Assessment of Hardy-Weinberg equilibrium and detection of chromosomal deletions in exome-wide sequencing data from large datasets : exploiting large exome datasets to improve identification of clinically relevant genetic variants

Benedetta Bigio

▶ **To cite this version:**

Benedetta Bigio. Assessment of Hardy-Weinberg equilibrium and detection of chromosomal deletions in exome-wide sequencing data from large datasets : exploiting large exome datasets to improve identification of clinically relevant genetic variants. Bioinformatics [q-bio.QM]. Université Paris Cité, 2020. English. NNT : 2020UNIP5200 . tel-04210907

HAL Id: tel-04210907

https://theses.hal.science/tel-04210907

Submitted on 19 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université de Paris

Pierre Louis de Santé Publique: Epidemiologie et Sciences de l'Information Biomédicale

Laboratory UMR_S 1163 IHU IMAGINE Institute of genetic diseases

Assessment of Hardy-Weinberg equilibrium and detection of chromosomal deletions in exome-wide sequencing data from large datasets

Exploiting large exome datasets to improve identification of clinically relevant genetic variants

Par Benedetta Bigio

Thèse de doctorat de Épidémiologie Génétique

Dirigée par Laurent Abel

Présentée et soutenue publiquement le 4 Décembre 2020

Devant un jury composé de :

Hervé Perdry, MCU, rapporteur

Emmanuelle Génin, Directeur de recherche, rapporteur

Lluis Quintana-Murci, Professeur au Collège de France, examinateur

Anne-Louise Leutenegger, Chargée de Recherche, examinateur

Laurent Abel, Directeur de recherche, Directeur de thèse

Except where otherwise noted, this work is licensed under
http://creativecommons.org/licenses/by-nc-nd/3.0/

Title : **Assessment of Hardy-Weinberg equilibrium and detection of chromosomal deletions in exome-wide sequencing data from large datasets**

Abstract :

A major focus of human genetics is on the identification of variants that may contribute to human diseases or adaptive traits. Next-generation sequencing (NGS) approaches, including whole exome sequencing (WES), provide unprecedent opportunities for discovering novel variants that may underlie susceptibility or resistance to disease. The basic principle of WES is the sequencing of coding regions, whereby DNA probes or baits are used to hybridize with the protein-coding portion of the genome, isolating it from the non-coding portions. After sequencing, millions of DNA sequences, known as reads, are aligned to a reference genome and undergo many types of downstream analysis, whereby the common goal is to identify novel targets underlying the scientific question that is being asked. Since its inception, NGS methods, including WES, have been providing an enormous amount of data at sustainable costs but also posing considerable challenges for the analysis and interpretation of the results. These technological advances increasingly require the development of sophisticated computational approaches, thus generating new research avenues in order to appropriately analyze and interpret enormous amounts of data. In turn, the wealth of exome data accumulated over the years has given the opportunity to pose scientific questions in ways that could not be possible earlier. My thesis took advantage from both these aspects.

First, I developed a computational approach that allows filtering of false positive variants that cannot be discarded with traditional bioinformatic approaches. We collectively referred to these variants as 'blacklist' and characterized them computationally and experimentally, discovering that a subset is out of Hardy-Weinberg (HW) equilibrium, a fundamental population genetic principle typically used as a filtering criterion in large-scale genotyping studies (e.g. GWAS). Based on these initial findings, we are currently studying HW equilibrium systematically and at a larger scale to determine whether HW equilibrium could be used not only to detect technical errors but also to inform about important phenomena relevant to population genetics. Our preliminary data focusing on variants with an excess or loss of homozygotes for the minor allele revealed promising candidate variants that could be indicative of protection (eg in *FUT2*, *SMN2*) or disadvantage (eg in *FANCD2*) to disease.

Second, I tackled the question of detection of copy number variants (CNVs) in WES data. CNVs are a specific class of variants traditionally difficult to detect in exome data of typical laboratory cohorts that are generated over time. In my thesis, I developed HMZDelFinder-opt, an algorithm that allows identification of partial exon homozygous and hemizygous deletions. Using HMZDelFinder_opt with both validated disease-causing deletions and simulated data, we demonstrated that the *a priori* selection of a reference control set with a coverage profile similar to that of the WES sample studied reduced the number of deletions detected, while improving the ranking of the true homozygous deletion. HMZDelFinder_opt also fills the gap in the study of deletions spanning less than an exon, by providing the first tool for the systematic identification of partial exon deletions. Collectively, these projects tackle heretofore-unexamined topics and hold promise to discover novel causal determinants of human diseases or traits.

Keywords : Hardy-Weinberg, whole exome sequencing, homozygous and hemizygous deletions, disease-causing mutations, blacklist.

Titre: **Intérêt de l'équilibre de Hardy-Weinberg et détection des délétions chromosomiques dans les données de séquençage d'exome à partir de grands ensembles de données**

Résume :

Un des principaux centres d'intérêt de la génétique humaine est l'identification des variants qui peuvent contribuer aux maladies humaines ou aux traits adaptatifs. Les approches de séquençage de nouvelle génération (NGS), y compris le séquençage de l'exome entier (WES), offrent des opportunités sans précédent pour découvrir de nouveaux variants impliqués dans la sensibilité ou la résistance à une pathologie. Le principe de base du WES est le séquençage des régions codantes, grâce auquel des sondes ADN sont utilisées pour s'hybrider avec la partie codante du génome. Après le séquençage, des millions de séquences d'ADN, appelées reads, sont alignées sur un génome de référence et sont analysées par différents outils, avec l'objectif d'identifier de nouvelles cibles pertinentes pour la question scientifique posée. Depuis leur création, les méthodes NGS, y compris le WES, ont fourni une énorme quantité de données qui posent des défis considérables pour leur analyse et l'interprétation des résultats correspondants. Ces avancées technologiques nécessitent de plus en plus le développement d'approches méthodologiques sophistiquées, générant ainsi de nouvelles questions de recherche afin d'optimiser l'analyse de ces données. Ainsi, les volumes de données d'exome accumulées au fil des ans permet de poser des questions scientifiques nouvelles. Ma thèse a porté sur ces aspects.

Tout d'abord, j'ai développé une approche qui permet de filtrer les variants qui sont des faux positifs et qui n'étaient pas éliminés avec les approches bioinformatiques classiques. Nous avons regroupé ces variants dans une « blacklist » et les avons caractérisés *in silico* et de façon expérimentale. Nous avons en particulier montré qu'un sous-ensemble de ces variants ne respectaient pas l'équilibre de Hardy-Weinberg (HW), un principe fondamental de génétique des populations généralement utilisé comme critère de filtre dans les études de génotypage à grande échelle (par exemple les études d'association génome entier). Sur la base de ces résultats initiaux, nous avons débuté une étude plus systématique de l'équilibre HW à plus grande échelle pour déterminer si ce test pourrait être utilisé non seulement pour détecter des erreurs techniques, mais aussi pour informer sur des phénomènes importants et pertinents en termes de génétique des populations. Nos données préliminaires se concentrant sur les variants avec un excès ou une perte d'homozygotes pour l'allèle mineur ont révélé certains variants candidats prometteurs qui pourraient indiquer un effet protecteur (dans *FUT2*, et *SMN2*) ou désavantageux (dans *FANCD2*) vis-à-vis ce certaines pathologies.

Au cours de cette thèse, j'ai également abordé la question de la détection des variations du nombre de copies (CNV) dans les données WES. Les CNV sont une classe spécifique de variants traditionnellement difficiles à détecter dans les données d'exome de cohortes de laboratoire qui sont générées au fil du temps. Dans ma thèse, j'ai développé HMZDelFinder-opt, un algorithme qui permet d'optimiser la détection de délétions homozygotes et hémizygotes et d'identifier des délétions partielles d'exons. En utilisant HMZDelFinder_opt avec à la fois des délétions pathogènes validées et des données simulées, nous avons démontré que la sélection optimisée d'un ensemble d'exomes contrôles de référence avec un profil de couverture similaire à celui de l'échantillon WES étudié réduisait le nombre de délétions faussement détectées, tout en améliorant l'identification des véritables délétions homozygotes. HMZDelFinder_opt permet également de fournir un nouvel outil pour l'identification systématique des délétions partielles d'exon. Au total, les questions traités dans ma thèse ont permis de proposer des approches nouvelles afin d'améliorer l'identification de nouveaux déterminants génétiques de pathologies humaines.

Mots clés: Equilibre de Hardy-Weinberg, séquençage d'exome entier, délétions homozygotes et hémizygotes, mutations pathogènes, liste noire.

Essai

Un des principaux centres d'intérêt de la génétique humaine est l'identification des variants qui peuvent contribuer aux maladies humaines ou aux traits adaptatifs. Les approches de séquençage de nouvelle génération (NGS), y compris le séquençage de l'exome entier (WES), offrent des opportunités sans précédent pour découvrir de nouveaux variants impliqués dans la sensibilité ou la résistance à une pathologie. Le principe de base du WES est le séquençage des régions codantes, grâce auquel des sondes ADN sont utilisées pour s'hybrider avec la partie codante du génome. Après le séquençage, des millions de séquences d'ADN, appelées reads, sont alignées sur un génome de référence et sont analysées par différents outils, avec l'objectif d'identifier de nouvelles cibles pertinentes pour la question scientifique posée. Depuis leur création, les méthodes NGS, y compris le WES, ont fourni une énorme quantité de données qui posent des défis considérables pour leur analyse et l'interprétation des résultats correspondants. Ces avancées technologiques nécessitent de plus en plus le développement d'approches méthodologiques sophistiquées, générant ainsi de nouvelles questions de recherche afin d'optimiser l'analyse de ces données. Ainsi, les volumes de données d'exome accumulées au fil des ans permettent de poser des questions scientifiques nouvelles. Ma thèse a porté sur ces aspects.

Tout d'abord, j'ai développé une approche computationnelle qui permet de filtrer les variants génétiques faussement positifs (FP) dans les données d'exome qui ne peuvent pas être éliminés avec les approches bioinformatiques traditionnelles. Les analyses informatiques des exomes de patients humains visent à filtrer autant de F que possible, sans supprimer les véritables mutations pathogènes. Cela implique de comparer l'exome du patient avec des bases de données publiques pour supprimer les variants rapportés incompatibles avec la prévalence de la maladie, le mode d'hérédité ou la pénétrance clinique. Cependant, des variants fréquents dans une cohorte donnée d'exomes, mais absents ou rares dans les bases de données publiques, ont également été rapportés et traités comme des FP, sans exploration rigoureuse. Nous avons rassemblé ces variants et nous les appelons la «liste noire». Cette liste noire n'a pas éliminé les mutations pathogènes connues des exomes de 129 patients et a diminué le nombre de FP restant dans un panel de 3 104 exomes d'une médiane de 62%. Nous avons démontré que les variants sur liste noire ne diffèrent pas des variants non sur liste noire en termes de scores de qualité et de métriques de prédiction des effets délétères. En outre, les approches de filtrage standard, telles que le classificateur Random Forest (RF) et le recalibrage du score de qualité des variantes (VQSR), n'ont pas réussi à faire la distinction entre les variantes sur liste noire et les variants vraiment positives (VP), suggérant que la méthode de la liste noire peut être utilisée en parallèle avec ces filtrages. L'efficacité de l'élimination d'une grande proportion de FP a été reproduite dans trois panels indépendants. Ensuite, nous avons caractérisé les variants de la liste noire de manière informatique et expérimentale. La plupart des variants de la liste noire correspondaient à de faux signaux générés par l'assemblage incomplet du génome de référence, la localisation dans les régions de faible complexité du génome et un mauvais traitement bioinformatique. La liste noire peut être utilisée comme une approche rapide et efficace pour filtrer les FP à partir des données d'exome.

Dans la suite de ce travail, j'ai découvert qu'un sous-ensemble spécifique de variants de la liste noire était en déséquilibre de Hardy-Weinberg (HW), un principe génétique fondamental de la population affirmant que les fréquences des allèles et des génotypes dans une population donnée sont constantes de génération en génération, en l'absence d'influences évolutives (ex: pas de migration, pas de mutation, pas de sélection naturelle). Dans le cas le plus simple d'un locus avec deux allèles, l'équilibre HW est utilisé pour estimer le nombre attendu de génotypes pour les génotypes homozygotes de type sauvage, hétérozygotes et homozygotes sur la base des fréquences alléliques. Ces génotypes attendus sont ensuite comparés aux génotypes observés dans la population pour évaluer si le locus donné est en équilibre ou en déséquilibre HW. Étant donné que les conditions d'absence d'influences évolutives sont généralement considérées comme valables, les écarts par rapport à l'équilibre HW dans les échantillons témoins ont été traditionnellement considérés comme indicatifs d'erreurs techniques. Ce principe a été à l'origine utilisé comme critère de filtrage dans les études de génotypage à grande échelle (par exemple les études d'association à l'échelle du génome ou GWAS) et utilisé dans les études d'exomes sans investigation rigoureuse. Sur la base des résultats initiaux du projet de liste noire, nous étudions actuellement l'équilibre HW de manière systématique et à plus grande échelle pour déterminer si l'équilibre HW pourrait être utilisé non seulement pour

détecter des erreurs techniques, mais aussi pour informer sur des phénomènes importants liés à la génétique des populations. Nos données préliminaires se concentrant sur les variants avec un excès ou une perte d'homozygotes pour l'allèle mineur ont révélé des variants candidats prometteurs qui pourraient indiquer une protection (par exemple dans FUT2, SMN2) ou un désavantage (par exemple dans FANCD2) à la maladie.

Enfin, j'ai abordé la question de la détection des variants du nombre de copies (CNV) dans les données WES. Les CNV sont des réarrangements déséquilibrés, couvrant classiquement plus de 50 paires de bases (pb), qui augmentent ou diminuent le nombre de copies de régions d'ADN spécifiques. Les méthodes basées sur le WES pour la détection des CNV ont rencontré un succès limité, principalement en raison de la nature des protocoles d'enrichissement ciblés. Les méthodes NGS courantes utilisent des points d'arrêt, les régions dans lesquelles les réarrangements se produisent, pour détecter les CNV. En revanche, le WES se concentre sur des cibles génomiques non contiguës (les exons), et la plupart des points de rupture ne sont pas séquencés. Par conséquent, les approches actuelles basées sur WES pour détecter les CNV utilisent la couverture comme un proxy pour les informations sur le nombre de copies. Cependant, étant donné le problème connu de non-uniformité de la couverture, les méthodes basées sur WES sont confrontées à des défis importants. Ce problème est encore exacerbé dans les panels de laboratoire typiques, qui comprennent des données d'exome générées au fil du temps, souvent dans des conditions différentes. Dans ma thèse, j'ai développé HMZDelFinder-opt, un algorithme qui permet d'identifier des délétions partielles d'exons homozygotes et hémizygotes. En utilisant HMZDelFinder_opt avec à la fois des délétions pathogènes validées et des données simulées, nous avons démontré que la sélection a priori d'un jeu de contrôle de référence avec un profil de couverture similaire à celui de l'échantillon WES étudié réduisait le nombre de délétions détectées, tout en améliorant le classement des véritables délétions homozygotes. HMZDelFinder_opt permet également l'étude des délétions s'étendant sur moins d'un exon, en fournissant le premier outil pour l'identification systématique des délétions partielles d'exon. HMZDelFinder_opt est une approche rapide et puissante pour détecter les délétions de HMZ, en particulier les délétions partielles d'exons, dans des panels de laboratoire, qui sont généralement hétérogènes.

Les nouvelles méthodes développées dans cette thèse fourniront à la communauté scientifique des outils utiles pour faciliter l'analyse des données WES. L'utilisation d'approches WES a considérablement alimenté la découverte de la base génétique de maladies rares (et principalement monogéniques). Cependant, il est toujours difficile d'identifier efficacement tous les différents types de variations génétiques (SNP, Indels et CNV) à partir des données d'exome, et aussi de les réduire à une courte liste de variants candidats pour l'inspection manuelle et la validation fonctionnelle. Un défi majeur dans l'analyse des données d'exome est dû à l'évolution continue de la technologie (séquençage et outils bio-informatiques correspondants) qui se traduit à la fois par des ensembles de données d'exomes hétérogènes avec des fluctuations extrêmes de la couverture et des faux signaux dépendants de la technologie. Nous démontrons que l'approche de la liste noire peut détecter de tels FP et les filtrer de manière rapide, efficace et personnalisable. Cette approche peut être utilisée en combinaison avec d'autres méthodes de pointe (telles que les outils VQSR et RF) car nous constatons qu'elles sont mutuellement exclusives dans la capture des FP. L'autre aspect critique dans l'analyse des données d'exome est la capacité des outils bioinformatiques actuels à identifier tout le spectre des variations génétiques.. Nous avons proposé une méthode (HMZDelFinder-opt) qui résout ces problèmes et sert de premier outil pour l'identification systématique des délétions partielles d'exons. Le génome humain contient environ 235 000 exons, dont environ 20% sont supérieurs à 200 pb. Par conséquent, HMZDelFinder_opt rend possible la découverte systématique de délétions HMZ actuellement inconnues dans ~ 47 000 exons qui ne sont pas détectables avec d'autres outils.

Une conséquence de l'adoption généralisée des approches WES en génétique humaine est l'accumulation rapide de données d'exome. Plusieurs groupes, dont le Broad Institute, ont entrepris la collecte et l'harmonisation de milliers de données d'exome dans le but de fournir aux chercheurs un référentiel public qui pourrait être utilisé pour faciliter l'interprétation médicale et fonctionnelle de la variation génétique. Ces grands ensembles de données sont non seulement aujourd'hui essentiels dans l'analyse des exomes (par exemple: pour évaluer les fréquences dans la population générale), mais peuvent également être réutilisés pour permettre une enquête systématique sur des questions théoriques spécifiques, ce qui n'était pas possible auparavant en raison de la puissance statistique limitée.

Un exemple est l'étude de l'équilibre HW que j'ai commencé à aborder au cours de ma thèse. Bien que nos résultats préliminaires sur les variants candidats prometteurs qui pourraient sous-tendre la sensibilité (par exemple FANCD2) ou la résistance (par exemple: FUT2 et SMN2) à certaines pathologies nécessitent des recherches et des preuves plus approfondies, ils justifient fortement l'utilisation de données d'exome de grande taille et facilement disponibles.

Le travail décrit dans cette thèse se prête à un certain nombre de directions futures. Par exemple, dans HMZDelFinder-opt, nous nous sommes concentrés jusqu'à présent sur les délétions homozygotes dans les chromosomes autosomiques et les délétions hémizygotes chez les hommes sur le chromosome X; dans les travaux futurs, il sera intéressant d'adapter HMZDelFinder-opt à la détection des délétions hétérozygotes. Étant donné que la couverture des délétions hétérozygotes devrait être la moitié de celle sans délétion, cette direction impliquera probablement le réglage fin du seuil pour appeler une délétion et l'inclusion d'autres mesures, en plus de la couverture. Une approche similaire pourrait également être appliquée à la détection des duplications. En outre, le WGS devient de plus en plus attrayant comme alternative, en raison de la couverture plus homogène, du coût en baisse constante et de la possibilité d'étudier des variants situés en dehors des régions codant les protéines du génome. Il sera donc intéressant d'évaluer et d'adapter les méthodes proposées ici aux données WGS. Enfin, pour le projet HW, il sera essentiel d'utiliser les données WGS pour confirmer et reproduire les résultats, et il sera intéressant d'étudier plus en profondeur les événements de sélection. Collectivement, ces projets abordent des sujets non encore examinés et promettent d'aider à la découverte de nouveaux déterminants causaux de maladies ou de traits humains. Ils jetteront également les bases de recherches futures pour étudier les classes spécifiques de rôle de la variation génétique (c'est-à-dire: délétions partielles et variants en fort excès / déplétion d'homozygotes pour l'allèle mineur) dans les maladies humaines.

LIST OF PUBLICATIONS

Publications in the period of the thesis (* indicates equal contribution)

First author:

1. Maffucci P*, **Bigio B***, Rapaport F, Cobat A, Borghesi A, Lopez M, Patin E, Bolze A, Shang L, Bendavid M, Scott EM, Stenson PD, Cunningham-Rundles C, Cooper DN, Gleeson JG, Fellay J, Quintana-Murci L, Casanova JL, Abel L, Boisson B, Itan Y. Blacklisting variants common in private cohorts but not in public databases optimizes human exam analysis. Proc Natl Acad Sci U S A 2018 doi: 10.1073/pnas.180840311

2. **Bigio B**, Seeleuthner Y, Kerner K, Migaud M, Rosain J, Boisson B, Nasca C, Puel A, Bustamante J, Casanova JL, Abel L, Cobat A. Detection of homozygous and hemizygous partial exon deletions by whole-exome sequencing. Under review.

Collaborative papers:

3. Zhang P, **Bigio B**, Rapaport F, Zhang SY, Casanova JL, Abel L, Boisson B, Itan Y. PopViz: a webserver for visualizing minor allele frequencies and damage prediction scores of human genetic variations. Bioinformatics. 2018;34(24):4307-4309. doi: 10.1093/bioinformatics/bty536

4. Requena D, Maffucci P, **Bigio B,** Shang L, Abhyankar A, Boisson B, Stenson PD, Cooper DN, Cunningham-Rundles C, Casanova JL, Abel L, Itan Y. CDG: An Online Server for Detecting Biologically Closest Disease-Causing Genes and its Application to Primary Immunodeficiency. Front Immunol. 2018; 9:1340. doi: 10.3389/fimmu.2018.01340.

5. Kerner G, Bouaziz M, Cobat A, **Bigio B**, Timberlake AT, Bustamante J, Lifton RP, Casanova JL, Abel L. A genome-wide case-only test for the detection of digenic inheritance in human exomes. PNAS 117(32):19367-19375. doi: 10.1073/pnas.1920650117

6. Zhang Q, Bastard P*, Liu Z*, Le Pen J*, Moncada-Velez M*, Chen J*, Ogishi M*, Sabli IKD*, Hodeib S*, Korol C*, Rosain J*, Bilguvar K*, Ye J, Bolze A*, **Bigio B***, et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science. 370(6515):eabd4570.

7. Bastard P, Rosen LB, Zhang Q, Michailidis E, Hoffmann HH, Zhang Y, Dorgham K, Philippot Q, Rosain J, Béziat V, Manry J, Shaw E, Haljasmägi L, Peterson P, Lorenzo L, Bizien L, Trouillet-Assant S, Dobbs K, de Jesus AA, Belot A, Kallaste A, Catherinot E, Tandjaoui-Lambiotte Y, Le Pen J, Kerner G, **Bigio B**, et al. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. Science. 370(6515):eabd4585.

8. Drutman SB , Mansouri D, Mahdaviani SA, Neehus A, Hum D , Bryk R, Hernandez N, Belkaya S, Rapaport F , **Bigio B**, et al. Fatal Cytomegalovirus Infection in an Adult With Inherited NOS2 Deficiency. N Engl J Med. 2020 382(5):437-445

9. Li J, Ritelli M, Ma CS, Rao G, Habib T, Corvilain E, Bougarn S, Cypowyj S, Grodecká L, Lévy R, Béziat V, Shang L, Payne K, Avery DT, Migaud M, Boucherit S, Boughorbel S, Guennoun A, Chrabieh M, Rapaport F, **Bigio B**, et al. Chronic Mucocutaneous Candidiasis and Connective Tissue Disorder in Humans With Impaired JNK1-dependent Responses to IL-17A/F and TGF-β . Sci Immunol 2019 4(41).

10. Nasca C, Menard C*, Hodes G*, **Bigio B***, Pena C, Lorsch Z, Zelli D, Ferris A, Kana V, Purushothaman I, Dobbin J, Nassim M, DeAngelis P, Merad M, Rasgon N, Meaney M, Nestler EJ, McEwen BS, Russo SJ. Multidimensional Predictors of Susceptibility and Resilience to Social Defeat Stress. Biol Psychiatry. 2019 86(6):483-491

11. Lim HK, Huang SXL, Chen J, Kerner G, Gilliaux O, Bastard P, Dobbs K, Hernandez N, Goudin N, Hasek ML, García Reino EJ, Lafaille FG, Lorenzo L, Luthra P, Kochetkov T, **Bigio B**, Boucherit S, Rozenberg F, Vedrinne C, Keller MD, Itan Y, García-Sastre A, Celard M, Orange JS, Ciancanelli MJ, Meyts I, Zhang Q, Abel L, Notarangelo LD, Snoeck HW, Casanova JL, Zhang SY. Severe influenza pneumonitis in children with inherited TLR3 deficiency. J Exp Med. 2019 2;216(9):2038-2056

12. Martínez-Barricarte R, Markle JG, Ma CS, Deenick EK, Ramírez-Alejo N, Mele F, Latorre D, Mahdaviani SA, Aytekin C, Mansouri D, Bryant VL, Jabot-Hanin F, Deswarte C, Nieto-Patlán A, Surace L, Kerner G, Itan Y, Jovic S, Avery DT, Wong N, Rao G, Patin E, Okada S, **Bigio B,** et al. Human IFN-γ immunity to mycobacteria is governed by both IL-12 and IL-23. Sci Immunol. 2018 ;3(30). pii: eaau6759. doi: 10.1126/sciimmunol.aau6759.

13. Nasca C, Watson-Lin K, **Bigio B**, Robakis TK, Myoraku A, Wroolie TE, McEwen BS, Rasgon N. Childhood trauma and insulin resistance in patients suffering from depressive disorders. Exp Neurol. 2019 315:15-20. doi: 10.1016/j.expneurol.2019.01.005.

14. Guérin A, Kerner G, Marr N, Markle JG, Fenollar F, Wong N, Boughorbel S, Avery DT, Ma CS, Bougarn S, Bouaziz M, Béziat V, Della Mina E, Oleaga-Quintas C, Lazarov T, Worley L, Nguyen T, Patin E, Deswarte C, Martinez-Barricarte R, Boucherit S, Ayral X, Edouard S, Boisson-Dupuis S, Rattina V, **Bigio B**, et al. IRF4 haploinsufficiency in a family with Whipple's disease. Elife. 2018 ;7. pii: e32340. doi: 10.7554/eLife.32340.

15. Boisson B, Honda Y, Ajiro M, Bustamante J, Bendavid M, Gennery AR, Kawasaki Y, Ichishima J, Osawa M, Nihira H, Shiba T, Tanaka T, Chrabieh M, **Bigio B**, Hur H, Itan Y, Liang Y, Okada S, Izawa K, Nishikomori R, Ohara O, Heike T, Abel L, Puel A, Saito MK, Casanova JL, Hagiwara M, Yasumi T. Rescue of recurrent deep intronic mutation underlying cell type-dependent quantitative NEMO deficiency. J Clin Invest. 2019; 129(2):583-597. doi: 10.1172/JCI124011.

16. Robakis TK, Watson-Lin K, Wroolie TE, Myoraku A, Nasca C, **Bigio B**, McEwen B, Rasgon NL. Early life adversity blunts responses to pioglitazone in depressed, overweight adults. Eur Psychiatry. 2019;55:4-9. doi: 10.1016/j.eurpsy.2018.09.009.

17. Nasca C, **Bigio B**, Lee FS, Young SP, Kautz MM, Albright A, Beasley J, Millington DS, Mathé AA, Kocsis JH, Murrough JW, McEwen BS, Rasgon N. Acetyl-l-carnitine deficiency in patients with major depressive disorder. Proc Natl Acad Sci U S A. 2018;115:8627-8632.

Index

1 Introduction

The purpose of this chapter is to introduce the reader to the context in which the thesis evolves, that is the study of genetic variants and their deleterious or protective role in human diseases, particularly rare diseases. The chapter also summarizes the state-of-the-art in the field, highlighting not only the opportunities, but also some of the challenges that my thesis tackled. Finally, the chapter ends with a short synopsis of the main objectives of the thesis.

1.1 Human genetics and genetic variation

Human genetics entails the study of the human genome and its basic unit, the gene (1, 2). Nowadays the field encompasses a variety of subdisciplines, including classical or Mendelian genetics, which focuses on how specific genes are involved in the etiology of human diseases or traits (1-3). It is fascinating noting that theories and studies in human genetics originate from simple observations that date back to ancient times. From Hippocrates's "pangenesis" theory and Aristotle's basic inheritance principle, classical genetics had a major shift with the studies of inheritance in garden peas by Gregor Mendel(1) (Figure 1). A very important step in the development of this discipline came with Archibald Garrod's demonstration of the first known Mendelian inborn error of humankind, alkaptonuria, in 1902 (4). This breakthrough discovery suggested, for the first time, that genetic traits are a predisposing factor for human diseases, and paved the way for decades of research in understanding the genetic basis of human traits and diseases (5).

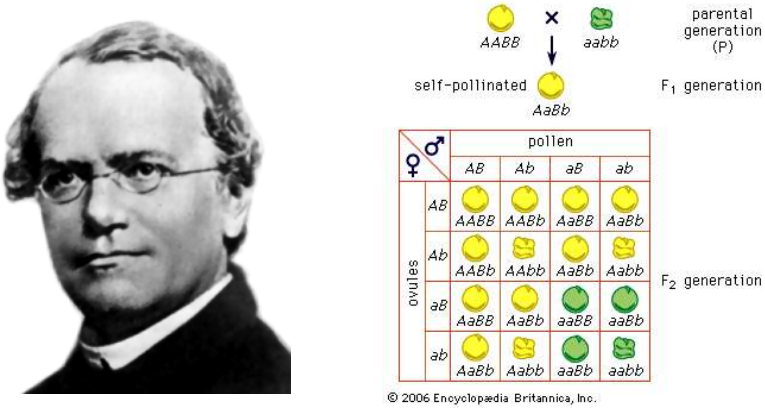


Figure 1. Classical or Mendelian genetics.
The studies of inheritance in garden peas by Gregor Mendel paved the way to understand how specific genes contribute to human diseases or traits.

Genetic traits are transmitted within every organism, including the human species (1, 2). The human genome includes 23 pairs of chromosomes: 22 pairs of homologous chromosomes and one pair of sex chromosomes (XX in women or XY in men). Each chromosome contain several genes, which are sequences of DNA that code for specific proteins. On all homologous chromosome pairs, there are two forms of the same gene that are known as alleles, with one allele inherited from each parent. Each pair of alleles represents a specific genotype for a given gene. In humans, the vast majority of the genome is the same across individuals. The human genetic variation accounts for less than 1% of each person's DNA and contributes to the huge phenotypic variation between individuals (6). The non-mutant form of a gene, encoding the normal genetic function, is called wild-type (WT) allele. For a given locus/loci, individuals can be homozygous WT (both non-mutant alleles referred, as AA), heterozygous (one mutant allele and one non-mutant allele, referred as Aa) or homozygous alternate (both mutant alleles, referred as aa).

Genetic variation occurs in many forms (Figure 2), which include base differences known as single nucleotide polymorphisms (SNPs), small insertions/deletions (Indels), or large variation in structure of chromosomes (structural variations), the latter including differences in the number of copies of a given sequence or gene (copy number variations or CNVs) (7-10). Genetic variation is very common in humans, with a typical difference between the genomes of two individuals being ~4 million base pairs, although most genetic variation is expected to have no consequence on the phenotype (6, 7). Each given variation may affect a different proportion of individuals, with frequencies (minor allele frequency or MAF) ranging from 0 (private variation) to 50%. Occurrence and frequency of variations vary also as function of the population. Ethnic groups such as Africans have a greater gene diversity (11, 12) than other more homogeneous ethnicities, including Europeans and Asians. In addition to within-population variability, there is also between-population genetic variability. For example, populations that are more geographically and ancestrally remote tend to differ more, and a given variant that is common in one geographical or ethnic group may be much rarer in another (13, 14). A small proportion of genetic variations has been linked to a given phenotypic trait, and only a few variants are known causal determinants of human diseases (7).

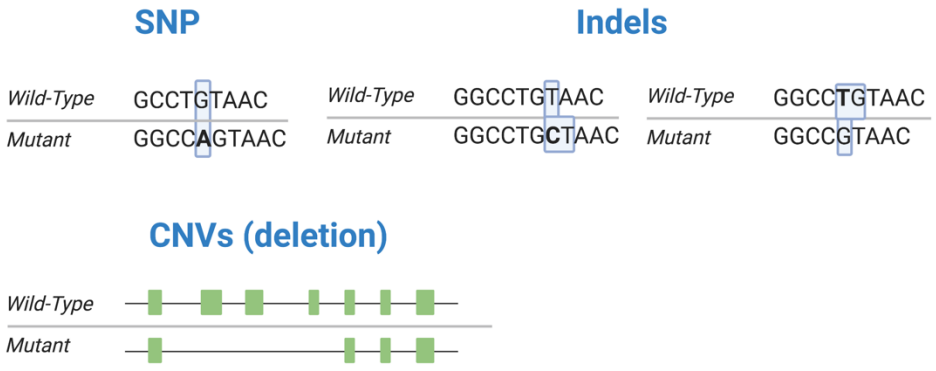


Figure 2. Types of genetic variations

Genetic variation occurs in many forms, including single nucleotide polymorphisms or SNPs, small insertions/deletions or Indels, copy number variations or CNVs (differences in the number of copies of a given sequence/gene).

An increasing number of rare diseases have now an identified genetic cause (3). Albeit not exclusively, rare diseases are often attributable to single gene mutations in a Mendelian (monogenic with complete penetrance, i.e., all of the individuals in a population who carry a specific genotype express the corresponding phenotype) or non-mendelian (monogenic with incomplete penetrance) manner (15, 16). The MAF of these mutations is typically less than 1% and depends on the prevalence, mode of inheritance, and clinical penetrance of the disease (17). As often occurs in science, methodological advances led the way to discover monogenic inborn errors underlying a variety of rare human conditions. Most notably, the discovery of the chain-termination technique to sequence DNA by Sanger and the introduction of positional cloning in 1986 have significant fueled gene discovery (18-21). However, these techniques were mainly based on a candidate approach and remained not scalable. Only in the 2000s with the introduction of next generations sequencing (NGS), DNA sequencing became accessible at large scale (18, 22). Today, NGS-based approaches are the state-of-the-art to discover genetic mutations underlying human traits or diseases.

1.2 Whole exome sequencing (WES): opportunities and challenges

The advent of NGS approaches has revolutionized genomic research, allowing interrogation of the genome at single-base resolution with limited time and affordable costs, and posing the basis for personalized medicine (18, 23). In parallel, this technological advance has resulted in important initiatives for the field of classical genetics, including the launch of the Human Genome Project, that culminated in the release of the sequencing of the entire human genome

(24). Unsurprisingly, these innovations have been paralleled by an accelerating pace of discoveries in genetics of human diseases, particularly of mutations underlying rare diseases (Figure 3) (3). The two main NGS approaches are whole genome and whole exome sequencing (WGS and WES, respectively) that differ for sequencing strategy (entire genome versus coding regions); in addition, costs and computational load are higher in WGS than WES (18, 25). The basic principle of the NGS approaches is the concurrent generation of millions of DNA sequences, known as *reads*, that are then aligned to a reference genome. Following alignment, *variant calling* analysis is used to determine the portions of the genome (e.g.: SNPs, indels, CNVs) that deviate from the reference genome. Next, investigators employ a variety of downstream analyses, including at patient- and cohort-level, to answer specific scientific questions with the ultimate goal to identify novel candidate variants or genes for the phenotype under study.

