abecasis, G.R., Adams, D.R., Altman, R.B., Antonarakis, S.E., Ashley, E.A. *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature*, **508**, 469–476.
11. Kapplinger, J.D., Tester, D.J., Alders, M., Benito, B., Berthet, M., Brugada, J., Brugada, P., Fressart, V., Guerchicoff, A., Harris-Kerr, C. *et al.* (2010) An international compendium of mutations in the SCN5A-encoded cardiac sodium channel in patients referred for Brugada syndrome genetic testing. *Heart Rhythm*, **7**, 33–46.
12. Maury, P., Rollin, A., Sacher, F., Gourraud, J.-B., Raczka, F., Pasquié, J.-L., Duparc, A., Mondoly, P., Cardin, C., Delay, M. *et al.* (2013) Prevalence and prognostic role of various conduction disturbances in patients with the Brugada syndrome. *Am. J. Cardiol.*, **112**, 1384–1389.
13. Probst, V., Allouis, M., Sacher, F., Pattier, S., Babuty, D., Mabo, P., Mansourati, J., Victor, J., Nguyen, J.-M., Schott, J.-J. *et al.* (2006) Progressive cardiac conduction defect is the prevailing phenotype in carriers of a Brugada syndrome SCN5A mutation. *J. Cardiovasc. Electrophysiol.*, **17**, 270–275.
14. Smits, J.P.P., Eckardt, L., Probst, V., Bezzina, C.R., Schott, J.J., Remme, C.A., Haverkamp, W., Breithardt, G., Escande, D., Schulze-Bahr, E. *et al.* (2002) Genotype–phenotype relationship in Brugada syndrome: electrocardiographic features differentiate SCN5A-related patients from non-SCN5A-related patients. *J. Am. Coll. Cardiol.*, **40**, 350–356.
15. Hu, D., Barajas-Martinez, H., Burashnikov, E., Springer, M., Wu, Y., Varro, A., Pfeiffer, R., Koopmann, T.T., Cordeiro, J.M., Guerchicoff, A. *et al.* (2009) A mutation in the beta 3 subunit of the cardiac sodium channel associated with Brugada ECG phenotype. *Circ. Cardiovasc. Genet.*, **2**, 270–278.
16. Kapa, S., Tester, D.J., Salisbury, B.A., Harris-Kerr, C., Pungliya, M.S., Alders, M., Wilde, A.A.M. and Ackerman, M.J. (2009) Genetic testing for long-QT syndrome: distinguishing pathogenic mutations from benign variants. *Circulation*, **120**, 1752–1760.
17. Shimizu, W. (2014) Importance of clinical analysis in the new era of molecular genetic screening. *J. Am. Coll. Cardiol.*, **64**, 80–82.
18. Burashnikov, E., Pfeiffer, R., Barajas-Martinez, H., Delpón, E., Hu, D., Desai, M., Borggrefe, M., Häissaguerre, M., Kanter, R., Pollevick, G.D. *et al.* (2010) Mutations in the cardiac L-type calcium channel associated with inherited J-wave syndromes and sudden cardiac death. *Heart Rhythm*, **7**, 1872–1882.
19. Fukuyama, M., Ohno, S., Wang, Q., Kimura, H., Makiyama, T., Itoh, H., Ito, M. and Horie, M. (2013) L-type calcium channel mutations in Japanese patients with inherited arrhythmias. *Circ. J.*, **77**, 1799–1806.
20. Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
21. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
22. Bezzina, C.R., Barc, J., Mizusawa, Y., Remme, C.A., Gourraud, J.-B., Simonet, F., Verkerk, A.O., Schwartz, P.J., Crotti, L., Dagradi, F. *et al.* (2013) Common variants at SCN5A–SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. *Nat. Genet.*, **45**, 1044–1049.
23. Lindenbaum, P., Le Scouarnec, S., Portero, V. and Redon, R. (2011) Knime4Bio: a set of custom nodes for the interpretation of next-generation sequencing data with KNIME. *Bioinformatics*, **27**, 3200–3201.
24. 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). *Mutat. Res.*, **615**, 28–56.

Annexe II
LE CALCUL DU F_{ST} SELON LA MÉTHODE DE WEIR ET COCKERHAM (1984)

Le calcul du F_{ST} selon la méthode de Weir et Cockerham (1984)

Dans le calcul du F_{ST}, trois composantes de la variance des fréquences alléliques sont estimées :

- a, composante inter-population de la variance des fréquences alléliques
- b, composante, entre individus à l'intérieur de chaque population, de la variance des fréquences alléliques
- c, composante de la variance au sein des individus.

Le F_{ST} est la proportion de la variance totale liée à la stratification de la population.

$$F_{ST} = \frac{a}{a+b+c}$$

Soient les populations $1, \dots, i, \dots, r$, avec n_i le nombre d'individus dans une population i. Si on considère un variant à deux allèles A et B, on note h_i la proportion de personnes hétérozygotes AB, et p_i la fréquence de l'allèle A dans la population i.

$$a = \frac{\bar{n}}{n_c}\left(s^2 - \frac{1}{\bar{n}-1} * \left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{1}{4}\bar{h}\right]\right)$$

$$b = \frac{\bar{n}}{\bar{n}-1}\left[\bar{p}(1-\bar{p}) - \frac{r-1}{r}s^2 - \frac{2\bar{n}-1}{4\bar{n}}\bar{h}\right]$$

$$c = \frac{1}{2}\bar{h}$$

avec

- $\bar{n} = \frac{1}{r}\sum_{i=1}^{r} n_i$ le nombre moyen d'individus sur l'ensemble des populations
- $n_c = \frac{\left(r\bar{n} - \sum_{i=1}^{r}\frac{n_i^2}{r\bar{n}}\right)}{r-1}$ la variance du nombre d'individus dans les populations
- $\bar{p} = \sum_{i=1}^{r}\frac{n_i p_i}{r\bar{n}}$ la fréquence moyenne de l'allèle A sur l'ensemble des populations

- $s^2 = \sum_{i=1}^{r} \frac{n_i (p_i - \bar{p})^2}{(r-1)\bar{n}}$ la variance de la fréquence de l'allèle A sur l'ensemble des populations
- $\bar{h} = \sum_{i=1}^{r} \frac{n_i h_i}{r\bar{n}}$ la fréquence moyenne du génotype hétérozygote

La valeur du F_{ST} pour un ensemble de variants $j \in \{1, \ldots, P\}$est alors :

$$F_{ST} = \frac{\sum_{j=1}^{P} a_j}{\sum_{j=1}^{P} a_j + b_j + c_j}$$

Annexe III
TABLEAUX D'ERREURS DE TYPE I ET DE PUISSANCES POUR LES SIMULATIONS BASÉES SUR LES TRAVAUX DE BASU ET PAN (2011)

Tableau S 1. Erreurs de type I au seuil α=5% des tests d'association pour variants rares.

Test	CAST	Sum	wSum	aSum	VT	CALPHA	SKAT	SKATO
Erreur de type I	0,035	0,041	0,05	0,052	0,053	0,046	0,042	0,041

Test	KBAC	ADA	DBM	CLUSTER	KERNEL	BOMP	PODKAT	DoEstRare
Erreur de type I	0,052	0,053	0,045	0,059	0,051	0,051	0,047	0,053

Tableau S 2. Puissances au seuil α=5% des tests d'association pour le scénario 1.

	OR2_0	**OR2_4**	**OR2_8**	**OR2_16**	**OR2_32**
CAST	0,945	0,842	0,744	0,559	0,367
Sum	0,96	0,861	0,797	0,616	0,455
wSum	0,976	0,9	0,853	0,704	0,569
aSum	0,935	0,839	0,792	0,637	0,503
VT	0,918	0,764	0,665	0,422	0,255
CALPHA	0,762	0,717	0,7	0,598	0,482
SKAT	0,771	0,714	0,692	0,581	0,469
SKATO	0,941	0,856	0,821	0,671	0,549
KBAC	0,968	0,91	0,886	0,767	0,611
ADA	0,392	0,631	0,616	0,48	0,327
DBM	0,903	0,739	0,632	0,412	0,247
CLUSTER	0,424	0,695	0,688	0,53	0,372
KERNEL	0,091	0,224	0,313	0,364	0,345
BOMP	0,542	0,51	0,469	0,406	0,332
PODKAT	0,859	0,702	0,595	0,442	0,297
DoEstRare	0,868	0,85	0,861	0,789	0,69

Tableau S 3. Puissances au seuil α=5% des tests d'association pour le scénario 2.

	ORs_0	ORs_4	ORs_8	ORs_16	ORs_32
CAST	0,618	0,481	0,415	0,298	0,185
Sum	0,627	0,502	0,437	0,311	0,223
wSum	0,729	0,607	0,55	0,442	0,335
aSum	0,839	0,741	0,671	0,514	0,422
VT	0,536	0,397	0,323	0,187	0,138
CALPHA	0,871	0,837	0,823	0,721	0,643
SKAT	0,875	0,833	0,821	0,708	0,634
SKATO	0,853	0,79	0,766	0,653	0,567
KBAC	0,888	0,816	0,762	0,638	0,482
ADA	0,677	0,536	0,511	0,409	0,269
DBM	0,64	0,493	0,433	0,285	0,177
CLUSTER	0,71	0,551	0,493	0,367	0,242
KERNEL	0,67	0,696	0,706	0,591	0,509
BOMP	0,784	0,722	0,712	0,593	0,499
PODKAT	0,81	0,688	0,614	0,502	0,353
DoEstRare	0,917	0,88	0,881	0,792	0,718

Annexe IV
INTERPRÉTATION DE L'AFM DES PROFILS DE PUISSANCE DES TESTS POUR LES SIMULATIONS BASÉES SUR LES TRAVAUX DE BASU ET PAN (2011)

Les valeurs propres des analyses séparées sont très élevées pour la première composante principale (premier facteur). Cependant on peut noter que pour le scénario 2, le deuxième facteur explique une part de l'inertie non négligeable. Ceci montre que pour le scénario 1, deux sources de variabilité sont observées pour la comparaison des puissances des tests pour les différents nombres de variants non causaux ajoutés. Tandis que pour le scénario 2, il y a surtout une source de variabilité, quelque soit le nombre de variants non causaux. Il est alors important de pondérer chacun des groupes de variables selon l'inertie du premier facteur pour faire une analyse globale (AFM).

Tableau S 4. Valeurs propres et pourcentages d'inertie des analyses partielles.

	Scénario 1 (OR2)			**Scénario 2 (ORs)**		
	Valeur propre	**Pourcentage d'inertie**	**Pourcentage d'inertie cumulé**	**Valeur propre**	**Pourcentage d'inertie**	**Pourcentage d'inertie cumulé**
Dim 1	4.02	80.44	80.44	4.83	96.69	96.69
Dim 2	0.84	16.89	97.34	0.15	2.91	99.60
Dim 3	0.13	2.57	99.90	0.01	0.22	99.83
Dim 4	0.00	0.06	99.96	0.01	0.14	99.97
Dim 5	0.00	0.04	100.00	0.00	0.03	100.00

La corrélation entre les facteurs des analyses séparées ne sont pas très élevées, montrant que les sources de variabilités entre les tests ne sont pas les mêmes selon les scénarios OR2 et ORs. Cependant le premier facteur du scénario OR2 semble plus corrélé avec le deuxième facteur du scénario ORs, et inversement.

Tableau S 5. Corrélations entre facteurs partiels.

	Dim.1.OR2	Dim.2.OR2	Dim.1.ORs	Dim.2.ORs
Dim.1.OR2	1.00			
Dim.2.OR2	0.00	1.00		
Dim.1.ORs	0.31	0.68	1.00	
Dim.2.ORs	0.52	-0.36	0.00	1.00

En se basant sur les indices de liaison Lg entre les deux groupes de variables (Tableau S 6), les profils des tests selon les deux groupes OR2 et ORs ne se ressemblent pas.