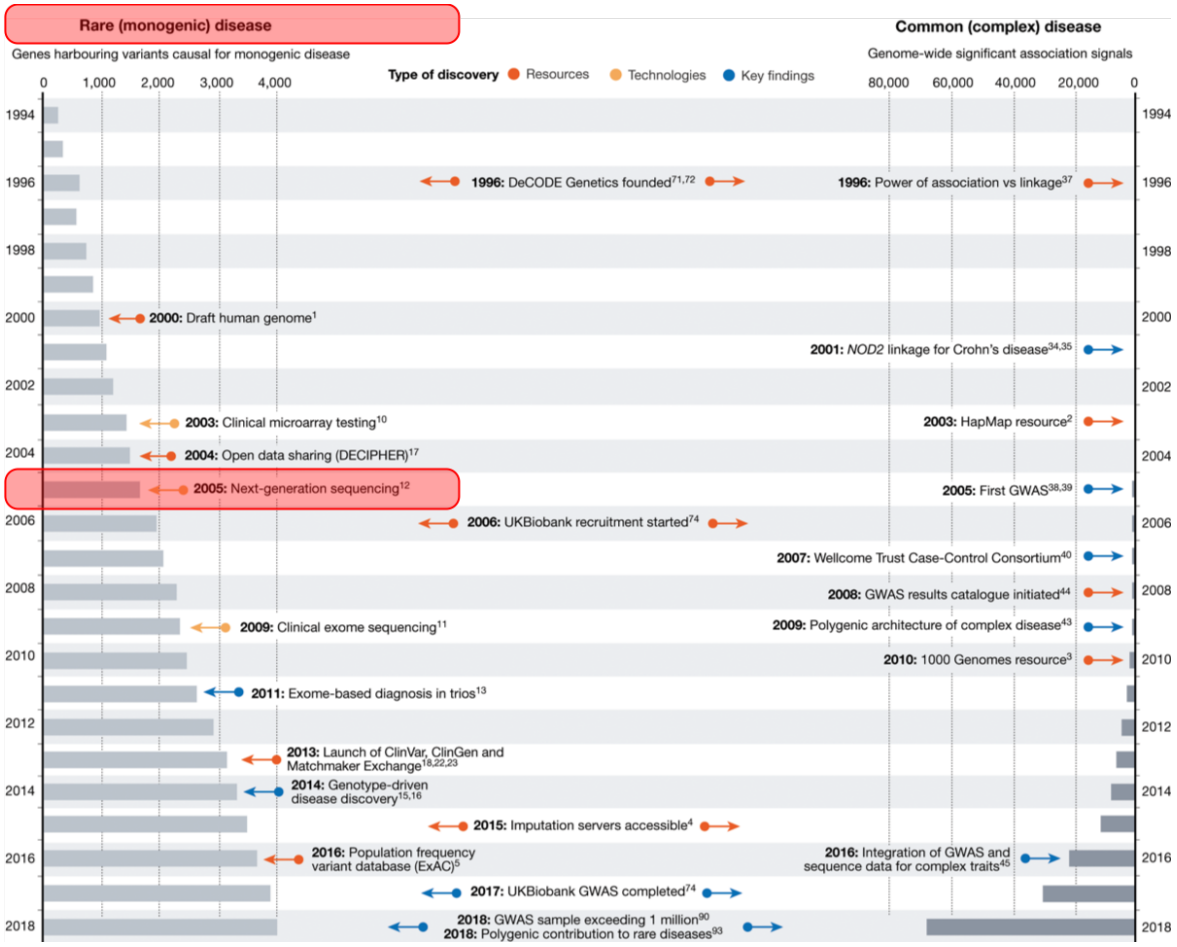


Figure 3. NGS and the growth in the discovery of disease-associated genetic variations
The advent of NGS approaches has revolutionized genomic research, allowing interrogation of the genome at single-base resolution with limited time and affordable costs, accelerating pace of discoveries in genetics of human diseases (especially rare diseases) and ultimately posing the basis for personalized medicine. Figure from Claussnitzeret al (3).

Whole exome sequencing (WES) is an NGS approach that has been recently optimized for sequencing of coding regions (~200,000 exons of the human genome)(26-28). In WES, a set of DNA probes or baits, called *capture kit*, is used to hybridize with the protein-coding portion of the genome, isolating it from the non-coding portion, in order to selectively capture the coding regions of the genome (29, 30). This target enrichment strategy is then followed by sequencing procedures and bioinformatic pipelines (29). After the alignment and calling procedures are performed, variants are annotated with various level of information (e.g., gene and gene function, predictive damage, quality metrics), that are used for the interpretation of the resulting detailed catalogues of genetic variations (31) (Figure 4). The candidate causing mutations detected by exome sequencing are validated by Sanger sequencing. The exome accounts for ~2% of the genome thus limiting the sequencing load, time and cost as compared to whole genome sequencing (18). Furthermore, it has been shown that the vast majority of exonic variants is evolutionary recent, rare and enriched for deleterious alleles, thus likely contributing significantly to phenotypic variation and diseases (32). Lastly it remains difficult to interpret variants lying outside the protein-coding regions of the genome (25). For these reasons, WES is nowadays the reference method used to discover genetic causes of rare diseases.

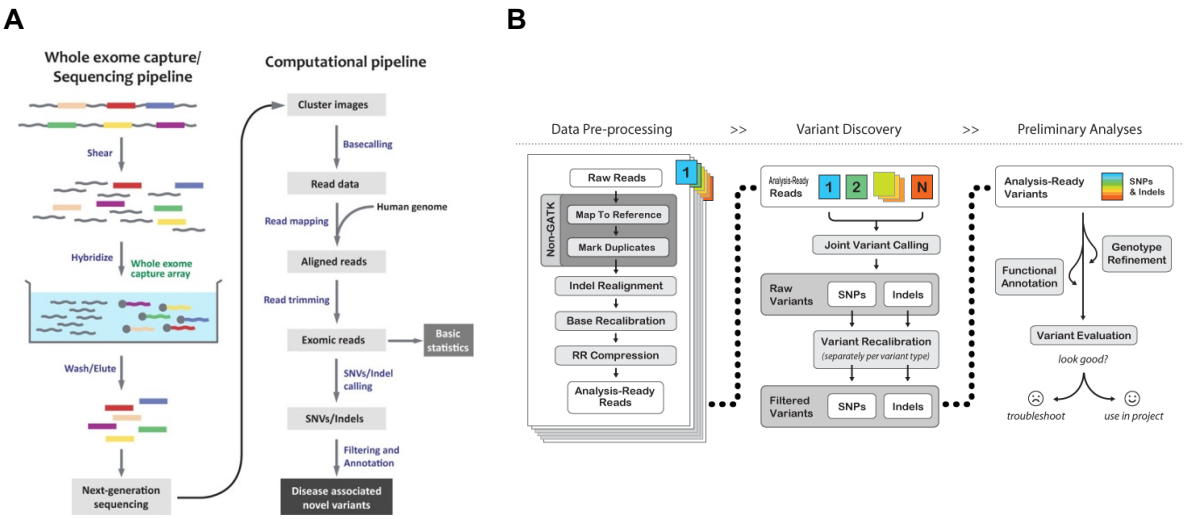


Figure 4. General workflow for whole exome sequencing (WES) approaches

(A) *Whole exome sequencing (WES) is an NGS approach optimized for sequencing of coding regions (~200,000 exons of the human genome). A set of DNA probes or baits, called capture kit, is used to hybridize with the protein-coding portion of the genome, isolating it from the non-coding portion, in order to selectively capture the coding regions of the genome. This target enrichment strategy is then followed by sequencing procedures and bioinformatic pipelines. After the alignment and calling procedures are performed, variants are annotated with various level of information (e.g., gene and gene function, predictive damage, quality metrics), that are used for the interpretation of the resulting detailed catalogues of genetic variations.* ***(B)*** *The GenomeAnalysisToolkit (GATK) Best practice pipeline is the state-of-the-art for the analysis of SNPs and Indels. Figures from Goh et al.(29) and from the GATK(33) website (https://gatk.broadinstitute.org/hc/, tab pipeline).*

1.2.1 False positive variants and filtering tools: the Blacklist

Despite its extensive use, the analysis of WES data still presents considerable challenges (18, 31). The mean number of exonic coding variants per individual relative to the reference human genome is about 100,000-150,000, of which 20,000 are high quality variants (25, 34). Thus, there is a real need for computational tools to effectively filter out as many false positive (FPs) as possible and also to prioritize the remaining true variants to efficiently separate nonpathogenic variants from candidate disease-causing mutations (31). A number of technical issues affects the sequencing data thus reducing the reliability of the called variants with consequences on the proportions of erroneous calls (FPs), and missed variants (false negative or FN). In addition, data analysis needs to address variants that are true signals but are not actually causative for the disease or phenotype under investigation. Typical prioritization approaches for the latter aspect include variant- and gene-levels tools that address population frequency, conservation and predicted damaging effects (31) (Figure 5); for example, the first step usually involves comparing the individual's exome with public databases to remove reported variants inconsistent with disease prevalence, mode of inheritance, or clinical penetrance (17).

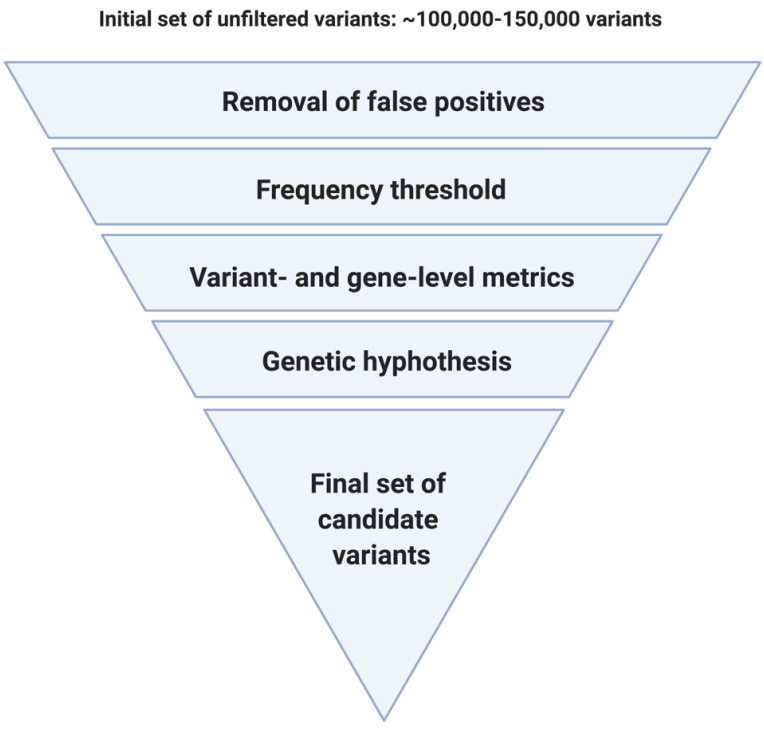


Figure 5. Filtering strategy of variants in WES data
Typical filtering strategy in exome data includes computational tools to filter out as many false positive (FPs) as possible and also to prioritize the remaining true variants to efficiently separate nonpathogenic variants from candidate disease-causing mutations.

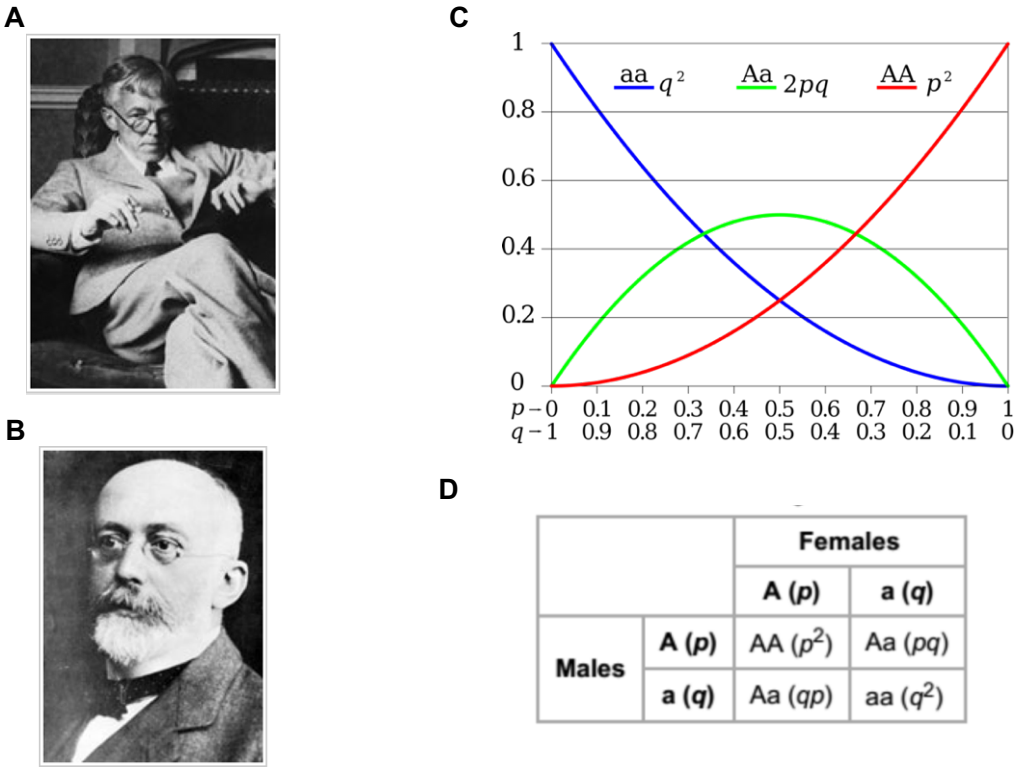
Among the technical issues affecting the reliability of exome data, the most important is the non-uniformity of coverage (35). The coverage is defined as the number (or median number) of reads that cover a given nucleotide (or a given portion of the exome). Exome data of good quality have a typical median coverage of 40-60X. Within an exome, the coverage may span from abnormally low covered regions (~10X) to 'hot spots' with abnormally high coverage (>90X). One of the factors determining this unevenness of coverage is the enrichment strategy employed in the first steps of the WES protocol (36, 37). For example, a region dense with SNPs can interfere with the capture process, as the hybridization of the enrichment of the probes may not occur as efficiently. The coverage issue is further complicated by the between-exome differences generated when using different target enrichment strategies(38-40). While the basic preparation steps are similar among the various platforms, there are major differences in the design of the DNA probes, including selection of target genomic regions, relative location of the probes (e.g.: overlapping, tiling or gapped probes), and the exome capture mechanisms(38, 41). Another major technical issue, which also affects the exome coverage, originates from the inherent structure of the human genome. The presence of low-complexity regions in the genome (i.e., repeats of single amino acids or short amino acid motifs) greatly influences the exome capture and calling process resulting in bioinformatic misprocessing. Furthermore, the human reference genome is still not completely assembled and annotated, thus complicating the alignment procedures(42).

Several computational approaches have been proposed to alleviate these technical issues with the goal to filter FP variants(31). Two well-known methods are the locus-specific variant quality score recalibration (VQSR) approach from the GATK suite(33, 43) and the variant-specific random forest classifier used in the Genome Aggregation Database (gnomAD) (44). They are both machine learning-based methods which use a clustering score to determine whether a called variant is true. Variants can also be flagged based on hard filtering, which use hard cutoffs for specific quality metrics. These computational approaches are not mutually exclusive but are often used in combination, although machine-learning approaches are time-consuming and are based on the entire sample, thus presenting important limitations in typical laboratory panels that constantly evolve over time. **The first part of my thesis's work** focused on the development of a time-effective method to filter FP variants that could not be filtered with other available tools(45). These FP variants were systematically investigated, characterized and collected for use as a 'blacklist' in WES analysis. As elaborated in chapter 2, these variants were mostly artifacts generated as a consequence of incomplete reference genome

assembly, location in low-complexity regions or deviation from Hardy-Weinberg (HW) equilibrium, a basic principle of genetics discussed in the next paragraph. Based on these initial findings, we decided to study HW equilibrium systematically and at a larger scale.

1.2.2 Hardy-Weinberg disequilibrium: opportunities beyond its utility as a filtering tool

The occurrence of Hardy-Weinberg (HW) disequilibrium is another metric that signals FP variants (44, 46, 47). The HW law or equilibrium is a basic principle of genetics; it states that allele and genotype frequencies in a given population are constant from generation to generation, in the absence of evolutionary influences (e.g.: no migration, no mutation, no natural selection, very large population and random mating)(48, 49). In the simplest case of a locus with two alleles, the HW equilibrium is used to estimate the expected number of genotypes for homozygous wild type, heterozygous and homozygous alternate genotypes based on the allele frequencies (Figure 6).



		Females	
		A (*p*)	**a (*q*)**
Males	**A (*p*)**	AA (p^2)	Aa (pq)
	a (*q*)	Aa (qp)	aa (q^2)

Figure 6. Hardy-Weinberg equilibrium

(A-B) *Hardy and Weinberg independently elaborated the HW law or equilibrium. It is a basic principle of genetics stating that allele and genotype frequencies in a given population are constant from generation to generation, in the absence of evolutionary influences.****(C-D)*** *In the simplest case of a locus with two alleles, the HW equilibrium is used to estimate the number of genotypes for homozygous wild type (AA), heterozygous (Aa) and homozygous alternate genotypes (aa) based on the allele frequencies (p and q).*

These expected genotypes are then compared with the genotypes observed in the population to assess if the given locus is in HW equilibrium or disequilibrium. Given the conditions of absence of evolutionary influences are usually considered to hold, deviations from HW equilibrium (or HW disequilibrium) in control samples have been traditionally considered indicative of technical errors (50-52). This principle was originally employed as filtering criterion in large-scale genotyping studies (e.g. genome-wide association studies or GWAS) and "lent" to exome studies without rigorous investigation. In WES approaches, variants in HW disequilibrium due to very extreme excess heterozygosity (i.e.: the observed number of heterozygous genotypes is statistically greater than the expected number of heterozygous genotypes) are filtered in large population databases, including gnomAD, the largest available dataset that includes 125,748 exomes (44, 47). This assumption is reasonable because it has been shown that variants with extreme heterozygosity are enriched in low-complexity regions of the genome, which are especially prone to sequencing and alignment errors.

While HW disequilibrium may truly indicate technical errors in specific circumstances, some studies cautioned against blinded exclusion of loci deviating from HW equilibrium that could instead signal causative mutations. For example, a population-based study designed to investigate the causes of deviation from HWE failed to find an explanation for about 30% of loci found to be in disequilibrium, suggesting there may be other reasons beyond actual errors to cause deviations from HWE(52). Another report investigating HW equilibrium in a Japanese sample of 104 individuals from 1000Genome (a large dataset collected by EMBL-EBI) suggested that HW disequilibrium in NGS data seems to be a major indicator for CNV(53). In line with these findings, a separate study has used deviations from HW equilibrium, and particularly loss of heterozygosity, as indicator of a specific class of CNVs, common deletions(54). Lastly, a recent report investigated HW disequilibrium in the whole set of exome data in gnomAD (cases and controls) (55). Authors mainly focused on excess heterozygosity with the main objective to identify variants and genes associated with autosomal recessive disorders. With the exception of very few classical examples (rs334 in *HBB*, which causes recessive sickle cell disease in homozygous status and confers protection from malaria in heterozygous status; rs1801178 in *CFTR*, which causes recessive cystic fibrosis disease in homozygous status and is hypothesized to be protective from cholera in heterozygous status), this study did not find candidates mutations. Furthermore, the authors recognized that the significance cutoff used in their study was lenient (0.05 without correction for multiple testing) in contrast to previous studies(53), and therefore the results from this study should be taken

with caution. Nevertheless, this last study strongly supports the timeliness of our project. **A part of my thesis's work,** for which I present preliminary results in the last chapter, was aimed to investigate the distribution of genotypes across populations in gnomAD, the largest available dataset of exome data, with the two-fold goal to determine variants that are in true HW disequilibrium possibly underlying susceptibility or resistance to disease, and to investigate the underlying origin in relation to specific population events (e.g.: natural selection)(56-58).

1.2.3 An overlooked type of variants in WES data: copy number variations (CNVs)

As mentioned above, prior studies have suggested HW disequilibrium as a major indicator of CNVs (53, 54). For example, common deletions (a specific class of CNVs) results in an apparent loss of heterozygosity (and, by symmetry, excess of homozygosity) thus violating HW equilibrium. CNVs are unbalanced rearrangements, classically covering more than 50 base pairs (bp), that increase or decrease the number of copies of specific DNA regions (59, 60). There is growing evidence to implicate CNVs in disease states (59, 61, 62). While other genetic variants, such as small variations (SNPs and Indels) have been well-studied as contributors of human diseases or traits, especially after the introduction of WES approaches, CNVs have received little attention in human genetics. This in part due to the limited availability of computational approaches to detect CNVs from exome data. It has been recently estimated that CNVs affect ~5–10% of the genome, suggesting that a number of potentially disease-causing CNVs have yet to be discovered (59, 63). The development of improved WES-based tools to identify novel CNVs would be pivotal to harness already-available large datasets of exome data for discovery of novel determinants of human diseases. This is especially important considering that, even after a detailed analysis of SNPs and indels as candidate disease-causing mutations, some patients with suspected syndromic conditions are left without a conclusive diagnosis.

In contrast to computational tools using data from WGS, WES-based methods for detection of CNVs have met with more limited success, mostly due to the nature of targeted enrichment protocols (64-66). Common WGS-based methods use breakpoints, the regions in which the rearrangements occur, to detect CNVs. By contrast, WES focuses on noncontiguous genomic targets (the exons), and most breakpoints are not sequenced. Hence, current WES-based approaches for detecting CNVs use the coverage as a proxy for copy number information. However, given the issue of non-uniformity of coverage, WES-based methods face important challenges. The exome coverage is heavily dependent on sequencing conditions, which are

continually evolving in typical laboratories that recruit patients and perform exome sequencing over several years. Thus, the exome data generated over time are inevitably heterogeneous, complicating the discovery of CNVs. In addition, widely used and actively maintained detection tools focus on detection of CNVs spanning one or more exons (67, 68), while no tool to date can detect smaller CNVs. Within the CNVs, rare homozygous and hemizygous (HMZ) deletions are of particular relevance for rare diseases because they may result in null alleles and a complete loss of gene function (Figure 7). **The second part of my thesis's work** focused on developing and testing HMZDelFinder_opt, a method that improves the performance the calling of HMZ deletions in typical laboratory panels, which are generated over time, and allows the systematic detection of partial exon deletions (i.e. deletions spanning less than one exon) (69).

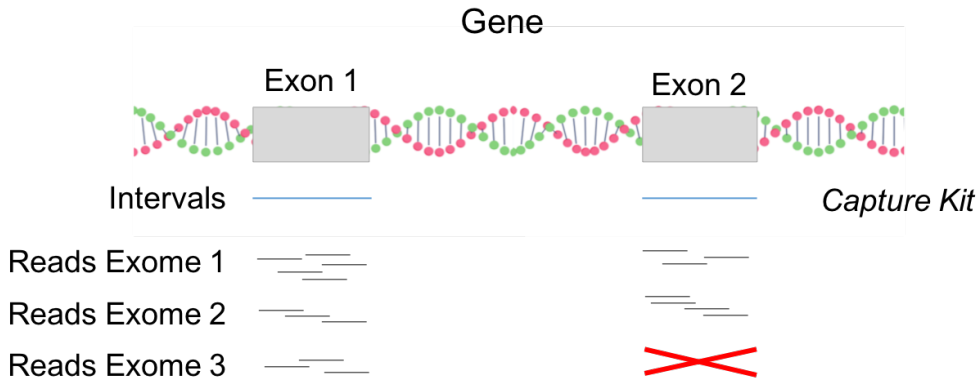


Figure 7. CNVs, and particularly HMZ deletions, in WES data
CNVs have received little attention in human genetics, in part due to the limited availability of computational tools to detect CNVs from WES data. Within the CNVs, rare homozygous and hemizygous (HMZ) deletions are of particular relevance for rare diseases because they may result in null alleles and a complete loss of gene function. Exome data are expected to show no coverage (red cross) in correspondence of HMZ deletions.

1.3 Aims of the thesis

The field of classical genetics has been experiencing a vigorous momentum since the advent of WES approaches. These technological advances increasingly require the development of sophisticated computational approaches, thus generating new research avenues in order to appropriately analyze and interpret enormous amounts of data. In turn, the wealth of exome data accumulated over the years has given the opportunity to pose scientific questions in ways that could not be possible earlier. My thesis took advantage from both these aspects and set out to:

1. Develop a novel filtering approach to blacklist FP variants for prioritizing candidate disease-causing variants in WES analysis (**presented in Chapter 2, article published in PNAS in 2019 (45)**)
2. Develop a novel WES-based algorithm to detect CNVs, particularly homozygous and hemizygous deletions spanning less than one exon (**presented in Chapter 3, article under review and available in BioRxiv (69),** https://doi.org/10.1101/2020.07.23.217976)
3. Investigate HW equilibrium across different ethnicities in gnomAD to determine variants in true HW disequilibrium possibly underlying susceptibility or resistance to disease, and their underlying origin in relation to specific population events (**preliminary findings presented in Chapter 4).**

Collectively, these projects tackle heretofore-unexamined topics and hold promise to discover novel causal determinants of human diseases or traits.

2 Blacklisting false positive variants

2.1 Introduction: reducing the number of false positives in exome data

NGS approaches, particularly WES and WGS, are increasingly being used for the discovery and diagnosis of human genetic disorders(34, 70, 71). The number of new disease-causing genetic variants logged by the Human Gene Mutation Database (HGMD) is currently increasing at a rate of ~10% per annum(72). This increase has coincided with an expansion of the use of WES and WGS(70, 71). The mean number of high-quality exonic coding variants per individual relative to the reference human genome is about 20,000(34, 71), but monogenic disease in any given individual is generally driven by at most two variants. The remaining variants may be real (rare or common, deleterious or neutral), or false/low-quality signals [sequencing artifacts, bioinformatic misprocessing of raw sequencing data, or resulting from limitations to the performance of current quality control (QC) methods]. In practice, analyses of individual exomes aim to generate a short list of high-quality candidate variants by filtering out as many FP as possible, while minimizing the risk of false negatives (FNs) due to the removal of true disease-causing mutations. The first step in this process typically involves the use of public databases to identify and remove variants through comparisons of their frequency in the general population with the prevalence of the disease considered, its proposed mode of inheritance, and its estimated clinical penetrance. The largest public database available at the time this project was undertaken (2017-2018) was the Genome Aggregation Database (gnomAD), which includes 123,136 exomes and 15,496 genomes from a total of 138,632 individuals (73). For the remaining variants, including those not reported in public databases, various variant-level and gene-level metrics can be used to predict deleteriousness and to select a smaller set of candidate variants for further experimental analysis (74-78). For example, the combined annotation dependent depletion (CADD) score is a variant-level metric to predict the impact of a given variant and the gene damage index (GDI) is a gene-level metric to assess the mutational load in each protein-coding gene.

In studies of rare genetic diseases, public databases are widely used for the initial elimination of common variants [minor allele frequency (MAF) > 0.01] (71, 79). However, some common variants within private databases may be absent from public databases, and most such variants are likely FPs (LFP) (31, 71). The efficacy with which such LFP variants are identified and used for analyses of exomes from panels of patients studied by a particular

research group has never been assessed in detail. An approach (defined as DFS) for detecting false-positive signals based on an internal panel of 118 whole-exome sequences from different individuals generated a shortlist of variants found to be in Hardy–Weinberg (HW) disequilibrium due to excess hetero- zygosity (the DFS list; 23,389 variants) (80). However, most of these variants (68%) had already been reported in dbSNP (80). Machine learning-based methods, such as variant quality score recalibration (VQSR), which uses a clustering score to determine whether a called variant is true(81), can limit the number of FPs in exome data. However, these methods are subject to several limitations: (i) they are computationally intensive and time-consuming; (ii) they often require a large number of samples; (iii) parameter optimization requires extensive testing; and (iv) the addition of new samples requires reprocessing of the entire panel. These methods are therefore little used by most researchers, who have small- or medium-sized exome panels evolving over time and may not have access to powerful computing resources. It has been suggested that variants common within a homogeneous panel and absent from public databases could be filtered out(71), but this approach has not been validated and there are currently no tools for the easy identification and compilation of such variants. In this context, we sought to establish a "blacklist" of LFP variants too frequent in our panel of 3,104 exomes to be causative from patients with severe infectious diseases(5, 82, 83).

2.2 Methods

2.2.1 Description of the samples: PID, Neuro, Infection and Africa

In this study, we used WES data from four different sets of samples, here referred as PID, Neuro, Infection and Africa. The PID panel was the main sample which was investigated and consisted of 3,104 individuals samples of diverse ancestral origins (North African: n=1,053; Caucasian: n=1,150; African: n=297; Middle Eastern: n=395; Asian: n=55; American: n=145; unknown: n=9) obtained by our laboratories and recruited with the help of clinicians. Most of the individuals had a wide range of different infectious diseases and immune deficiency phenotypes, and probands' family members accounted for the rest. All study participants provided written informed consent for the use of their DNA in studies aiming to identify genetic risk variants for disease. IRB approval was obtained from The Rockefeller University and Necker Hospital for Sick Children, along with a number of collaborating institutions. The exomes of 3,869 individuals suffering from neurological disease ("Neuro"

panel) were obtained from the GME Consortium, with recruitment according to a similar protocol (84). The exomes of 902 individuals suffering from severe infectious diseases ("Infection" panel) were obtained from patients enrolled in studies coordinated by Dr. Jacques Fellay's laboratory at EPFL, Switzerland. The exomes of 400 individuals in the "Africa" panel were provided by Dr. Lluis Quintana-Murci's laboratory at the Pasteur Institute, Paris, France. Table 1 describes the main sequencing and bioinformatic parameters of the four panels.

2.2.2 WES, bioinformatics analysis and quality control

Rockefeller PID exome sequences: genomic DNA from peripheral blood mononuclear cells was extracted and sheared with a Covaris S2 Ultrasonicator. An adaptor-ligated library (Illumina) was generated, and exome capture was performed with SureSelect Human All Exon 37, 50, or 71 Mb kits (Agilent Technologies). Massively parallel whole-exome sequencing was performed on a HiSeq 2000 or 2500 machine (Illumina), generating 72-, 100- or 125-base reads. Quality controls were applied at the lane and fastq levels. Specifically, the cutoff used for a successful lane is Pass Filter > 90%, with over 250 M reads for the high-output mode. The fraction of reads in each lane assigned to each sample (no set value) and the fraction of bases with a quality score >Q30 for read 1 and read 2 (above 80% expected for each) were also checked. In addition, the FASTQC tool kit (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to review base quality distribution, representation of the four nucleotides of particular k-mer sequences (adaptor contamination). We used the Genome Analysis Software Kit (GATK, version 3.4-46) best-practice pipeline to analyze our WES data(81). Reads were aligned with the human reference genome (GRCh37), using the maximum exact matches algorithm in Burrows-Wheeler Aligner (BWA) (85). PCR duplicates were removed with Picard tools (http://picard.sourceforge.net/). The GATK base quality score recalibrator was applied to correct sequencing artifacts. GATK HaplotypeCaller was used to identify variant calls at the individual level (one VCF per individual). DP >=5 and MQ >=30 were used as standard hard filtering criteria (86). Variants were annotated with SnpEff (http://snpeff.sourceforge.net/). Exomes were annotated for PASS and non-PASS variants in gnomAD r2.0.2 (Exome Aggregation Consortium, Broad Institute) and the 1000 Genomes Project Phase 3 (http://www.1000genomes.org/) databases. Joint genotyping followed by VQSR filtering was not used because there have been reports of fractions of variants unique to individual samples being missed (http://gatkforums.broadinstitute.org/gatk/discussion/4150/should-i-analyze-my-samples-

alone-or-together), rendering this approach unsuitable for our studies. For the purpose of comparison between the blacklist and VQSR approaches, VQSR was calculated with VariantRecalibrator and ApplyRecalibration for both SNPs and indels, with ts_filter_level set to 99.0 and other settings as specified by GATK recommendations. We did not use the InbreedingCoeff as this is discouraged in situations in which the sample includes members of the same family, as in our sample. Similarly, we did not include DP among the parameters of the VQSR, as it is not suitable for targeted exome sequencing samples.

Neuro (GME Consortium neurological exome sequences): whole-exome sequencing for the GME Consortium was performed as previously described (84). Briefly, genomic DNA was extracted from peripheral blood mononuclear cells with Qiagen reagents and captured with the Agilent SureSelect Human All Exome 50 Mb kit. WES was performed on an Illumina HiSeq 2000. The GATK best-practice pipelines were used to analyze WES data(81). BWA was used to align reads with human reference genome GRCh37(85). The variant-call format files generated were annotated with the Rockefeller pipeline, as described above.

"Infection" exome sequences: whole-exome sequencing for the Infection panel was performed as previously described(87, 88). In brief, genomic DNA was extracted from whole blood with the QIAamp DNA blood kit and captured with the Agilent SureSelect Human All Exome 50 Mb kit (Agilent SureSelect Human all exon V4 or V5) or Illumina Truseq 65 Mb enrichment kit. WES was performed on an Illumina HiSeq 2000 or Illumina HiSeq 2500 machine. BWA-MEM was used to map reads onto the human reference genome GRCh37 decoy, and GATK v3.8 (or an earlier version of this software) was used for data processing and analysis, according to GATK best practice.

"Africa" exome sequences: whole-exome sequences were obtained for 300 African samples(89), and these data were processed together with those for 100 European individuals(90). All samples were sequenced with the Nextera Rapid Capture Expanded Exome kit, which delivers 62 Mb of genomic content per individual, including exons, untranslated regions (UTRs), and microRNAs. Using the GATK Best Practice recommendations(91), we first mapped read-pairs onto the human reference genome (GRCh37) with BWA v.0.7.7(85), and reads duplicating the start position of another read were marked as duplicates with Picard Tools v.1.94 (http://picard.sourceforge.net/). GATK v.3.5(81) was then used for base quality score recalibration ("BaseRecalibrator"), insertion/deletion (indel) realignment ("IndelRealigner"), and SNP and indel discovery for each sample ("Haplotype Caller").

Panel	Size	Kit	Sequencer	Aligner	Reference Genome	Caller	Annotator
PID	3,104	Agilent 37, 50, 71 Mb	Hiseq 2000, 2500	bwa(v0.7.12)	hg19	GATK (v3.4-46)	snpEff
Neuro	3,869	Agilent 50 Mb	Hiseq 2000	bwa (v0.7.5)	GRCh37	GATK (v.3.1-1)	snpEff
Africa	400	Nextera Rapid Capture Expanded Exome 61 Mb	Hiseq 2500	bwa (v0.7.7)	GRCh37	GATK (v.3.5)	snpEff
Infection	902	Agilent 50 Mb, Illumina 65Mb	Hiseq 2000, 500	bwa (v0.7.10)	hg19 decoy	GATK (v3.8)	snpEff

Table 1: Summary of the technology employed for each panel of the Blacklist.

2.2.3 Algorithm and statistics

2.2.3.1 Blacklist creation and Refine algorithm

The blacklists used in and provided with this manuscript were created by first collecting unique variants from 3,104 patient exomes and counting the occurrence of each variant (the number of patients reported to have the variant). The QC criteria used to collect these variants were equivalent to those used in gnomAD (MQ≥30). However, we used a lower DP (DP≥5), compatible with research approaches in which investigators wish to retain as much information as possible. These criteria correspond to a high degree of QC despite low coverage, but may allow the discovery of true disease-causing variants, as illustrated by the example of the deletion of *ISG15,* which was initially identified by exome analysis despite a low DP of 4(92). We did not use the QD value as a QC criterion due to the erroneous calls for some variants (https://gatkforums.broadinstitute.org/gatk/discussion/8912/most-variants-called). We explored the FN rate of the blacklists in the HGMD database and excluded variants that were present in the set of true disease-causing variants in HGMD according to further analyses(93). The measurement of variation at multiallelic sites was rendered more effective by separating variants into biallelic and multiallelic variant groups. Multiallelic variants represent a very specific challenge for the elimination of nonpathogenic variants from exomes, as variants at multiallelic positions may occur individually in a small number of samples. Collectively, however, these variants may occur in a large proportion of the members of the panel (*i.e.* many individuals may contain one of a number of variants at the position). The variants at multiallelic sites are often similar (*e.g.* G in the reference and an alternative of GA, GAA, GAAA, GAAAA, GAAAAA, *etc.*) but have remained resistant to removal from exomes by bioinformatic methods. For the capture of these variants, we collapsed all variants at multiallelic sites to a

single value by calculating the total number of patients with any variant at the multiallelic position. When this number exceeded 1% of our panel, all variants at the position concerned were included in the full blacklist. This procedure can thus identify variants present in only a few individuals but nevertheless occurring at positions with a high cumulative burden of variation in a panel. We then considered biallelic variants. If the number of patients with any individual biallelic variant exceeded 1% of our panel, the variant concerned was included in the full blacklist. For a schematic diagram of this pipeline, see Figure 8.

Figure 8. Methodology for blacklist generation

The blacklist was generated by first collecting unique high- quality variants (DP>=5, MQ>=30) from patient exomes and counting the occurrence of each variant. These variants were assembled into two classes: (1) biallelic, with a single alternative allele in our panel; and (2) multiallelic, with two or more alternative alleles in the panel, for which we collapsed all variants at a unique chromosomal position and summed the total number of patients containing these variants. We then collected the variants that had a frequency >=1% in the panel (the Blacklist: "Common in-house variants"). Of these variants, 21.4% (167,144) were absent from gnomAD exome and genome databases. We considered these 167,144 variants to be "blacklist-annotated" (BL-A).

We designed ReFiNE (Reducing False Positives in NGS Elucidation) software, an easy-to-use tool for extracting a blacklist of LFP variants from internal panels of WES or WGS data on

the basis of a user-defined frequency cutoffusing Python programming language (version 2.7.14, https://www.python.org/) and R, using both default and publicly available libraries. The Python Tkinter module was used to design and implement the graphical interface for ReFiNE. ReFiNE is available as a graphical interface program (including a command-line option) that can be run on a standard laptop and is compatible with comma-separated (CSV) files. ReFiNE can also generate blacklists from WGS data, although this application has yet to be extensively tested. ReFiNE includes an optional parameter for the exclusion of a list of variants from the blacklist regardless of their frequency in the in-house database. This option can be used to remove a small number of known true disease-causing HGMD variants, for example. We also provide precalculated blacklists generated from our panel of 3,104 PID exomes with cutoffs of 1%, 3%, 5% and 10%. These blacklists can be used for small panels for which it may not be possible to generate custom blacklists. We also provide the PID, Neuro, Infection, Africa and combined blacklists used in this manuscript, annotated with gnomAD MAFs. Finally, we have constructed a public server (http://lab.rockefeller.edu/casanova/BL) containing all the supplemental files, the ReFiNE program, and a user-friendly online tool that can be used to query whether a variant is included in our blacklist or to annotate lists of variants in a similar manner. ReFiNE and pre-calculated blacklists: http://lab.rockefeller.edu/casanova/BL

2.2.3.2 Simulating minimum sample size and sample size saturation for blacklists

We determined the minimum number of samples required for the creation of safe blacklists by generating random blacklists based on 10, 50, 100, 250, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000, 5500, 6000, or 6500 individuals from the PID and Neuro panels. We weighted the random selection of individuals for the blacklists by project size (*i.e.* for a sample size of 10, we picked 4 individuals at random from the PID panel and 6 at random from the Neuro panel). The selection of individuals for each sample size was repeated 30 times, and full blacklists for each iteration were generated with ReFiNE. The median number of blacklist-annotated variants and a 99% confidence interval based on a normal distribution were calculated for each sample size and plotted. The number of samples required to reach saturation for blacklist variants was predicted by fitting a logarithmic trendline to the blacklist dataset based on the coefficient of determination (R^2). The equation for this line was:

$$y = 2801.1 \times \ln(x) + 3466.3$$

where $R^2 = 0.7088$. We defined saturation as the number of samples for which less than one cohort-specific variant was added to the blacklist per new exome. Based on the best-fit equation, we calculated the saturation point as 2,801 individuals.

2.2.3.3 Statistics and figures

The Scipy library (https://www.scipy.org/) was used for statistical analyses performed in Python. Seaborn (https://web.stanford.edu/~mwaskom/software/seaborn/) was used to generate figures in Python, together with matplotlib (https://matplotlib.org). Venn diagrams were generated with jvenn software(94). Wordclouds were generated with the WordCloud library (https://github.com/amueller/word_cloud). Prism (Graphpad) was also used for figure generation and statistical analysis.

2.2.4 Characterization of blacklisted variants and Sanger sequencing

The blacklisted variants were characterized according to various metrics, including by HW equilibrium/disequilibrium, occurrence in low-complexity regions, and allelic distribution across genetic ancestries. HW disequilibrium was calculated for the blacklisted variants found to be present in the European population (n=1150), which constituted the largest population of the PID panel. Chi-squared tests were used to assess HW equilibrium. Given the large number of tests performed and the heterogeneity of European origins in our European panel, a stringent threshold of 10^{-8} for significance was used for significance. A total of 106 variants with a p-value below 10^{-8} were considered to be in HW disequilibrium and were stratified by excess genotype as follows: excess of heterozygotes (observed no. of heterozygotes > expected no. of heterozygotes, 57 variants), excess wild-type homozygotes (observed no. of wild-type homozygotes > expected no. of wild-type homozygotes, and chi-squared for the wild-type homozygote > chi-squared for the alternative homozygote, 13 variants), excess alternative homozygotes (observed no. of alternative homozygotes > expected no. of alternative homozygotes and chi-squared for alternative homozygotes > chi-squared for wild-type homozygotes, 36 variants).

The occurrence of the variants in low-complexity regions was assessed with the following tracks from the UCSC Genome Browser: RepeatMasker and Simple Repeats (group: Repeats), and GC percent (group: Mapping and Sequencing). RepeatMasker was created from the RepeatMasker program, which screens DNA sequences for interspersed repeats and low-

complexity DNA sequences; Simple Repeats reports simple tandem repeats located by Tandem Repeats Finder (TRF), which was designed especially for this purpose. Variants were considered to occur in GC-rich regions in which the G+C content exceeded 80%.

The heterogeneity of ethnicity was assessed in the four largest genetic ancestry groups in our panel (European (E), African (A), North African (nA) and Middle Eastern (ME)), for the variants found to be in HW equilibrium in the European population. Chi-squared tests were used to test the allelic distribution. In total, 203 variants with a *p*-value below 10^{-8} were considered to be heterogeneous across ancestries. The ancestry driving heterogeneity was unequivocally determined for 67 variants, by testing the allelic distributions of four combinations of three populations from those mentioned above (E-A-nA, E-A-ME, E-nA-ME, A-nA-ME), determining the combination among these four combinations that did not reach significance, and identifying the population that was missing in the non-significant combination. For example, if a variant was found to be in allelic heterogeneity in E-A-nA, E-A-ME and A-nA-ME but not in E-nA-ME, the ancestry driving heterogeneity was determined to be A. For Sanger sequencing, DNA was extracted from 10 SV40-fibroblast cell lines from patients included in our panel. PCR amplification was performed with Hot-Start Taq Blue DNA Polymerase (Denville Scientific, Inc.), 85 ng of template genomic DNA and specific primers. Sanger sequencing was performed with the BigDye Terminator kit (Perkin Elmer).

2.2.5 Analysis of variation in patient exomes

We identified the disease-causing mutation in patient D2 from a previous study(95), using a standard filtration pipeline. In brief, we removed variants with low-quality metrics (DP<4, MQ<40, QD<2) that were common in public databases (variant frequency in gnomAD < 0.0001), variants of high-GDI genes (74), and variants with CADD scores below their gene-specific mutation significance cutoff (77). Gene burden was analyzed in our chronic mucocutaneous candidiasis (CMC) panel by first filtering each exome, as described above. We then compared the numbers of individuals with at least one variant for each mutated gene in the patient group between the patient (n=208) and control (n=960) groups in a one-tailed Fisher's exact test. The resulting p-values were used to rank genes, to identify those with the highest levels of enrichment in patients.

2.3 Results

2.3.1 Generating the blacklist

We observed that numerous candidate variant calls (see Materials & Methods) (96) predicted to be damaging to the corresponding transcript or protein were present in >1% of our panel of 3,104 in-house exomes from primary immune deficiency (PID) patients with heterogeneous ancestral background(97) (i.e. too common to cause PID) but absent from public databases (e.g. 1KG, ExAC, gnomAD). These variants are poor candidates for involvement in rare diseases and likely false positive variants but are impossible to eliminate by current methods based on variant frequencies in public databases(71). We therefore sought to classify and characterize these variants in a rigorous and comprehensive manner, to enable users to remove them from their WES/WGS analyses. First, we determined a statistical cutoff frequency above which in-house variants should be considered too frequent to cause rare diseases. We found that the MAF of all experimentally validated disease-causing mutations in HGMD followed a Gilbrat distribution(98). We then calculated the 99% Gilbrat distribution confidence interval (CI) for these frequencies and found that the upper boundary of the CI for the frequency of known disease-causing mutations was 0.01 (1%). We therefore used this cutoff as a criterion for LFP variants occurring in too many patients in our database to explain a rare monogenic illness. The MAF>0.01 cutoff used here is an example of the blacklist approach to removing LFP variants in studies of rare genetic disorders. The cutoff can be adjusted according to the mode of inheritance and genetic architecture, assumed penetrance, disease prevalence, and the phenotypic homogeneity of the panel (17). For example, assuming complete penetrance and allelic homogeneity, a rare recessive genetic disorder with a prevalence of 1 in 100,000 could be analyzed with a MAF cutoff of 0.0033, whereas a more common recessive genetic disorder with a prevalence of 1 in 1,000 should be analyzed with a MAF cutoff of 0.033. Some caution may be needed for specific cases. For example, if the sample contains an high proportions of patients with the same disease and therefore could potentially be enriched with the same allele (hypothesis of strong allelic homogeneity), one should consider higher cutoffs. Similarly, the assumption of incomplete penetrance may lead to the definition of higher cutoffs, whereas the assumption of allelic/genetic heterogeneity may lead to the use of lower cutoffs.

We first designed the ReFiNE (Reducing False Positives in NGS Elucidation) software, an easy-to-use tool for extracting a blacklist of LFP variants from internal panels of WES or WGS data on the basis of a user-defined frequency cutoff (see Materials and Methods for details).

ReFiNE creates a blacklist consisting of the full set of variants occurring in >1% (or any user-defined cutoff) of an investigated panel, which can then be further filtered separately by the user, using MAF cutoffs from a population genetic database of choice. Using ReFiNE, we first collated all variants present at a frequency >1% in our PID WES panel of 3,104 exomes (Fig. 8, Materials & Methods) with a depth of coverage (DP) ≥ 5 and mapping quality (MQ)≥ 30 (see Materials and Methods(73, 86)). A large number of multiallelic variants in our panel were absent from gnomAD for specific chromosomal positions. ReFiNE therefore collapsed all variants at a unique chromosomal position and summed the total number of patients at each of these positions. This generated a list of 780,956 LFP variants, defined as the blacklist. This blacklist is the full list of variants occurring at single chromosomal positions for which >1% of patients had an alternative allele. These LFP variants belonged to two classes: (1) biallelic, with a single alternative allele in our panel; and (2) multiallelic, with two or more alternative alleles in our panel. The blacklist includes variants already reported in public databases, so we needed to extract the subset of variants unique to our method for further analysis. We thus annotated the blacklist with gnomAD, currently the most extensive public population genetics database available (6, 73). We found that 21.4% (167,144) of these 780,956 variants were absent from the gnomAD full exome and genome databases. As these 167,144 LFP variants are not captured by the most extensive public database available, we focused the analysis of our method on this subset of variants, which, for simplicity, we will refer to as blacklist-annotated (BL-A): common in-house LFP variants absent from gnomAD that cannot, therefore, be filtered out of analyses based on gnomAD.

2.3.2 Efficacy of the blacklist filtering

We then assessed the efficacy of BL-A for filtering out LFP variants from patient exome data. We first applied the standard procedure for rare genetic disorders, by removing variants with a MAF>0.01 in gnomAD from our 3,104 exomes (31, 34). This reduced the median number of variants in the patients' exomes by 90% (Fig. 9A). Subsequent filtering with BL-A removed 62% of the remaining variants that could not be removed by other means (Fig. 9A, a median of 9,056 variants removed per exome). By comparison, the DFS list (80) decreased the median number of these variants by only 1.8% (median of 260 variants removed per exome). BL-A filtering was effective for both coding sequences (CDS), including indel, exon-deleted, non-synonymous, synonymous and essential splicing variants, and for non-CDS variants, including UTR, non-essential splicing, intronic, downstream and upstream variants, and for all

three exome kits available for our panel (37 Mb, 50 Mb, and 71 Mb). We then assessed the performance of BL-A filtering for variants absent from the gnomAD database (i.e. variants private to the PID database), which would be considered among the strongest candidates for a causal role in disease. This approach decreased the number of cohort-private variants potentially associated with PID in each exome by 86%, versus only 2.2% for the DFS list, and was similarly effective for CDS and non-CDS variants (Fig. 9B). Thus, when used as a filtering tool, our blacklist was able to remove LFP variants absent from public databases and to decrease the number of candidate variants per exome considerably.

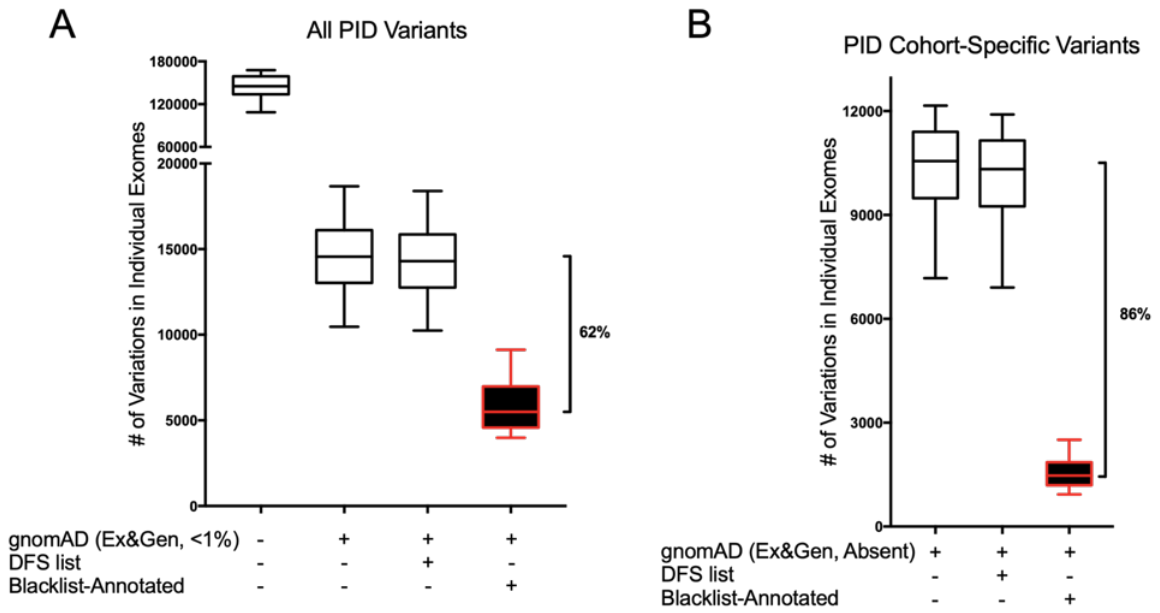


Figure 9. Blacklist filtering of 3,104 PID exomes with the PID blacklist.
(A) Filtering of all variants in each exome by first removing those common in gnomAD exome and genome databases (MAF greater than 0.01). The remaining variants were subsequently filtered with the blacklist. (B) Filtering of cohort- specific variants in each exome with the blacklist. Filtering with the DFS list is shown for comparison. Error bars represent the 10th to 90th percentiles.

We then explored whether the quality control (QC) scores for BL-A variants were similar to those for polymorphic variants (MAF>0.01) reported in gnomAD. By comparing the median MQ and DP scores for blacklisted variants and polymorphic variants from our panel (Fig. 10A-B), we demonstrated that none of these QC metrics could differentiate between these two sets of variants (especially when considering commonly used criteria for hard filtration, see Materials and Methods for further details). We then investigated whether machine learning QC metrics could classify these variants. With variant quality score recalibration (VQSR), only 25% of BL-A variants were annotated as "non-pass" (not shown). One of the key goals of the blacklist approach is providing an efficient tool for researchers who cannot easily perform

VQSR. We therefore retained these VQSR "non-pass" variants in the blacklist. We also assessed the ability of a random forest classifier trained on polymorphic variants from the gnomAD dataset well-characterized by different methods to separate true variants from FP artifacts called by the variant-calling pipeline(73). We then used the same method to construct a new scoring function with the gnomAD dataset. We applied both scoring functions to the LFP blacklisted variants and a set of variants present in both the gnomAD dataset and our panel, with a minor allele frequency of more than 1% in each dataset. The score distributions obtained were almost identical (Fig. 10C), demonstrating an inability of this standard classification method to distinguish between the LFP blacklisted variants and true positive (TP) variants.

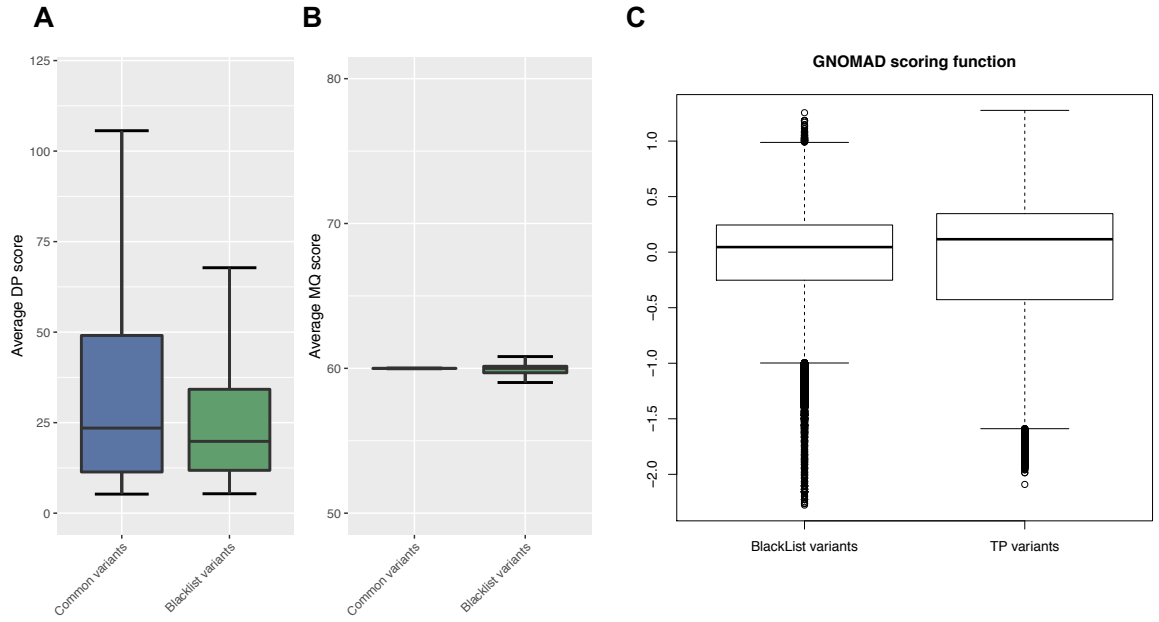


Figure 10. Comparison of quality metrics and machine learning-based filtering methods.
Quality metrics for blacklisted and non-blacklisted variants: mean (A) read depth (DP) and (B) mapping quality (MQ) were calculated for common variants present in gnomAD with a MAF>1% (blue bar), and for blacklist-annotated variants (green bar). Error bars represent the upper and lower limits of 1.5 times the interquartile range. Score distributions at random forest scoring functions for blacklist-annotated variants and for a set of true-positive (TP) variants present in both the gnomAD dataset and our panel with a MAF exceeding 1% in each dataset.

We then characterized the variants and genes included in BL-A with computational damage prediction metrics. A variant-level analysis revealed that the combined annotation-dependent depletion (CADD) scores for LFP blacklisted variants were not significantly different from those for variants not included in the blacklist. A gene-level analysis (74) of all genes with blacklist variants (*n*=13,665 genes) showed them to have low gene damage index (GDI) values. However, some genes with a high GDI have many BL-A variants (e.g. *HLA-DRB1*: 658

variants, *MUC16*: 455 variants). Filtration methods based on QC and variant- and gene-level damage prediction metrics would not efficiently detect and remove the LFP blacklisted variants absent from gnomAD. These results demonstrate the value of blacklisting as a complementary approach to analyses based on standard public databases, including gnomAD, QC filtering, and damage prediction metrics.

We estimated the proportion of TP disease-causing mutations removed by the blacklist approach, by screening 129 exomes from patients in our panel for whom the TP mutations had been validated experimentally. Filtering these exomes with the complete blacklist did not remove any of the known TP mutations (0% FN rate). Even though most variants in any patient are not pathogenic, our analysis indicates that it is very safe to apply the blacklist to patient exomes. We also compared the complete blacklist with the list of 144,641 disease-causing mutations in HGMD and noted an overlap of only 263 variants (0.18% FN rate). These variants are listed as disease-causing in the HGMD dataset, but 47% have a MAF > 0.01 in the gnomAD exome or genome databases, suggesting that are unlikely to be the cause of a rare disorder. These findings indicate that our FN rate is probably lower than the rate of 0.18% for HGMD in the context of rare disorders. Only eight BL-A variants were present in HGMD (0.001% FN rate), indicating that the FN rate for our specific BL-A list was lower than that for gnomAD. Together, these results suggest that the FN rate is very low for this technique. We also screened 3,731,152 somatic cancer-causing or cancer-associated variants available from TCGA (http://cancergenome.nih.gov/). We found that 59,151 of these TCGA variants (1.5%) were present in the complete blacklist and 2,471 (0.07%) were present in BL-A. As our blacklist was derived from germline exome data, the presence of these blacklist variants in the TCGA database suggests that they may be FPs that could be removed, as previously reported(99). Together, these data indicate that the blacklist approach results in an extremely low FN rate when applied to patients with rare diseases, and that it is therefore safe to use this approach to remove LFP variants from patient exome data.

2.3.3 Practical applications of the blacklist to the analysis of exome data

We assessed the use of blacklisting for practical analyses of patient exomes. We selected a case from our panel with an autosomal dominant disease-causing mutation described in a previous study (Patient D2 from(95)). We filtered this patient's exome with a standard pipeline to identify disease-causing mutations (Fig. 11). This standard approach reduced the number of candidate variants from 142,473 to 3,526. Taking known mode of inheritance into account and

restricting the analysis to CDS variants (excluding synonymous alterations), the number of candidate variants was reduced further, to 231. The inclusion of BL-A in the pipeline decreased the final number of candidate variants to 109 (Fig. 11), with retention of the known *IKZF1* mutation. Overall, this corresponds to a 53% decrease in the number of variants from this patient's exome to be considered. The remaining variants were high-quality candidates that would probably merit rigorous analysis in exome analyses for patients with diseases of unknown etiology. Thus, blacklisting greatly decreases the number of candidate variants for further study in practice, in exome analyses on individual patients.

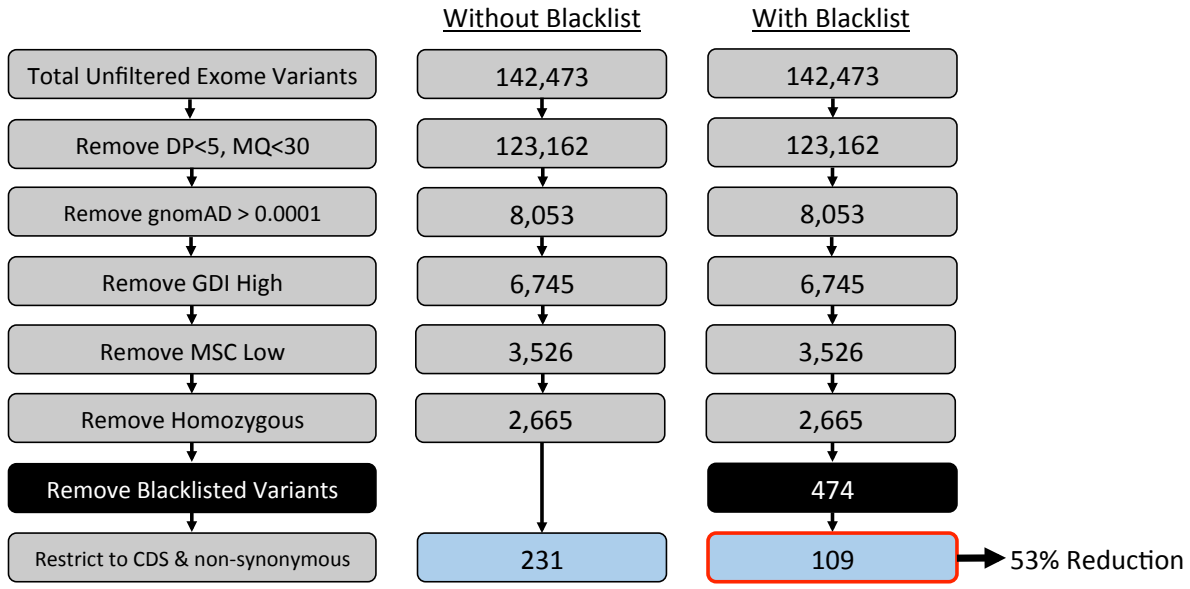


Figure 11. Practical analysis of a single patient exome by blacklisting.
The practical utility of the blacklist approach was demonstrated with the exome of a patient with a published disease-causing mutation. The patient's exome was filtered with a standard pipeline with and without application of the blacklist-annotated. The numbers in each box represent the number of variants remaining in the exome after each filtering step. GDI: gene damage index; MSC: mutation significance cutoff.

We then explored the use of our blacklist for gene burden analysis for genetic homogeneity at the population level. We compared the number of patients with at least one variant of any given gene between a panel of 202 patients suffering from CMC and 852 phenotypically unrelated controls(100). When standard filtering with public databases was applied in the absence of blacklisting, the enrichment observed for the known disease-causing gene in the CMC panel, *STAT1* (p-value=3.32 x10^{-6}), was not significant considering the corrected threshold at the genome-wide level (p-value$_{threshold}$ = 0.05 ÷ 20,554 = 2.43 x10^{-6}, Fig. 12A). However, following the addition of BL-A to the pipeline, *STAT1* was correctly identified as a gene displaying strong and significant genome-wide enrichment in the disease panel (p-value

= 4.63x10^{-10}; Fig. 12B). In this instance, our blacklist removed two variants present in a large proportion of our PID exomes (both cases and controls) that confounded the statistical comparison between the CMC and control groups. Together with the previous practical example, these analyses demonstrate the power of blacklisting for removing LFP variants from patient exomes, both to simplify candidate variant identification in patients and for other large-scale statistical analyses of patient groups.

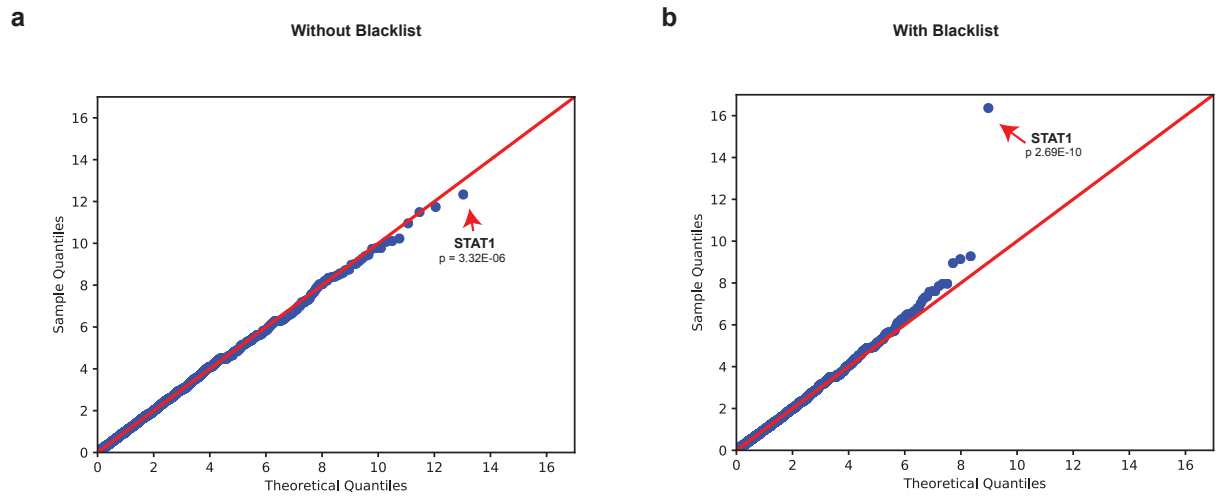


Figure 12. Application of the blacklisting approach to enrichment analysis
Quantile–quantile plots depicting the analysis of genetic homogeneity for a panel of 202 patients with chronic mucocutaneous candidiasis (CMC) before (A) and after (B) application of the blacklist. The control panel consisted of 852 unrelated individuals. In each panel, the red arrows indicate STAT1, the known cause of CMC in our panel, before and after blacklist application.

2.3.4 Characterization and experimental validation of the LFP blacklisted variants

We then characterized the PID panel BL-A variants (*n*= 167,144). Most of the variants (91.5%) in the blacklist were multiallelic (Fig. 13A). The cohort-specific variants present in the blacklist were therefore due to multiallelic sites displaying high levels of variation in our panel. We began by hypothesizing that the multiallelic variants might lie in low-complexity regions of the human genome, leading to sequencing errors. The annotation of all these variants with RepeatMasker, Simple Repeats, and GC percent tracks from UCSC Genome confirmed that 118,154 of the 152,915 variants (77.3%) occurred in repetitive or GC-rich regions, and that most (65,646; 56%) were located in short tandem repeat (STR) regions (Fig. 13A-C).

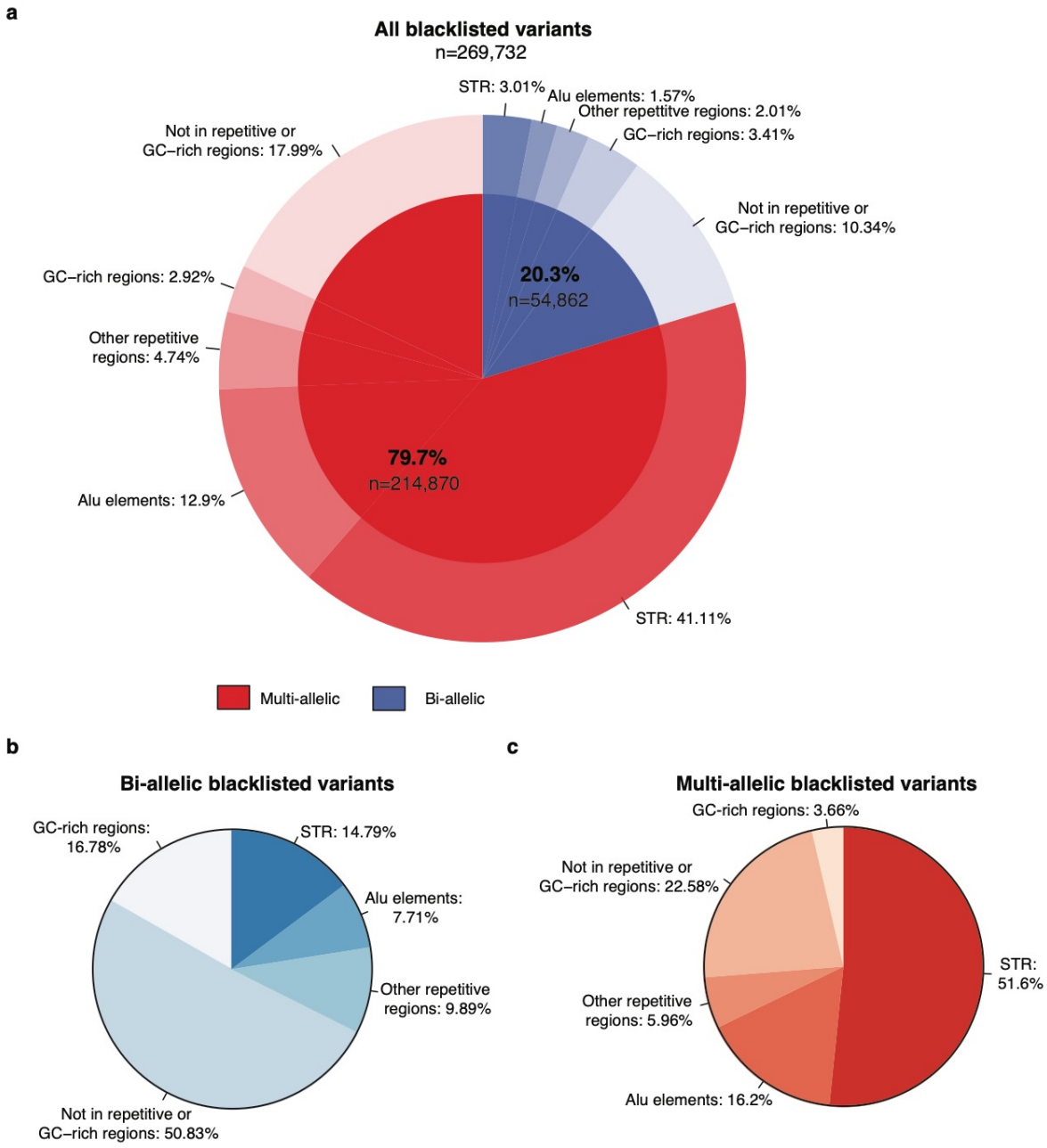


Figure 13. Characterization of the blacklisted variants in low-complexity regions of the genome
Occurrence of the blacklisted multiallelic (red) and biallelic (blue) variants in repetitive [short tandem re- peats (STRs), Alu elements, other repetitive regions] and GC-rich regions; percent relative to the total number of blacklisted variants (A) or the total number of biallelic (B) or multiallelic (C) blacklisted variants.

We analyzed the biallelic variants, which were also found to be located in repetitive or GC-rich regions, albeit to a lesser extent (6,711; 47.2%) (Fig. 13A-B). We also characterized these biallelic variants, focusing on those located in CDS regions, in the 1,150 individuals of European origin according to PCA analysis(97), to determine whether these variants were under Hardy-Weinberg (HW) equilibrium. In total, 388 CDS variants were found to be located in

repetitive or GC-rich regions; 339 (87.4%) of these variants were in HW equilibrium and 49 (13.6%) were in HW disequilibrium (threshold of $p<10^{-8}$; Table 2). An investigation of the biallelic variants not present in repetitive regions (7,518; 52.8%) yielded a similar distribution, with 209 (89.3%) and 25 (10.7%) of the 234 CDS variants in HW equilibrium and disequilibrium, respectively. Overall, 74 CDS variants were in HW disequilibrium, and, in 39 of these variants (52.7%), the cause was an excess of homozygous wild-type (28, 14.9%) or homozygous alternative (11, 37.8%) genotypes (Table 2). Most of these 39 variants had low coverage (wild-type=15.6x, alternative=20.5x; Table 2), which may have led to miscalls for a homozygous genotype. Most of the variants (35, 47.3%) in HW disequilibrium presented heterozygote excess, with high mean coverage rates of 163x (much higher than the 42.5x coverage of the 548 CDS variants in HW equilibrium), suggesting an excess of reads wrongly mapped to the region (Table 2).

Hardy-Weinberg equilibrium in CDS bi-allelic variants in Caucasian Individuals (n = 1150)			
Total	**<10^-8**	**>=10^-8**	**% Disequilibrium**
622	74	548	12
Cause of HW disequilibrium by excess genotype			
	excess het	**excess hom alt**	**excess hom WT**
Counts	35	28	11
%	47.3	37.8	14.9
DP	163.0	20.5	15.6

Table 2. Hardy-Weinberg of Bi-allelic CDS Variants in Caucasian Individuals.
Breakdown of the 662 CDS bi-allelic variants in the Caucasian individuals by HW disequilibrium and cause of disequilibrium.