Tableau S 6. Coefficients Lg de liaison entre groupes.

	OR2	ORs	MFA
OR2	1.05		
ORs	0.21	1.00	
MFA	0.92	0.89	1.33

Pour l'interprétation des résultats de l'analyse globale, nous nous sommes focalisés sur les deux premiers axes de l'AFM qui expliquent 93.77% de l'inertie totale des données.

Tableau S 7.Valeurs propres et pourcentages d'inertie de l'analyse globale.

	AFM		
	Valeur propre	**Pourcentage d'inertie**	**Pourcentage d'inertie cumulé**
Dim 1	1.36	59.89	59.89
Dim 2	0.77	33.88	93.77
Dim 3	0.10	4.46	98.23
Dim 4	0.03	1.28	99.51
Dim 5	0.01	0.36	99.8

D'après les corrélations entre les variables canoniques et les facteurs de l'analyse globale (Tableau S 8), le premier facteur est commun aux deux groupes de variables. Le deuxième facteur est surtout corrélé à la variable canonique pour le groupe OR2. Le premier axe correspond à une dimension d'inertie importante pour chacun des deux groupes tandis que le deuxième axe l'est que pour le premier groupe OR2 (Figure S 1, Tableau S 9).

Pour aller plus loin, on peut voir les corrélations entre les facteurs partiels et les facteurs de l'analyse globale dans la Figure S 2. Les premiers facteurs des analyses partielles sur chacun des groupes semblent corrélés au premier facteur de l'analyse globale, bien que ceux-ci soient non corrélés entre eux (ce qui rejoint le commentaire effectué plus haut sur les corrélations entre les facteurs partiels). Les deux premiers facteurs partiels pour l'analyse du groupe OR2 sont corrélés au deuxième facteur de l'analyse globale.

Tableau S 8. Corrélations entre les variables canoniques et les facteurs de l'analyse globale.

	Dim.1	**Dim.2**
OR2	0.82	0.77
ORs	0.86	0.52

Tableau S 9. Coordonnées et aides à l'interprétation des groupes.

	coord		**contrib**		**cos2**	
	Dim.1	**Dim.2**	**Dim.1**	**Dim.2**	**Dim.1**	**Dim.2**
OR2	0.62	0.51	45.8	66.21	0.37	0.25
ORs	0.74	0.26	54.2	33.79	0.55	0.07

Interprétation du graphe des individus

En se basant sur la Figure S 3 et la Figure S 4, on peut constater que le premier axe de l'AFM permet de distinguer des tests qui ont globalement une puissance élevée pour chacun des scénarios envisagés (effets OR2 ou ORs et nombres de variants causaux allant de 0 à 32), à des tests qui présentent une puissance faible. DoEstRare, SKAT-O, MiST et KBAC présentent des puissances élevées pour chacun des scénarios envisagés. Les tests KERNEL, DBM, CLUSTER présentent des puissances globalement faibles. Le deuxième axe de l'AFM permet de distinguer des tests qui présentent des puissances élevées dans les scénarios où tous les variants sont à risque avec les mêmes effets. C'est le cas des tests *burden* VT, CAST, Sum, wSum. Cependant les tests KERNEL et BOMP présentent des puissances faibles pour ces scénarios.

Dans le graphe des individus (Figure S 3), sont aussi tracés les individus partiels pour les trois tests DoEstRare, BOMP et KERNEL présentant la plus forte inertie intra, c'est-à-dire s'écartant de l'individu moyen. DoEstRare présente une puissance élevée, quelque soit le

nombre de variants non causaux, dans le cadre de l'analyse du groupe de variables pour le scénario OR2 avec tous les variants causaux à risque. C'est également le cas pour l'analyse du groupe de variables pour le scénario ORs. Concernant les tests KERNEL et BOMP, ils présentent une puissance faible pour les variables du groupe OR2, mais présentent une puissance ni faible ni élevée pour les variables du groupe ORs.

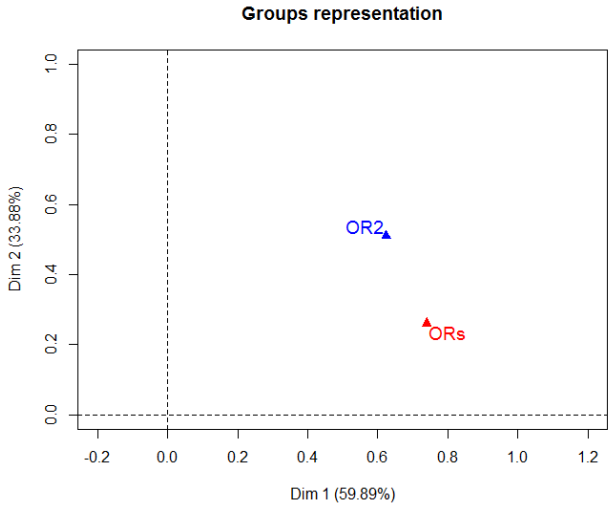


Figure S 1. Représentation des groupes.

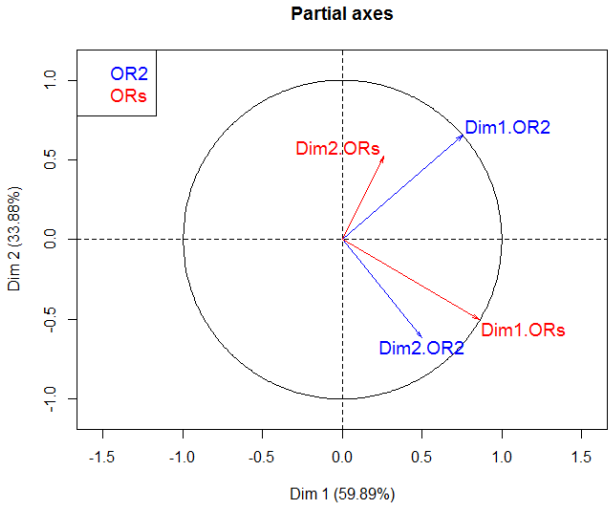


Figure S 2. Corrélation entre les facteurs partiels et les facteurs de l'analyse globale.

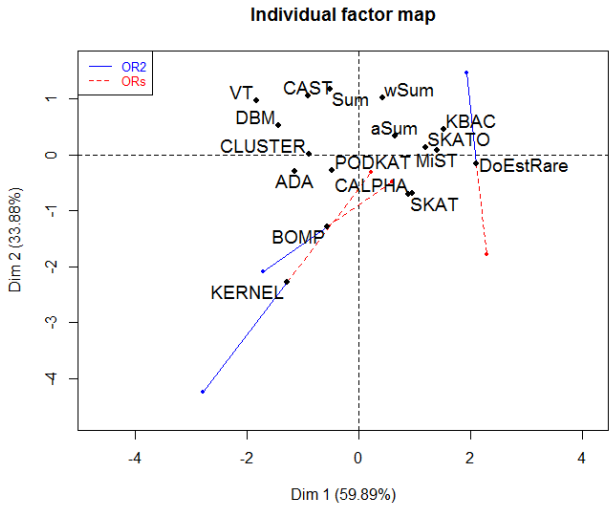


Figure S 3. Graphe des individus.

Sont tracés les individus partiels pour les trois tests KERNEL, BOMP et DoEstRare présentant les inerties intra les plus élevées.

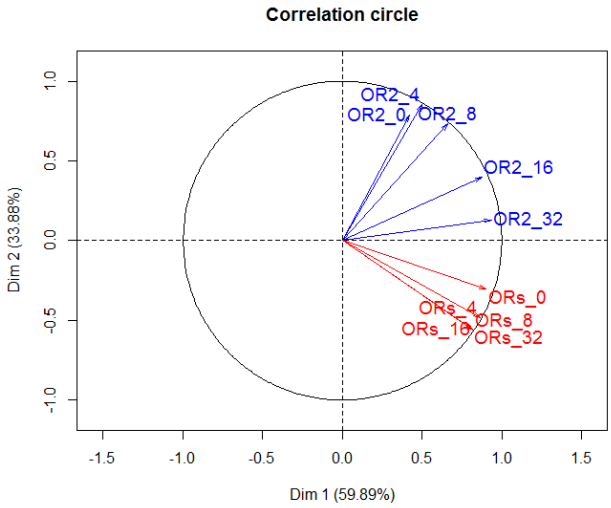


Figure S 4. Corrélations entre les variables et les facteurs de l'analyse globale.

Annexe V
TABLEAUX D'ERREURS DE TYPE I ET DE PUISSANCES POUR LES SIMULATIONS DE REGROUPEMENTS DE VARIANTS À RISQUE

Tableau S 10. Erreurs de type I au seuil α=5% des tests d'association.

CAST	Sum	wSum	aSum	VT_score	CALPHA	SKAT
0,0439	0,0445	0,0516	0,0518	0,0477	0,0464	0,0444

SKATO	MiST	KBAC	ADA	Fier	CLUSTER	KERNEL
0,0478	0,0467	0,0518	0,0508	0,0484	0,0536	0,0475

PODKAT	BOMP	DoEstRare
0,0471	0,0504	0,0515

Tableau S 11. Puissances au seuil α=5% des tests d'association pour le scénario 1 : pas de cluster de variants rares à risque.

	P5	P10	P15	P20
CAST	0,178	0,395	0,61	0,792
Sum	0,218	0,46	0,649	0,788
wSum	0,292	0,538	0,733	0,885
aSum	0,246	0,528	0,705	0,829
VT	0,173	0,379	0,563	0,732
CALPHA	0,36	0,636	0,791	0,888
SKAT	0,35	0,632	0,798	0,892
SKATO	0,328	0,625	0,793	0,895
MiST	0,361	0,659	0,815	0,929
KBAC	0,328	0,62	0,822	0,941
ADA	0,189	0,391	0,607	0,763
DBM	0,165	0,352	0,546	0,726
CLUSTER	0,177	0,363	0,583	0,728
KERNEL	0,286	0,526	0,682	0,761
PODKAT	0,25	0,508	0,663	0,8
BOMP	0,321	0,62	0,795	0,914
DoEstRare	0,369	0,657	0,828	0,92

Tableau S 12. Puissances au seuil α=5% des tests d'association pour le scénario 2 : présence d'un cluster de variants rares à risque.

	P5	P10	P15	P20
CAST	0,163	0,411	0,638	0,795
Sum	0,203	0,447	0,669	0,812
wSum	0,254	0,523	0,75	0,864
aSum	0,253	0,526	0,724	0,832
VT	0,15	0,384	0,584	0,74
CALPHA	0,348	0,639	0,792	0,89
SKAT	0,335	0,633	0,797	0,887
SKATO	0,317	0,612	0,799	0,904
MiST	0,347	0,655	0,831	0,923
KBAC	0,33	0,64	0,83	0,927
ADA	0,167	0,375	0,592	0,739
DBM	0,155	0,353	0,57	0,716
CLUSTER	0,181	0,413	0,635	0,783
KERNEL	0,313	0,592	0,768	0,868
PODKAT	0,267	0,626	0,831	0,931
BOMP	0,333	0,619	0,798	0,912
DoEstRare	0,378	0,707	0,865	0,944

Tableau S 13. Puissances au seuil α=5% des tests d'association pour le scénario 2 : présence de deux clusters de variants rares à risque.