We also studied the 548 biallelic CDS variants in HW equilibrium, to evaluate their distribution across ethnicities. We focused the analysis on the four largest genetic ancestry groups in our panel: European (n=1,150), African (n=297), North African (n=1,053), and Middle Eastern (395), as determined by PCA analysis. In total, 200 (36.5%) of these variants were heterogeneously distributed across genetic ancestries (threshold of $p<10^{-8}$; Table 3). The observed heterogeneous distribution was probably due to one specific genetic ancestry in 46 (23%) of the variants (Table 3). In 20 variants (43.5%), the individual genetic ancestry was Middle Eastern (Table 3), which is poorly represented in public databases (84) suggesting that these variants are true variants that are more common in this population.

Ethnicity Distribution of CDS bi-allelic variants in HW equilibrium				
	Total	**$<10^{-8}$**	**$>10^{-8}$**	**Ethnical Disequilibrium (%)**
Counts	548	200	348	36.5
Causal Ethnicity for Disequilibrium				
	Middle Eastern		**African**	**Caucasian**
Counts	20		20	6
%	43.5		43.5	13.0

Table 3. Ethnicity distribution of Bi-allelic CDS Variants in HW equilibrium

We further investigated the features of BL-A variants. We first focused on biallelic blacklist CDS variants in HW disequilibrium displaying excess heterozygosity and absent from repetitive regions in individuals of European ancestry (*n*=35). We found that 48.6% of these variants (*n*=17) mapped to four chromosomal regions, in the *HLA-DRB1*, *MUC6*, *OR8U1*, *TAS2R43* genes with consecutive blacklist variants (less than 300 base pairs). Most of these regions contain flagged variants annotated in gnomAD (47% in Exome and 65% in Genome, annotated as AC0, RF and/or InbreedingCoeff). For the remaining variants (referred to as "unique"), we found that the blacklist variants were at the same location (but with different genotypes) as flagged variants annotated in gnomAD, like the consecutive variants (28% in Exome and 50% in Genome, annotated as AC0, RF and/or InbreedingCoeff). Integrative Genomics Viewer (IGV) (101) showed that the consecutive variants in these regions belonged to the same reads, suggesting the existence of an "alternative" sequence (referred to as a segmental duplication by gnomAD or as an alternative haplotype). These observations strongly suggest that some blacklist biallelic variants define alternative haplotypes belonging to unmapped regions absent from the human reference genome. These variants were probably incorrectly mapped to the region of the reference genome for which the best match was obtained, leading to a mixture of wild-type and alternative alleles in these regions, resulting in higher coverage and a final erroneous heterozygous call. In a second analysis, we focused on multiallelic variants. Most of these variants (77%) were located in low-complexity regions (short tandem repeats, Alu elements, GC-rich regions, or other repetitive regions; Fig. 13). IGV analysis of three multiallelic variants absent from these regions and common in our panel (MAF>0.9) revealed that they were located in the vicinity of a small stretch of repeated nucleotides. Extending the analysis to the 23% of multiallelic variants not previously detected in low-complexity regions (*n*=34,761), we found that 83.3% were also located close to mononucleotide repeats (26,165; 75.3%) or to small repetitive stretches (two or more

nucleotides; 2,802; 8.1%). Attempts to confirm these variants by Sanger sequencing failed, due to the mononucleotide repeat, strongly suggesting that the WES approach may have been affected by a polymerase artifact similar to that reported in previous studies(102, 103). This exploration of blacklist variants suggests that the multiallelic variants probably resulted from sequencing/calling errors during WES on low-complexity regions, whereas a proportion of the blacklist biallelic variants, particularly those in HW disequilibrium, were due to mapping errors resulting from the incomplete nature of the GRCh37/GRCh38 genome assembly. Overall, these analyses demonstrated that the majority of blacklisted variants are FP.

2.3.5 Testing the blacklist approach in three unrelated panels

We assessed the suitability of the blacklist approach for filtering in other private databases. We used three unrelated independently processed exome panels (from DNA preparation to VCF data): (1) 3,869 exomes from patients suffering from neurological diseases ("Neuro") (84); (2) 902 exomes from patients suffering from diseases with an infectious phenotype ("Infection")(87, 88); and (3) 400 exomes (100 from Europeans and 300 from Africans) from a study on the demographic history of Central Africans ("Africa")(89),(90). We first generated separate blacklists for the Neuro, Infection, and Africa panels, according to the pipeline described above. After filtering on the basis of MAF>1% (in the specific panel) in gnomAD, the application of the cohort-specific blacklists for the Neuro, Infection, and Africa panels decreased the number of variants retained by 35%, 57% and 51%, respectively (a median of 3,160, 3,462 and 7,905 variants per exome, respectively; Fig. 14A,C,E). Considering only cohort-private variants (i.e. those appearing in the specific panel but absent from gnomAD exomes and genomes), applying the cohort-specific blacklists to the Neuro, Infection and Africa panels reduced the number of variants in each exome by 90%, 92% and 93%, respectively, eliminating a median of 3,195, 3,418 and 7,861 variants per exome, respectively (Fig. 14B,D,F). This filtering was effective for both CDS and non-CDS variants.

A comparison of the four blacklists revealed that a substantial number of variants were unique to each blacklist (Fig. 15), demonstrating the panel specificity of the blacklisted variants, particularly for the Africa panel, probably due to ancestry. Specifically, each blacklist contained 63% to 91% of the unique biallelic variants (Fig. 15A) and 46% to 92% of the unique multiallelic variants (Fig. 15B). A similar pattern was observed when the analysis was restricted to biallelic and multiallelic CDS variants (Fig. 15C-D). Thus, the efficacy of blacklist filtering in our PID panel was not due to specific pipeline settings or enrichment within our exomes.

Instead, our results suggest that the blacklist method should effectively remove a substantial proportion of the FP variants not already removed by public database analysis from any panel of exomes considered.

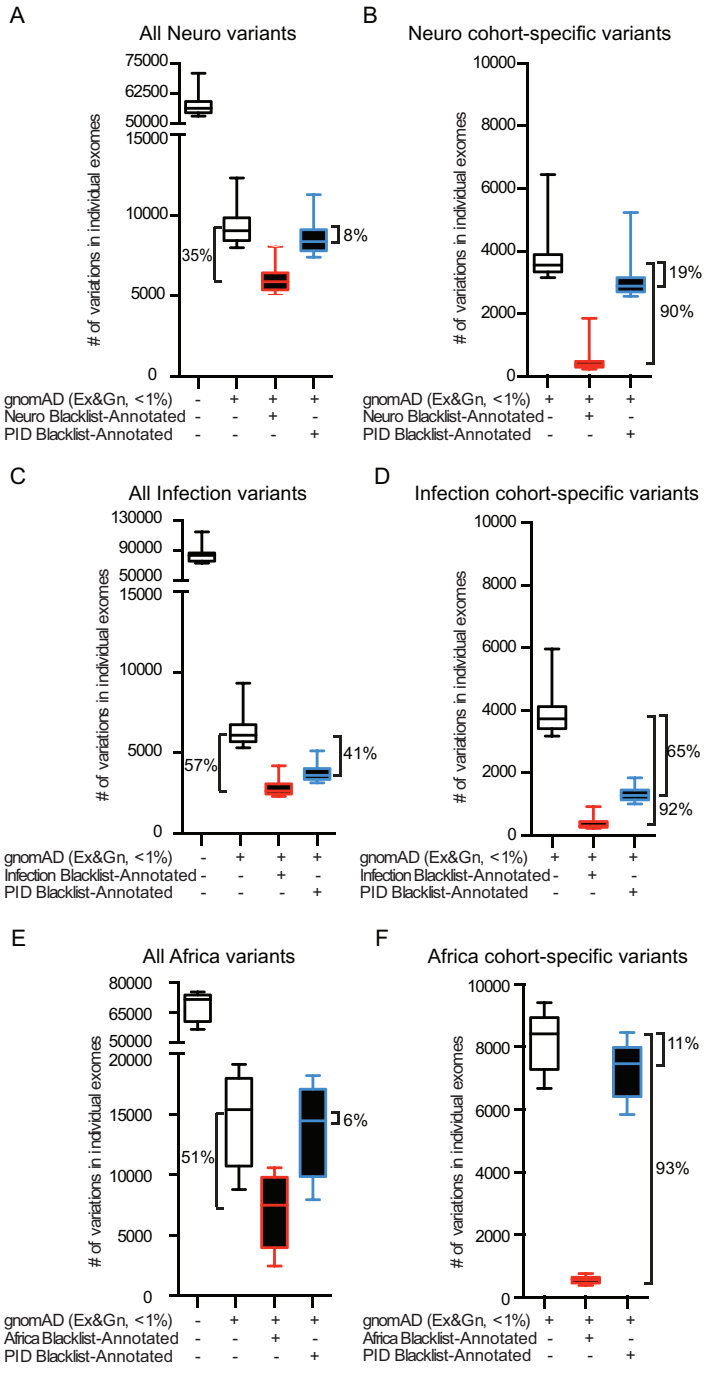


Figure 14. Blacklist filtering of unrelated panel exomes

(A, C, and E) Filtering of all variants in the neurological (A), infectious disease (C), and central African (E) exomes by first removing those common in gnomAD exome and genome databases (MAF greater than 0.01). The remaining variants were subsequently filtered with the Neuro (A), Infection (C), or Africa (E) blacklists (red boxes), or the PID blacklist (blue boxes). (B, D, and F) Filtering of exomes restricted to cohort-specific variants, with the Neuro (B), Infection (D), or Africa (F) blacklists (red boxes), or the PID blacklist (blue boxes). Error bars represent the 10th to 90th percentiles.

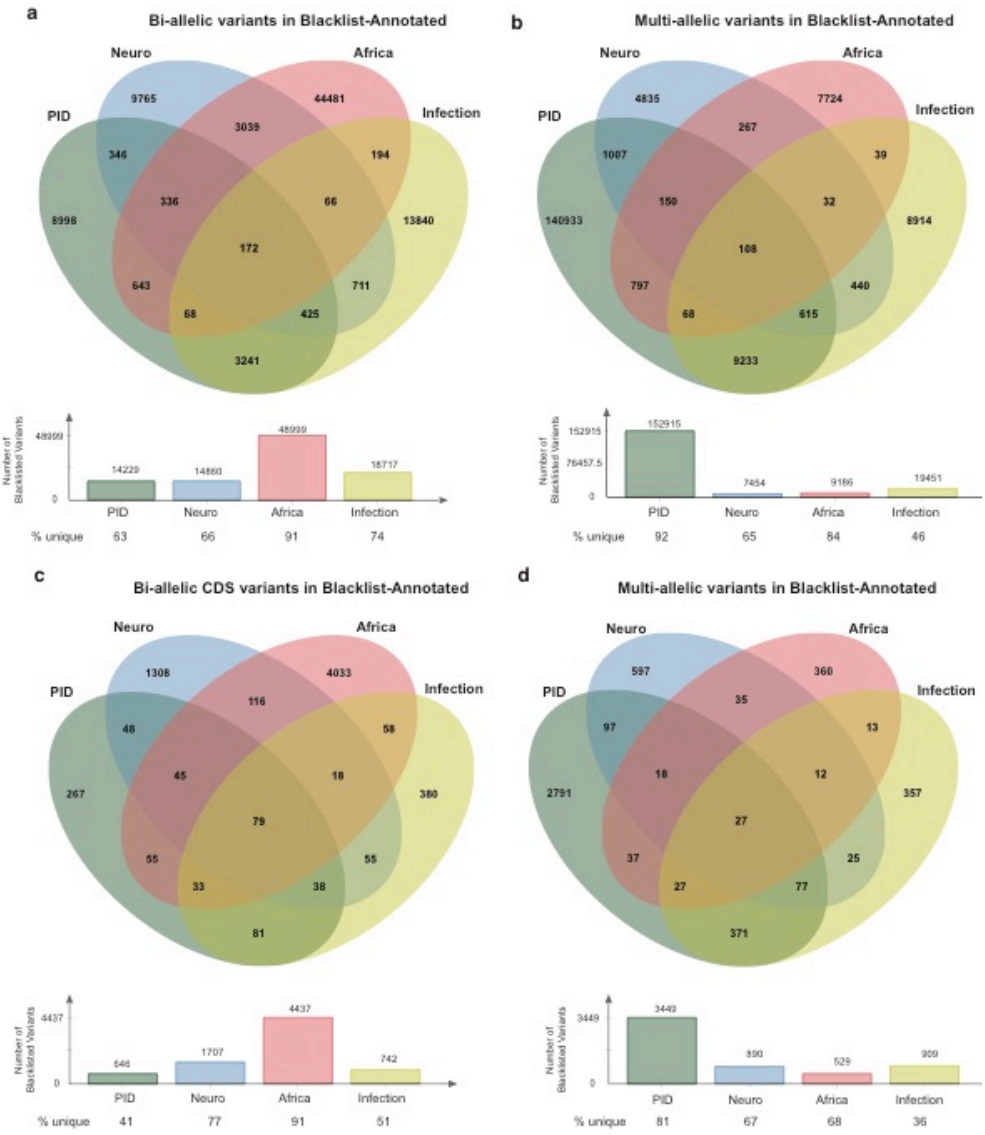


Figure 15. Relationship between the four blacklists

Common and unique biallelic (A), multiallelic (B), biallelic restricted to CDS (C), and multiallelic restricted to CDS (D) variants from the Blacklist-Annotated in the PID, Neuro, Africa and Infection panels.

We then assessed whether the originally generated PID blacklist would effectively filter exomes from the unrelated Neuro, Infection, and Africa panels used above. We removed variants with a MAF>0.01 in gnomAD from the Neuro, Infection, and Africa exomes and then applied the PID BL-A. This reduced the median number of remaining variants in the Neuro, Infection and Africa exomes by 8%, 41%, and 6%, respectively (median of 715, 2,487, and 947 variants per exome, respectively; Fig. 14A,C,E, blue box). When the analysis was restricted to cohort-private variants in the Neuro, Infection, and Africa exomes, the PID blacklist decreased the number of variants in individual exomes by 19%, 65%, and 11%, respectively (median of 673, 2,439, and 957 variants per exome, respectively, Fig. 14B,D,F blue box). The superior

efficiency of the PID blacklist for the Infection panel may reflect the library preparation technique (SureSelect) and sequencing technology (HiSeq sequencer) used. Nevertheless, the PID blacklist was shown to be a useful filtering approach in unrelated panels in which exomes were captured with different kits and sequencing technologies (SureSelect or Nextera kits and HiSeq 2000 or HiSeq 2500 sequencing, respectively). We also found that filtering our PID exomes with the blacklist from the Neuro panel did not remove any TP variants from the 129 PID exomes with proven disease-causing mutations. Blacklists are, therefore, effective for filtering exomes other than those with which they were developed and including cohort-private FP variants. However, generating internal blacklists from the panel under investigation was found to be the most effective approach to removing FP variants.

We sought to determine the minimum sample size appropriate for the generation of a custom blacklist for a panel of interest. We combined the two largest panels studied here — our PID panel (3,104) and the Neuro panel (3,869) — and simulated blacklists by randomly sampling various numbers of individuals relative to panel size, with 30 iterations for each sample size. As the Neuro panel was captured with the 50 Mb kit, which targets CDS, we focused this analysis exclusively on CDS variants. The number of CDS variants in the simulated BL-A increased rapidly with sample size between 10 and 500 individuals, whereas the number of variants increased more slowly when sample size exceeded 500 individuals. We therefore propose the use of samples of at least 500 heterogeneous unrelated individuals, to ensure the reliable capture of common cohort-specific variants. We estimated the saturation point for the blacklist's CDS variants (less than 1 new variant added per new individual) at a sample size of approximately 2,801 individuals. Thus, a blacklist generated with the pipeline described here could be considered "saturated" for the purpose of capturing most of the cohort-specific CDS variants that cannot be removed by public database analysis.

2.3.6 Efficacy of the combined blacklist

Finally, we explored the efficacy of a "universal" blacklist generated by combining the four BL-As presented in this study. We reasoned that the aggregation of blacklists obtained from different panels (and different samples/data processing methods) would result in a "universal blacklist" with the number of filtered variants eventually converging. We tested this hypothesis by aggregating either a) the four blacklists (PID, Neuro, Infection, and Africa blacklists) into a single 'combined blacklist'; or b) four combinations from the set of blacklists (Neuro, Infection, Africa) into four combined blacklists (i.e.: Neuro-Africa, Neuro-Infection, Africa-Infection,

Neuro-Africa-Infection), and applying the combined blacklists obtained in a) and b) to the PID panel. As the PID blacklist was not included in the four combined blacklists in b), we refer to these blacklists as 'non-cohort-specific combined blacklists'. These blacklists removed a decreasing number of variants with increasing size of the sets making up the blacklists (Fig. 16). After standard filtering with public databases, the 'Neuro-Africa' non-cohort-specific blacklist removed a median of 1,102 (8%) variants, the 'Neuro-Infection' non-cohort-specific blacklist removed a median of 3,833 (26%) variants, the 'Africa-Infection' non-cohort-specific blacklist removed a median of 3,886 (27%) of variants, and the 'Neuro-Africa-Infection' non-cohort-specific blacklist removed a slightly larger number of variants (median of 4,078, or 28% of variants). By contrast, the PID blacklist removed a median of 9,056 variants. The 'four combined' blacklist removed a median of 25 (0.45%) additional variants not captured by the PID blacklist alone (Fig. 16). Overall, these findings suggest that the number of variants filtered by the blacklist approach converges with the inclusion of blacklists from additional panels, consistent with the results for blacklist saturation. This universal filtering by blacklisting can be effectively applied to other individuals/panels, provided that the sequencing technology used, and the genetic ancestries of the panel are homogeneous. Moreover, the efficiency of a cohort-specific panel applied to a different panel (e.g. PID and infection panels) was greater for panels similar in terms of ethnic background and sequencing procedure (both mostly European and capture with similar kits), consistent with the results in Fig. 14C. Finally, although cohort-specific blacklists maximize the efficiency of this approach, the use of non-cohort-specific combined blacklists is nevertheless a very useful approach for filtering out a large number of unwanted variants, reinforcing the power of blacklist filtering even in the absence of a custom blacklist for the panel.

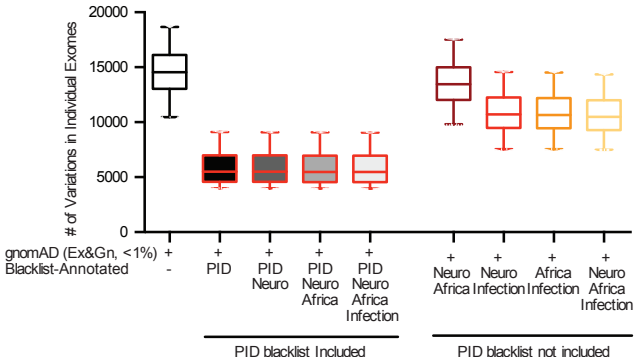


Figure 16. Efficiency of various combinations of the four blacklists.
Filtering of all variants in each PID exome with combinations of the various blacklists, with and without inclusion of the PID blacklist. Error bars represent the 10th to 90th percentiles.

2.4 Discussion

An essential step in the analysis of exomes from patients with rare genetic disorders is the removal of FP variants common in public databases (such as gnomAD, Bravo, and TopMed) at frequencies inconsistent with the prevalence, mode of inheritance, and penetrance of the disease(79). In principle, variants found to be common in a private panel but absent from public databases should also be filtered out. However, only one other previous study has explored the generation of filtering lists based on internal panels(80). Moreover, there are currently no tools available for filtering based on allele frequencies in internal panels. We report here the identification of in-house variants too common to cause rare monogenic illnesses (typically with a population prevalence of $<10^{-4}$) in a panel of 3,104 exomes. We assembled these variants into a blacklist and subsequently explored the use of this blacklist for filtering FPvariants from exome sequencing data, using the subset of variants that makes our approach unique (blacklist-annotated: those that are absent from public databases). These variants had high-quality metrics and 75% of them would not be captured by the rigorous application of available software, such as VQSR. We further validated this approach in three other independently processed and unrelated panels, demonstrating that our blacklist approach is generally, and perhaps universally, effective for filtering variants, and that the generation of blacklists specific to a given panel significantly increases the number of variants filtered out. We provide a computational tool (ReFiNE) for automatically generating in-house cohort-specific blacklists. We show that our blacklist can be used in synergy with standard public database filtering, to remove variants displaying disproportionate enrichment in an internal panel.

Public databases such as gnomAD, which represent major population groups (about half of individuals are of European ancestry and the others are a mixture of Admixed Americans, Africans/African Americans, South Asians, East Asians and Others), are an invaluable resource for estimating the frequency of variants in the general population and in different genetic ancestry groups. However, cohort-specific exomes may contain common variants (e.g. >1%) that are absent from or rare in public databases, partly because they are population-specific variants less represented in gnomAD (as observed for African(89) and Middle-Eastern individuals(84)). Moreover, public databases, such as gnomAD, make considerable efforts to ensure the rigorous removal of FP variants to ensure that they provide high-quality, high-stringency information about variants. However, these public databases do not provide a list of filtered FP variants and their summary statistics for filtration purposes. We demonstrated this with 113 1KG genomes generated by our in-house pipeline, showing that 23% of the variants

were absent from the public 1KG database, highlighting discrepancies between the analyzed and released data due to different bioinformatic procedures. Moreover, resources such as dbSNP are difficult to use for FP filtering because their FP variant rate is high(104). Therefore, even when using the latest versions of public databases and gene-level filtration(74, 75), ReFiNE is an effective tool for collecting data independently from external resources.

The technology associated with the NGS analyses (sequencing platform, targeting procedures and software) is strongly associated with the calling of the variants. We and others have previously observed biases specific for WES and WGS(96) or variant-calling pipelines(105). Differences in technology can therefore lead to the mis-annotation of variants in a given panel. The main sources of mis-annotation are as follows: (i) variants in gnomAD collected by different technologies (PCR for WES and PCR-free + PCR for WGS) apply rigorous QC cutoffs based on high-quality technologies, resulting in higher proportions of variants from lower-quality technologies being removed; (ii) despite the presence of 15,496 genomes in gnomAD, some genomic regions remain poorly covered or not covered at all, whereas these regions are covered by our panel and contain variants (2% of our BL-A); (iii) a recent comparative studied revealed strong discrepancies between the variant callers used in NGS analyses(106); these discrepancies have been highlighted by the differences between the gnomAD and ExAC databases (https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/); and (iv) the annotation of NGS variants in multiallelic positions is often problematic(107) because current annotation software (SNPeff (108), VEP(109), ANNOVAR(110)) cannot identify these variants efficiently. Indeed, 91.5% of our blacklist variants were located at multiallelic sites according to gnomAD's genome annotation. Each panel is unique (in terms of technology, quality, ethnicities). Our blacklisting resource is intended to fill this gap, particularly for researchers without the large exome or genome databases required for filtering with computationally intensive methods, such as VQSR. ReFiNE can, thus, overcome anomalies in sequence alignment or variant-calling processes, such as large indel events(111).

We show here that analyses of variant frequency within internal panels constitute an additional method for filtering out variants too common to cause rare disease. The blacklists generated by ReFiNE are easy to use and rapidly identify FPs that may confound the dissection of patient exomes. As WES and WGS are increasingly widely used for the investigation of genetic disorders in patients, it will be possible to apply the blacklisting approach described here and ReFiNE software to larger panels of patients, facilitating the effective identification of FPs in these panels. However, caution is required when generating blacklists with ReFiNE

from phenotypically homogeneous panels, particularly if of the same underrepresented ethnic origin, as this approach may remove TP variants in such conditions. Finally, such extensive, rapidly generated blacklists (1 hour for 3,104 exomes) should increase the efficiency of FPs elimination from exomes and genomes, without the need for the large computer clusters required by current machine-learning algorithms, such as VQSR (a month for 3,104 exomes). As exome capture kits become increasingly efficient, and with the widespread adoption of WGS, the blacklists generated by ReFiNE will facilitate efficient noise reduction in NGS data, independently of the technology used, making it easy to find the needles in increasingly large haystacks of genetic variants in patients.

3 Identification of homozygous and hemizygous (HMZ) partial exon deletions

3.1 Introduction: identification of CNVs from WES data

Copy number variations (CNVs) are unbalanced rearrangements, classically covering more than 50 base pairs (bp), that increase or decrease the number of copies of specific DNA regions(59, 60). There is growing evidence to implicate CNVs in common and rare diseases(59, 61, 62, 112). CNVs have also been linked to adaptive traits, in environmental contexts for example(112). It has been recently estimated that CNVs affect ~5–10% of the genome, suggesting that a number of potentially disease-causing CNVs have yet to be discovered(59, 63). Next-generation sequencing (NGS) techniques, such as whole-genome and whole-exome sequencing (WGS and WES), provide unprecedent opportunities for studying CNVs. Computational tools using data from WGS have been successfully used to detect CNVs(25, 113-115), but WES-based methods have met with more limited success, mostly due to the nature of targeted enrichment protocols(64-66). Common WGS-based methods use breakpoints, the regions in which the rearrangements occur, to detect CNVs. By contrast, WES focuses on noncontiguous genomic targets (the exons), and most breakpoints are not sequenced. Hence, current WES-based approaches for detecting CNVs use the read depth (or coverage information) as a proxy for copy number information.

The HMZDelFinder algorithm is a recently developed coverage-based method for detecting rare homozygous and hemizygous (HMZ) deletions(67). This subset of CNVs may result in null alleles and a complete loss of gene function. Their identification may, therefore, lead to the discovery of novel genes or variations underlying Mendelian diseases. HMZDelFinder jointly evaluates the normalized per-interval coverage of all the samples of the entire dataset, making it possible to detect rare exonic HMZ deletions while minimizing the number of false-positive calls due to low-coverage regions. HMZDelFinder outperformed other CNV-calling tools, such as CONIFER(116), CoNVex(117), XHMM(118), ExonDel(119), CANOES(120), CLAMMS(121) and CODEX(122), particularly for the detection of single-exon deletions (i.e. deletions spanning only one exon)(67). However, two major limitations remain to be addressed. First, HMZDelFinder has been optimized to detect HMZ deletions from an entire dataset (>500) of homogeneous exome data. Its performance for typical laboratory panel, which include exome

data generated over time, often under different conditions, is, therefore, not optimal. Second, HMZDelFinder was not designed for the systematic detection of partial exon deletions (i.e. deletions spanning less than one exon). Here, we provide HMZDelFinder_opt, a method that extends the scope of HMZDelFinder by improving the performance of the algorithm for the calling of HMZ deletions in typical laboratory panels, which are generated over time, and by allowing the systematic detection of partial exon deletions.

3.2 Methods

3.2.1 Description of the panel

The 3,954 individuals used in this study were recruited in collaborations with clinicians, and most of them present different severe infectious diseases. Probands' family members account for the rest. Although these individuals do not form a random sample, they were ascertained through a number of distinct phenotypes and in different countries. Cohort-specific effects are, therefore, not expected to bias patterns of variation. All study participants provided written informed consent for the use of their DNA in studies aiming to identify genetic risk variants for disease. IRB approval was obtained from The Rockefeller University and Necker Hospital for Sick Children, along with a number of collaborating institutions.

WES and bioinformatics analysis were performed as previously described(45). Briefly, genomic DNA was extracted and sheared with a Covaris S2 Ultra-sonicator. An adaptor-ligated library (Illumina) was generated, and exome capture was performed with either SureSelect Human All Exon kits (V5-50Mb, V4-50Mb, V4-71Mb, or V6-60Mb) from Agilent Technologies, or xGen Exome Research 39Mb Panel from Integrated DNA Technologies (IDT xGen) (Table 4). Massively parallel WES was performed on a HiSeq 2500 machine (Illumina), generating 100- or 125-base reads. Quality controls were applied at the lane and fastq levels. Specifically, the cutoff used for a successful lane is Pass Filter > 90%, with over 250 M reads for the high-output mode. The fraction of reads in each lane assigned to each sample (no set value) and the fraction of bases with a quality score > Q30 for read 1 and read 2 (above 80% expected for each) were also checked. In addition, the FASTQC tool kit (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to review base quality distribution, representation of the four nucleotides of particular k-mer sequences (adaptor contamination). We used the Genome Analysis Software Kit (GATK) (version 3.2.2 or 3.4-46)

best-practice pipeline to analyze our WES data(33). Reads were aligned with the human reference genome (hg19), using BWA(123). PCR duplicates were removed with Picard tools (picard.sourceforge.net/). The GATK base quality score recalibrator was applied to correct sequencing artifacts.

Kit	Kit (full name)	Number (Percentage) of Exomes	Median Coverage (SD)	% bases above 10X
IDT-xGen	xGen Exome Research Panel v2 from Integrated DNA Technologies	188 (4.8%)	41.7 (9.5)	91.4
V4-50Mb	Agilent SureSelect Human All Exon V4	354 (9.0%)	50.0 (15.5)	83.2
V4-71Mb	Agilent SureSelect Human All Exon V4+UTRs	3095 (78.3%)	47.4 (10.2)	81.0
V5-50Mb	Agilent SureSelect Human All Exon V5	101 (2.6%)	72.4 (43.7)	70.3
V6-60Mb	Agilent SureSelect Human All Exon V6	216 (5.5%)	125.9 (38.6)	99.0

Table 4. Distribution of the capture kit in the 3,954 exomes and corresponding coverage metrics

3.2.2 Positive controls

The five WES samples used as positive controls carry rare HMZ disease-causing deletions that were confirmed with state-of-the-art molecular approaches(124-126). Specifically, these HMZ deletions comprise one or more exons and have different lengths as follows (Table 5). P1 carries a deletion of exons 21 to 23 in *DOCK8* (10,800 bp) that was validated by multiplex ligation-dependent probe amplification (MLPA). The deletion in *DOCK8* was functionally linked to staphylococcus infection(124). P2 had a deletion of exon 5 in *NCF2* (134 bp) that was also validated by MLPA and found to be causal in chronic granulomatous disease (manuscript in preparation). P3's deletion spanned exons 2 to 8 in *IL12RB1* (13,000 bp) and was validated by sanger sequencing. This deletion was demonstrated to be causal for a Mendelian susceptibility to mycobacterial disease(125). P4 has a deletion of the entire *CYBB* (3,400,000 bp) validated by MLPA and CGH array that resulted in chronic granulomatous disease(126). Finally, P5 is a patient with hyper IgE syndrome carrying a deletion of exons 7 to 15 in entire *DOCK8* (28,000 bp) that was validated by Sanger sequencing. *CYBB* is on the X chromosome while all other genes are autosomal.

Patient	Confirmed Homozygous Deletion				Exome		
	Location	Gene	Size (kbp)	Validation method	Mean	%	Bases
P1	Chr 9, Exons 21 to 23	DOCK8	10.8	MLPA	23	68.9	
P2	Chr 1, Exon 5	NCF2	0.13	MLPA	115	99.5	
P3	Chr 19, Exons 2 to 8	IL12RB1	13	Sanger sequencing	206	99.5	
P4	Chr X, Whole gene	CYBB	3,400	MLPA and CGH array	156	99.2	
P5	Chr 9, Exons 7 to 15	DOCK8	28	Sanger sequencing	66	99.5	

Table 5. Validated rare HMZ disease-causing deletions and exome coverage in the five exomes used as positive controls

3.2.3 HMZDelFinder-opt

The general workflow used in HMZDelFinder-opt is depicted in Figure 17. First, HMZDelFinder_opt computes coverage profiles from the BAM files of the entire dataset. Second, the Principal component analysis (PCA) is calculated from a covariance matrix based on standardized coverage profiles and a k nearest neighbors algorithm is used to select the reference control set. Third, the BAM file of a given sample and the BAM files of the reference control set are used as input of HMZDelFinder to detect HMZ deletions. Fourth, when HMZDelFinder_opt is provided with the parameter -sliding_window_size and the related size, it will employ a sliding window approach for identification of partial deletions of exons. Each of these steps is described in the following paragraphs.

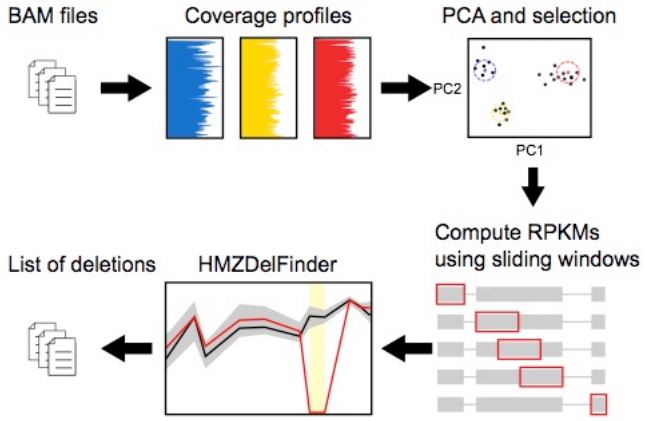


Figure 17. Schematic representation of the method employed by HMZDelFinder_opt to detect partial-exon homozygous and hemizygous deletions

First, HMZDelFinder_opt computes coverage profiles from the BAM files. The PCA is then calculated from a covariance matrix based on standardized coverage profiles and a k nearest neighbors algorithm is used to select the reference control set. The BAM file of a given sample and the BAM files of the reference control set are used as input of HMZDelFinder to detect homozygous and hemizygous deletions. In addition, HMZDelFinder_opt accepts a parameter (-sliding_window_size) to employ a sliding window approach for identification of partial-exon deletions.

3.2.3.1 Principal component analysis (PCA) and k nearest neighbors algorithm

The PCA was performed on the coverage profile of the 3,954 WES using per-exon coverage. Specifically, for each sample, the coverage profile was calculated using the mean depth of coverage of the 194,528 exons from the consensus coding sequences (CCDS) annotation of GRCh37 obtained using biomaRt(127). The PCA was then performed using the 'prcomp' function from R 3.5.1 on the scaled coverage profiles. To select the reference control set for a given sample, we computed pairwise weighted Euclidean distances between

individuals i and j based on the first 10 principal components from the PCA using the 'dist' function of R 3.5.1, using the formula:

$$dist(i,j) = \sqrt{\sum_{k=1}^{10} \lambda_k \left(PC_{ki} - PC_{kj}\right)^2}$$

where PC is the matrix of principal components (PCs) calculated on the coverage of common variants and λ_k the eigenvalue corresponding to the k-th principal component PC_k.

3.2.3.2 HMZDelFinder

We used the HMZDelFinder algorithm as described(67). In brief, HMZDelFinder calculates per-exon read depth (reads per thousand base pairs per million reads; RPKM) to detect HMZ deletions. For our purpose of covering all the coding regions, we employed an interval file containing all coding sequences from Gencode. For a given interval, the criteria to call a deletion are as follows: 1) RPKM < 0.65 and 2) frequency of the deletion within the dataset ≤ 0.5%. Filtering criteria at the interval and sample levels include removal of low quality intervals (RPKM median < 7 across all samples) and removal of low quality samples (2% with highest number of calls). When using the optional absence of heterozygosity (AOH) step, HMZDelFinder uses VCF files to filter out deletions not falling in AOH regions, assuming that rare and pathogenic homozygous deletions are likely to be located within larger AOH regions due to the inheritance of a shared haplotype block from both parents. Finally, to prioritize deletions, z-scores are computed. The z-score of a deletion measures the number of standard deviations between the coverage of the deleted interval in a given sample compared to the mean coverage of the same interval in the rest of the dataset. A very low z-score indicates high mean coverage with low variance in the dataset and very low (or no coverage at all) in a given sample. Hence, lower z-scores denote higher confidence in a given deletion.