	P5	P10	P15	P20
CAST	0,154	0,396	0,618	0,793
Sum	0,204	0,449	0,651	0,812
wSum	0,254	0,538	0,732	0,88
aSum	0,256	0,51	0,707	0,839
VT	0,165	0,374	0,588	0,758
CALPHA	0,361	0,623	0,78	0,876
SKAT	0,352	0,623	0,782	0,881
SKATO	0,339	0,601	0,775	0,89
MiST	0,351	0,656	0,816	0,92
KBAC	0,324	0,65	0,823	0,931
ADA	0,196	0,394	0,593	0,736
DBM	0,192	0,394	0,569	0,744
CLUSTER	0,195	0,393	0,606	0,782
KERNEL	0,317	0,566	0,714	0,83
PODKAT	0,28	0,591	0,787	0,909
BOMP	0,341	0,624	0,795	0,906
DoEstRare	0,387	0,689	0,841	0,933

Annexe VI
INFORMATIONS SUPPLÉMENTAIRES SUR L'ANALYSE DES DONNÉES EOAD

Tableau S 14. Nombre de cas EOAD et de témoins FREX selon le design de capture (après filtre qualité)

Design de capture	Nombre de témoins	Nombre de cas
Agilent SureSelect Human All Exons V1	0	11
Agilent SureSelect Human All Exons V2	0	2
Agilent SureSelect Human All Exons V3	10	0
Agilent SureSelect Human All Exons V4	12	31
Agilent SureSelect Human All Exons V4UTR	6	9
Agilent SureSelect Human All Exons V5	2	431
Agilent SureSelect Human All Exons V5UTR	553	0
Total	**583**	**494**

Annexe VII
ACP SUR LES RÉSULTATS DE SIGNIFICATIVITÉ POUR LES DONNÉES BRS

Dimensions images : 1325 698

L'ACP normée est effectuée sur les résultats d'association de 58 gènes candidats pour 12 tests considérés en actif et 4 tests mis en illustratif. Les résultats d'association sont sous la forme de -log10(p-value).
Les 12 tests en actif sont : CAST, Sum, wSum, aSum, MiST, C-alpha, SKAT, SKAT-O, KBAC, DoEstRare, KERNEL et PODKAT.
Les 4 tests en illustratif sont : VT, ADA, CLUSTER et BOMP.

Les valeurs propres de l'ACP sont présentées dans le Tableau S 15. Les deux premiers axes ont une valeur propre supérieure à 1 et expliquent chacun respectivement 63.18% et 18.96% de l'inertie totale. Si nous utilisons la règle de Kaiser, nous pouvons nous concentrer sur ces deux axes. Nous étudions aussi la troisième composante avec une valeur propre proche de 1 qui explique 7.73% de l'inertie totale, pour comparer avec les résultats d'ACP pour les données EOAD. Ces trois composantes principales expliquent 89.86% de la variabilité totale des données, ce qui donne une bonne représentation des résultats d'association.

Tableau S 15. Valeurs propres de l'ACP pour les données BrS.

	Valeur propre	Pourcentage d'inertie	Pourcentage d'inertie cumulé
Dim 1	7.58	63.18	63.18
Dim 2	2.28	18.96	82.14
Dim 3	0.93	7.73	89.86
Dim 4	0.45	3.78	93.64

Tableau S 16. Corrélations entre les variables et les composantes principales de l'ACP pour les données BrS.

	Dim 1	Dim 2	Dim 3
CAST	0.83	-0.35	0.37
Sum	0.90	-0.33	0.24
wSum	0.88	-0.34	-0.24
aSum	0.90	-0.30	0.14
KBAC	0.81	-0.24	-0.40
CALPHA	0.28	0.75	0.40
SKAT	0.58	0.76	0.01
SKATO	0.95	0.18	0.08
MiST	0.97	-0.04	0.15
KERNEL	0.64	0.68	-0.07
PODKAT	0.84	-0.14	-0.02
DoEstRare	0.71	0.36	-0.55
VT	0.87	-0.35	0.14
ADA	0.53	0.52	0.17
CLUSTER	0.45	0.48	0.23
BOMP	0.90	0.03	-0.21

Annexe VIII
ACP SUR LES RÉSULTATS DE SIGNIFICATIVITÉ POUR LES DONNÉES EOAD

L'ACP normée est effectuée sur les résultats d'association de 17409 gènes autosomaux pour 12 tests en actif. Les résultats d'association sont sous la forme de -log10(p-value).
Les 12 tests en actif sont : CAST, Sum, wSum, aSum, MiST, C-alpha, SKAT, SKAT-O, KBAC, DoEstRare, KERNEL et PODKAT.

Les trois premières valeurs propres de l'ACP sont supérieures à 1 (Tableau S 17). En appliquant la règle de Kaiser, nous étudions ces trois premières composantes principales. Elles expliquent chacune 60.87%, 15.93% et 9.67% de l'inertie totale, soit 86.47%.

Tableau S 17. Valeurs propres de l'ACP pour les données EOAD.

	Valeur propre	Pourcentage d'inertie	Pourcentage d'inertie cumulé
Dim 1	7.30	60.87	60.87
Dim 2	1.91	15.93	76.80
Dim 3	1.16	9.67	86.47
Dim 4	0.44	3.70	90.17

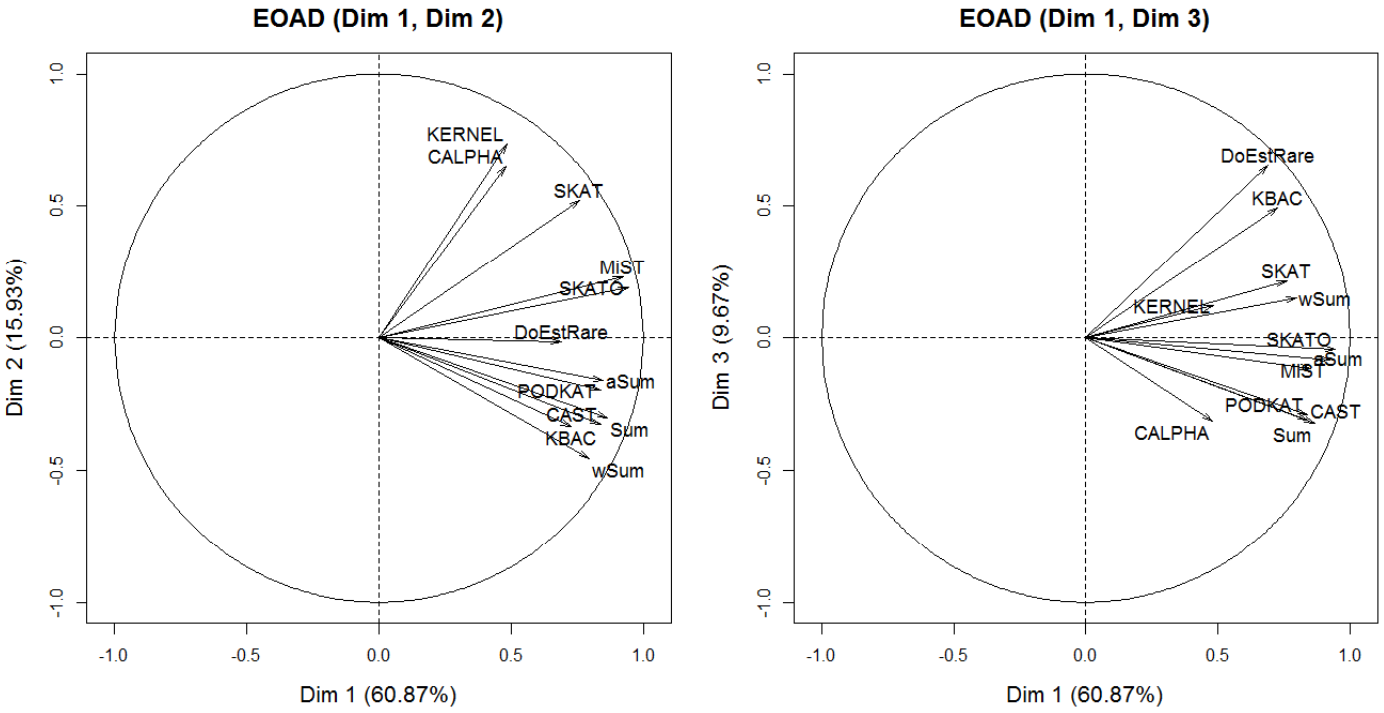


Figure S 5. Graphes des corrélations de l'ACP pour les données EOAD.

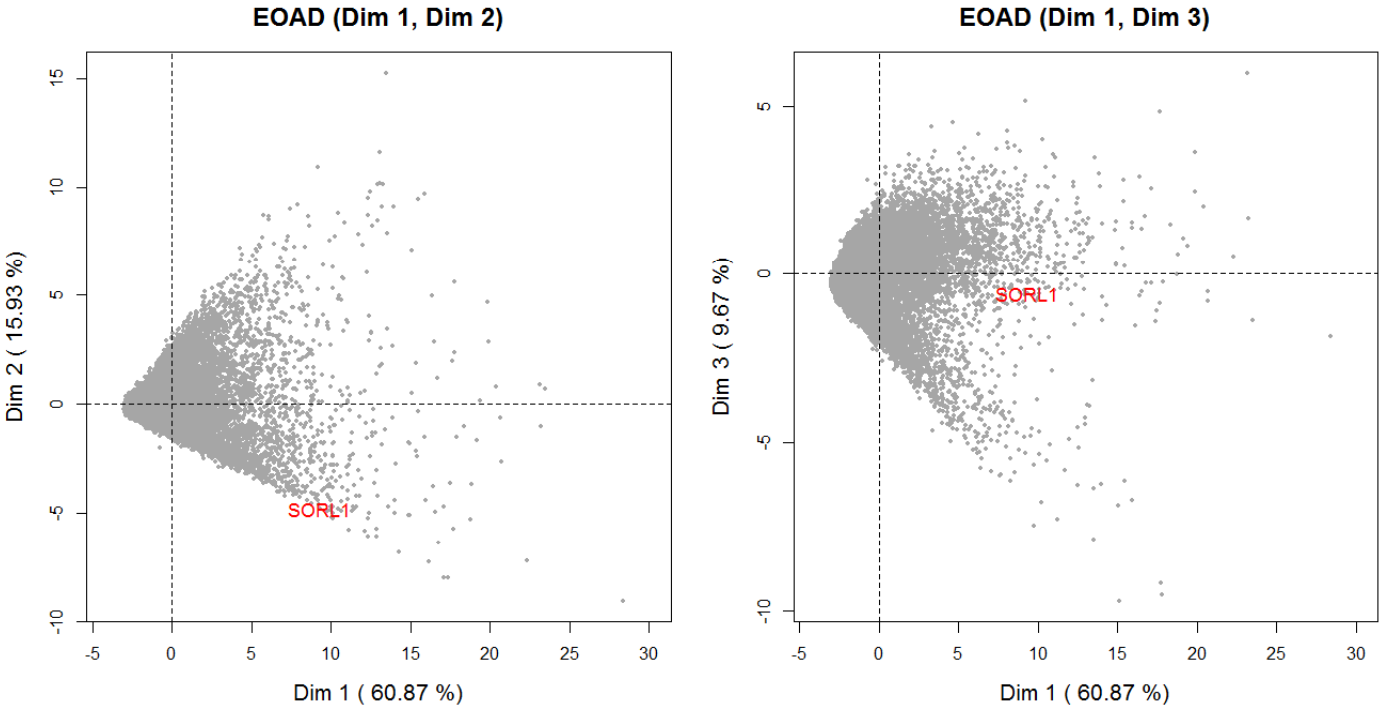


Figure S 6. Graphe des individus de l'ACP pour les données EOAD.

Le gène SORL1 identifié avec l'étude de Nicolas et al. (2015) est indiqué en rouge.

Tableau S 18. Corrélations entre les variables et les composantes principales de l'ACP pour les données EOAD.

	Dim 1	Dim 2	Dim 3
CAST	0.84	-0.33	-0.31
Sum	0.87	-0.30	-0.32
wSum	0.80	-0.46	0.15
aSum	0.85	-0.16	-0.12
KBAC	0.73	-0.34	0.49
CALPHA	0.48	0.65	-0.31
SKAT	0.76	0.52	0.21
SKATO	0.94	0.19	-0.04
MiST	0.92	0.23	-0.08
KERNEL	0.48	0.73	0.12
PODKAT	0.84	-0.20	-0.29
DoEstRare	0.69	-0.01	0.65

Annexe IX
RÉSULTATS DES TESTS D'ASSOCIATION POUR LES DONNÉES BRS

Tableau S 19. P-values des gènes candidats spécifiques pour le BrS.

	CAST	Sum	wSum	aSum	VT
SCN5A	0,0025	0,0010	0,0001	0,0010	0,0005
KCNH2	0,0609	0,0579	0,1988	0,0480	0,0529
ABCC9	0,0649	0,0879	0,3526	0,9990	0,0919
SCN10A	0,0729	0,1219	0,0360	0,1139	0,0350
CACNA1C	0,2328	0,3636	0,9471	0,9980	0,4955
CACNB2	0,4136	0,3297	0,9361	0,5495	0,4635
TRPM4	0,4675	0,7243	0,2967	0,7632	0,6943
CACNA2D1	0,9990	0,8212	0,3946	0,0729	0,9041
HCN4	0,9990	0,9990	0,5275	0,9441	0,4046
PKP2	0,9990	0,9990	0,5704	0,9391	1,0000

	CALPHA	SKAT	SKATO	KBAC	ADA
SCN5A	0,1873	0,0215	0,0003	0,0005	0,0771
KCNH2	1,0000	0,4840	0,0534	0,9800	0,7522
ABCC9	0,0949	0,0540	0,0448	0,9341	0,1409
SCN10A	0,8292	0,5905	0,0452	0,0539	0,8551
CACNA1C	1,0000	0,3906	0,5781	0,8322	0,2358
CACNB2	0,0929	0,0677	0,1090	0,5894	0,0980
TRPM4	0,6763	0,3647	0,5534	0,1518	0,6214
CACNA2D1	0,0749	0,0352	0,0535	0,1658	0,2449
HCN4	0,1908	0,9280	0,7518	0,5944	0,9540
PKP2	0,1968	0,8440	0,9193	0,6284	0,9640

	CLUSTER	KERNEL	DoEstRare	PODKAT	BOMP
SCN5A	0,1477	0,0098	0,0036	0,0005	0,0001
KCNH2	0,7542	0,4436	0,9890	0,4411	0,9836
ABCC9	0,1518	0,0080	0,5784	0,3022	0,8556
SCN10A	0,8711	0,2757	0,2278	0,4226	0,0792
CACNA1C	0,2448	0,6893	0,8342	0,5359	0,3524
CACNB2	0,1688	0,0759	0,5674	0,4726	0,3097
TRPM4	0,5884	0,1768	0,1219	0,4234	0,2960
CACNA2D1	0,2500	0,0190	0,0330	0,2367	0,1207
HCN4	0,9540	0,9600	0,8551	1,0000	0,6820
PKP2	0,9550	0,8881	0,7253	0,6586	0,8739

Annexe X
RÉSULTATS DES TESTS D'ASSOCIATION POUR LES DONNÉES EOAD

Q-Q plots et Manhattan plots

De la Figure S 7 à la Figure S 18, sont représentés les résultats de significativité des 17 409 gènes autosomaux testés avec les 12 méthodes. Le graphe de gauche est le Q-Q plot permettant de vérifier s'il y a une inflation des p-values. Le graphe de droite est un Manhattan plot permettant d'identifier des gènes significatifs. Seuls les gènes *KRTAP5-5* et *CELA3B*, sont indiqués sur les Manhattan plots. La ligne rouge correspond au seuil de significativité $\alpha=2.87.10^{-6}$ (correction de Bonferroni du seuil de 5% par le nombre de gènes testés).

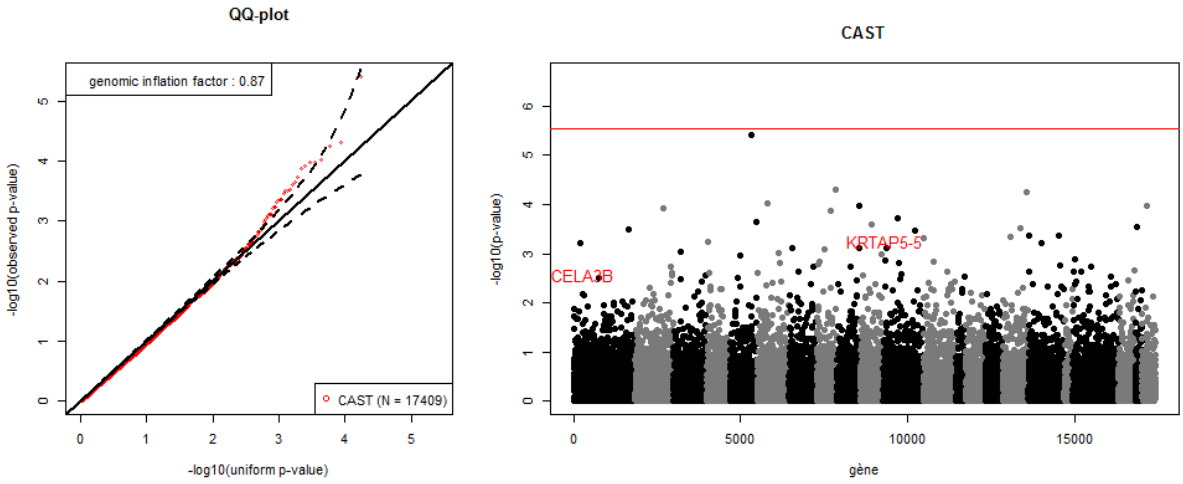


Figure S 7. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec CAST.

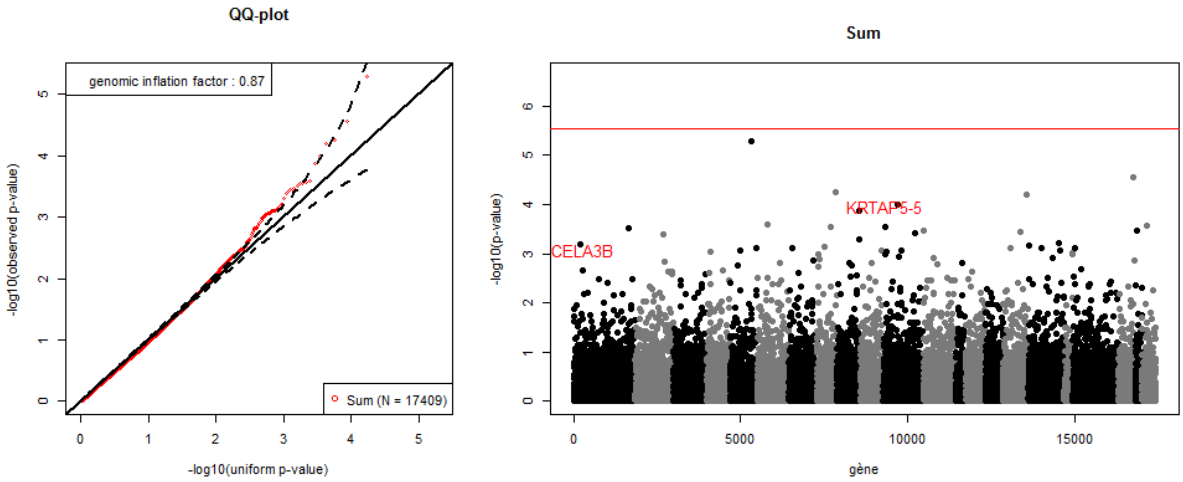


Figure S 8. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec Sum.

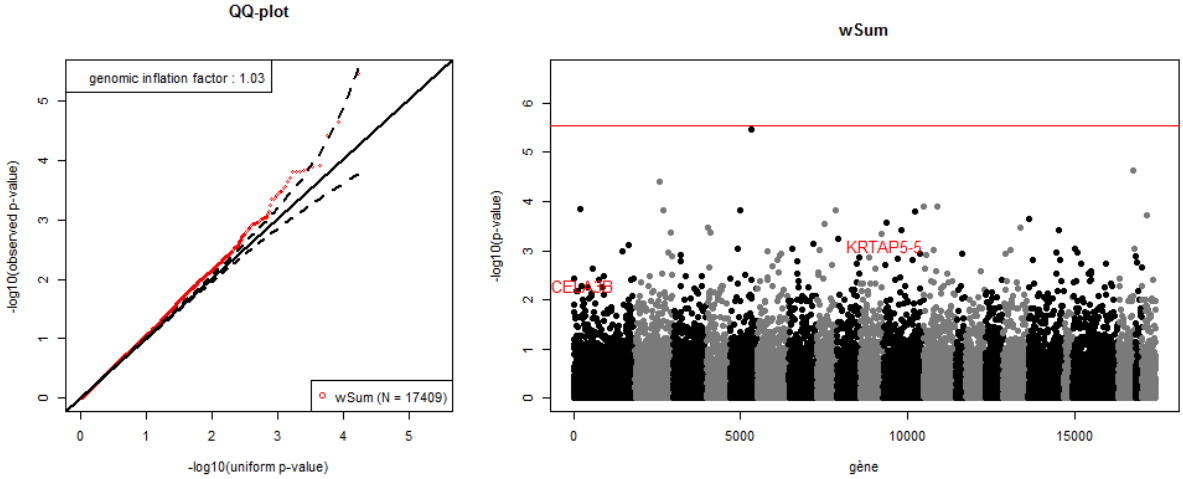


Figure S 9. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec wSum.

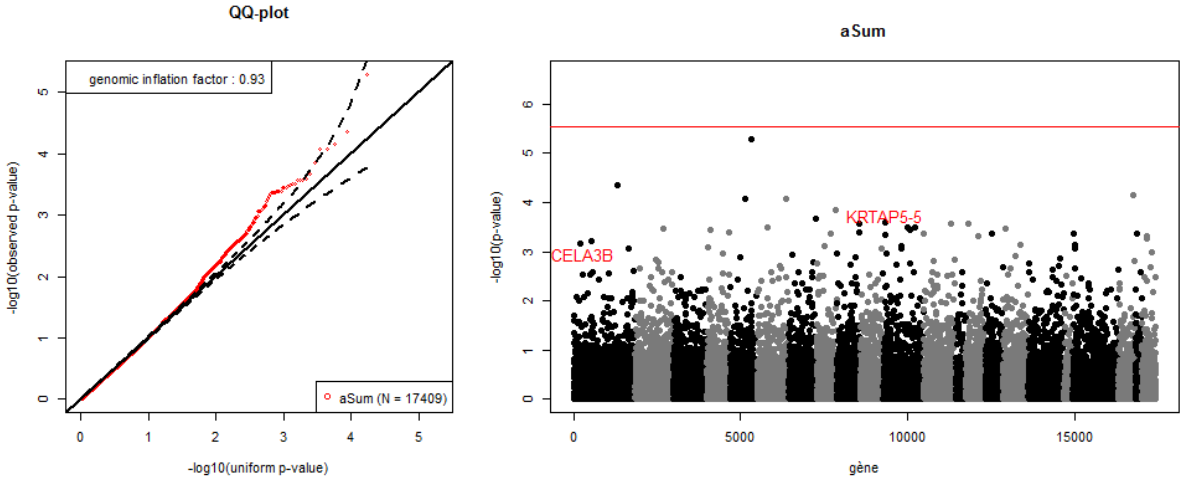


Figure S 10. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec aSum.