3.2.3.3 Sliding window approach and simulated data

We simulated deletions of variable size in 200 randomly selected individuals among our in-house panel but excluding the oldest samples (V4-50Mbp capture kit), due to a lower quality than present standards. Two different exons were selected to undergo simulated deletions: a favorable case, exon 11 from LIMCH1 gene (409bp) with a mean coverage of approximately 85X in our samples, and an unfavorable case, exon 4 from RPL15 gene (406 bp) with a mean

coverage of 15X in our samples. For both exons, we deleted a segment of 25%, 50%, 75% or 100% of the exon size, using the '-v' argument of the 'bedtools intersect' command (bedtools v1.9) on the BAM file to remove all reads overlapping the segment. We then ran HMZDelFinder and HMZDelFinder_opt (with and without the --sliding_windows parameter) on the whole BAM files. Specifically, we applied a sliding window approach, in which each exon was divided into 100 bp windows, with 50 bp overlaps, and BAM files for individual exomes were transformed into per-window read depths. In a separate analysis, we used 50 bp windows, with 25 bp overlaps.

3.2.4 Analysis of common deletions

To determine whether some of the called deletions were previously reported as common deletions, we utilized the CNVs from the Gold Standard track (hg19 version dated 2016-05-15) of the Database of Genomic Variants (DGV), a highly curated resource that collects CNVs in the human genome(128). We retained only entries with field 'variant_sub_type' equal to 'Loss' and frequency greater than 1%. We then crossed the retained entries with the deletions called by HMZDelFinder and HMZDelFinder_opt in the positive controls. Deletions were considered common in the DGV database when they overlapped at least 50% with the retained entries from the DGV database.

3.3 Results

3.3.1 Determination of the reference control set

We first aimed to improve the performance of HMZDelFinder for detecting HMZ deletions in typical heterogeneous laboratory panels, which were generated over time and in different experimental settings (e.g. capture kit). We reasoned that comparing a given sample with an optimized reference control set would limit the impact of the background variability intrinsic to exome data, thereby improving the performance of HMZDelFinder. We designed the optimized reference control set as a selection of samples with similar coverage profiles (Figure 17). We did this by first performing a principal component analysis (PCA) of the depth of coverage for consensus coding sequences (CCDS) for 3,954 exomes from our in-house panel, including mostly patients with severe infectious diseases. As expected, given the different

sequencing conditions used for whole-exome sequencing (Table 4), the coverage profiles of the samples were highly variable (Figure 18). The first two principal components (PCs) of the PCA identified six distinct clusters, mostly reflecting the capture kit used (Figure 18).

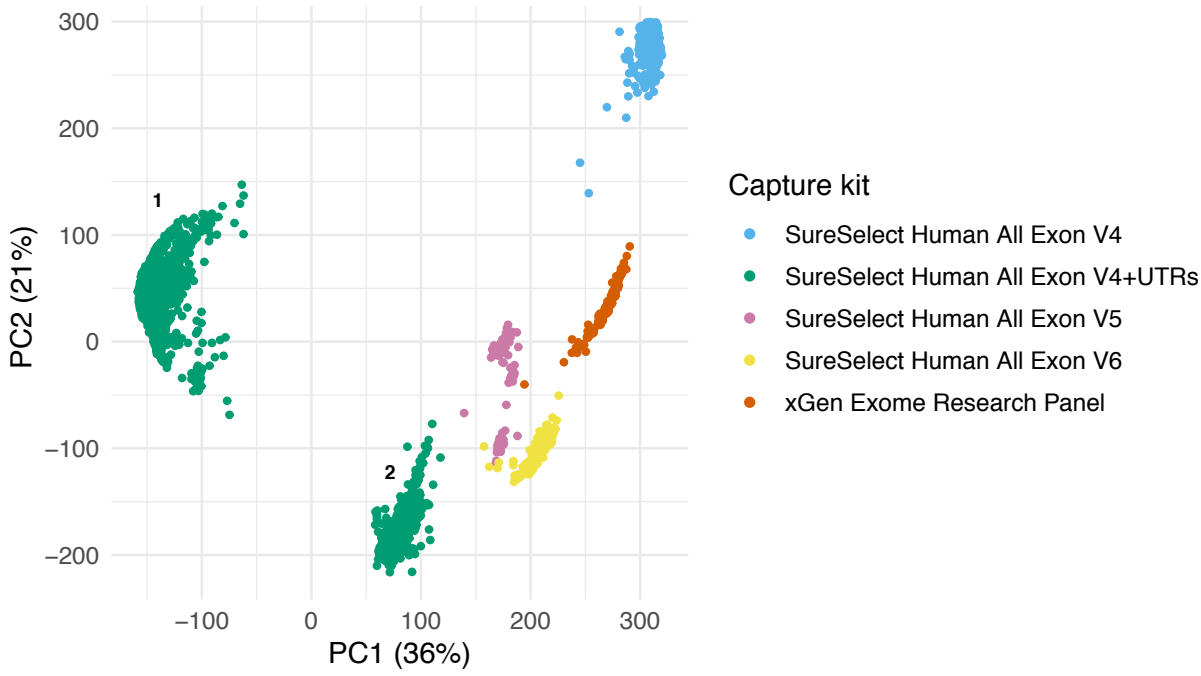


Figure 18. Principal Component Analysis (PCA) of the WES coverage
The PCA was computed from the coverage profiles of consensus coding sequences (CCDS) from 3,954 individuals. Dots are color-coded by the type of the capture kit used for sequencing. Two different clusters (clusters 1 and 2) corresponded to the V4-71Mb capture kit.

Interestingly, two different clusters (clusters 1 and 2 on Figure 1) corresponded to the V4-71Mb capture kit, the difference between these clusters being associated mostly with a minor change in the sequencing chemistry of the kit, leading to a significant improvement in coverage profile for the more recently generated exome data (not shown). We then used the first 10 PCs to calculate the pairwise weighted Euclidean distances between all samples(97) (see methods). We used this metric to determine, for each sample of interest, the closest neighbors, for use as the reference control set in HMZDelFinder_opt.

3.3.2 Optimization of the reference control set in HMZDelFinder_opt

We then compared the performances of HMZDelFinder_opt and HMZDelFinder, using five WES samples carrying validated rare HMZ disease-causing deletions of different lengths as positive controls (Table 5, methods). Specifically, we tested the ability of HMZDelFinder_opt

and HMZDelFinder to detect the validated deletions, and we also compared the total numbers of deletions called and their *z*-scores (see Methods). In HMZDelFinder_opt, we compared reference control sets of different size (ranging from 50 to 500, Figure 19), selected for each sample as described above. In HMZDelFinder, we used the entire dataset, consisting of 3,954 WES samples. For both approaches, the final set of called deletions for each sample was narrowed down to the capture kit corresponding to the patient WES data. We chose to benchmark HMZDelFinder because it has been shown to perform at least as well as, and sometimes better than several widely used and actively maintained detection tools(67).

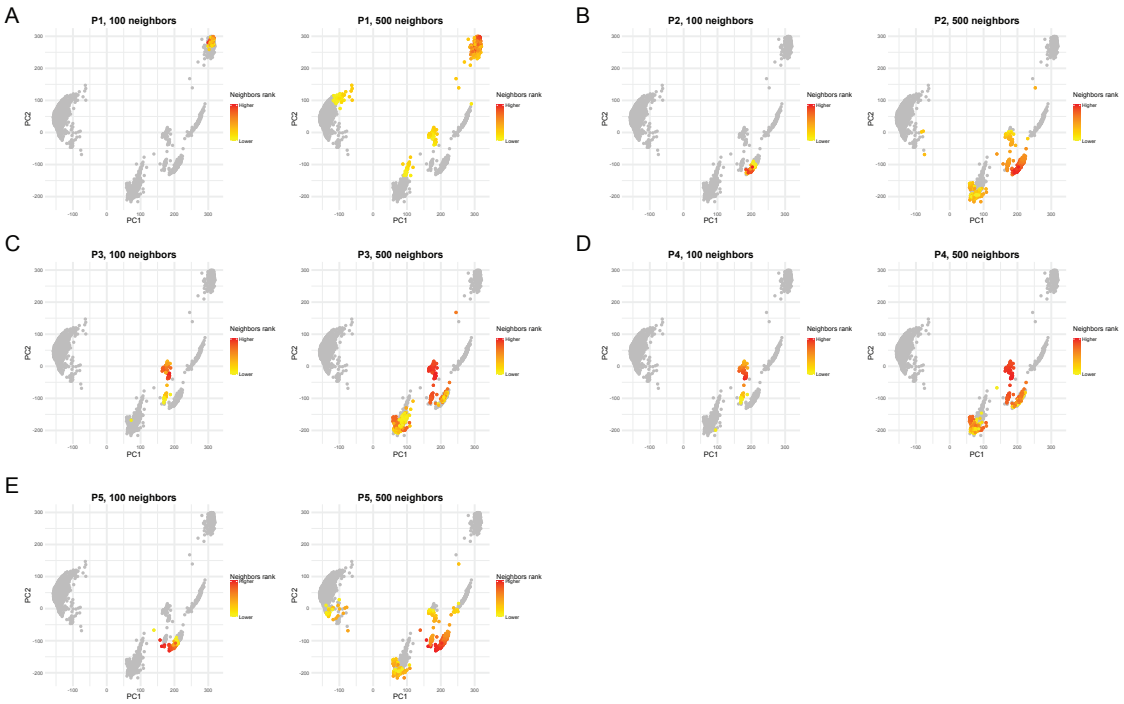


Figure 19. Closest neighbors of the positive controls as function of the size of the reference control set
A total of 100 and 500 neighbors are showed for P1 (A), P2 (B), P3 (C), P4 (D), and P5 (E).

Both HMZDelFinder and HMZDelFinder_opt successfully detected all five confirmed HMZ deletions in the positive controls, regardless of the size of the reference control set (Table 6). However, HMZDelFinder_opt detected a smaller total number of deletions than HMZDelFinder (Table 6). Specifically, the total number of deletions ranged from one to 21 deletions for HMZDelFinder_opt, and from 11 to 2,586 for HMZDelFinder, suggesting that a smaller number of false-positive calls were obtained with HMZDelFinder_opt. Using the optional filtering step based on the absence of heterozygosity (AOH) information for HMZDelFinder (see methods) decreased the number of deletions detected, but this number nevertheless remained much higher than that for HMZDelFinder_opt (Table 6). We hypothesized that the large difference between the two methods for P1 reflected the low quality

of exome data for this patient. Indeed, the mean coverage and the proportion of bases with coverage above 10x were much lower for P1 than for the other four patients (e.g. only 68.9% of bases had a coverage above 10x for P1, versus >99% for the other patients) (Table 4), leading to a large number of likely false positive deletions detected when not using an appropriate reference control set with similar coverage. Consistently, the number of deletions detected for P1 with HMZDelFinder_opt was larger with the largest reference sample size (500) (Table 6). We therefore performed subsequent HMZDelFinder_opt analyses with a reference sample size of 100, which provided a good compromise between the algorithm performance and computation time.

		P1	**P2**	**P3**	**P4**	**P5**
	KIT	V4-50MB	V6-60MB	V5-50MB	V5-50MB	V6-60MB
METHOD	**N NEIGHBORS**	**Confirmed deletion (Rank/Total number of deletions)**				
HMZDelFinder_opt	50	DOCK8 (1/11)	NCF2 (1/2)	IL12RB1 (1/1)	CYBB (3/5)	DOCK8 (1/3)
	100	DOCK8 (1/11)	NCF2 (1/2)	IL12RB1 (1/1)	CYBB (4/5)	DOCK8 (1/2)
	200	DOCK8 (1/11)	NCF2 (1/3)	IL12RB1 (1/1)	CYBB (4/5)	DOCK8 (1/3)
	500	DOCK8 (4/21)	NCF2 (1/2)	IL12RB1 (1/3)	CYBB (3/5)	DOCK8 (1/2)
HMZDelFinder	All	DOCK8 (1/2586)	NCF2 (120/120)	IL12RB1 (4/11)	CYBB (7/13)	DOCK8 (1/163)
HMZDelFinder AOH	All	DOCK8 (1/457)	NCF2 (37/37)	IL12RB1 (2/5)	CYBB (4/7)	DOCK8 (1/46)

Table 6. Comparison of the deletions called by HMZDelFinder_opt and HMZDelFinder.
Comparison between HMZDelFinder_opt and HMZDelFinder by using five positive controls carrying validated rare HMZ disease-causing deletions. The table reports the confirmed deletion for each positive control with the rank and total number of called deletion in parenthesis.

We then compared the rankings of the confirmed deletions between the two algorithms, using the *z*-score provided by HMZDelFinder (see method). While the two approaches ranked the confirmed disease-causing deletions for P1 and P5 first, HMZDelFinder_opt ranked higher the confirmed disease-causing deletions for P2, P3 and P4 than HMZDelFinder (Table 6; Figure 20). Moreover, *z*-scores were consistently better with HMZDelFinder_opt (Figure 20) than with HMZDelFinder, leading to a more specific discovery of true HMZ deletions. Again, using the AOH option for HMZDelFinder slightly removed the number of deletions, thus slightly improving the ranking (Table 6), but not changing the absolute the *z*-score value. Together,

these results suggest that HMZDelFinder_opt gives better z-scores for deletions than HMZDelFinder, which should lead to higher sensitivity in the general case.

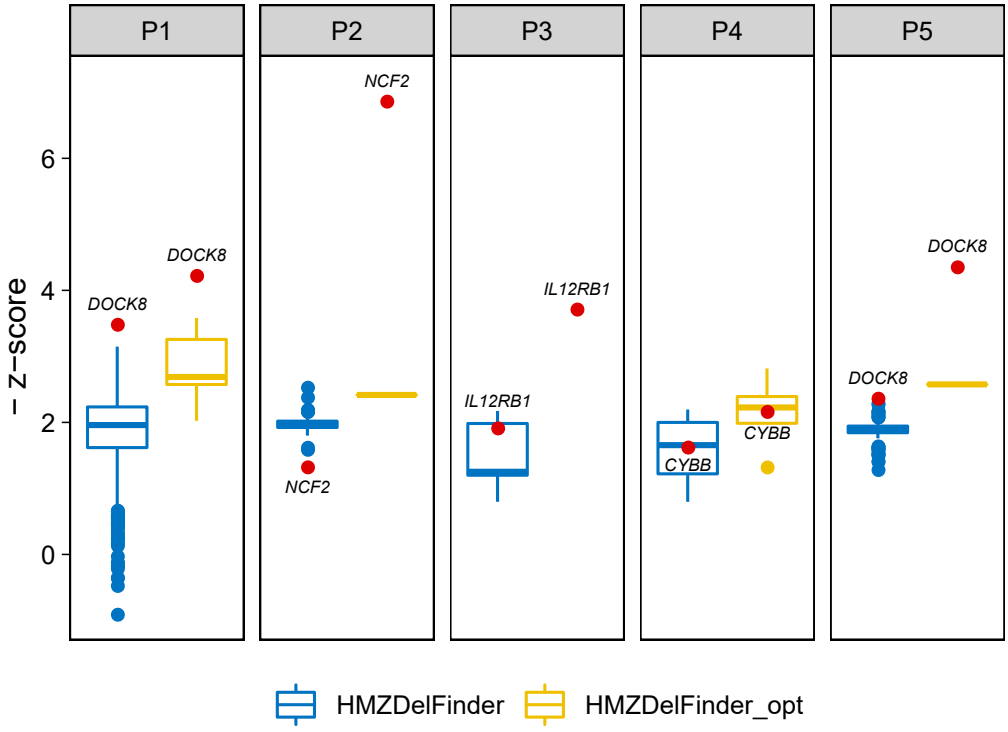


Figure 20. Comparison of the ranking of the deletions called by HMZDelFinder_opt and HMZDelFinder. *The ranking is expressed as - z-score. Lower z-scores (and higher ranking) indicate more confidence in a given deletion. The confirmed deletions ranked 1st in P1, P2, P3, P5 with HMZDelFinder_opt while they ranked 1st only in P1 and P5 with HMZDelFinder as shown by the red dots in the blue (HMZDelFinder) and yellow (HMZDelFinder_opt) distributions. The ranking was consistently higher with HMZDelFinder_opt than with HMZDelFinder. Results are shown for HMZDelFinder_opt using 100 as size of the reference control set.*

Finally, we studied the HMZ deletions called by both approaches, in addition to the validated ones, to determine whether some of the deletions identified were reported as common deletions. We used the CNVs from the gold standard track of the Database of Genomic Variants (DGV), a highly curated resource containing CNVs from the human genome(128). We focused on the positive controls with high data quality (P2, P3, P4 and P5), and found that the HMZ deletions called by HMZDelFinder_opt were more enriched in common deletions (frequency > 1%) than those called by HMZDelFinder (Table 7). Among the 6 and 303 additional HMZ deletions called by HMZDelFinder–opt (with the reference control set of 100 exomes) and HMZDelFinder, 50% and 1%, respectively, were present in the DGV database (Table 7), suggesting that the deletions called by HMZDelFinder_opt were enriched in true deletions. Overall, these findings demonstrate that the use of an appropriate reference control set of WES data based on a PCA-derived coverage distance improves the performance of HMZDelFinder.

These results also provided a first validation of HMZDelFinder_opt for five confirmed disease-causing HMZ deletions.

		P2	P3	P4	P5	TOTAL
	KIT	V6-60MB	V5-50MB	V5-50MB	V4-71MB	
METHOD	N NEIGHBORS	(COMMON DELETIONS/NUMBER OF OTHER DETECTED DELETIONS)				
HMZDelFinder_opt	50	0/1 (0%)	0/0 (-)	2/4 (50%)	2/2 (100%)	4/7 (60%)
	100	0/1 (0%)	0/0 (-)	2/4 (50%)	1/1 (100%)	3/6 (50%)
	200	0/2 (0%)	0/0 (-)	2/4 (50%)	2/3 (67%)	4/9 (44%)
	500	0/1 (0%)	0/2 (0%)	2/4 (50%)	1/1 (100%)	3/8 (38%)
HMZDelFinder	all	0/119 (0%)	0/10 (0%)	1/12 (8%)	2/162 (1%)	3/303 (1%)

Table 7. Comparison of the number and percentage of common deletions.
Number and percentage of common deletions (>1% frequency in DGV) among the detected deletions (other than the confirmed deletion).

3.3.3 Detection of HMZ partial exon deletions by HMZDelFinder_opt

In HMZDelFinder, individual exome BAM files are transformed into per-exon read depths, facilitating a more efficient detection of single-exon HMZ deletions than can be achieved with other classical CNV-calling algorithms(67). Here, we aimed to address the need for the identification of even smaller HMZ deletions, spanning less than an exon (partial exon deletions). To this end, we used HMZDelFinder_opt with a sliding window approach, in which each exon was divided into 100 bp windows, with 50 bp overlaps, and BAM files for individual exomes were transformed into per-window read depths. We tested this approach by simulating deletions in two exons of similar size (~400 bp) but with different mean coverages in a randomly selected dataset of 200 WES samples from our in-house panel. The deletions spanned 100%, 75%, 50% or 25% of either exon 11 of *LIMCH1* (409 bp, ~85x mean coverage) or exon 4 of *RPL15* (406 bp, ~15x mean coverage). We used these datasets to compare the performances of HMZDelFinder_opt with sliding windows of 100 bp (HMZDelFinder_opt+sw100), HMZDelFinder_opt without sliding windows (HMZDelFinder_opt), and the original HMZDelFinder. For HMZDelFinder_opt+sw100 and HMZDelFinder_opt ,we used reference control sets of size 100.

For deletions spanning the full exon (100%), we confirmed that HMZDelFinder_opt had a detection rate (98% and 93% for exons with higher and lower coverage, respectively; Figure

21) similar to that of HMZDelFinder (98% and 93% for exons with higher and lower coverage, respectively). However, the total number of HMZ deletions called by HMZDelFinder_opt was only one eighth the total number of HMZ deletions called by HMZDelFinder (median number of HMZ deletions: 2 vs. 13 Figure 22) The detection rate was slightly higher when sliding windows were used (detection rate for HMZDelFinder_opt+sw100 of 99% and 94% for exons with a higher and lower coverage, respectively), but at the cost of a slightly larger total number of HMZ deletions called than for HMZDelFinder_opt (median number of deletions: 5 vs. 2). Nevertheless, the total number of HMZ deletions called by HMZDelFinder_opt+sw100 remained lower than the total number of HMZ deletions called by HMZDelFinder.

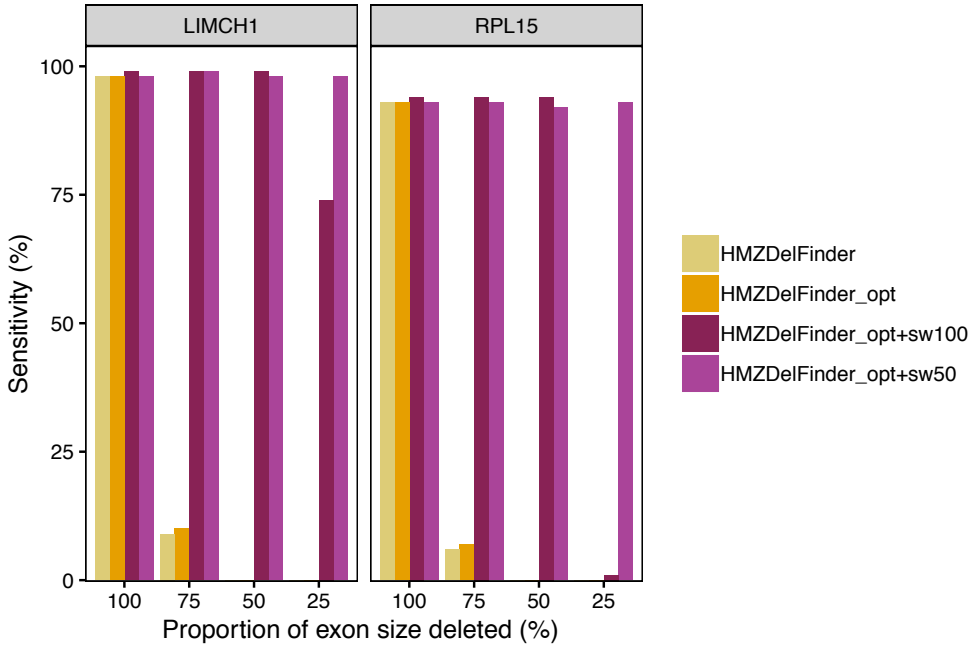


Figure 21. Comparison of the proportion of deletions detected between HMZDelFinder_opt with or without sliding windows and HMZDelFinder.
Proportions of deletions detected in simulated data in the higher (LIMCH1) or lower (RPL15) covered exons by using HMZDelfinder (yellow), HMZDelFinder_opt (orange), HMZDelFinder_opt+sw100 (red), HMZDelFinder_opt+sw50 (pink).

For partial exon deletions, the detection rates of HMZDelFinder and HMZDelFinder_opt were much lower, at less than 10% for deletions spanning 75% of the exon and 0% for deletions spanning 25% or 50% of the exon. Conversely, HMZDelFinder_opt+sw100 succeeded in detecting simulated deletions spanning 50% or 75% (200 bp or ~300 bp) of both exon 11 of *LIMCH1* and exon 4 of *RPL15* in 99% of the samples, with a median number of called HMZ deletions of 5 (Figure 21 and 22). For deletions spanning 25% of the exon (~100 bp), HMZDelFinder_opt+sw100 had a detection rate of 74% for the exon with the highest coverage

in *LIMCH1,* but it failed to detect the deletions in the exon with the lowest coverage in *RPL15*. We assessed the performance of this method further, using a smaller sliding window of 50 bp in size, and a step size of 25 bp, to improve granularity. We found that the use of smaller sliding windows with HMZDelFinder_opt+sw50 greatly increased the detection rate for deletions spanning 25% of the exon with the lowest coverage, exon 4 of *RPL15* (93% for sw50 vs. 1% for sw100) and of the exon with the highest coverage in *LIMCH1* (98% for sw50 vs. 74% for sw100) (Figure 21). Thus, the use of a sliding window makes it possible to detect HMZ partial exon deletions that would otherwise be missed, and the use of simulated data further validated HMZDelFinder_opt.

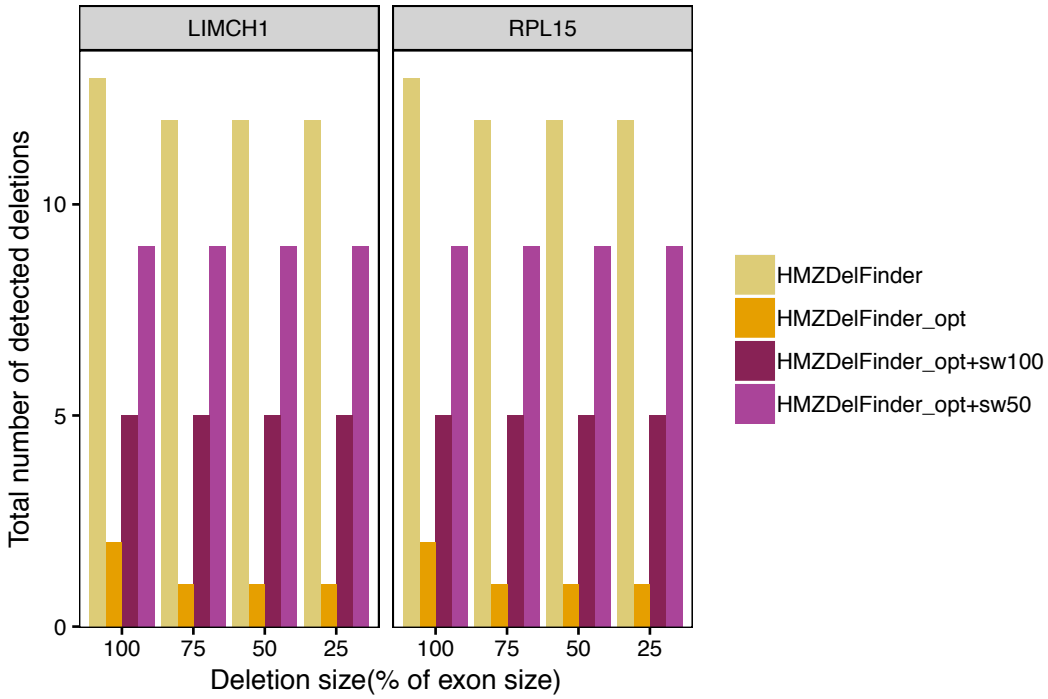


Figure 22. Comparison of the number of deletions detected between HMZDelFinder_opt with or without sliding windows and HMZDelFinder.

Median number of detected deletions in the simulated data in the higher (LIMCH1) or lower (RPL15) covered exons by using HMZDelfinder (yellow), HMZDelFinder_opt (orange), HMZDelFinder_opt+sw100 (red), HMZDelFinder_opt+sw50 (pink).

3.4 Discussion

WES offers unprecedent opportunities for identifying HMZ deletions as novel causal determinants of human diseases, but it poses a number of computational challenges. Most current methods for detecting HMZ deletions compare the depth of coverage between a given exome and the rest of the exomes in the dataset. However, coverage depth is heavily dependent

on sequencing conditions, which are continually evolving in typical laboratory settings. Thus, the exome data generated over time are inevitably heterogeneous, complicating the discovery of deletions. Using HMZDelFinder_opt with both validated disease-causing deletions and simulated data, we demonstrated that the *a priori* selection of a reference control set with a coverage profile similar to that of the WES sample studied reduced the number of deletions detected, while improving the ranking of the true HMZ deletion. These results are consistent with a recent report showing that the selection of an appropriate reference control set with multidimensional scaling significantly improves the sensitivity of various CNV callers(129). In further support for our findings, the ranking of the known deletion and the number of additional deletions detected by HMZDelFinder_opt start worsening with increasing numbers of controls in the reference set, including neighbors with a less similar coverage profile, as illustrated, for P1, in Fig 19A.

In addition to providing an optimized tool for detecting deletions in typical laboratory panels, HMZDelFinder_opt also fills the gap in the study of deletions spanning less than an exon, by providing the first tool for the systematic identification of partial exon deletions. Existing CNV callers are optimized for the detection of either large deletions (usually spanning more than three exons), or deletions of full single exons(67, 68). Other established callers, such as GATK, are not designed to detect CNVs and can therefore identify deletions of only a few dozen base pairs (typically up to 50 bp, https://gatkforums.broadinstitute.org/gatk/discussion/5938/using-gatk-tool-how-long-insertion-deletion-could-be-detected and (130)). The human genome contains ~235,000 exons, about 20% of which are larger than 200 bp(131). HMZDelFinder_opt therefore makes possible the systematic discovery of currently unknown HMZ deletions in ~47,000 exons that are not detectable with other tools. In sum, we describe HMZDelFinder_opt, a method for improving the detection of HMZ deletions in heterogeneous exome data that can be used to identify partial exon deletions that would otherwise be missed, through an extension of the scope of HMZDelFinder.

4 HW equilibrium and implications for population genetics events: preliminary findings

4.1 Introduction: what HW equilibrium hints beyond the technical errors

The Hardy-Weinberg (HW) law or equilibrium is a basic principle of genetics(48, 49). It states that allele and genotype frequencies in a given population are constant from generation to generation, in the absence of evolutionary influences (e.g.: no migration, no mutation, no natural selection, very large population and random mating). Given the conditions of absence of evolutionary influences are usually considered to hold, deviations from HW equilibrium (or HW disequilibrium) have been traditionally considered indicative of technical errors(50-52). This principle was originally employed as filtering criterion in large-scale genotyping studies (e.g. genome-wide association studies or GWAS) and "lent" to exome studies without rigorous investigation.

In the simplest case of a locus with two alleles, the HW equilibrium is used to estimate the expected number of genotypes for homozygous wild type (or major allele), heterozygous and homozygous alternate (or minor allele) genotypes based on the allele frequencies. These expected genotypes are then compared with the genotypes observed in the population to assess if the given locus is in HW equilibrium or disequilibrium. In WES approaches, variants in HW disequilibrium due to very extreme excess heterozygosity (i.e.: the observed number of heterozygous genotypes is significantly greater than the expected number of heterozygous genotypes) are filtered in large population databases, including gnomAD, the largest available dataset that includes 125,748 exomes (44, 47). This assumption is reasonable because it has been shown that variants with extreme heterozygosity are enriched in low-complexity regions of the genome, which are especially prone to sequencing and alignment errors.

While HW disequilibrium may truly indicate technical errors in specific circumstances, some studies cautioned against blinded exclusion of loci deviating from HW equilibrium that could instead signal causative mutations. For example, a population-based study designed to investigate the causes of deviation from HWE failed to find an explanation for about 30% of loci found to be in disequilibrium, suggesting there may be other reasons beyond actual errors to cause deviations from HWE(52). Another report investigating HW equilibrium in a Japanese sample of 104 individuals from 1000Genome (a large dataset collected by EMBL-EBI)

suggested that HW disequilibrium in NGS data seems to be a major indicator for CNV(53). In line with these findings, a separate study has used deviations from HW equilibrium, and particularly loss of heterozygosity, as indicator of a specific class of CNVs, common deletions(54). Lastly, a recent report investigated HW disequilibrium in the whole set of exome data in gnomAD (cases and controls)(55). Authors mainly focused on excess heterozygosity with the main objective to identify variants and genes associated with autosomal recessive disorders. With the exception of very few classical examples (rs334 in *HBB*, which causes recessive sickle cell disease in homozygous status and confers protection from malaria in heterozygous status; rs1801178 in *CFTR*, which causes recessive cystic fibrosis disease in homozygous status and is hypothesized to be protective from cholera in heterozygous status), this study did not find candidates mutations. Furthermore, the authors recognized that the significance cutoff used in their study was lenient (0.05 without correction for multiple testing) in contrast to previous studies(53), and therefore the results from this study should be taken with caution. Nevertheless, this last study strongly supports the timeliness of our project.

Here I present very preliminary findings from a study aimed at investigating the distribution of genotypes across populations in the control set of gnomAD, with the two-fold goal to determine variants that are in true HW disequilibrium possibly underlying susceptibility or resistance to disease, and to investigate the underlying origin in relation to specific population events (e.g.: natural selection). We elected to use the control set of gnomAD to avoid the heterogeneity that results from the aggregation of different categories of cases in the complete gnomAD database. In addition, we use stringent thresholds with particular attention to the number of conditions being tested. Our preliminary data focus on two categories that we think are especially important for the goal of the study to identify variants and gene underlying susceptibility or resistance to diseases. Specifically, after investigating for possible technical errors, we focus on variants in HW disequilibrium due to strong excess or depletion of homozygotes for the minor allele (as defined in the methods) as they might have a protective or deleterious role, respectively. The rationale for this specific focus stems from evidence that homozygous mutations can have a broad range of effect on the phenotype, spanning from deleterious effects on multiple phenotypes (low redundancy) or on a single phenotype (high redundancy), to no detectable effect on the phenotype (complete redundancy), or to advantageous effects conferring resistance to given phenotypes (beneficial redundancy)(56).

4.2 Methods

4.2.1 Description of the panel: gnomAD

The Genome Aggregation Database (gnomAD)(44) consists of 125,748 exomes from seven populations. In the most recent version of the database, authors made available a subset of control exomes only (no cases from common disease case/control studies that contributed data to gnomAD). This subset of controls consists of 54,704 exomes. The breakdown by the seven populations is as follows: African/African American (AFR, n = 3,582), Latino/Admixed American (AMR, n = 8,556), Ashkenazi Jewish (ASJ, n = 1,160), East Asian (EAS, n = 4,523), Finnish (FIN, n = 6,697), Non- Finnish European (NFE, n = 21,384), and South Asian (SAS, n = 7,845) (Figure 23). We elected to use the control-only subset for our study to avoid the heterogeneity that results from the aggregation of different categories of cases in the complete gnomAD database and thus aiding the interpretation of the final set of variants in HW disequilibrium. The initial variant dataset consisted of more than 17 million variants.

Population	gnomAD		controls		non-cancer		non-neuro		non-TOPMed	
	exomes	genomes	exomes	genomes	exomes	genomes	exomes	genomes	exomes	genomes
African/African American	8,128	4,359	3,582	1,287	7,451	4,359	8,109	1,694	6,013	4,278
Latino	17,296	424	8,556	123	17,130	424	15,262	277	17,229	405
Ashkenazi Jewish	5,040	145	1,160	19	4,786	145	3,106	123	4,999	69
East Asian	9,197	780	4,523	458	8,846	780	6,708	780	9,195	761
Finnish	10,824	1,738	6,697	581	10,816	1,738	8,367	582	10,823	1,738
Non-Finnish European	56,885	7,718	21,384	2,762	51,377	7,718	44,779	6,813	55,840	5,547
South Asian	15,308	*	7,845	*	15,263	*	15,304	*	15,308	*
Other	3,070	544	957	212	2,810	544	2,433	367	3,032	506
Female	57,787	6,967	25,645	2,508	53,850	6,967	47,831	4,799	55,662	6,299
Male	67,961	8,741	29,059	2,934	64,629	8,741	56,237	5,837	66,777	7,005
Total	**125,748**	**15,708**	**54,704**	**5,442**	**118,479**	**15,708**	**104,068**	**10,636**	**122,439**	**13,304**

Figure 23. Breakdown of the control population in the gnomAD database.
The gnomAD dataset is the largest available collection of exome data and includes a breakdown by population. The latest release comprises a control-only subset (n=54,704) no cases from common disease case/control studies that contributed data to gnomAD)

4.2.2 Determination of the high-quality subsets of variants for the HW analysis

We determined the high-quality sets of variants to be used in the study as follows (Figure 24): 1) bi-allelic variants restricted to the coding regions (including all isoforms) with 2 base pairs (bp) padding to include splicing variants (Gencode); 2) variants deemed as high-quality

("PASS" status in the gnomAD dataset) defined as passing the random forest (RF), allele count, and inbreeding coefficient filtering; 3) variants in an autosomal chromosome; 4) MAF > 0.001 in at least one population, 4) ethnic-specific call rate >85%.

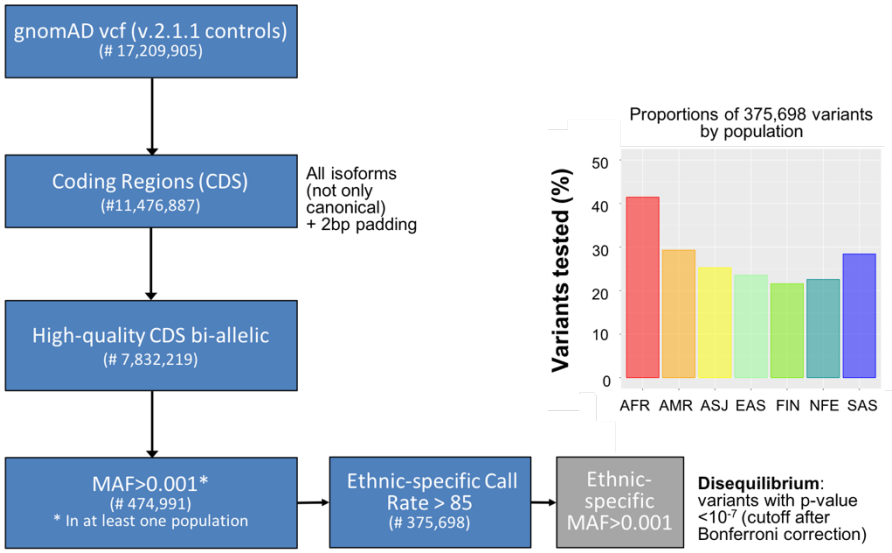


Figure 24. Schematic of the workflow

We determined the high-quality sets of variants to be used in the study by 1) restricting our study to exonic or splicing variants in autosomal chromosomes (using the Gencode interval list), 2) reducing the number of false positives (random forest and call rate filtering), 3) retaining variants with genotype distributions to enable sufficient power (MAF filtering). These steps reduced the number of variants from 17,209,905 to 375,698. This high-quality set was tested for Hardy-Weinberg equilibrium, using a p-value cutoff of 10-7.

4.2.3 Methodological and statistical approach

4.2.3.1 Hardy-Weinberg equilibrium

Within each ethnic group, we computed the observed genotype counts (homozygous wild-type or HOM_WT, heterozygous or HET) from the appropriate fields provided in the gnomAD dataset (homozygous alternate or HOM_ALT, alternate allele count or AC, total number of alleles or AN) with the following formulae:

$$\mathrm{HET} = \mathrm{AC} - 2\,\mathrm{HOM\_ALT}$$

$$\mathrm{HOM\_WT} = \mathrm{AN} - (\mathrm{HOM\_ALT} + \mathrm{HET})$$

Next, we used the *HardyWeinberg* package(132) in R to calculate the expected genotypes (from the MAF) and to assess HW disequilibrium. Specifically, the HW equilibrium was calculated using the exact test for loci with at least one expected genotype less or equal to 5 because the exact test is more robust with low number of expected genotypes, and chi-square

test otherwise. To determine HW disequilibrium, we set the significance threshold to 10^{-7} to account for multiple testing using the Bonferroni correction (0.05/375,698).

4.2.3.2 Cause of Hardy-Weinberg disequilibrium and annotation

For all variants found to be in HW disequilibrium, we first defined two categories: variants in which the number of observed heterozygous is greater than the number of expected heterozygous (excess heterozygosity or E_HET), and, conversely, variants in which the number of observed heterozygous is lower than the number of expected heterozygous (loss of heterozygosity or L_HET). Next, for these preliminary analyses, we focused on two categories in which we are particularly interested: loss of homozygotes for the minor allele (within the E_HET) and excess of homozygotes for the minor allele (within the L_HET), which are especially important for the goal of the study to identify HW disequilibrium underlying susceptibility or resistance to diseases. Specifically: i) within the variants in E_HET, we define variants for which there is loss of the homozygotes for the minor allele alleles (retaining only variants where the observed number of minor allele was lower or equal to 5), ii) within the variants in L_HET, we define variants for which there is excess of the homozygotes for the minor allele.

Subsequently, variants were annotated according to functional significance, selection scores and known associations in genetic diseases. As in our previous reports (45, 69), we used SNPEff and a custom script to annotate for functional significance and other variant- and gene-level scores (such as GDI). For selection scores we used a newly developed selection score, called CoNeS(133), which integrate known negative selection scores through principal component projection. CoNeS is a standardized metric in which negative scores correspond to negative selection(133). For the known associations with genetic diseases, we used the Human Gene Mutation Database (HGMD), the largest catalogue of mutations associated to disease (134). HGMD reports both disease-causing and protective variants. These annotation were used for further filtering and prioritization of variants. Specifically, among the excess and loss of homozygotes for the minor allele, we retained only missense and predicted loss of function (LOF, defined as falling in one of the following categories: "frameshift", "stop-gained", "stop-lost", "start-lost", "splice-donor", "splice-acceptor", "indel-frameshift", or “essential_splicing”), and variants present in HGMD and/or sorted by CoNeS score.

4.3 Results

4.3.1 Proportions of HW disequilibrium within ethnic groups by MAF

We first determined the initial set of high quality variants that could be meaningfully interpreted in the subsequent study of HW equilibrium. Specifically (see also methods), we aimed at 1) restricting our study to bi-alleic exonic or splicing variants in autosomal chromosomes (using the Gencode interval list), 2) reducing the number of false positives (random forest and call rate filtering), 3) retaining variants with genotype distributions to enable sufficient power (MAF filtering). These steps reduced the number of variants from 17,209,905 to 375,698 (2.2% of the starting set), with a total of 26,856 common variants (tested in all seven populations). The proportion of variants tested in each population ranged from 21.6% (FIN) to 41.5% (AFR) (Table 8). The majority of variants tested were rare (Fig. 25), especially in the AFR population.

	Variants with MAF>0.001 in each population		Variants in HW disequilibrium	
ethnicity	**counts**	**%**	**counts**	**%**
AFR	155,778	41.5	894	0.57
AMR	110,132	29.3	1,850	1.68
ASJ	94,857	25.2	395	0.42
EAS	88,469	23.5	847	0.96
FIN	81,060	21.6	866	1.07
NFE	84,799	22.6	1,484	1.75
SAS	106,729	28.4	1,391	1.30

Table 8. Number and percentage of variants tested for HW equilibrium within each population
Summary statistics of the variants tested for HW equilibrium in each of the seven populations.

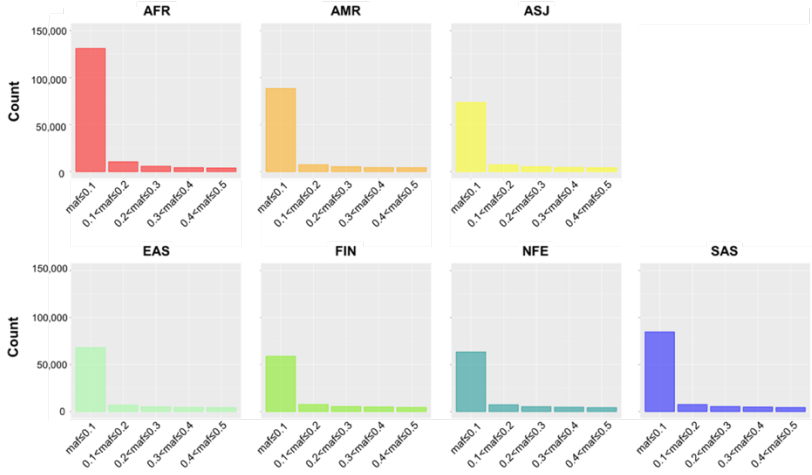


Figure 25. MAF Distribution of the variants tested for HW disequilibrium by ethnic group
Distribution of MAF of variants tested for HW disequilibrium by each of the seven populations.

Depending on the number of expected genotypes (see methods), we used either the chi-square or exact test to determine the variants in HW disequilibrium and used a p-value cutoff of 10^{-7} (Bonferroni correction). We found 7,727 variants in HW disequilibrium in the seven populations. As some variants are in disequilibrium in more than one population, we also calculated the number of *unique* variants in HW disequilibrium that is 3,878. The proportions of variants in HW disequilibrium ranged from 0.42% in ASJ to 1.75% in NFE, overall reflecting the expected inherent structure of the population (Table 8, Fig. 26A). For example, we found one of the greatest proportions of variants in HW disequilibrium (1.68%) in the AMR population, which is composed by several subpopulations (i.e.: Colombians, Mexicans, Peruvians and Puerto Ricans); this high rate of disequilibrium might be in part due to Wahlund effect. Conversely, the proportion of variants in HW disequilibrium was smaller in ASJ, EAS and FIN, which are relatively homogeneous populations. However, we found a minor proportion of variants in HW disequilibrium than expected in AFR (0.57%), a population also composed by several subgroups. These same trends were observed when dichotomizing by MAF (below or above 0.01), and were especially evident for common variants (MAF>1%) (Fig. 26B-C). The unexpected proportion of variants in HW disequilibrium in AFR could be due to limited power resulting from the relatively smaller sample size of AFR as well as to a larger number of rare variants in AFR (Fig. 25) as compared to other populations.

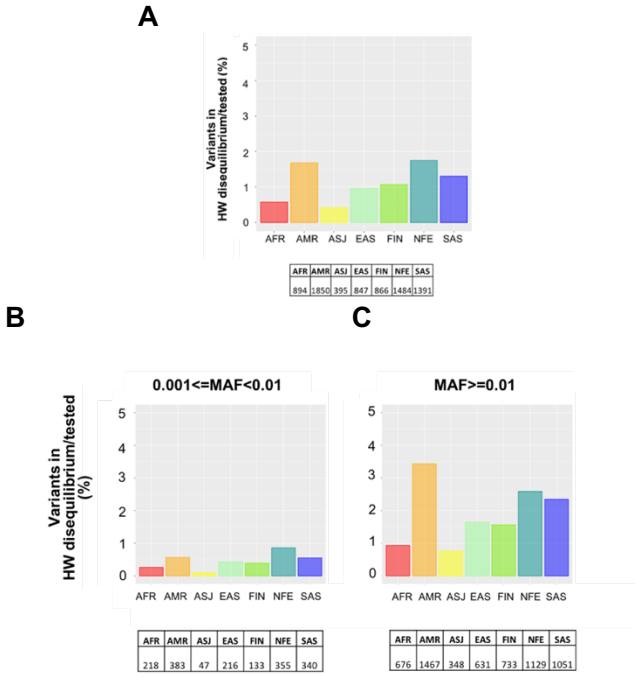


Figure 26. Proportions of variants in HW disequilibrium by ethnic group and MAF
Proportions of variants in HW disequilibrium by population (A) and dichotomized by MAF (<0.01 in B and > 0.01 in C).

4.3.2 Classification by type of HW disequilibrium

For the 3,878 unique variants found to be in HW disequilibrium, we determined the driving cause of disequilibrium according to the two main categories of excess or loss of heterozygosity (E_HET or L_HET, respectively). The cause of HW disequilibrium was common across the seven populations for almost all variants (99.4%, n=3,853). Of the 3,853 variants, the vast majority (n=3,184, 83%) were in HW disequilibrium due to L_HET, while the remaining 669 (17%) were in HW disequilibrium due to E_HET (Fig. 27A). Our general strategy is to separate HW disequilibrium due to technical errors from potentially true HW disequilibrium in order to help identifying variants that could underlie resistance or susceptibility to diseases. To this end, we started to look for indicators of technical errors and hence determined the coverage in these two categories of HW disequilibrium. The coverage is main indicator of the reliability of variants in exome data. Exome data of good quality have a typical median coverage of 40-60X, spanning from abnormally low covered regions (~10X) to 'hot spots' with abnormally high coverage (>90X). We found that the median DP in variants in HW disequilibrium due to E_HET was larger than that of variants in HW disequilibrium due to L_HET (73.5X vs 57.9X), and even larger than the coverage of variants in HW equilibrium (EQ, DP=43.4X) (Fig. 27B). Albeit to a lesser extent, we also observe greater median coverage in L_HET than in variants in HW equilibrium (variants with a p-value > 10^{-2}) (Fig. 27B); more in depth investigations will be needed to determine if this observation could be indicative of a specific cause.

There might be different explanations for the observed higher coverage in E_HET variants. First, similarly to what we found for the blacklisted variants in excess heterozygosity(45), the higher coverage could indicate issues in the mapping process. Specifically, it could indicate the occurrence of alternative haplotypes belonging to unmapped regions absent from the human reference genome that are incorrectly mapped to the region of the reference genome for which the best match is obtained, leading to a mixture of wild-type and alternative alleles in these regions, and resulting in higher coverage and a final erroneous heterozygous call. Second, the higher coverage in E_HET could be linked to issues in sequencing due to low-complexity regions of the genome. We are currently testing this hypothesis by determining if variants in HW disequilibrium due to E_HET are enriched in variants located in low complexity regions of the genome (which are known to be problematic in WES analysis) as compared to variants in HW disequilibrium due to L_HET. In regard to L_HET, we expect that a possible reason for HW disequilibrium will be the presence of common deletions. Indeed, a visual inspection of

the list of variants in L_HET revealed hits in the Complement Factor H–Related Genes (*CFHR1*), a gene known to carry deletions common in the general population. Specifically we find 7 variants in L_HET in *CFHR1* and that are located within a deletion present in 12.45% of the general population(128).

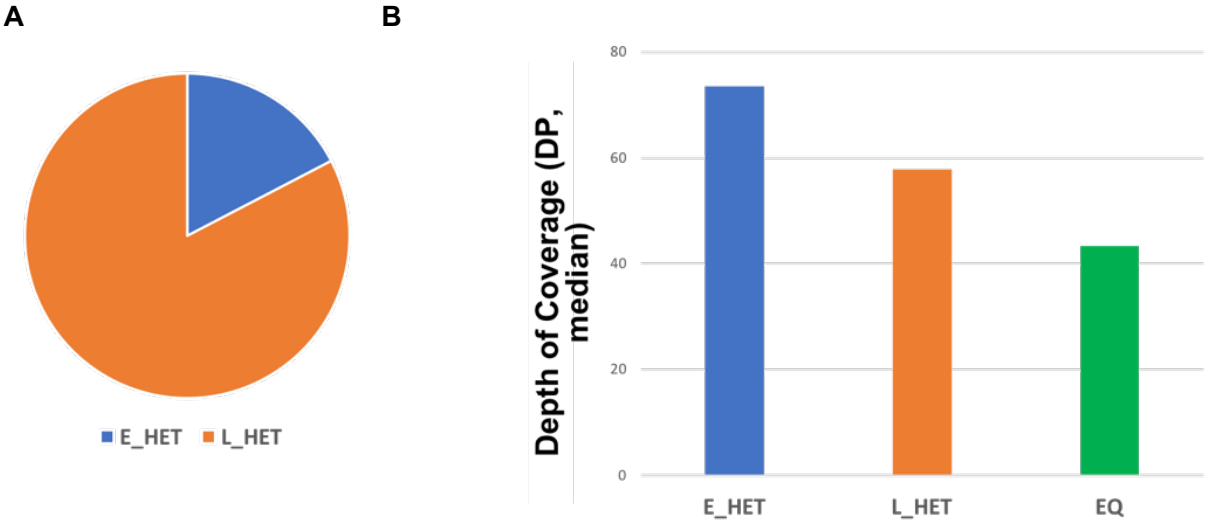


Figure 27. HW disequilibrium by excess or loss of heterozygosity

Proportions (A) and coverage (B) of variants in HW disequilibrium due to excess or loss of heterozygosity. Coverage of variants in HW equilibrium (EQ) is reported for comparison. p < 2.2e-16 at t-test (pairwise comparison among the 3 groups). Effect size (Cohen's d) ranges from 0.6 (medium/large) to 1.15 (very large).

Next, we performed a more in depth analysis of the 3,853 unique variants in HW disequilibrium. We determined a 7x7 matrix with the number of population in which the variants were tested as columns and number of populations in which we detected HW disequilibrium as rows (Table 9). This analysis showed that most of the variants (n=2,262, 59%) are in HW disequilibrium in one population, although this could be due to limited power to detect a significant effect in some populations own to sample size (Table 9). To limit this confounding factor, we focused our analysis on two specific categories at the end of the spectrum: variants tested in one population and in HW disequilibrium in one population (1/1, n=812) and variants tested in on all seven populations and in HW disequilibrium in all seven populations (7/7, n=62). For the category 1/1, we cannot exclude that some variants were present in other populations but were filtered in the QC steps described above and could not be tested for HW equilibrium. We hypothesized that variants in disequilibrium in all seven populations (the 7/7 category) were more likely to be false positives. Our analysis of the DP coverage supports our hypothesis by showing a greater proportion of E_HET versus L_HET in 7/7 as compared to 1/1 and a greater average DP in E_HET versus L_HET within the 7/7 category (Fig. 28). Additional analyses of

enrichment in low complexity regions of the genome are warranted to further support this postulate.

1/1

ALL		# of Populations Tested							
		1	2	3	4	5	6	7	
# of Population Detected Disequilibrium	1	812	254	131	76	90	203	696	2262
	2		230	97	54	64	71	136	652
	3			69	43	56	58	95	321
	4				51	50	63	67	231
	5					52	78	69	199
	6						46	80	126
	7							62	62
		812	484	297	224	312	519	1205	3853

7/7

Table 9. Focus on specific categories of HW disequilibrium

Matrix of variants in HW disequilibrium by number of population tested (columns) and number of populations in which HW disequilibrium was detected (rows). 1/1 indicates variants tested in one population and in HW disequilibrium in one population and 7/7 indicates variants tested in on all seven populations and in HW disequilibrium in all seven populations

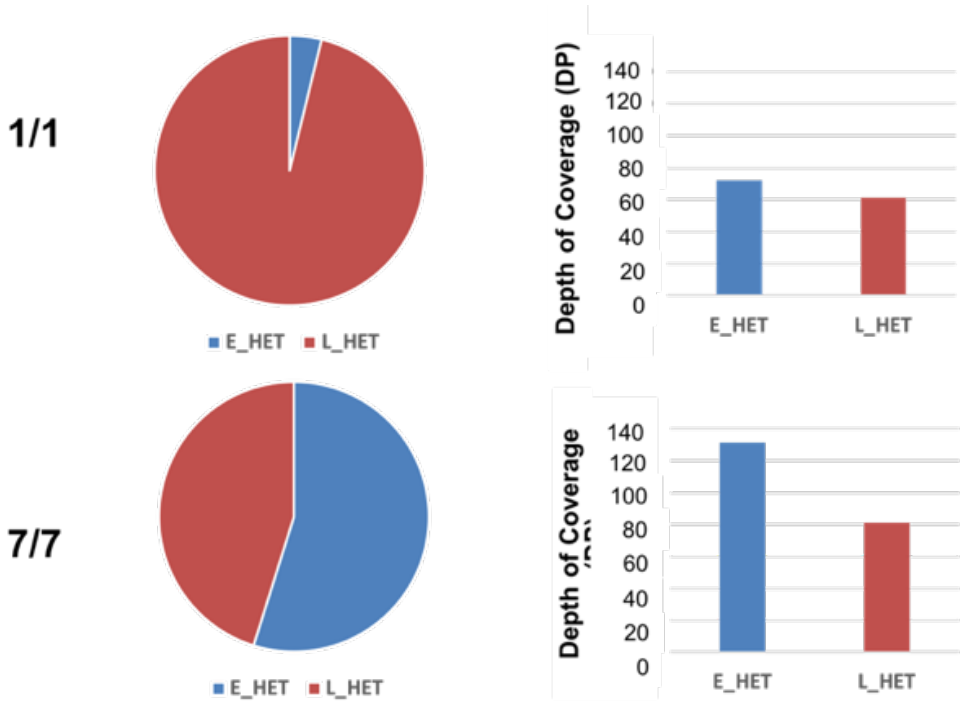


Figure 28. Preliminary analysis of specific categories of HW disequilibrium

Proportions (left panels) and coverage (right panels) of variants in HW disequilibrium falling in 1/1 and 7/7 categories.

4.3.3 Focus on excess of homozygotes for the minor allele

Motivated by our interest in identifying mutations that could underlie resistance or susceptibility to diseases, we focused our attention on two specific subsets within the main categories of L_HET and E_HET. Specifically we studied excess of homozygotes for the minor allele within the L_HET and loss of homozygotes for the minor within the E_HET, and annotated these two subsets of variants according to known genetic diseases using HGMD and selection scores (see methods). Variants with excess of homozygotes for the minor allele could indicate mutations that are protective or are in genes with favorable redundancy; conversely, variants with loss of homozygotes for the minor allele could indicate mutations that are deleterious in genes with low, high or complete redundancy(56). Protection and disadvantage to disease could also be reflected in selection events, as variants with a protective role may be associated with positive selection, and variants with a deleterious role may be associated with negative selection (58, 135, 136).

First, we started with the category of excess of homozygotes for the minor allele. A total of 2,703 (85%) variants among the 3,184 variants in HW disequilibrium due to L_HET fell in the category. Given the expectation that at least a sizeable proportion of homozygous pre-mature stop mutations are likely to result in a complete deficiency of all encoded isoforms in all tissues, we initially focused on predicted loss of function (LOF) variants. Among the 2,703 variants in excess of homozygotes for the minor allele, we found that 75 unique variants (3%) or 120 (counting the same variant in disequilibrium in more than one population) in 56 genes were predicted to be LOF. We ranked the LOF variants in excess of homozygotes for the minor allele by p-value and chi-square term for the heterozygous status. In the top 10 variants(Table 10), we found a variant in *FUT2* (rs601338, W154X, in SAS, ranked 7/120), well-known to confer resistance to viral infection(137). *FUT2* regulates the expression of antigens on the surface of epithelial cells and mucosal secretions, and is responsible for the secretor phenotype(138). The W154X nonsense mutation results in a non-secretor phenotype(139). Non-secretor individuals have been shown to be resistant to infections with norovirus(140, 141), and rotavirus(142, 143). We also found the W154X *FUT2* variant in another population (AMR, not ranked in the top 10) as well as other *FUT2* variants when enlarging to missense mutations (Table 11). Specifically, we found the missense rs1047781 (I140F) in SAS, which has been also linked to the non-secretor phenotype in HGMD, as well as other two variants (p.Pro112Leu and p.Ala104Val) that might also underlie resistance to infection (Pro112Leu was recently shown to be causing non-secretor phenotype and associated to enterotoxigenic Escherichia coli

infection, and Ala104Val was recently shown to reduce A antigen in mutant-transfected COS-7 cells)(144, 145). Consistently, the CoNeS score for *FUT2* is positive, indicating that this gene is not under negative selection. The other variants in the top 10 are currently under investigation; the CoNeS scores in these genes are positive, hence not expected to be under negative selection which is in line with the possible hypothesized protective role.

Variant			Genetic phenotype		Selection	Hardy-Weinberg				
function	AAChange	gene	HGMD	Disease	CoNeS	eth	p.val	EXP HOM_MINOR	OBS HOM_MINOR	maf
indel-frameshift	p.Ala461fs	ZNF626			-	nfe	<E-100	15	514	0.03
						eas	<E-100	16	232	0.06
						sas	<E-100	9	214	0.03
						fin	<E-100	8	181	0.03
						afr	<E-100	2	70	0.02
						asj	1.01E-31	0	17	0.02
indel-frameshift	p.Val234fs	PCDHB8			1.84	fin	<E-100	8	193	0.04
						sas	<E-100	21	214	0.06
						nfe	<E-100	14	207	0.03
						asj	1.36E-10	0	7	0.01
stop-gained	p.Tyr63*	OR4P4			1.07	fin	<E-100	47	265	0.09
						afr	<E-100	44	125	0.11
						asj	<E-100	23	64	0.15
indel-frameshift	p.Gln502fs	MAML3			-0.78	sas	<E-100	428	666	0.24
						fin	<E-100	273	477	0.21
						asj	<E-100	584	641	0.27
indel-frameshift	p.Leu187fs	POMZP3			0.57	fin	<E-100	774	976	0.35
stop-gained (W154X)	p.Trp154*	FUT2	X	Non-secretor phenotype	1.26	sas	<E-100	783	979	0.32
						amr	4.25E-12	587	711	0.26
indel-frameshift	p.Ile83fs	OR4L1			2.1	amr	<E-100	1223	1425	0.40
						nfe	2.49E-14	2592	2838	0.37
stop-gained	p.Tyr269*	RHD	X	Rhesus negative blood group	3.29	afr	<E-100	7	48	0.05
						amr	2.31E-14	0	7	0.00
indel-frameshift	p.Ser177fs	KRTAP5-5			1	fin	<E-100	14	67	0.05
						afr	<E-100	6	28	0.04
						nfe	<E-100	7	42	0.02
indel-frameshift	p.Gln175fs					fin	0	13	66	0.04
						afr	0	7	42	0.02
						nfe	0	6	28	0.04

Table 10. Top 10 variants in excess of homozygotes for the minor allele

Top 10 variants in excess of homozygotes for the minor allele (sorted by p-value and chi-square term for the heterozygous status). For each variant ranking in the top 10, the results of excess of homozygotes for the minor allele in the other populations (even if not in the top 10) are also reported. For example, p.Ala461fs is among the top 10 only in NFE, EAS, SAS and FIN but the results for AFR and ASJ are also reported.

Variant			Genetic Phenotype		Selection	Hardy Weinberg results				
function	AAChange	gene	HGMD	Disease	CoNeS	eth	p.val	EXP HOM_MINOR	OBS HOM_MINOR	maf
missense (I140F)	p.Ile140Phe	FUT2	X	Non-secretor phenotype	1.26	SAS	2.7E-10	1	10	0.01
missense	p.Pro112Leu	FUT2			1.26	SAS	<E-100	138	254	0.13
missense	p.Ala104Val	FUT2			1.26	SAS	1.8E-13	3	23	0.02

Table 11. Other FUT2 variants

Missense variants in FUT2 found to be in excess of homozygotes for the minor allele.

Next, we focused on variants known to be associated to disease and reported in the HGMD database, the largest collection of demonstrated (or possible) disease-causing and protective variants. To be more inclusive in terms of impact, we included both LOF and missense variants. We found a total of 69 unique variants in 41 genes listed as associated to disease in HGMD, but only six variants listed as possible pathological/protective (labelled as DM?) and only two variants listed as disease-causing/protective (labelled as DM) (Table 12). The two DM variants are especially interesting. The first variant was the p.Pro112Leu in *FUT2* that we describe in the paragraph above and is linked to fucosyltransferase deficiency. The second variant was found in *SMN2* in NFE, SAS, AMR and ASJ. *SMN2* belong to the family of survival motor neuron (SMN) genes that have been shown to be the primary determining gene of Spinal Muscular Atrophy (SMA). An homozygous deletion in *SMN2* has been reported to result in the SMA phenotype(146). However, the variant we found in *SMN2* (rs121909192, G287R) is listed in the HGMD database as a positive modifier of the SMA phenotype. Specifically, it was reported that in a patient with a mild form of SMA who carried the SMN1 genotype (predicted to lead to a more severe form of the disorder), this *SMN2* variant increases the amount of full-length *SMN2* transcripts, thus resulting in less severe phenotypes(147).

Variant			Genetic phenotype		Selection	Hardy-Weinberg				
function	AAChange	gene	Disease	Category	CoNeS	eth	p.val	EXP HOM_MINOR	OBS HOM_MINOR	maf
missense	p.Pro112Leu	FUT2	Fucosyltransferase deficiency	DM	1.26	sas	<E-100	138	254	0.133
missense	p.Ala104Val					sas	1.84E-13	3	23	0.021
missense	p.Gly287Arg	SMN2	Spinal muscular Atrophy modifier	DM		nfe	3.28E-39	0	23	0.004
						sas	2.73E-23	0	15	0.006
						amr	5.69E-12	0	5	0.002
						asj	4.80E-08	0	3	0.004
stop-lost	p.Ter342Argext*?	KIR2DL3	KIR2DL3 variant	DM?	1.48	amr	1.43E-25	0	10	0.002
missense	p.Arg142Cys	CFHR3	Haemolytic uraemic syndrome atypical	DM?	1.5	sas	1.53E-19	0	7	0.002
						eas	1.02E-09	2	14	0.020
missense	p.Cys208Tyr	CFHR1	Haemolytic uraemic syndrome atypical	DM?	1.38	sas	3.59E-16	0	7	0.002
missense	p.Tyr355Cys	CYP2D6	Hypercholesterolaemia	DM?	1.87	sas	2.39E-13	0	5	0.001
missense	p.Arg265Gly	ABCC6	Pseudoxanthoma elasticum autosomal recessive	DM?	0.57	nfe	3.13E-13	71	129	0.059
missense	p.Pro426Leu	MYO1A	Hearing loss	DM?	0.91	amr	2.18E-10	410	510	0.219

Table 12. Variants in excess of homozygotes for the minor allele and reported in HGMD
Variants in excess of homozygotes for the minor reported in HGMD as disease-causing/protective mutations (DM) or probable/possible pathological/protective mutation (DM?).

4.3.4 Focus on depletion of homozygotes for the minor allele

Second, we looked at the subset of variants in HW disequilibrium due to depletion of homozygotes for the minor allele. A total of 281 variants (42%) among the 669 variants in HW disequilibrium due to E_HET fell in the category of depletion of homozygotes for the minor allele; we retained only variants where the number of observed minor alleles is lower or equal to an arbitrary cutoff of 5 given our interest in identifying variants with a possible deleterious role in disease that are expected to be rare. Only 4 LOF variants fell in this category, hence we retained missense variants in addition to LOF, and we found that 125 unique variants (45%) or 280 (counting the same variant in disequilibrium in more than one population) were missense or LOF. We ranked these variants by p-value and chi-square term for the heterozygous status. In the top 20 variants (Table 13), we found a variant in *IRF8* (rs79518337, p.Glu74Asp, in EAS, ranked 18/280; the same trends for this variant were also found in NFE, ranked 21/280; AMR ranked 123/280, SAS ranked 233/280) that seemed promising. Interferon regulatory factor 8 (*IRF8*) regulates the expression of genes stimulated by IFN-α/β, is expressed in macrophages and dendritic cells, and plays an important role in several aspects of myeloid cells(148). Mutations of the human *IRF8* gene underlie mycobacterial disease(148). Given the counts were very extreme (0 observed versus 186 expected homozygotes for the minor allele), we decided to look at this variant in a separate database (the genome dataset of gnomAD). We found that the variant is much less frequent in the second database (global frequency: 0.12 in the exome database; 0.003 in the genome database) and also flagged as false positive by the random forest classifier, suggesting a specific problem with this variant. This example suggests that the comparison of the results between the exome and genome dataset of gnomAD will be very important to remove false positives. In addition, these findings should be taken with caution because this category also includes some variants in genes known to be polymorphic (e.g.: *MUC17*, *FLG*) and presumably technical artifacts as suggested by the high number of variants and/or high coverage.

Finally, we focused on variants known to be associated to disease and reported in the HGMD database. We found a total of 4 unique variants in 4 genes listed as disease-associated in HGMD, but only one variant in *FANCD2* (p.Leu456Arg in AFR) listed as possible pathological/protective (labelled as DM?)(Table 14). FANCD2 belongs to the family of Fanconi anemia complementation group and it is required for maintenance of chromosomal stability. This variant is reported in the HGMD database as “possible pathological” and as

pathogenic recessive(149) in the Fanconi Anemia Mutation Database hosted by the Leiden Open Variation Database although as hypomorphic.

Variant			Selection	Hardy-Weinberg					Quality	
function	AAChange	gene	CoNeS	eth	p.val	EXP HOM_MINOR	OBS HOM_MINOR	maf	coverage	number of variants
missense	p.Thr1686Ser	MUC17	4.52	sas	<E-100	298	1	0.21	201.5	29
missense	p.Asn23Asp	IGKV1D-17		nfe	<E-100	430	3	0.15	150.5	2
				fin	8.88E-16	55	1	0.09	150.5	2
missense	p.Val187Ile	SKA3	1	sas	<E-100	298	0	0.20	74.7	3
				nfe	<E-100	335	0	0.13	74.7	3
				amr	<E-100	217	0	0.16	74.7	3
missense	p.Ile226Asn	PRKRA	-0.29	nfe	<E-100	416	0	0.14	83.2	5
				amr	<E-100	231	0	0.17	83.2	5
				fin	<E-100	148	0	0.15	83.2	5
				sas	<E-100	93	0	0.11	83.2	5
				asj	1.91E-14	38	0	0.19	83.2	5
start-lost	p.Met1?	AL133481.1		sas	<E-100	276	0	0.20	42.5	2
				asj	<E-100	55	0	0.24	42.5	2
missense	p.Gly30Arg	HLA-DRB5	2.49	fin	<E-100	261	2	0.20	95.7	2
				nfe	<E-100	83	0	0.07	95.7	2
missense	p.Lys4Glu	TTLL1	-0.16	afr	<E-100	191	0	0.24	78.8	3
				eas	<E-100	137	0	0.18	78.8	3
				amr	<E-100	109	0	0.11	78.8	3
missense	p.Ser4Pro	TNXB	0.05	sas	<E-100	255	0	0.19	59.4	4
				nfe	<E-100	198	0	0.10	59.4	4
				amr	<E-100	73	0	0.10	59.4	4
				eas	2.39E-08	26	0	0.08	59.4	4
missense	p.Asp2936Gly	FLG	10.23	nfe	<E-100	376	0	0.14	185.1	14
missense	p.Gln24Leu	PPIAL4G	0.82	eas	<E-100	211	0	0.22	139.3	11
missense	p.Val930Ala	POTEF	0.41	eas	<E-100	205	2	0.22	61.8	4
				amr	<E-100	162	2	0.15	61.8	4
				nfe	<E-100	149	0	0.09	61.8	4
				sas	<E-100	62	1	0.09	61.8	4
missense	p.Arg427Cys	PDPR	0.09	fin	<E-100	221	0	0.19	50.3	4
				nfe	<E-100	315	0	0.13	50.3	4
				amr	<E-100	109	0	0.12	50.3	4
				sas	<E-100	96	0	0.12	50.3	4
missense	p.Ile48Phe	IGHV1OR21-1		eas	<E-100	2673	2477	0.21	241.4	4
				sas	<E-100	5743	5580	0.14	241.4	4
				amr	<E-100	6443	6297	0.13	241.4	4
				afr	<E-100	2551	2466	0.15	241.4	4
missense	p.Thr2795Pro	AKAP13	0.08	eas	<E-100	188	0	0.22	43.7	2
				amr	<E-100	115	0	0.12	43.7	2
missense	p.Glu74Asp	IRF8	-1.25	eas	<E-100	186	0	0.22	67.8	4
				nfe	<E-100	323	0	0.13	67.8	4
				amr	<E-100	95	0	0.11	67.8	4
				sas	3.55E-10	34	0	0.07	67.8	4

Table 13. Top 20 variants in depletion of homozygotes for the minor allele

Top 20 variants in depletion of homozygotes for the minor allele (sorted by p-value and chi-square term for the heterozygous status). For each variant ranking in the top 20, the results of depletion of homozygotes for the minor allele in the other populations (even if not in the top 20) are also reported. Genes in red font are potential technical artifacts. No variant was found in HGMD.

While these findings are very preliminary and need to be taken further in terms of filtering and annotation, they warrant an in depth study of variants in HW disequilibrium due to excess or depletion of homozygotes for the minor allele as it could reveal interesting, novel variants that underlie resistance or susceptibility to diseases.

Variant			Genetic Phenotype			Selection	Hardy Weinberg results				
function	AAChange	gene	HGMD	Disease	Category	CoNeS	eth	p.val	EXP HOM_MINOR	OBS HOM_MINOR	maf
missense	p.Leu456Arg	FANCD2	X	Fanconi Anemia	DM?	0.6	AFR	6.1E-10	35	2	0.10

Table 14. Variants in loss of homozygotes for the minor allele and reported in HGMD

Examples of variants in HW disequilibrium due to excess of heterozygosity, and specifically with a loss of the minor allele.