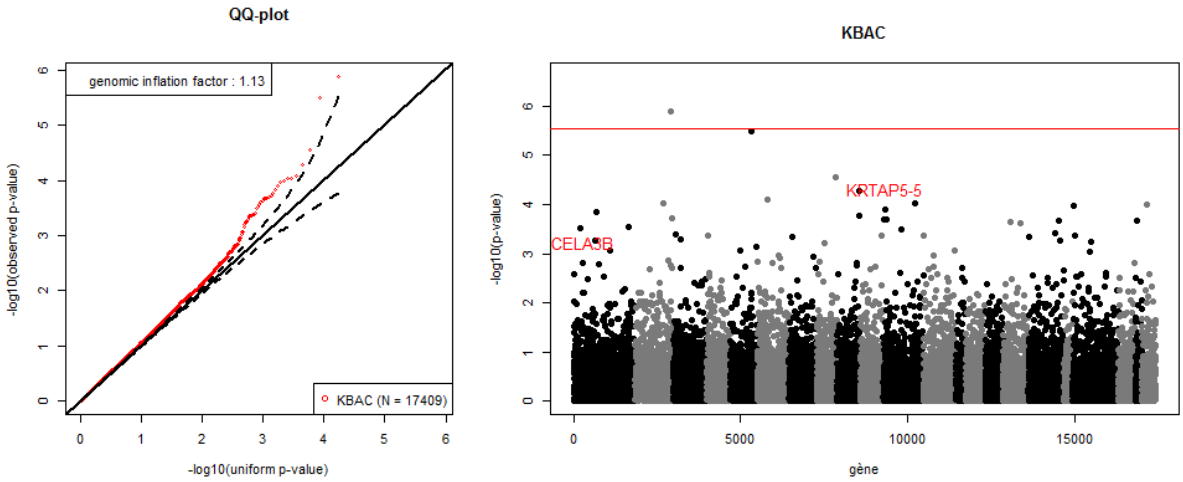


Figure S 11. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec KBAC.

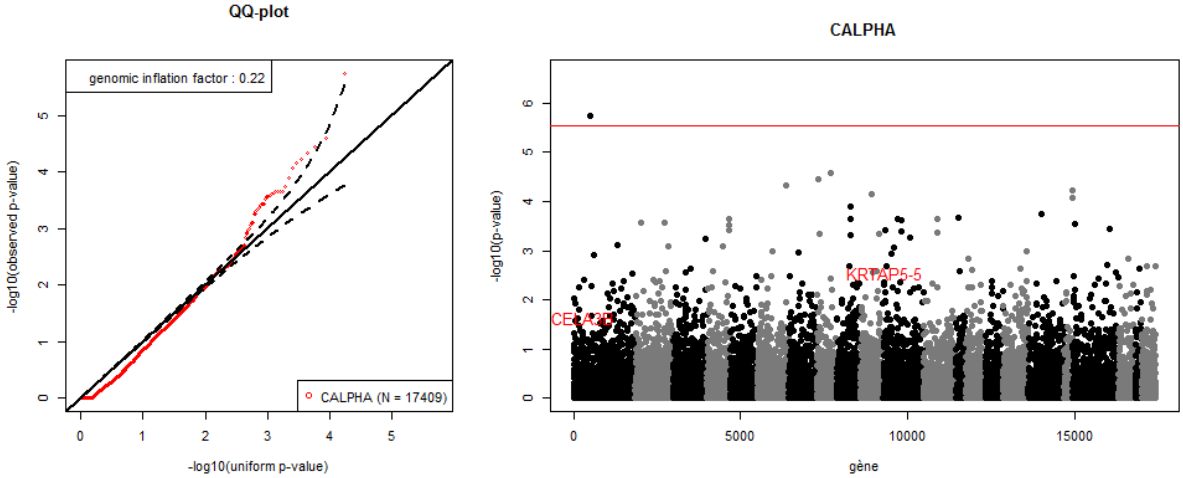


Figure S 12. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec C-alpha.

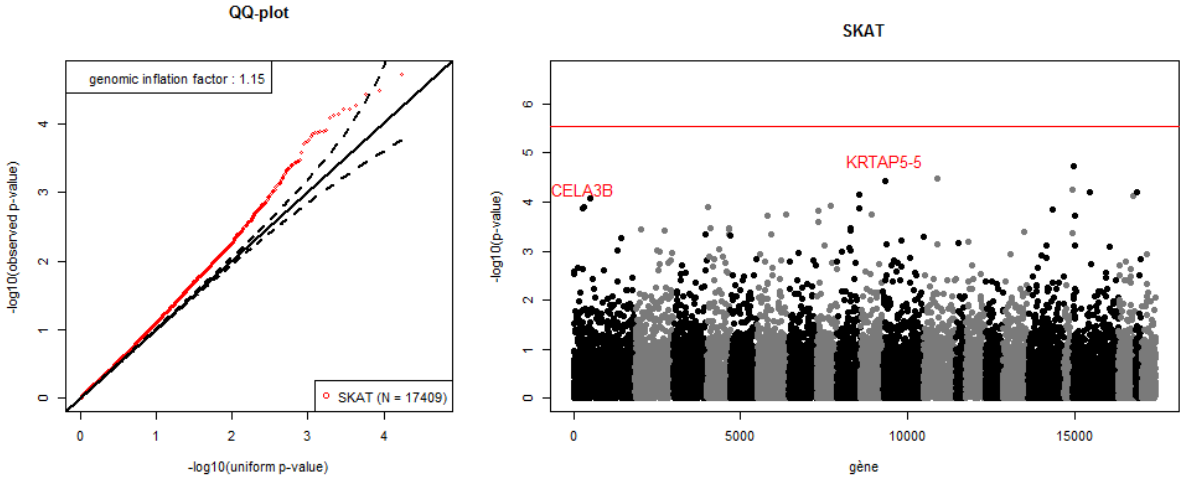


Figure S 13. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec SKAT.

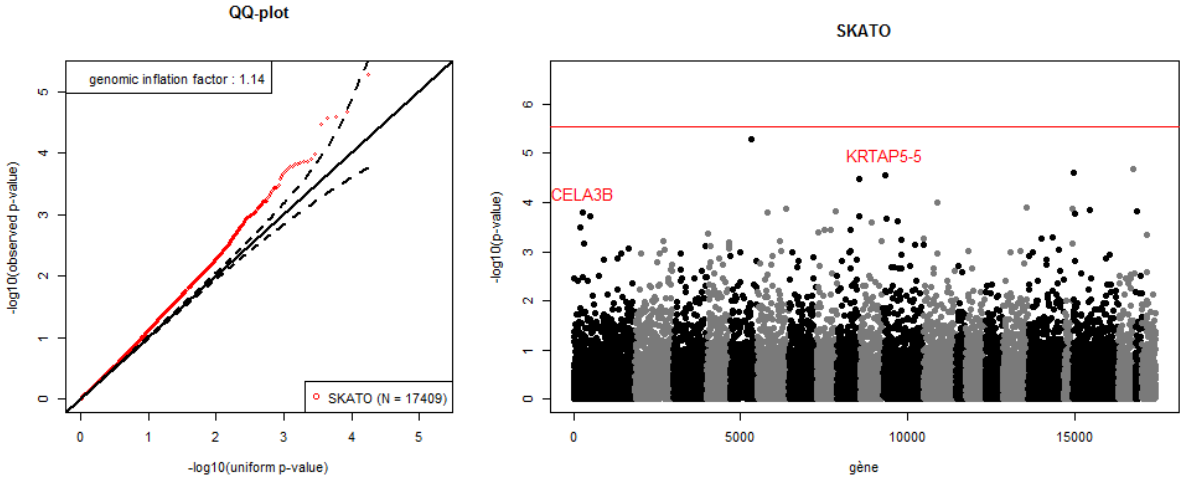


Figure S 14. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec SKAT-O.

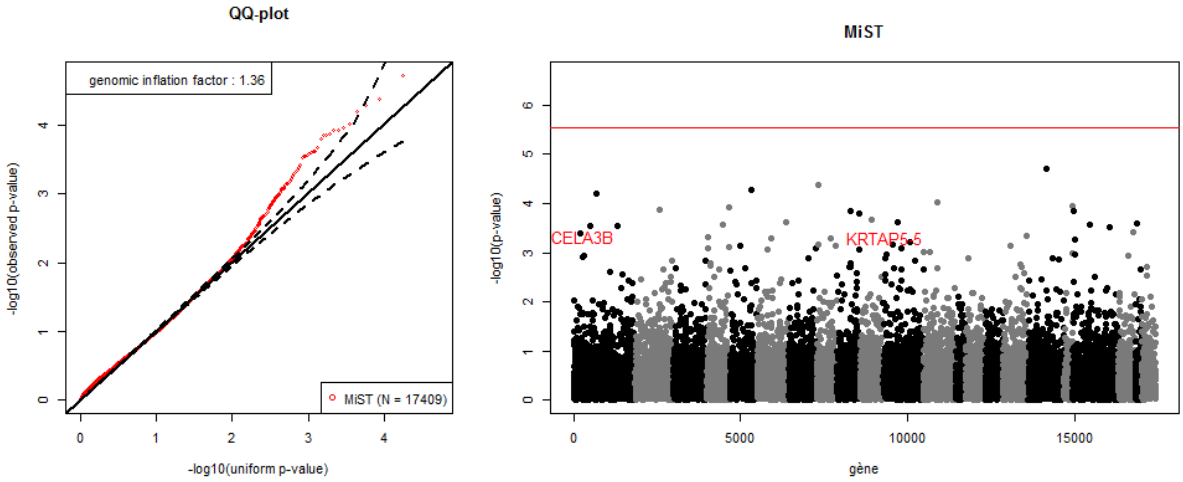


Figure S 15. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec MiST.

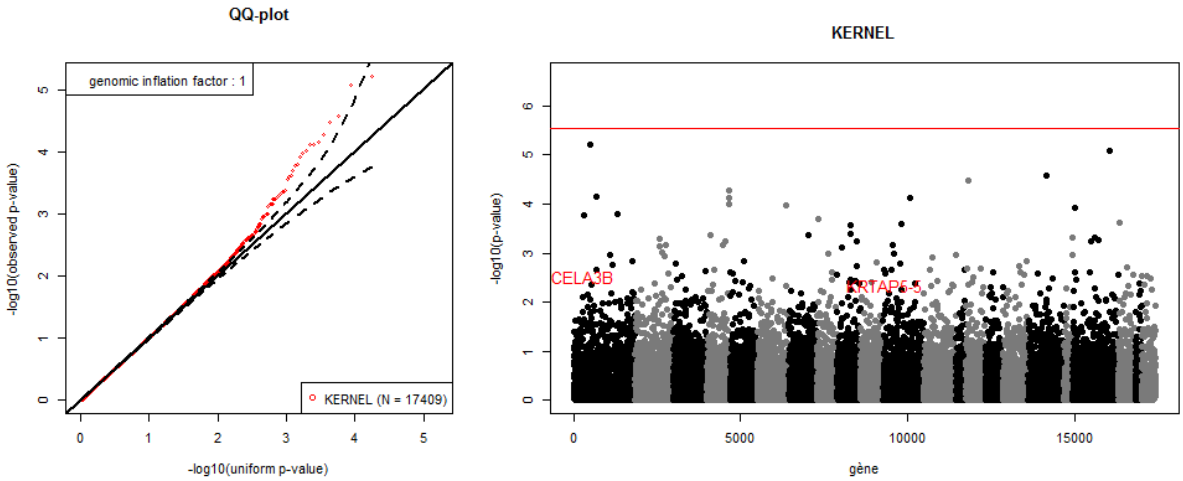


Figure S 16. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec KERNEL.