4.4 Conclusions and perspective

The study of HW equilibrium in large exome datasets may reveal interesting candidate variants that could underlie susceptibility or resistance to disease. We found that overall 1% variants in the control set of the gnomAD database (n=54,704) are in HW disequilibrium with a greater proportion of disequilibrium due to L_HET than E_HET (83% vs 17%, although this difference could be more marked because gnomAD filters out variants that show very extreme excess heterozygosity). A major step to determine true HW disequilibrium is to separate technical errors from potential hits as much as possible. Our preliminary findings suggest that variants in HW disequilibrium due to E_HET and variants in disequilibrium in all seven populations are more likely to be technical errors. While these initial findings are plausible, we are continuing to curate the annotation of variants in HW disequilibrium to filter out potential artifacts from our analysis of HW disequilibrium. For example, we will use information on low complexity regions of the genome to better define (and filter) the technical errors. Furthermore, given common deletions result in an apparent loss of heterozygosity thus violating HW equilibrium, we are currently investigating whether HW disequilibrium in the L_HET category is due to common deletions in a subset of variants in loss of heterozygosity.

Our preliminary data also support the notion of investigating HW disequilibrium to identify candidate variants underlying human diseases, especially variants with an excess or loss of homozygotes for the minor allele that could be indicative of protection or disadvantage to disease. These two categories might also be linked to positive and negative selection, and our plan is to use selection scores to investigate this aspect. Our findings confirm known mutations in *FUT2* that lead to the non-secretor phenotype that is resistant to norovirus and rotovirus (140-143). We also found other *FUT2* variants that might also be involved in resistance to viral infections as well as that a mutation in *SMN2* (known to lead to a less severe SMA phenotype)(147) is in strong excess homozygosity for the minor allele, raising the hypothesis it may have a protective effect. Similarly, the mutation found in *FANCD2* (149) support the investigation of HW disequilibrium due to loss of the minor allele to inform about recessive disorders. We confirmed the same pattern of HW disequilibrium for the variants in these three promising genes in an independent dataset (i.e.: the genome dataset part of gnomAD), suggesting that they are likely to be true hits. Conversely, we could not find confirmation for the variant in IRF8 highlighting the importance to independently replicate the findings from this project. Given the preliminary nature of these findings, it will be important to perform more in depth analysis, determine if some of the variants in HW disequilibrium due to excess or loss of homozygotes for the minor allele are present in public databases such as HGMD (93) (as we presented here), ClinVar (150) or OMIM (151).

5 Final remarks and future directions

My thesis leveraged the opportunities set out by WES approaches for both the development of sophisticated computational approaches to analyze exome data and the formulation of novel scientific questions with the use of readily available large exome datasets. Specifically, we provide two novel methods to facilitate the identification of disease-causing mutations in human diseases, and we investigate HW equilibrium using the largest available dataset of exome data. The ***first*** method, the Blacklist, considerably decreases the number of false positives in exome data thus facilitating the prioritization of the remaining variants as candidate disease-causing mutations (45). The ***second*** method, HMZDelFinder-opt, provides a timely approach to identify partial homozygous and hemizygous deletions, which is a specific class of mutations traditionally difficult to detect in exome data of typical, heterogeneous laboratory panels that are generated over time (69). ***Lastly***, our preliminary findings of HW disequilibrium in the gnomAD database are very promising in supporting the investigation of HW equilibrium as a way to determine variants that could underlie resistance or susceptibility to diseases and that may be under selection.

The novel methods we present in this thesis will provide the scientific community with timely tools to facilitate the analysis of WES data. The use of WES approaches, which are focused on sequencing of the exome (the coding region of the genome), has significantly fueled discovery of the genetic basis of rare (and mostly monogenic) diseases (3, 18). There is a solid rationale to focus on the study of the exome in rare diseases. The vast majority of exonic variants is evolutionary recent, rare and enriched for deleterious alleles, thus likely to contribute significantly to diseases(18, 32). In addition, WES approaches are less expensive, faster and simpler to analyze as compared to approaches studying the whole genome (WGS) (18, 25). However it is still challenging to efficiently determine all of the different types of genetic variation (SNPs, Indels and CNVs) from exome data, and also to narrow them to a short list of candidate variants for manual inspection and functional validation.

A major challenge in the analysis of exome data is due to the continuous evolution of technology (sequencing and corresponding bioinformatic tools) that results in both heterogenous exome datasets with extreme fluctuations in coverage (Fig. 18) and technology-dependent false signals(31, 35-40). We demonstrate that the Blacklist approach (Chapter 2) can detect such FP and filtering them in a fast, efficient and customizable manner (45). This

approach can be used in combination with other state-of-the-art methods (such as the VQSR and RF tools) as we find they are mutually exclusive in the capture of FP. The other critical aspect in analysis of exome data is the ability of current bioinformatics tools to identify the whole spectrum of genetic variations. While state-of-the-art approaches such as GATK are well calibrated to detect SNPs as well as short (up to 50bp) insertion/deletions (or Indels) (130), the development of computations tools to detect larger duplications/deletions (CNVs) is challenged by i) the nature of targeted exome data (hence the breakpoints, the parameter commonly used to determine CNVs in WGS data, can't be used because it is not systematically sequenced) and ii) the unevenness of exome coverage. We tackle these issues by providing a method (HMZDelFinder-opt) that *a priori* selects a reference control set with a coverage profile similar to that of the WES sample under study to reduce the number of called HMZ deletions and improve the ranking of the true HMZ deletion (69). Our method also fills the gap in the study of deletions spanning less than an exon, by providing the first tool for the systematic identification of partial exon deletions. The human genome contains ~235,000 exons, about 20% of which are larger than 200 bp (131). Therefore HMZDelFinder_opt makes possible the systematic discovery of currently unknown HMZ deletions in ~47,000 exons that are not detectable with other tools.

An intriguing consequence of the unprecedent and widespread adoption of WES approaches in human genetics is the rapid accumulation of exome data. Several groups, including the Broad Institute, have undertaken the collection and harmonization of thousands of exome data in an effort to provide investigators with a public repository that could be used to aid the medical and functional interpretation of genetic variation (6, 44, 152). These large datasets not only are nowadays pivotal in exome analysis (e.g.: to assess frequencies in the general population), but can also be repurposed to enable systematic investigation of specific theoretical questions, something that was not possible before because of the limited statistical power in small/medium panels. One example is the study of HW equilibrium that I started tackling during my thesis. While our preliminary results of promising candidate variants that could underlie susceptibility (e.g. *FANCD2*) or resistance (e.g.: *FUT2* and *SMN2*) to disease need further investigation and proof, they strongly warrant the use of large and readily available exome data to investigate HW equilibrium.

The work described in this thesis lend itself to a number of future directions. For example, in HMZDelFinder-opt, we focused so far on homozygous deletions in autosam chromosomes

and hemizygous deletions in males on the X chromosome; in future work it will be interesting to adapt HMZDelFinder-opt to the detection of heterozygous deletions. Given the coverage of heterozygous deletions is expected to be half of that with no deletion, this direction will likely entail the fine-tuning of the cutoff to call a deletion and the inclusion of other measures, in addition to coverage. A similar approach could also be applied to detection of duplications. Furthermore, WGS is becoming increasingly attractive as an alternative, due to the more homogenous coverage, steadily decreasing cost, and the opportunity to study variants lying outside the protein-coding regions of the genome (25). Thus it will be interesting to evaluate and adapt the methods proposed here to WGS data. We expect that the Blacklist will still be highly valuable in the filtering of FP in WGS data, while HMZDelFinder-opt will likely benefit from the addition of breakpoint information to improve the sensitivity in detecting HMZ deletions. Lastly, for the HW project, it will be critical to use WGS data to confirm and replicate the findings, and it will be interesting to study selection events more in depth and try to apply the findings of candidate variants/genes to our in-house panel.

Collectively, these projects tackle heretofore-unexamined topics and hold promise to aid the discovery of novel causal determinants of human diseases or traits. They will also lay the foundation for future research to investigate the role specific classes of genetic variation (i.e.: partial deletions and variants in strong excess/depletion of homozygotes for the minor allele) in human diseases.

6 Index of the figures

7 Index of the tables

8 Bibliography

1. Speicher M, Antonarakis SE, & Motulsky AG (2010) Vogel and Motulsky's Human Genetics Problems and Approaches.
2. Russell PJ (2000) *Fundamentals of genetics* (Addison Wesley Longman, San Francisco).
3. Claussnitzer M*, et al.* (2020) A brief history of human disease genetics. *Nature* 577(7789):179-189.
4. Garrod AE (2002) The incidence of alkaptonuria: a study in chemical individuality. 1902 [classical article]. *The Yale journal of biology and medicine* 75(4):221-231.
5. Casanova JL (2015) Human genetic basis of interindividual variability in the course of infection. *Proc Natl Acad Sci U S A* 112(51):E7118-7127.
6. Auton A*, et al.* (2015) A global reference for human genetic variation. *Nature* 526(7571):68-74.
7. Frazer KA, Murray SS, Schork NJ, & Topol EJ (2009) Human genetic variation and its contribution to complex traits. *Nature reviews. Genetics* 10(4):241-251.
8. Eichler EE*, et al.* (2007) Completing the map of human genetic variation. *Nature* 447(7141):161-165.
9. Kruglyak L & Nickerson DA (2001) Variation is the spice of life. *Nat Genet* 27(3):234-236.
10. Mullaney JM, Mills RE, Pittard WS, & Devine SE (2010) Small insertions and deletions (INDELs) in human genomes. *Human molecular genetics* 19(R2):R131-136.
11. Manica A, Amos W, Balloux F, & Hanihara T (2007) The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 448(7151):346-348.
12. Bryc K*, et al.* (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A* 107(2):786-791.
13. Witherspoon DJ*, et al.* (2007) Genetic similarities within and between human populations. *Genetics* 176(1):351-359.
14. Novembre J*, et al.* (2008) Genes mirror geography within Europe. *Nature* 456(7218):98-101.
15. Casanova J-L (2015) Severe infectious diseases of childhood as monogenic inborn errors of immunity. in *Proc. Natl. Acad. Sci. U.S.A.*, pp E7128-7137.
16. Chong JX*, et al.* (2015) The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am J Hum Genet* 97(2):199-215.
17. Whiffin N*, et al.* (2017) Using high-resolution variant frequencies to empower clinical genome interpretation. *Genetics in medicine : official journal of the American College of Medical Genetics* 19(10):1151-1158.
18. Petersen BS, Fredrich B, Hoeppner MP, Ellinghaus D, & Franke A (2017) Opportunities and challenges of whole-genome and -exome sequencing. *BMC genetics* 18(1):14.
19. Vogel F & Motulsky AG (Vogel and Motulsky's Human Genetics. *Problems and Approaches.*
20. Sanger F & Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. in *J. Mol. Biol.*, pp 441-448.
21. Sanger F, Nicklen S, & Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74(12):5463-5467.

22. Gilissen C, Hoischen A, Brunner HG, & Veltman JA (2011) Unlocking Mendelian disease using exome sequencing. *Genome biology* 12(9):228.
23. Boycott KM, Vanstone MR, Bulman DE, & MacKenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. in *Nat. Rev. Genet.*, pp 681-691.
24. Venter JC*, et al.* (2001) The sequence of the human genome. in *Science* (American Association for the Advancement of Science), pp 1304-1351.
25. Belkadi A*, et al.* (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. in *Proc. Natl. Acad. Sci. U.S.A.*, pp 5473-5478.
26. Lander ES*, et al.* (2001) Initial sequencing and analysis of the human genome. in *Nature* (Nature Publishing Group), pp 860-921.
27. Choi M*, et al.* (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 106(45):19096-19101.
28. Hodges E*, et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 39(12):1522-1527.
29. Goh G & Choi M (2012) Application of whole exome sequencing to identify disease-causing variants in inherited human diseases. in *Genomics & informatics*, pp 214-219.
30. Ng SB*, et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. in *Nature*, pp 272-276.
31. MacArthur DG*, et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature* 508(7497):469-476.
32. Tennessen JA*, et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090):64-69.
33. DePristo MA*, et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. in *Nat. Genet.*, pp 491-498.
34. Goldstein DB*, et al.* (2013) Sequencing studies in human genetics: design and interpretation. *Nature reviews. Genetics* 14(7):460-470.
35. Hardwick SA, Deveson IW, & Mercer TR (2017) Reference standards for next-generation sequencing. *Nature reviews. Genetics* 18(8):473-484.
36. Sims D, Sudbery I, Ilott NE, Heger A, & Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics* 15(2):121-132.
37. Meienberg J*, et al.* (2015) New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res* 43(11):e76.
38. Wang Q, Shashikant CS, Jensen M, Altman NS, & Girirajan S (2017) Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. *Sci Rep* 7(1):885.
39. Clark MJ*, et al.* (2011) Performance comparison of exome DNA sequencing technologies. *Nature biotechnology* 29(10):908-914.
40. Parla JS*, et al.* (2011) A comparative analysis of exome capture. *Genome biology* 12(9):R97.
41. Phillippy AM, Deng X, Zhang W, & Salzberg SL (2009) Efficient oligonucleotide probe selection for pan-genomic tiling arrays. *BMC Bioinformatics* 10:293.
42. K HW (2020) A long read of the human genome. *Nature reviews. Genetics* 21(10):577.
43. Van der Auwera GA*, et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. in *Current protocols in bioinformatics* (John Wiley & Sons, Inc.), pp 11.10.11-33.
44. Karczewski KJ*, et al.* (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809):434-443.

45. Maffucci P, *et al.* (2019) Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis. *Proc Natl Acad Sci U S A* 116(3):950-959.
46. Ryckman K & Williams SM (2008) Calculation and use of the Hardy-Weinberg model in association studies. *Current protocols in human genetics* Chapter 1:Unit 1.18.
47. Karczewski KJ, Gauthier LD, & Daly MJ (2019) Technical artifact drives apparent deviation from Hardy-Weinberg equilibrium at CCR5-Δ32 and other variants in gnomAD. *bioRxiv*:784157.
48. Hardy GH (1908) MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science* 28(706):49-50.
49. Weinberg W (1908) *Uber den Nachweis der Vererbung beim Menschen* ([publisher not identified], [Place of publication not identified]).
50. Leal SM (2005) Detection of genotyping errors and pseudo-SNPs via deviations from Hardy-Weinberg equilibrium. *Genetic epidemiology* 29(3):204-214.
51. Gomes I, *et al.* (1999) Hardy-Weinberg quality control. *Annals of human genetics* 63(Pt 6):535-538.
52. Hosking L, *et al.* (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European journal of human genetics : EJHG* 12(5):395-399.
53. Graffelman J, Jain D, & Weir B (2017) A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data. *Human genetics* 136(6):727-741.
54. McCarroll SA, *et al.* (2006) Common deletion polymorphisms in the human genome. *Nat Genet* 38(1):86-92.
55. Abramovs N, Brass A, & Tassabehji M (2020) Hardy-Weinberg Equilibrium in the Large Scale Genomic Sequencing Era. *Frontiers in genetics* 11:210.
56. Casanova JL & Abel L (2018) Human genetics of infectious diseases: Unique insights into immunological redundancy. *Seminars in immunology* 36:1-12.
57. Quintana-Murci L (2016) Understanding rare and common diseases in the context of human evolution. *Genome biology* 17(1):225.
58. Quintana-Murci L & Barreiro LB (2010) The role played by natural selection on Mendelian traits in humans. *Annals of the New York Academy of Sciences* 1214:1-17.
59. Zarrei M, MacDonald JR, Merico D, & Scherer SW (2015) A copy number variation map of the human genome. *Nature reviews. Genetics* 16(3):172-183.
60. Collins RL, *et al.* (2019) An open resource of structural variation for medical and population genetics. *bioRxiv*:578674.
61. Zhang F, Gu W, Hurles ME, & Lupski JR (2009) Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics* 10:451-481.
62. Lee C & Scherer SW (2010) The clinical context of copy number variation in the human genome. *Expert Rev Mol Med* 12:e8-e8.
63. Sharp AJ, Cheng Z, & Eichler EE (2006) Structural variation of the human genome. *Annual review of genomics and human genetics* 7:407-442.
64. Kadalayil L, *et al.* (2015) Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform* 16(3):380-392.
65. Fromer M, *et al.* (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 91(4):597-607.
66. Tan R, *et al.* (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat* 35(7):899-907.
67. Gambin T, *et al.* (2017) Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res* 45(4):1633-1648.

68. de Ligt J, *et al.* (2013) Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum Mutat* 34(10):1439-1448.
69. Bigio B, *et al.* (2020) Detection of homozygous and hemizygous partial exon deletions by whole-exome sequencing. *bioRxiv*:2020.2007.2023.217976.
70. Casanova JL, Conley ME, Seligman SJ, Abel L, & Notarangelo LD (2014) Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies. *The Journal of experimental medicine* 211(11):2137-2149.
71. Meyts I, *et al.* (2016) Exome and genome sequencing for inborn errors of immunity. *The Journal of allergy and clinical immunology* 138(4):957-969.
72. Stenson PD, *et al.* (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Human genetics* 136(6):665-677.
73. Lek M, *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285-291.
74. Itan Y, *et al.* (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences of the United States of America* 112(44):13615-13620.
75. Petrovski S, Wang Q, Heinzen EL, Allen AS, & Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS genetics* 9(8):e1003709.
76. Kircher M, *et al.* (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* 46(3):310-315.
77. Itan Y, *et al.* (2016) The mutation significance cutoff: gene-level thresholds for variant predictions. *Nature methods* 13(2):109-110.
78. Itan Y, *et al.* (2013) The human gene connectome as a map of short cuts for morbid allele discovery. *Proceedings of the National Academy of Sciences of the United States of America* 110(14):5558-5563.
79. Bao R, *et al.* (2014) Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer informatics* 13(Suppl 2):67-82.
80. Fuentes Fajardo KV, *et al.* (2012) Detecting false-positive signals in exome sequencing. *Human mutation* 33(4):609-613.
81. DePristo MA, *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 43(5):491-498.
82. Alcais A, *et al.* (2010) Life-threatening infectious diseases of childhood: single-gene inborn errors of immunity? *Annals of the New York Academy of Sciences* 1214:18-33.
83. Casanova JL (2015) Severe infectious diseases of childhood as monogenic inborn errors of immunity. *Proceedings of the National Academy of Sciences of the United States of America* 112(51):E7128-7137.
84. Scott EM, *et al.* (2016) Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nature genetics* 48(9):1071-1076.
85. Wood P & Network UPI (2009) Primary antibody deficiencies: recognition, clinical diagnosis and referral of patients. in *Clin Med*, pp 595-599.
86. Guo Y, Ye F, Sheng Q, Clark T, & Samuels DC (2014) Three-stage quality control strategies for DNA re-sequencing data. *Briefings in bioinformatics* 15(6):879-889.
87. Asgari S, *et al.* (2017) Severe viral respiratory infections in children with IFIH1 loss-of-function mutations. *Proceedings of the National Academy of Sciences of the United States of America* 114(31):8342-8347.

88. Asgari S, *et al.* (2016) Exome Sequencing Reveals Primary Immunodeficiencies in Children with Community-Acquired Pseudomonas aeruginosa Sepsis. *Frontiers in immunology* 7:357.
89. Lopez M, *et al.* (2018) The demographic history and mutational load of African hunter-gatherers and farmers. *Nature Ecology & Evolution* 2(4):721-730.
90. Quach H, *et al.* (2016) Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations. *Cell* 167(3):643-656.e617.
91. Van der Auwera GA, *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics* 43:11.10.11-33.
92. Bogunovic D, *et al.* (2012) Mycobacterial disease and impaired IFN-gamma immunity in humans with inherited ISG15 deficiency. *Science (New York, N.Y.)* 337(6102):1684-1688.
93. Stenson PD, *et al.* (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 21(6):577-581.
94. Bardou P, Mariette J, Escudié F, Djemiel C, & Klopp C (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics* 15(1):293.
95. Kuehn HS, *et al.* (2016) Loss of B Cells in Patients with Heterozygous Mutations in IKAROS. *The New England journal of medicine* 374(11):1032-1043.
96. Belkadi A, *et al.* (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proceedings of the National Academy of Sciences of the United States of America* 112(17):5473-5478.
97. Belkadi A, *et al.* (2016) Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proceedings of the National Academy of Sciences of the United States of America* 113(24):6713-6718.
98. Jones E, Oliphant, T. & Pearu, P. (2001) Scipy: Open Source Scientific Tools for Python.
99. Buckley AR, *et al.* (2017) Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. *BMC genomics* 18(1):458.
100. Toubiana J, *et al.* (2016) Heterozygous STAT1 gain-of-function mutations underlie an unexpectedly broad clinical phenotype. *Blood* 127(25):3154-3164.
101. Robinson JT, *et al.* (2011) Integrative genomics viewer. *Nature Biotechnology* 29:24.
102. Fazekas A, Steeves R, & Newmaster S (2010) Improving sequencing quality from PCR products containing long mononucleotide repeats. *BioTechniques* 48(4):277-285.
103. Clarke LA, Rebelo CS, Gonçalves J, Boavida MG, & Jordan P (2001) PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Molecular Pathology* 54(5):351-353.
104. Mitchell AA, Zwick ME, Chakravarti A, & Cutler DJ (2004) Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics (Oxford, England)* 20(7):1022-1032.
105. Hwang S, Kim E, Lee I, & Marcotte EM (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports* 5:17875.
106. Sandmann S, *et al.* (2017) Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Scientific Reports* 7:43169.
107. Campbell IM, *et al.* (2016) Multiallelic Positions in the Human Genome: Challenges for Genetic Analyses. *Human mutation* 37(3):231-234.
108. Cingolani P, *et al.* (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6(2):80-92.

109. McLaren W, *et al.* (2016) The Ensembl Variant Effect Predictor. *Genome biology* 17(1):122.
110. Wang K, Li M, & Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 38(16):e164.
111. Ghoneim DH, Myers JR, Tuttle E, & Paciorkowski AR (2014) Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC research notes* 7:864.
112. Perry GH, *et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39(10):1256-1260.
113. Handsaker RE, Korn JM, Nemesh J, & McCarroll SA (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* 43(3):269-276.
114. Zhou B, *et al.* (2018) Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J Med Genet* 55(11):735-743.
115. Gross AM, *et al.* (2019) Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genetics in Medicine* 21(5):1121-1130.
116. Krumm N, *et al.* (2012) Copy number variation detection and genotyping from exome sequence data. *Genome research* 22(8):1525-1532.
117. Amarasinghe KC, Li J, & Halgamuge SK (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics* 14(2):S2.
118. Fromer M & Purcell SM (2014) Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Current protocols in human genetics* 81:7.23.21-21.
119. Guo Y, *et al.* (2014) Detection of internal exon deletion with exon Del. *BMC Bioinformatics* 15(1):332.
120. Backenroth D, *et al.* (2014) CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res* 42(12):e97.
121. Packer JS, *et al.* (2016) CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics (Oxford, England)* 32(1):133-135.
122. Jiang Y, Oldridge DA, Diskin SJ, & Zhang NR (2015) CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res* 43(6):e39.
123. Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. in *Bioinformatics (Oxford, England)* (Oxford University Press), pp 1754-1760.
124. Aydin SE, *et al.* (2015) DOCK8 deficiency: clinical and immunological phenotype and treatment options - a review of 136 patients. *Journal of clinical immunology* 35(2):189-198.
125. Rosain J, *et al.* (2018) A Variety of Alu-Mediated Copy Number Variations Can Underlie IL-12Rβ1 Deficiency. *Journal of clinical immunology* 38(5):617-627.
126. Blancas-Galicia L, *et al.* (2020) Genetic, Immunological, and Clinical Features of the First Mexican Cohort of Patients with Chronic Granulomatous Disease. *Journal of clinical immunology* 40(3):475-493.
127. Smedley D, *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 43(W1):W589-W598.

128. MacDonald JR, Ziman R, Yuen RKC, Feuk L, & Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res* 42(Database issue):D986-D992.
129. Kuśmirek W, Szmurło A, Wiewiórka M, Nowak R, & Gambin T (2019) Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers performance. *BMC Bioinformatics* 20(1):266-266.
130. Shigemizu D*, et al.* (2018) IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis. *Scientific Reports* 8(1):5608.
131. Sakharkar MK, Chow VTK, & Kangueane P (2004) Distributions of exons and introns in the human genome. *In Silico Biol* 4(4):387-393.
132. Hill WG (1996) Genetic Data Analysis II. By Bruce S. Weir, Sunderland, Massachusetts. Sinauer Associates, Inc.445 pages. ISBN 0-87893-902-4. *Genet. Res. Genetical Research* 68(2):187.
133. Rapaport F*, et al.* (2020) Negative selection on human genes causing severe inborn errors depends on disease outcome and both the mode and mechanism of inheritance. *bioRxiv*:2020.2002.2007.938894.
134. Stenson PD*, et al.* (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. in *Hum. Genet.*, p 68.
135. Quintana-Murci L (2019) Human Immunology through the Lens of Evolutionary Genetics. *Cell* 177(1):184-199.
136. Quintana-Murci L & Clark AG (2013) Population genetic tools for dissecting innate immunity in humans. *Nature reviews. Immunology* 13(4):280-293.
137. Le Pendu J, Ruvoën-Clouet N, Kindberg E, & Svensson L (2006) Mendelian resistance to human norovirus infections. *Seminars in immunology* 18(6):375-386.
138. Marionneau S*, et al.* (2001) ABH and Lewis histo-blood group antigens, a model for the meaning of oligosaccharide diversity in the face of a changing world. *Biochimie* 83(7):565-573.
139. Ferrer-Admetlla A*, et al.* (2009) A Natural History of FUT2 Polymorphism in Humans. *Molecular Biology and Evolution* 26(9):1993-2003.
140. Lindesmith L*, et al.* (2003) Human susceptibility and resistance to Norwalk virus infection. *Nature Medicine* 9(5):548-553.
141. Thorven M*, et al.* (2005) A homozygous nonsense mutation (428G→A) in the human secretor (FUT2) gene provides resistance to symptomatic norovirus (GGII) infections. *Journal of Virology* 79(24):15351-15355.
142. Nordgren J*, et al.* (2014) Both Lewis and Secretor Status Mediate Susceptibility to Rotavirus Infections in a Rotavirus Genotype–Dependent Manner. *Clinical Infectious Diseases* 59(11):1567-1573.
143. Payne DC*, et al.* (2015) Epidemiologic Association Between FUT2 Secretor Status and Severe Rotavirus Gastroenteritis in Children in the United States. *JAMA Pediatrics* 169(11):1040-1045.
144. Mottram L, Wiklund G, Larson G, Qadri F, & Svennerholm AM (2017) FUT2 non-secretor status is associated with altered susceptibility to symptomatic enterotoxigenic Escherichia coli infection in Bangladeshis. *Sci Rep* 7(1):10649.
145. Santos-Cortez RLP*, et al.* (2018) FUT2 Variants Confer Susceptibility to Familial Otitis Media. *Am J Hum Genet* 103(5):679-690.
146. Srivastava S*, et al.* (2001) SMN2-deletion in childhood-onset spinal muscular atrophy. *American journal of medical genetics* 101(3):198-202.

147. Prior TW*, et al.* (2009) A positive modifier of spinal muscular atrophy in the SMN2 gene. *Am J Hum Genet* 85(3):408-413.
148. Bustamante J, Boisson-Dupuis S, Abel L, & Casanova JL (2014) Mendelian susceptibility to mycobacterial disease: genetic, immunological, and clinical features of inborn errors of IFN-γ immunity. *Seminars in immunology* 26(6):454-470.
149. Kalb R*, et al.* (2007) Hypomorphic mutations in the gene encoding a key Fanconi anemia protein, FANCD2, sustain a significant group of FA-D2 patients with severe phenotype. *Am J Hum Genet* 80(5):895-910.
150. Landrum MJ*, et al.* (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. in *Nucleic Acids Res.*, pp D980-985.
151. Hamosh A, Scott AF, Amberger JS, Bocchini CA, & McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. in *Nucleic Acids Res.* (Oxford University Press), pp D514-517.
152. Lek M*, et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. in *Nature* (Nature Research), pp 285-291.

Appendix: articles resulting from the thesis work

Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis

Patrick Maffucci[a,b,c,1], Benedetta Bigio[a,d,e,1], Franck Rapaport[a], Aurélie Cobat[d,e], Alessandro Borghesi[f], Marie Lopez[g,h,i], Etienne Patin[g,h,i], Alexandre Bolze[j], Lei Shang[a], Matthieu Bendavid[a], Eric M. Scott[k], Peter D. Stenson[l], Charlotte Cunningham-Rundles[b,c], David N. Cooper[l], Joseph G. Gleeson[k,m], Jacques Fellay[f], Lluis Quintana-Murci[g,h,i], Jean-Laurent Casanova[a,d,e,m,n,3], Laurent Abel[a,d,e], Bertrand Boisson[a,d,e,2], and Yuval Itan[a,o,p,2,3]

[a]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065; [b]Immunology Institute, Graduate School, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [c]Department of Medicine, Division of Clinical Immunology, Icahn School of Medicine at Mount Sinai, New York, NY 10029; [d]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, 75015 Paris, France; [e]Imagine Institute, Paris Descartes University, 75015 Paris, France; [f]School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland; [g]Human Evolutionary Genetics Unit, Pasteur Institute, 75015 Paris, France; [h]CNRS UMR2000, 75015 Paris, France; [i]Center of Bioinformatics, Biostatistics and Integrative Biology, Pasteur Institute, 75015 Paris, France; [j]Helix, San Carlos, CA 94070; [k]Rady Children's Institute for Genomic Medicine, Department of Neurosciences, University of California, San Diego, La Jolla, CA 92093; [l]Institute of Medical Genetics, School of Medicine, Cardiff University, Cardiff CF14 4XW, United Kingdom; [m]Howard Hughes Medical Institute, New York, NY 10065; [n]Pediatric Hematology–Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France; [o]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029; and [p]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029

Contributed by Jean-Laurent Casanova, November 11, 2018 (sent for review May 17, 2018; reviewed by Harry Ostrer, Amalio Telenti, and Magdalena Walkiewicz)

Computational analyses of human patient exomes aim to filter out as many nonpathogenic genetic variants (NPVs) as possible, without removing the true disease-causing mutations. This involves comparing the patient's exome with public databases to remove reported variants inconsistent with disease prevalence, mode of inheritance, or clinical penetrance. However, variants frequent in a given exome cohort, but absent or rare in public databases, have also been reported and treated as NPVs, without rigorous exploration. We report the generation of a blacklist of variants frequent within an in-house cohort of 3,104 exomes. This blacklist did not remove known pathogenic mutations from the exomes of 129 patients and decreased the number of NPVs remaining in the 3,104 individual exomes by a median of 62%. We validated this approach by testing three other independent cohorts of 400, 902, and 3,869 exomes. The blacklist generated from any given cohort removed a substantial proportion of NPVs (11–65%). We analyzed the blacklisted variants computationally and experimentally. Most of the blacklisted variants corresponded to false signals generated by incomplete reference genome assembly, location in low-complexity regions, bioinformatic misprocessing, or limitations inherent to cohort-specific private alleles (e.g., due to sequencing kits, and genetic ancestries). Finally, we provide our precalculated blacklists, together with ReFiNE, a program for generating customized blacklists from any medium-sized or large in-house cohort of exome (or other next-generation sequencing) data via a user-friendly public web server. This work demonstrates the power of extracting variant blacklists from private databases as a specific in-house but broadly applicable tool for optimizing exome analysis.

exome | variant | blacklist | WES analysis | WES annotation

Next-generation sequencing (NGS), particularly whole-exome sequencing (WES) and whole-genome sequencing (WGS), is increasingly being used for the discovery and diagnosis of human genetic disorders (1–3). The number of new disease-causing genetic variants logged by the Human Gene Mutation Database (HGMD) is currently increasing at a rate of ∼10% per annum (4). This increase has coincided with an expansion of the use of WES and WGS (1, 2). The mean number of exonic coding variants per individual relative to the reference human genome is about 20,000 (2, 3), but monogenic disease in any given individual is generally driven by at most two variants. The remaining nonpathogenic variants (NPVs) may be real variants (rare or common, deleterious or neutral), or false/low-quality variants [sequencing artifacts, bioinformatic misprocessing of raw sequencing data, or resulting from limitations to the performance of current quality control (QC) methods]. In practice, analyses of individual exomes aim to generate a short list of high-quality candidate variants by filtering out as many NPVs as possible, while minimizing the risk of false negatives (FNs) due to the removal of true disease-causing mutations. The first step in this process typically involves the use of public databases to identify and remove NPVs through comparisons of their frequency in the

Significance

Whole-exome sequencing data from patients with monogenic inborn errors identify thousands of genetic variants in each patient, only a few of which are pathogenic. Identifying pathogenic mutations therefore requires the rigorous filtration of variants that are too common in public databases to cause a disease of the observed incidence. We report that a large proportion of the variants common in patient cohorts are paradoxically absent from public databases. We define these nonpathogenic, cohort-specific common variants that cannot be removed from the analysis as a "blacklist." We describe these blacklisted variants, demonstrate their usefulness for removing nonpathogenic variants, explain their origin experimentally, and provide a web server and software enabling researchers to automate the creation of their own blacklists.

Author contributions: P.M., J.-L.C., L.A., B. Boisson, and Y.I. designed research; P.M., B. Bigio, and B. Boisson performed research; P.M., B. Bigio, F.R., A.C., A. Borghesi, M.L., E.P., A. Bolze, L.S., M.B., E.M.S., P.D.S., C.C.-R., D.N.C., J.G.G., J.F., L.Q.-M., J.-L.C., L.A., B. Boisson, and Y.I. analyzed data; P.M., B. Bigio, J.-L.C., L.A., B. Boisson, and Y.I. wrote the paper; P.M. designed and wrote the ReFiNE software; and L.S. and B. Bigio processed and prepared exomes, and designed and implemented the blacklist webserver.

Reviewers: H.O., Albert Einstein College of Medicine; A.T., Scripps Research Institute; and M.W., NIH.

Conflict of interest statement: A.T. has coauthored multiple papers with J.F. and J.G.G. M.W. coauthored a 2017 paper with J.G.G.

Published under the PNAS license.

Data deposition: ReFiNE and precalculated blacklists are available at GitLab on https://gitlab.com/pmaffucci/refine.

[1]P.M. and B. Bigio contributed equally to this work.

[2]B. Boisson and Y.I. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: casanova@rockefeller.edu or yuval.itan@mssm.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1808403116/-/DCSupplemental.

Published online December 27, 2018.

general population with the prevalence of the disease considered, its proposed mode of inheritance, and its estimated clinical penetrance. The largest public database currently available is the Genome Aggregation Database (gnomAD), which includes 123,136 exomes and 15,496 genomes from a total of 138,632 individuals (5). For the remaining variants, including those not reported in public databases, various variant-level and gene-level metrics can be used to predict deleteriousness and to select a smaller set of candidate variants for further experimental analysis (6–10).

In studies of rare genetic diseases, public databases are widely used for the initial elimination of common variants [minor allele frequency (MAF) > 0.01] (2, 11). However, some common variants within private databases may be absent from public databases, and most such variants are likely to be NPVs (2, 12). The efficacy with which such variants are identified and used for analyses of exomes from cohorts of patients studied by a particular research group has never been assessed in detail. An approach for detecting false-positive signal (defined as DFS) based on an internal cohort of 118 whole-exome sequences from different individuals generated a shortlist of variants found to be in Hardy–Weinberg (HW) disequilibrium due to excess heterozygosity (the DFS list; 23,389 variants) (13). However, most of these variants (68%) had already been reported in dbSNP (13). Machine learning-based methods for removing false positives (FPs) from sequencing data, such as variant quality score recalibration (VQSR), which uses a clustering score to determine whether a called variant is true (14), can limit the number of NPVs in exome data. However, these methods are subject to several limitations: (*i*) they are computationally intensive and time-consuming; (*ii*) they often require a large number of samples; (*iii*) parameter optimization requires extensive testing; and (*iv*) the addition of new samples requires reprocessing of the entire cohort. These methods are therefore little used by most researchers, who have small- or medium-sized exome cohorts and may not have access to powerful computing resources. It has been suggested that variants common within a homogeneous cohort and absent from public databases could be filtered out (2), but this approach has not been validated and there are currently no tools for the easy identification and compilation of such variants. In this context, we sought to establish a "blacklist" of variants too frequent in our cohort of 3,104 exomes from patients with severe infectious diseases (15–17).