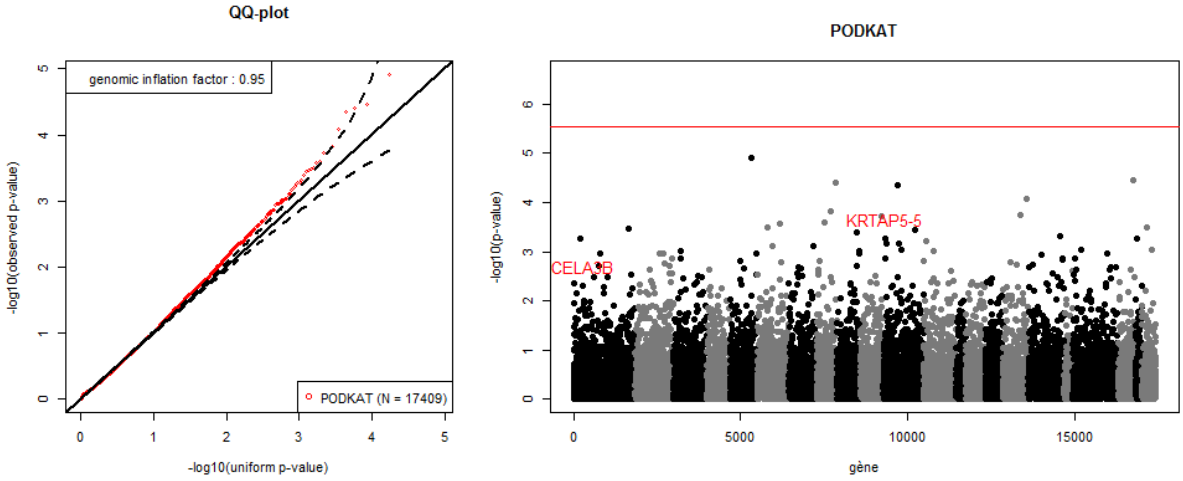


Figure S 17. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec PODKAT.

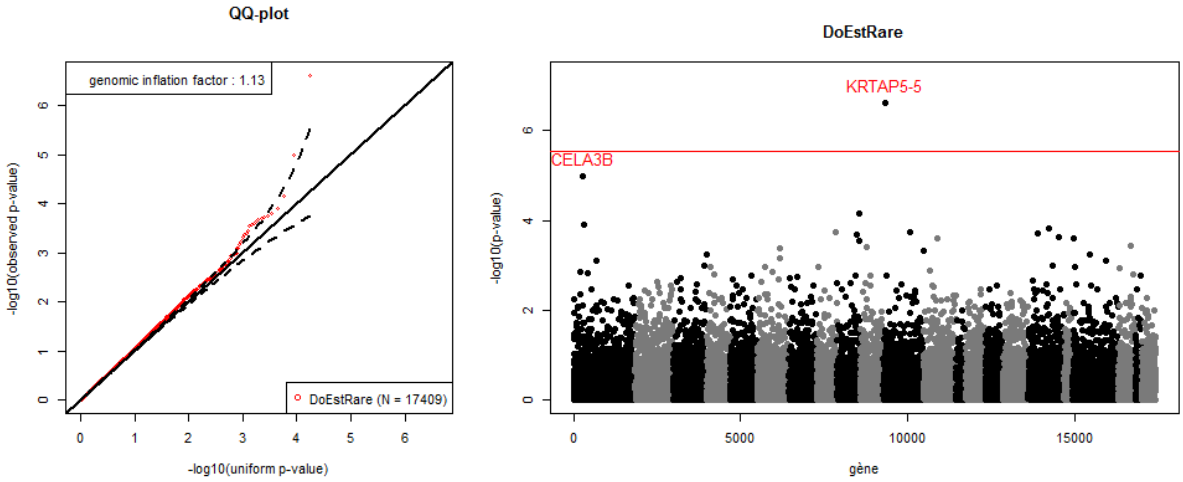


Figure S 18. Q-Q plot et Manhattan plot pour l'analyse de l'EOAD avec DoEstRare.

Les Figure S 19 Figure S 20 représentent les distributions des positions des variants rares pour les deux gènes les plus significatifs avec le gène DoEstRare *KRTAP5-5* et *CELA3B*. La ligne rouge correspond à la distribution chez les témoins et la ligne noire à celle chez les témoins. La densité est estimée avec le noyau gaussien.

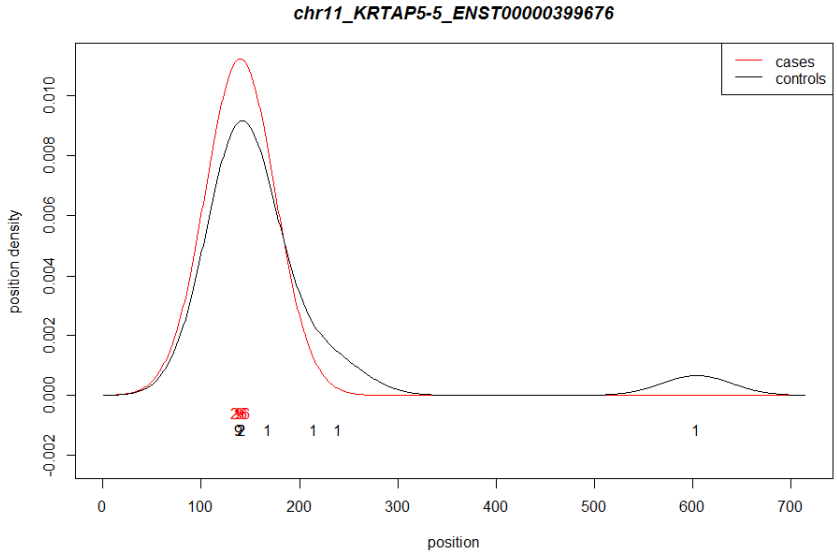


Figure S 19. Distribution des positions des mutations pour le gène *KRTAP5-5*.

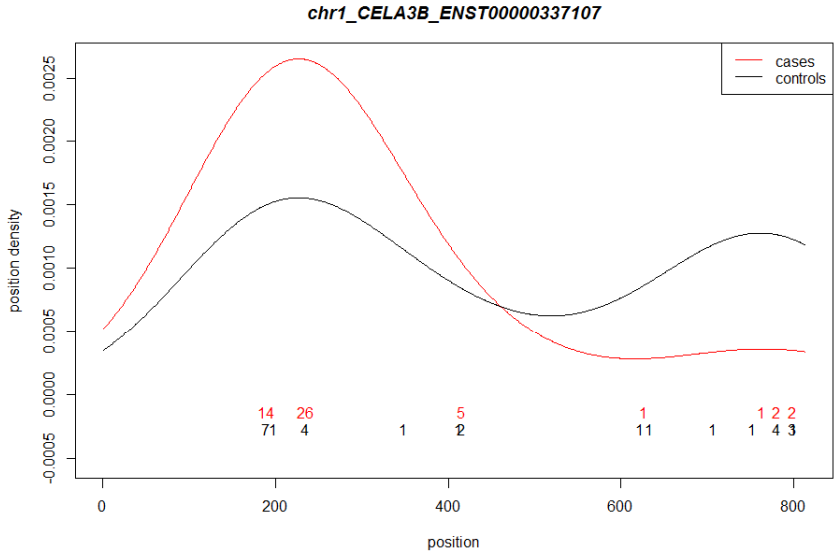


Figure S 20. Distribution des positions des mutations pour le gène *CELA3B*.

Annexe XI
VALEURS DE F_{ST} POUR LE PROJET FREX D'APRÈS GÉNIN ET AL. (2016)

Les valeurs de F_{ST} de FREX ont été fournies par Joanna Giemza, les résultats du projet ayant été présentés lors du congrès ASHG 2016 par Génin et al. (2016) [198].

Tableau S 20. Effectifs pour les différentes populations de FREX

	Effectif
LILLE	82
ROUEN	93
BORDEAUX	84
BREST	89
NANTES	93
DIJON	86

Tableau S 21. Valeurs de F_{ST} entre les différentes populations de FREX

	LILLE	ROUEN	BORDEAUX	BREST	NANTES	DIJON
LILLE	0	169	1069	1012	555	390
ROUEN	169	0	765	820	278	212
BORDEAUX	1069	765	0	1694	527	715
BREST	1012	820	1694	0	821	1172
NANTES	555	278	527	821	0	319
DIJON	390	212	715	1172	319	0

Annexe XII
SIMULATIONS DE 16 POPULATIONS

Pour tester le paramètre de migration *cosi* en lien avec la géographie des individus, nous avons d'abord dans un premier temps simulé 16 sous-populations avec la structure géographique présentée dans la Figure S 21.

Les sous-populations sont issues de la population européenne il y a 80 générations dans le modèle « *bestfit* » décrit par Schaffner et al. (2005).

Chaque sous-population est constituée de 1000 individus (2000 haplotypes). Nous avions simulé l'information pour le génome entier, i.e 3 milliards de paires de bases (bp), que nous avons découpé en simulations indépendantes de 6000 régions de 500 kb.

La structure géographique est renseignée par les taux de migration entre les sous-populations. Les taux de migration entre populations voisines est de 0.1. À chaque éloignement de population, le taux de migration est divisé par 10. Soit n le degré d'éloignement, et m_n le taux de migration pour le degré n. On a $m_1 = 0.1$, $m_2 = 0.1 \times m_1$, $m_3 = 0.1 \times m_2$, …, $m_{n+1} = 0.1 \times m_n$. Ceci correspond à une suite géométrique de raison 0.1 et donc $m_n = (0.1)^{n-1}0.1$.

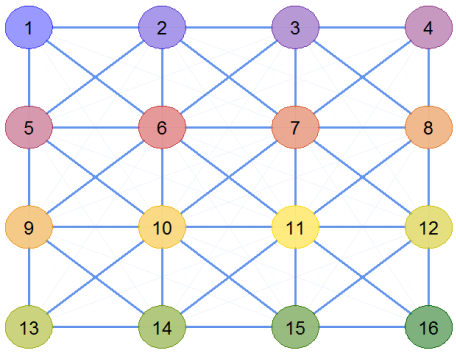


Figure S 21. Structure géographique simulée

Nous pouvions constater avec une analyse MDS réalisée avec PLINK [59] que la structure génétique des populations simulées correspond bien à la structure géographique voulue.

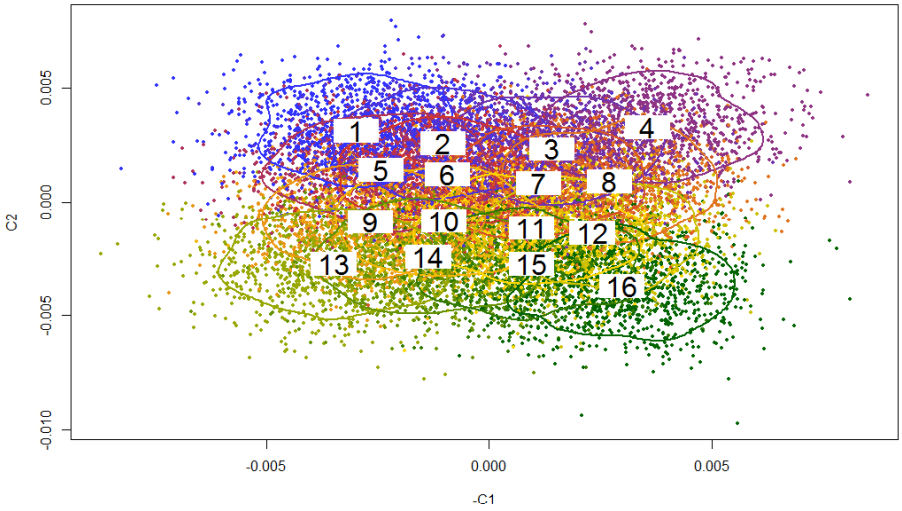


Figure S 22. Graphe MDS des individus des 16 populations.

L'analyse MDS a été effectuée à l'aide du logiciel PLINK [59] sur la matrice de distance IBS de taille 16000 individus x 100 000 variants fréquents indépendants (MAF≥10%, r^2<0.2).

Annexe XIII

TABLEAUX DES ERREURS DE TYPE I POUR L'ÉTUDE DE L'IMPACT DE LA STRUCTURE DE POPULATION

Tableau S 22. Erreurs de type I au seuil α=5% en présence d'une structure de population.

Migration	0.01				0.025			
% Pop B	25	50	75	100	25	50	75	100
CAST	0,0655	0,1463	0,2726	0,3969	0,0528	0,081	0,135	0,2091
Sum	0,0681	0,1392	0,2596	0,3745	0,0533	0,0806	0,1321	0,2019
wSum_betaMAFtot	0,0741	0,1519	0,2818	0,4149	0,0583	0,0905	0,1447	0,2239
wSum_MAFtot	0,0726	0,1727	0,3597	0,5585	0,0618	0,1015	0,1811	0,3031
wSum_MAFctrl	0,1033	0,2859	0,5732	0,8196	0,0797	0,155	0,3084	0,5184
aSum	0,1082	0,3328	0,6917	0,8884	0,0749	0,155	0,3154	0,531
KBAC	0,1247	0,3585	0,6867	0,8886	0,0816	0,1725	0,3441	0,5603
SKAT	0,0952	0,3238	0,6852	0,9044	0,0638	0,137	0,3154	0,5622
wSKAT_betaMAFtot	0,0955	0,3344	0,711	0,924	0,0647	0,1426	0,3301	0,5957
wSKAT_MAFtot	0,0866	0,3217	0,7455	0,9603	0,06	0,1404	0,3517	0,6771
SKATO	0,0959	0,3011	0,6652	0,8922	0,069	0,1314	0,2879	0,5316
wSKATO_betaMAFtot	0,0933	0,3155	0,6904	0,914	0,0686	0,1331	0,3043	0,5627
wSKATO_MAFtot	0,0899	0,3149	0,7335	0,9539	0,0652	0,1376	0,3369	0,6489
PODKAT	0,0788	0,1952	0,4114	0,6383	0,0586	0,0968	0,1881	0,3234
DoEstRare	0,0939	0,2423	0,5111	0,7598	0,0648	0,1179	0,2356	0,4058

Migration	0.05				0.1			
% Pop B	25	50	75	100	25	50	75	100
CAST	0,0459	0,0592	0,0909	0,1295	0,0428	0,0517	0,0671	0,0813
Sum	0,0469	0,0612	0,0885	0,1204	0,0452	0,0517	0,0669	0,0837
wSum_betaMAFtot	0,0522	0,0703	0,0967	0,1356	0,052	0,0586	0,0734	0,0912
wSum_MAFtot	0,0565	0,0745	0,116	0,1772	0,0503	0,0613	0,0805	0,1108
wSum_MAFctrl	0,0687	0,1114	0,1856	0,3106	0,0602	0,0846	0,1204	0,1795
aSum	0,0636	0,1088	0,182	0,3201	0,0564	0,0726	0,114	0,1705
KBAC	0,0688	0,1088	0,1906	0,3173	0,0628	0,0838	0,1187	0,1747
SKAT	0,0563	0,0878	0,1576	0,2833	0,0489	0,0643	0,0927	0,1447
wSKAT_betaMAFtot	0,0567	0,0902	0,1624	0,3016	0,0473	0,0642	0,0948	0,1517
wSKAT_MAFtot	0,0523	0,0848	0,1675	0,3348	0,0434	0,0577	0,0909	0,1564
SKATO	0,056	0,0885	0,1513	0,2618	0,0524	0,0663	0,0943	0,1389
wSKATO_betaMAFtot	0,0548	0,0896	0,1555	0,2778	0,052	0,0658	0,0957	0,1469
wSKATO_MAFtot	0,0544	0,0893	0,1702	0,3221	0,0487	0,0631	0,0979	0,1574
PODKAT	0,0539	0,0717	0,1109	0,1702	0,0484	0,0612	0,0751	0,1048
DoEstRare	0,054	0,0793	0,1222	0,2105	0,0522	0,0604	0,0816	0,1136

Tableau S 23. Erreurs de type I au seuil α=5% en absence de structure de population

CAST	0,0413
Sum	0,0417
wSum_betaMAFtot	0,0465
wSum_MAFtot	0,0459
wSum_MAFctrl	0,0477
aSum	0,0478
KBAC	0,0491
SKAT	0,047
wSKAT_betaMAFtot	0,0475
wSKAT_MAFtot	0,0411
SKATO	0,0497
wSKATO_betaMAFtot	0,0495
wSKATO_MAFtot	0,0469
PODKAT	0,0481
DoEstRare	0,0475

Annexe XIV
RÉSULTATS DE L'ANALYSE ACP AVEC *SMARTPCA*

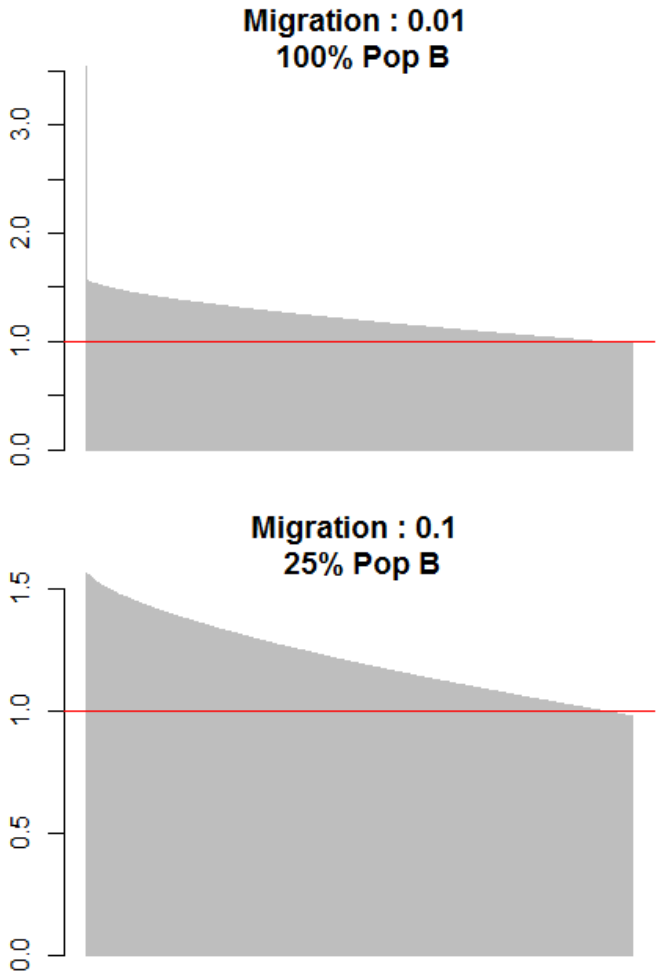


Figure S 23. Valeurs propres pour les 1000 premières composantes de l'ACP pour les scénarios les plus extrêmes.

La ligne rouge indique le seuil de 1 pour les valeurs propres.

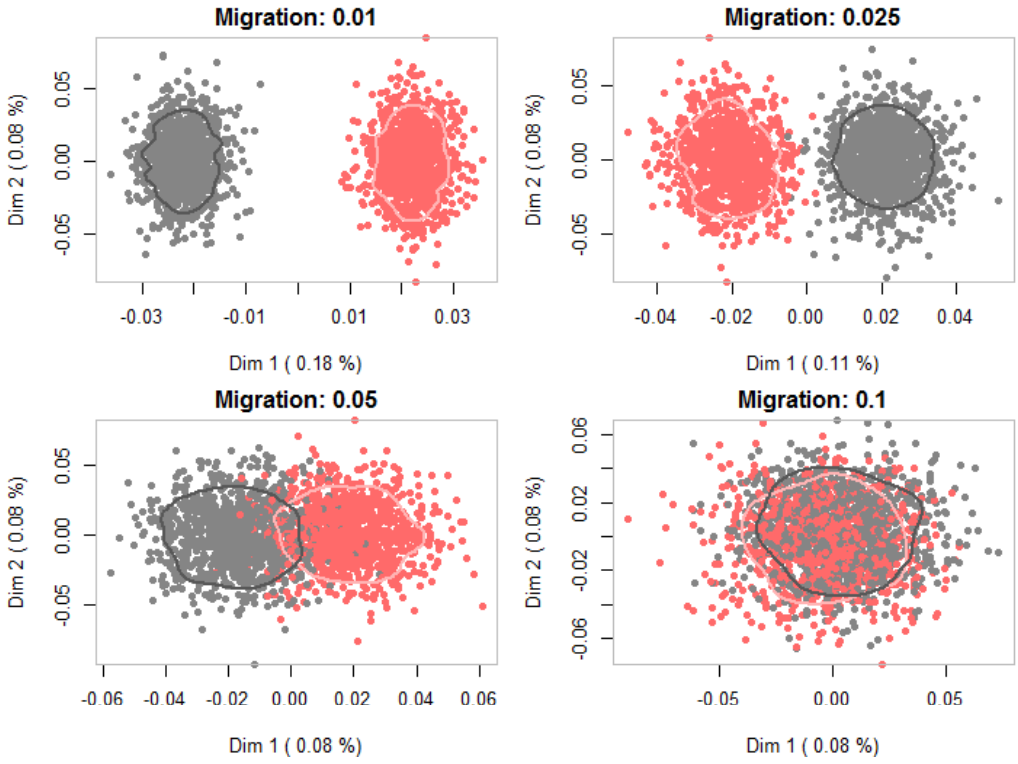


Figure S 24. Graphes des individus de l'ACP pour 100% des témoins de la population B.

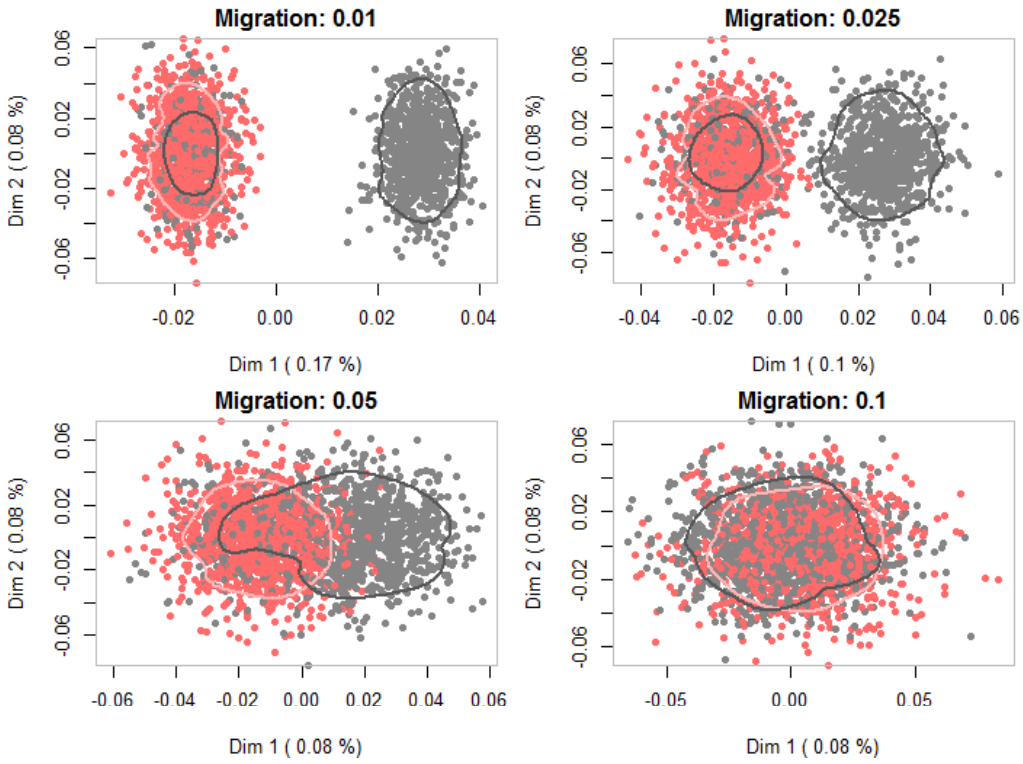


Figure S 25. Graphes des individus de l'ACP pour 75% des témoins de la population B.

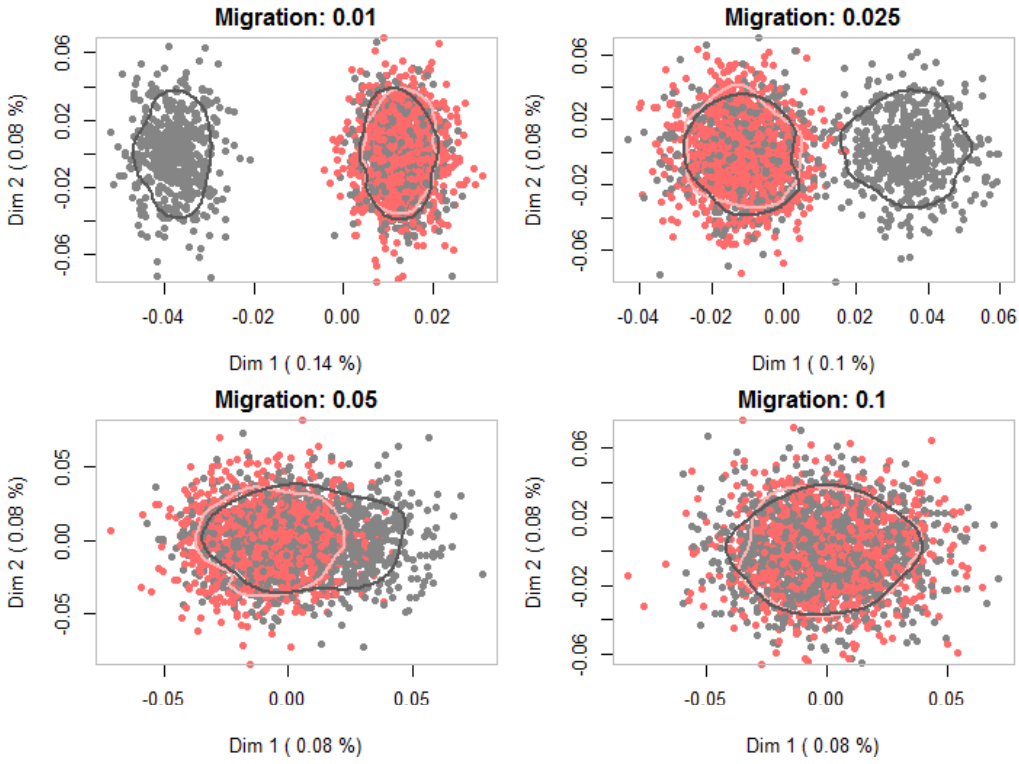


Figure S 26. Graphes des individus de l'ACP pour 50% des témoins de la population B.

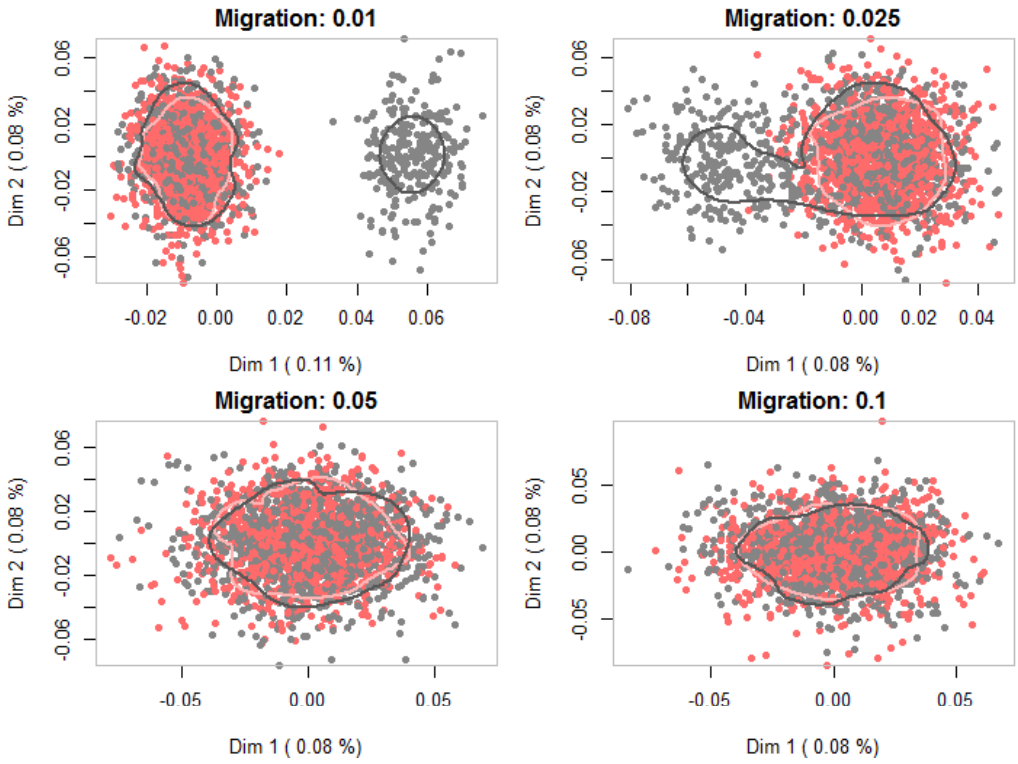


Figure S 27. Graphes des individus de l'ACP pour 25% des témoins de la population B.

Annexe XV

TABLEAUX D'ERREURS DE TYPE I POUR L'ÉTUDE DE LA CORRECTION DES TESTS POUR LA STRATIFICATION DE POPULATION

Tableau S 24. Erreurs de type I au seuil α=5% pour les tests après correction avec les 2PC.

Migration	0,01			0,025		
% Pop B	25	50	75	25	50	75
CAST	0,0488	0,0515	0,0552	0,0517	0,0476	0,0555
Sum	0,0483	0,0535	0,0529	0,0509	0,0487	0,0561
wSum_betaMAFtot	0,0499	0,0521	0,0547	0,0503	0,0484	0,058
wSum_MAFtot	0,0523	0,056	0,063	0,0521	0,0541	0,0603
wSum_MAFctrl	0,0528	0,0567	0,0571	0,0573	0,0549	0,0559
aSum	0,06	0,0764	0,1149	0,0582	0,071	0,0932
SKAT	0,0535	0,0592	0,0564	0,0539	0,058	0,0519
wSKAT_betaMAFtot	0,0543	0,0584	0,0575	0,0543	0,0591	0,0525
wSKAT_MAFtot	0,0513	0,0535	0,06	0,0533	0,059	0,0588
SKATO	0,054	0,0608	0,0576	0,0573	0,0569	0,0548
wSKATO_betaMAFtot	0,0537	0,061	0,0579	0,0573	0,059	0,0545
wSKATO_MAFtot	0,0538	0,0559	0,0591	0,0551	0,0607	0,0581
PODKAT	0,0531	0,0539	0,0501	0,0521	0,055	0,0495

Migration	0,05			0,1		
% Pop B	25	50	75	25	50	75
CAST	0,0533	0,0558	0,0587	0,05	0,06	0,0736
Sum	0,0536	0,0565	0,0635	0,0512	0,0567	0,0706
wSum_betaMAFtot	0,0534	0,0568	0,0653	0,0516	0,0588	0,0723
wSum_MAFtot	0,0565	0,0595	0,0707	0,0513	0,0593	0,0805
wSum_MAFctrl	0,0674	0,0798	0,0835	0,0598	0,0827	0,1152
aSum	0,0609	0,0769	0,0909	0,0545	0,0699	0,0987
SKAT	0,0579	0,0649	0,0594	0,0485	0,0651	0,0938
wSKAT_betaMAFtot	0,0565	0,0654	0,0596	0,0468	0,064	0,0943
wSKAT_MAFtot	0,0518	0,0651	0,0629	0,0429	0,0576	0,0906
SKATO	0,0565	0,0682	0,0632	0,0521	0,0666	0,0905
wSKATO_betaMAFtot	0,0562	0,067	0,0647	0,0516	0,0652	0,0927
wSKATO_MAFtot	0,0556	0,0689	0,0672	0,0483	0,0621	0,0965
PODKAT	0,0555	0,0599	0,0566	0,0489	0,0615	0,0751

UNIVERSITE BRETAGNE LOIRE

UNIVERSITÉ DE NANTES

Thèse de Doctorat

Elodie PERSYN

Analyse d'association de variants génétiques rares dans une population démographiquement stable

Association analysis of rare genetic variants in a fine geographical scale population

Résumé

Les études d'association sur génome entier ont permis d'identifier de nombreux facteurs de risque génétiques impliqués dans des maladies complexes. Il apparaît cependant que les variants fréquents n'expliquent qu'une faible partie de l'héritabilité des maladies. Une partie non négligeable serait due à la présence de variants rares avec des effets génétiques plus forts. Tester l'association de ces variants est problématique du fait de leur faible fréquence dans la population générale. De nombreuses méthodes statistiques ont été développées avec la stratégie commune d'agréger l'information pour un groupe de variants.

Cette thèse a pour objectif de comparer les principales stratégies à l'aide de simulations de différents scénarios génétiques et de l'application à de vraies données de séquençage. Nous avons aussi développé un test, appelé DoEstRare, comparant les distributions des positions des variants rares entre les cas et les témoins, afin de détecter des regroupements de variants dans des régions locales.

Enfin, il a été montré qu'une structure de population est un facteur de confusion pour l'interprétation des résultats d'analyse de variants rares. Avec le recrutement de témoins pour les analyses, avec des projets tels que French Exome et VACARME, il est alors nécessaire de comprendre l'impact d'une structure à fine échelle géographique (e.g. échelle de la France) pour les différentes stratégies statistiques. La seconde partie de cette thèse consiste à évaluer cet impact au moyen de simulations de données génétiques pour des structures géographiques locales.

Mots clés

Épidémiologie génétique, Statistique appliquée, Eudes d'association, Variants génétiques rares, Tests statistiques

Abstract

Genome-wide association studies have identified many common risk alleles for a wide variety of complex diseases. However these common variants explain a very small part of the heritability. A hypothesis is the presence of rare genetic variants with stronger effects. Testing the association of those rare variants is challenging due to their low frequency in populations. Many statistical methods have been developed with the strategy to aggregate the information for a group a rare variants.

This thesis aims to compare the main strategies through simulating under various genetic scenarios and the application to real sequencing data. We also developed a statistical test, called DoEstRare, which can detect clustered disease-risk variants in local genetic regions, by comparing the position distributions between cases and controls.

Moreover, it has been shown that population stratification represents a confounding factor in the analysis interpretations for rare variants. With the recruitment of controls, in the context of projects such as French Exome and VACARME, it is necessary to assess the impact of a very fine geographical structure (France) for different statistical strategies. The second part of this thesis consists in estimating this impact by simulating fine-scale population structures.

Key Words

Genetic epidemiology, Applied statistics, Association studies, Rare genetic variants, Statistical tests