Results

Determining a Frequency Cutoff for NPVs. We observed that numerous candidate variant calls (*Materials and Methods*) (18) predicted to be damaging to the corresponding transcript or protein were present in >1% of our cohort of 3,104 in-house exomes from primary immune deficiency (PID) patients with heterogeneous ancestral backgrounds (19) (i.e., too common to cause PID) but absent from public databases (e.g., 1KG, ExAC, gnomAD). These variants are poor candidates for involvement in rare diseases but are impossible to eliminate by current methods based on variant frequencies in public databases (2). We therefore sought to classify and characterize these variants in a rigorous and comprehensive manner, to enable users to remove them from their WES/WGS analyses. First, we determined a statistical cutoff frequency above which in-house variants should be considered too frequent to cause rare diseases. We found that the MAF of all experimentally validated disease-causing mutations in HGMD followed a Gilbrat distribution (20). We then calculated the 99% Gilbrat distribution confidence interval (CI) for these frequencies and found that the upper boundary of the CI for the frequency of known disease-causing mutations was 0.01 (1%). We therefore used this cutoff as a criterion for the nonpathogenicity of variants (occurring in too many patients in our database to explain a rare monogenic illness). The MAF > 0.01 cutoff used here is an example of the blacklist approach to removing FP variants in studies of rare genetic disorders. The cutoff can be adjusted according to the mode of inheritance and genetic architecture, assumed penetrance, and prevalence of the disease, and the phenotypic homogeneity of the cohort (21). For example, assuming complete penetrance and allelic homogeneity, a rare recessive genetic disorder with a prevalence of 1 in 100,000 could be analyzed with a MAF cutoff of 0.0033, whereas a more common recessive genetic disorder with a prevalence of 1 in 1,000 should be analyzed with a MAF cutoff of 0.033. The assumption of incomplete penetrance may lead to the definition of higher cutoffs, whereas the assumption of allelic/genetic heterogeneity may lead to the use of lower cutoffs.

Generating the Blacklist. We first designed the reducing FPs in NGS elucidation (ReFiNE) software, an easy-to-use tool for extracting blacklist variants from internal cohorts of WES or WGS data on the basis of a user-defined frequency cutoff (see *Materials and Methods* for details). ReFiNE creates a blacklist consisting of the full set of variants occurring in >1% (or any user-defined cutoff) of an investigated cohort, which can then be further filtered separately by the user, using MAF cutoffs from a population genetic database of choice. Using ReFiNE, we first collated all variants present at a frequency >1% in our PID WES cohort of 3,104 exomes (*Materials and Methods* and *SI Appendix*, Fig. S1) with a depth of coverage (DP) ≥ 5 and mapping quality (MQ) ≥ 30 (*Materials and Methods*) (5, 22). A large number of multiallelic variants in our cohort were absent from gnomAD for specific chromosomal positions. ReFiNE therefore collapsed all variants at a unique chromosomal position and summed the total number of patients at each of these positions. This generated a list of 780,956 variants, defined as the blacklist. This blacklist is the full list of variants occurring at single chromosomal positions for which >1% of patients had an alternative allele. These variants belonged to two classes: (*i*) biallelic, with a single alternative allele in our cohort; and (*ii*) multiallelic, with two or more alternative alleles in our cohort. The blacklist includes variants already reported in public databases, so we needed to extract the subset of variants unique to our method for further analysis. We thus annotated the blacklist with gnomAD, currently the most extensive public population genetics database available (5, 23). We found that 21.4% (167,144) of these 780,956 variants were absent from the gnomAD full exome and genome databases. As these 167,144 variants are not captured by the most extensive public database available, we focused the analysis of our method on this subset of variants, which, for simplicity, we will refer to as blacklist-annotated (BL-A): common in-house variants absent from gnomAD that cannot, therefore, be filtered out of analyses based on gnomAD.

Blacklist Filtering Removes 62% of the NPVs Remaining After Standard Public Database Filtering. We then assessed the efficacy of BL-A for filtering out NPVs from patient exome data. We first applied the standard procedure for rare genetic disorders, by removing variants with a MAF > 0.01 in gnomAD from our 3,104 exomes (3, 12). This reduced the median number of variants in the patients' exomes by 90% (Fig. 1*A*). Subsequent filtering with BL-A removed 62% of the remaining variants that could not be removed by other means (Fig. 1*A*, a median of 9,056 variants removed per exome). By comparison, the DFS list (13) decreased the median number of these variants by only 1.8% (median of 260 variants removed per exome). BL-A filtering was effective for both coding sequences [coding DNA sequences (CDSs)], including indel, exon-deleted, nonsynonymous, synonymous, and essential splicing variants, and for non-CDS variants, including untranslated region (UTR), nonessential splicing, intronic, downstream, and upstream variants, and for all three exome kits available for our cohort (37, 50, and

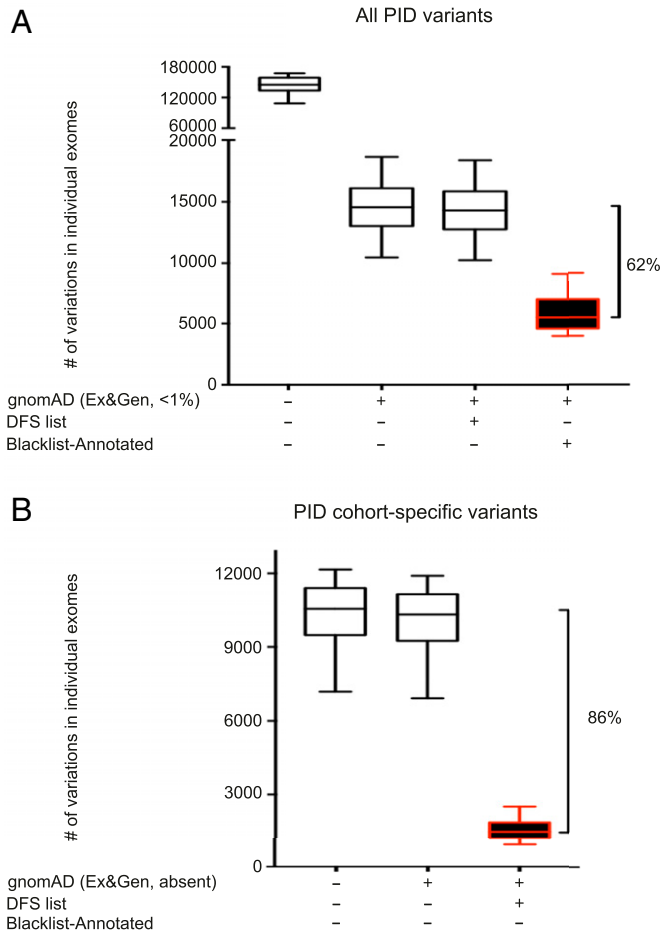


Fig. 1. Blacklist filtering of 3,104 PID exomes with the PID blacklist. (*A*) Filtering of all variants in each exome by first removing those common in gnomAD exome and genome databases (MAF greater than 0.01). The remaining variants were subsequently filtered with the blacklist. (*B*) Filtering of cohort-specific variants in each exome with the blacklist. Filtering with the DFS list is shown for comparison. Error bars represent the 10th to 90th percentiles.

71 Mb; *SI Appendix*, Figs. S2 and S3). We then assessed the performance of BL-A filtering for variants absent from the gnomAD database (i.e., variants private to the PID database), which would be considered among the strongest candidates for a causal role in disease. This approach decreased the number of cohort-private variants potentially associated with PID in each exome by 86%, versus only 2.2% for the DFS list, and was similarly effective for CDS and non-CDS variants (Fig. 1*B* and *SI Appendix*, Fig. S4). Thus, when used as a filtering tool, our blacklist was able to remove variants absent from public databases and to decrease the number of candidate variants per exome considerably.

Metric Characteristics of the Variants and Genes Included in the Blacklist. We then explored whether the QC scores for BL-A variants were similar to those for polymorphic variants (MAF > 0.01) reported in gnomAD. By comparing the median MQ and DP scores for blacklisted variants and polymorphic variants from our cohort (*SI Appendix*, Fig. S5), we demonstrated that none of these QC metrics could differentiate between these two sets of variants (especially when considering commonly used criteria for hard filtration; see *Materials and Methods* for further details). We then investigated whether machine learning QC metrics could classify these variants. With VQSR, only 25% of BL-A variants were annotated as "nonpass" (*SI Appendix*, Table S1). One of the key goals of this approach is providing an efficient tool for researchers who cannot easily perform VQSR. We therefore retained these VQSR nonpass variants in the blacklist. We also assessed the ability of a random forest classifier trained on polymorphic variants from the gnomAD dataset well-characterized by different methods to separate true variants from FP artifacts called by the variant-calling pipeline (5). We then used the same method to construct a new scoring function with the gnomAD dataset. We applied both scoring functions to the blacklist variants and a set of variants present in both the gnomAD dataset and our cohort, with a MAF of more than 1% in each dataset. The score distributions obtained were almost identical (*SI Appendix*, Fig. S6), demonstrating an inability of this standard classification method to distinguish between the blacklisted variants and true-positive (TP) variants. We then characterized the variants and genes included in BL-A with computational damage prediction metrics. A variant-level analysis revealed that the combined annotation-dependent depletion (CADD) (8) scores for blacklist variants were not significantly different from those for variants not included in the blacklist (*SI Appendix*, Fig. S7). A gene-level analysis (6) of all genes with blacklist variants (n = 13,665 genes) showed them to have low gene damage index (GDI) values (*SI Appendix*, Fig. S8). However, some genes with a high GDI have many BL-A variants (e.g., *HLA-DRB1*, 658 variants; *MUC16*, 455 variants). Filtration methods based on QC and variant- and gene-level damage prediction metrics would not efficiently detect and remove the blacklist variants absent from gnomAD. These results demonstrate the value of blacklisting as a complementary approach to analyses based on standard public databases, including gnomAD, QC filtering, and damage prediction metrics.

Determining the FN Rate Associated with Blacklist Use. We estimated the proportion of TP disease-causing mutations removed by the blacklist approach, by screening 129 exomes from patients in our cohort for whom the TP mutations had been validated experimentally. Filtering these exomes with the complete blacklist did not remove any of the known TP mutations (0% FN rate). Even though most variants in any patient are not pathogenic, our analysis indicates that it is very safe to apply the blacklist to patient exomes. We also compared the complete blacklist with the list of 144,641 disease-causing mutations in HGMD and noted an overlap of only 263 variants (0.18% FN rate). These variants are listed as disease-causing in the HGMD dataset, but 47% have a MAF > 0.01 in the gnomAD exome or genome databases, suggesting that are unlikely to be the cause of a rare disorder. These findings indicate that our FN rate is probably lower than the rate of 0.18% for HGMD in the context of rare disorders. Only eight BL-A variants were present in HGMD (0.001% FN rate), indicating that the FN rate for our specific BL-A list was lower than that for gnomAD. Together, these results suggest that the FN rate is very low for this technique (*SI Appendix*, Table S2). We also screened 3,731,152 somatic cancer-causing or cancer-associated variants available from TCGA (https://cancergenome.nih.gov). We found that 59,151 of these TCGA variants (1.5%) were present in the complete blacklist and 2,471 (0.07%) were present in BL-A. As our blacklist was derived from germline exome data, the presence of these blacklist variants in the TCGA database suggests that they may be FPs that could be removed, as previously reported (24). Together, these data indicate that the blacklist approach results in an extremely low FN rate when applied to patients with rare diseases, and that it is therefore safe to use this approach to remove NPVs from patient exome data.

Practical Application of the Blacklist to the Analysis of Patients' Exomes. We assessed the use of blacklisting for practical analyses of patient exomes. We selected a case from our cohort with an autosomal dominant disease-causing mutation described in a previous study (patient D2 from ref. 25). We filtered this patient's

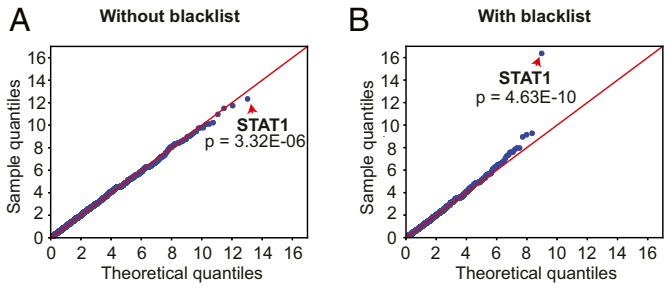


Fig. 2. Application of the blacklisting approach to enrichment analysis. Quantile–quantile plots depicting the analysis of genetic homogeneity for a cohort of 202 patients with chronic mucocutaneous candidiasis (CMC) before (*A*) and after (*B*) application of the blacklist. The control cohort consisted of 852 unrelated individuals. In each panel, the red arrows indicate *STAT1*, the known cause of CMC in our cohort, before and after blacklist application.

exome with a standard pipeline to identify disease-causing mutations (*SI Appendix*, Fig. S9). This standard approach reduced the number of candidate variants from 142,473 to 3,526. Taking known mode of inheritance into account and restricting the analysis to CDS variants (excluding synonymous alterations), the number of candidate variants was reduced further, to 231. The inclusion of BL-A in the pipeline decreased the final number of candidate variants to 109 (*SI Appendix*, Fig. S9), with retention of the known *IKZF1* mutation. Overall, this corresponds to a 53% decrease in the number of variants from this patient's exome to be considered. The remaining variants were high-quality candidates that would probably merit rigorous analysis in exome analyses for patients with diseases of unknown etiology. Thus, blacklisting greatly decreases the number of candidate variants for further study in practice, in exome analyses on individual patients.

Practical Application of Blacklisting to the Analysis of Population Exomes. We then explored the use of our blacklist for gene burden analysis for genetic homogeneity at the population level. We compared the number of patients with at least one variant of any given gene between a cohort of 202 patients suffering from chronic mucocutaneous candidiasis (CMC) and 852 phenotypically unrelated controls (26). When standard filtering with public databases was applied in the absence of blacklisting, the enrichment observed for the known disease-causing gene in the CMC cohort, *STAT1* (P value = 3.32×10^{-6}) was not significant considering the corrected threshold at the genome-wide level (P $\text{value}_{\text{threshold}} = 0.05 \div 20554 = 2.43 \times 10^{-6}$; Fig. 2*A*). However, following the addition of BL-A to the pipeline, *STAT1* was correctly identified as a gene displaying strong and significant genome-wide enrichment in the disease cohort (P value = 4.63×10^{-10}; Fig. 2*B*). In this instance, our blacklist removed two variants present in a large proportion of our PID exomes (both cases and controls) that confounded the statistical comparison between the CMC and control groups. Together with the previous practical example, these analyses demonstrate the power of blacklisting for removing NPVs from patient exomes, both to simplify candidate variant identification in patients and for other large-scale statistical analyses of patient groups.

Characterization of Multiallelic Variants from the Blacklist. We then characterized the PID cohort BL-A variants (n = 167,144). Most of the variants (91.5%) in the blacklist were multiallelic (*SI Appendix*, Table S3). The cohort-specific variants present in the blacklist were therefore due to multiallelic sites displaying high levels of variation in our cohort (*SI Appendix*, Table S4). We began by hypothesizing that the multiallelic variants might lie in low-complexity regions of the human genome, leading to sequencing errors. The annotation of all these variants with RepeatMasker, Simple Repeats, and GC percent tracks from University of California, Santa Cruz (UCSC) Genome confirmed that 118,154 of the 152,915 variants (77.3%) occurred in repetitive or GC-rich regions, and that most (65,646; 56%) were located in short tandem repeat (STR) regions (Fig. 3 and *SI Appendix*, Table S4).

Characterization of Biallelic Variants from the Blacklist. We analyzed the biallelic variants, which were also found to be located in repetitive or GC-rich regions, albeit to a lesser extent (6,711; 47.2%) (Fig. 3 and *SI Appendix*, Table S4). We also characterized these biallelic variants, focusing on those located in CDS regions, in the 1,150 individuals of European origin according to principal-component analysis (PCA) (19), to determine whether these variants were under HW equilibrium. In total, 388 CDS variants were found to be located in repetitive or GC-rich regions; 339 (87.4%) of these variants were in HW equilibrium and 49 (13.6%) were in HW disequilibrium (threshold of $P < 10^{-8}$; *SI Appendix*, Table S5). An investigation of the biallelic variants not present in repetitive regions (7,518; 52.8%) yielded a similar distribution, with 209 (89.3%) and 25 (10.7%) of the 234 CDS variants in HW equilibrium and disequilibrium, respectively. Overall, 74 CDS variants were in HW disequilibrium, and in 39 of these variants (52.7%), the cause was an excess of homozygous wild-type (14.9%) or homozygous alternative (37.8%) genotypes (*SI Appendix*, Table S5). Most of these 39 variants had low coverage (wild-type = 15.6×, alternative = 20.5×; *SI Appendix*, Table S5), which may have led to miscalls for a homozygous genotype. Most of the variants (35; 47.3%) in HW disequilibrium presented heterozygote excess, with high mean coverage rates of 163× (much higher than the 42.5× coverage of the 548 CDS variants in HW equilibrium), suggesting an excess

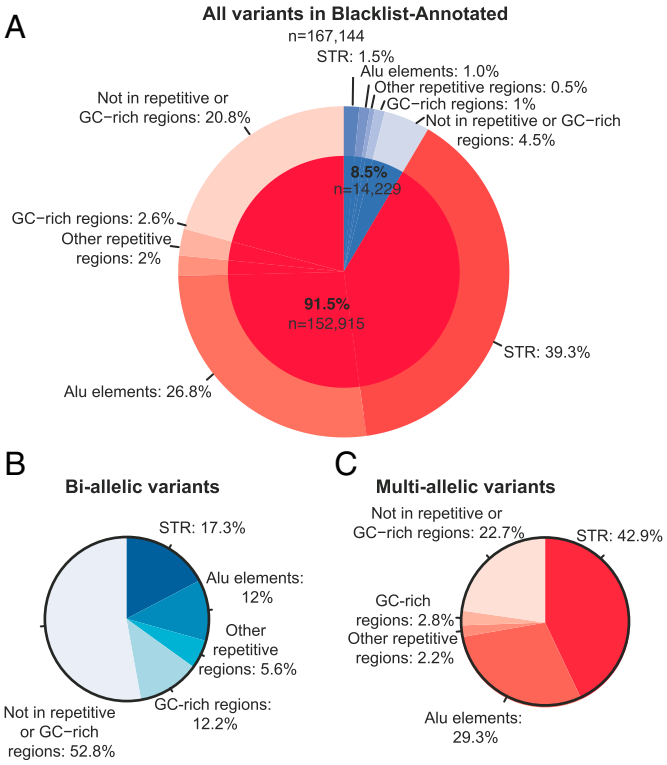


Fig. 3. Characterization of the blacklisted biallelic and multiallelic variants in low-complexity regions of the genome. Occurrence of the blacklisted multiallelic (red) and biallelic (blue) variants in repetitive [short tandem repeats (STRs), Alu elements, other repetitive regions] and GC-rich regions; percent relative to the total number of blacklisted variants (*A*) or the total number of biallelic (*B*) or multiallelic (*C*) blacklisted variants.

of reads wrongly mapped to the region (*SI Appendix*, Table S5). We also studied the 548 biallelic CDS variants in HW equilibrium, to evaluate their distribution across ethnicities. We focused the analysis on the four largest genetic ancestry groups in our cohort (*SI Appendix*, Fig. S10): European, African, North African, and Middle Eastern, as determined by PCA (19). In total, 200 (36.5%) of these variants were heterogeneously distributed across genetic ancestries (threshold of $P < 10^{-8}$; *SI Appendix*, Table S6). The observed heterogeneous distribution was probably due to one specific genetic ancestry in 46 (23%) of the variants (*SI Appendix*, Table S6). In 20 variants (43.5%), the individual genetic ancestry was Middle Eastern (*SI Appendix*, Table S6), which is poorly represented in public databases (27), suggesting that these variants are true variants that are more common in this population.

Experimental Investigation of the Blacklisted Variants. We further investigated the features of BL-A variants. We first focused on biallelic blacklist CDS variants in HW disequilibrium displaying excess heterozygosity and absent from repetitive regions in individuals of European ancestry ($n = 35$). We found that 48.6% of these variants ($n = 17$) mapped to four chromosomal regions, in the *HLA-DRB1*, *MUC6*, *OR8U1*, and *TAS2R43* genes with consecutive blacklist variants (less than 300 bp) (*SI Appendix*, Table S7). Most of these regions contain flagged variants annotated in gnomAD (47% in Exome and 65% in Genome, annotated as AC0, RF, and/or InbreedingCoeff; *SI Appendix*, Table S7). For the remaining variants (referred to as "unique"), we found that the blacklist variants were at the same location (but with different genotypes) as flagged variants annotated in gnomAD, like the consecutive variants (28% in Exome and 50% in Genome, annotated as AC0, RF, and/or InbreedingCoeff). Integrative genomics viewer (IGV) (28) showed that the consecutive variants in these regions belonged to the same reads, suggesting the existence of an "alternative" sequence (referred to as a segmental duplication by gnomAD or as an alternative haplotype; *SI Appendix*, Figs. S11–S13). These observations strongly suggest that some blacklist biallelic variants define alternative haplotypes belonging to unmapped regions absent from the human reference genome. These variants were probably incorrectly mapped to the region of the reference genome for which the best match was obtained, leading to a mixture of wild-type and alternative alleles in these regions, resulting in higher coverage and a final erroneous heterozygous call. In a second analysis, we focused on multiallelic variants. Most of these variants (77%) were located in low-complexity regions (STRs, Alu elements, GC-rich regions, or other repetitive regions; Fig. 3). IGV analysis of three multiallelic variants absent from these regions and common in our cohort (MAF > 0.9) revealed that they were located in the vicinity of a small stretch of repeated nucleotides (*SI Appendix*, Figs. S14–S16). Extending the analysis to the 23% of multiallelic variants not previously detected in low-complexity regions ($n =$ 34,761), we found that 83.3% were also located close to mononucleotide repeats (26,165; 75.3%) or to small repetitive stretches (two or more nucleotides; 2,802; 8.1%). Attempts to confirm these variants by Sanger sequencing failed, due to the mononucleotide repeat (*SI Appendix*, Table S8), strongly suggesting that the WES approach may have been affected by a polymerase artifact similar to that reported in previous studies (29, 30). This exploration of blacklist variants suggests that the multiallelic variants probably resulted from—to a large extent—sequencing/calling errors during WES on low-complexity regions, whereas a proportion of the blacklist biallelic variants, particularly those in HW disequilibrium, were due to mapping errors resulting from the incomplete nature of the GRCh37/GRCh38 genome assembly.

Testing the Blacklist Approach as a General Filtering Method in Three Unrelated Cohorts. We assessed the suitability of the blacklist approach for filtering in other private databases. We used three unrelated independently processed exome cohorts (from DNA preparation to VCF data): (*i*) 3,869 exomes from patients suffering from neurological diseases ("Neuro") (27); (*ii*) 902 exomes from patients suffering from diseases with an infectious phenotype ("Infection"); and (*iii*) 400 exomes (100 from Europeans and 300 from Africans) from a study on the demographic history of Central Africans ("Africa") (31). We first generated separate blacklists for the Neuro, Infection, and Africa cohorts, according to the pipeline described above. After filtering on the basis of MAF > 1% (in the specific cohort) in gnomAD, the application of the cohort-specific blacklists for the Neuro, Infection, and Africa cohorts decreased the number of variants retained by 35%, 57%, and 51%, respectively (a median of 3,160, 3,462, and 7,905 variants per exome, respectively; Fig. 4 *A, C*, and *E*). Considering only cohort-private variants (i.e., those appearing in the specific cohort but absent from gnomAD exomes and genomes), applying the cohort-specific blacklists to the Neuro, Infection, and Africa cohorts reduced the number of variants in each exome by 90%, 92%, and 93%, respectively, eliminating a median of 3,195, 3,418, and 7,861 variants per exome, respectively (Fig. 4 *B, D*, and *F*). This filtering was effective for both CDS and non-CDS variants (*SI Appendix*, Fig. S17). A comparison of the four blacklists revealed that a substantial number of variants were unique to each blacklist (*SI Appendix*, Fig. S18), demonstrating the cohort specificity of the blacklisted variants, particularly for the Africa cohort, probably due to ancestry. Specifically, each blacklist contained 63–91% of the unique biallelic variants (*SI Appendix*, Fig. S18*A* and Table S3) and 46–92% of the unique multiallelic variants (*SI Appendix*, Fig. S18*B*). A similar pattern was observed when the analysis was restricted to biallelic and multiallelic CDS variants (*SI Appendix*, Fig. S18 *C* and *D* and Table S3). Thus, the efficacy of blacklist filtering in our PID cohort was not due to specific pipeline settings or enrichment within our exomes. Instead, our results suggest that the blacklist method should effectively remove a substantial proportion of the NPVs not already removed by public database analysis from any cohort of exomes considered.

Application of the Blacklist to Unrelated Cohorts. We then assessed whether the originally generated PID blacklist would effectively filter exomes from the unrelated Neuro, Infection, and Africa cohorts used above. We removed variants with a MAF > 0.01 in gnomAD from the Neuro, Infection, and Africa exomes and then applied the PID BL-A. This reduced the median number of remaining variants in the Neuro, Infection, and Africa exomes by 8%, 41%, and 6%, respectively (median of 715, 2,487, and 947 variants per exome, respectively; Fig. 4 *A, C*, and *E*, blue box). When the analysis was restricted to cohort-private variants in the Neuro, Infection, and Africa exomes, the PID blacklist decreased the number of variants in individual exomes by 19%, 65%, and 11%, respectively (median of 673, 2,439, and 957 variants per exome, respectively; Fig. 4 *B, D*, and *F*, blue box). The superior efficiency of the PID blacklist for the Infection cohort may reflect the library preparation technique (SureSelect) and sequencing technology (HiSeq sequencer) used. Nevertheless, the PID blacklist was shown to be a useful filtering approach in unrelated cohorts in which exomes were captured with different kits and sequencing technologies (SureSelect or Nextera kits and HiSeq 2000 or HiSeq 2500 sequencing, respectively). We also found that filtering our PID exomes with the blacklist from the Neuro cohort did not remove any TP variants from the 129 PID exomes with proven disease-causing mutations. Blacklists are, therefore, effective for filtering exomes other than those with which they were developed and including cohort-private NPVs. However, generating internal blacklists from the cohort under

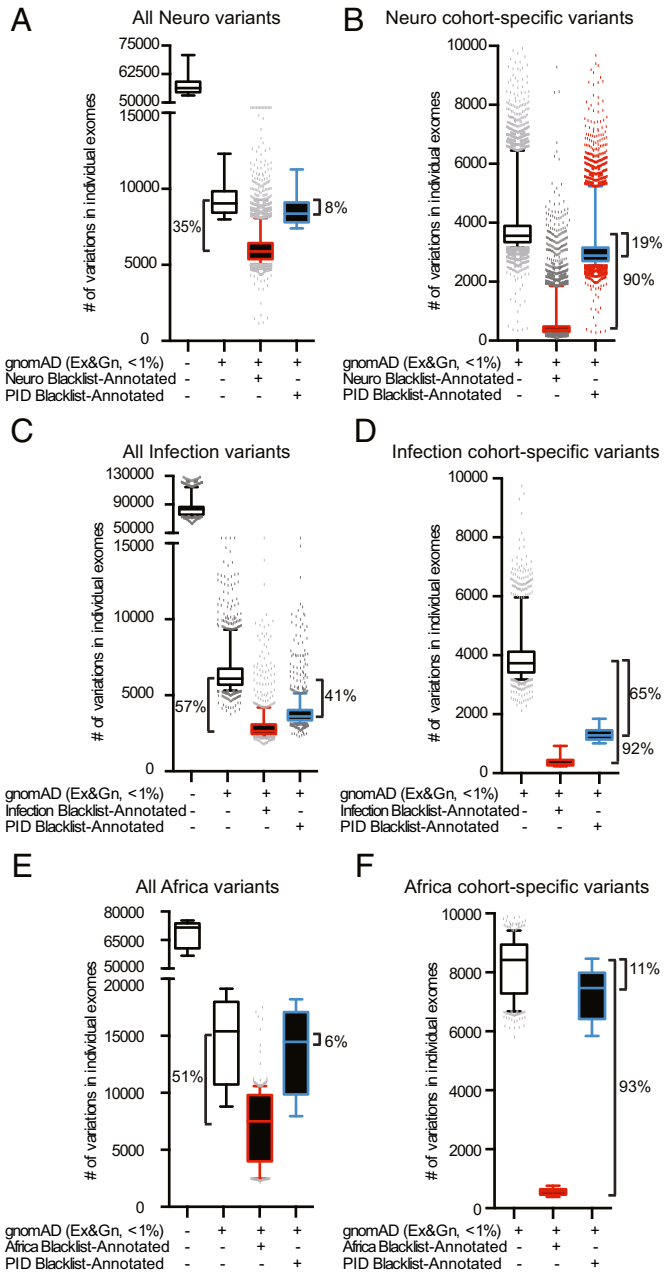


Fig. 4. Blacklist filtering of unrelated cohort exomes. (*A*, *C*, and *E*) Filtering of all variants in the neurological (*A*), infectious disease (*C*), and central African (*E*) exomes by first removing those common in gnomAD exome and genome databases (MAF greater than 0.01). The remaining variants were subsequently filtered with the Neuro (*A*), Infection (*C*), or Africa (*E*) blacklists (red boxes), or the PID blacklist (blue boxes). (*B*, *D*, and *F*) Filtering of exomes restricted to cohort-specific variants, with the Neuro (*B*), Infection (*D*), or Africa (*F*) blacklists (red boxes), or the PID blacklist (blue boxes). Error bars represent the 10th to 90th percentiles.

investigation was found to be the most effective approach to removing NPVs.

Determining the Minimum Cohort Size and Saturation Point for the Blacklist. We sought to determine the minimum sample size appropriate for the generation of a custom blacklist for a cohort of interest. We combined the two largest cohorts studied here—our PID cohort (3,104) and the Neuro cohort (3,869)—and simulated blacklists by randomly sampling various numbers of individuals relative to cohort size, with 30 iterations for each sample size (*SI Appendix*, Fig. S19). As the Neuro cohort was captured with the 50-Mb kit, which targets CDS, we focused this analysis exclusively on CDS variants. The number of CDS variants in the simulated BL-A increased rapidly with sample size between 10 and 500 individuals, whereas the number of variants increased more slowly when sample size exceeded 500 individuals. We therefore propose the use of samples of at least 500 heterogeneous unrelated individuals, to ensure the reliable capture of common cohort-specific variants. We estimated the saturation point for the blacklist's CDS variants (less than one new variant added per new individual) at a sample size of ∼2,801 individuals (*SI Appendix*, Fig. S19). Thus, a blacklist generated with the pipeline described here could be considered "saturated" for the purpose of capturing most of the cohort-specific CDS variants that cannot be removed by public database analysis.

Efficacy of the Combined Blacklist. Finally, we explored the efficacy of a "universal" blacklist generated by combining the four BL-As presented in this study. We reasoned that the aggregation of blacklists obtained from different cohorts (and different samples/data-processing methods) would result in a "universal blacklist" with the number of filtered variants eventually converging. We tested this hypothesis by aggregating either (*i*) the four blacklists (PID, Neuro, Infection, and Africa blacklists) into a single "combined blacklist"; or (*ii*) four combinations from the set of blacklists (Neuro, Infection, Africa) into four combined blacklists (i.e., Neuro–Africa, Neuro–Infection, Africa–Infection, Neuro–Africa–Infection), and applying the combined blacklists obtained in (*i*) and (*ii*) to the PID cohort. As the PID blacklist was not included in the four combined blacklists in (*ii*), we refer to these blacklists as "non–cohort-specific combined blacklists." These blacklists removed a decreasing number of variants with increasing size of the sets making up the blacklists (Fig. 5). After standard filtering with public databases, the "Neuro–Africa" non–cohort-specific blacklist removed a median of 1,102 (8%) variants, the "Neuro–Infection" non-cohort-specific blacklist removed a median of 3,833 (26%) variants, the "Africa–Infection" non–cohort-specific blacklist removed a median of 3,886 (27%) of variants, and the "Neuro–Africa–Infection" non–cohort-specific blacklist removed a slightly larger number of variants (median of 4,078, or 28% of variants). By contrast, the PID blacklist removed a median of 9,056 variants. The "four combined" blacklist removed a median of 25 (0.45%) additional variants not captured by the PID blacklist alone (Fig. 5). Overall, these findings suggest that the number of variants filtered by the blacklist approach converges with the inclusion of blacklists from additional cohorts, consistent with the results for blacklist saturation. This universal filtering by blacklisting can be effectively applied to other individuals/cohorts. It is most efficient when the sequencing technology used, and the genetic ancestries of the individuals/cohorts under analysis, are similar to the universal blacklist (*SI Appendix*, Fig. S19). Moreover, the efficiency of a cohort-specific cohort applied to a different cohort (e.g., PID and infection cohorts) was greater for cohorts similar in terms of ethnic background and sequencing procedure (both mostly European and capture with similar kits), consistent with the results in Fig. 4*C*. Finally, although cohort-specific blacklists maximize the efficiency of this approach, the use of non–cohort-specific combined blacklists is nevertheless a very useful approach for filtering out a large number of unwanted variants, reinforcing the power of blacklist filtering even in the absence of a custom blacklist for the cohort.

Discussion

An essential step in the analysis of exomes from patients with rare genetic disorders is the removal of NPVs common in public databases (such as gnomAD, Bravo, and TopMed) at frequencies inconsistent with the prevalence, mode of inheritance, and penetrance of the disease (11). In principle, variants found to be

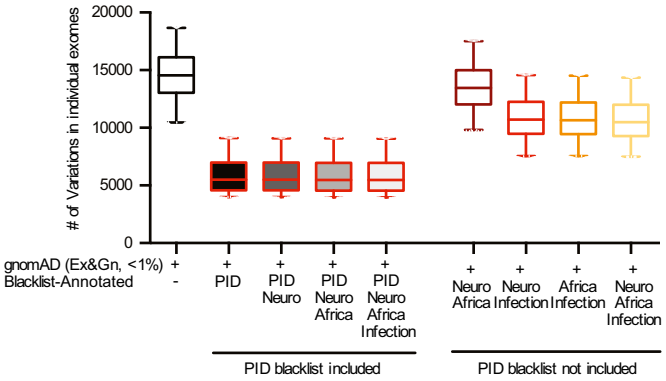


Fig. 5. Efficiency of various combinations of the four blacklists. Filtering of all variants in each PID exome with combinations of the various blacklists, with and without inclusion of the PID blacklist. Error bars represent the 10th to 90th percentiles.

common in a private cohort but absent from public databases should also be filtered out. However, only one other previous study has explored the generation of filtering lists based on internal cohorts (13). Moreover, there are currently no tools available for filtering based on allele frequencies in internal cohorts. We report here the identification of in-house variants too common to cause rare monogenic illnesses (typically with a population prevalence of $<10^{-4}$) in a cohort of 3,104 exomes. We assembled these variants into a blacklist and subsequently explored the use of this blacklist for filtering NPVs from exome sequencing data, using the subset of variants that makes our approach unique (BL-A: those that are absent from public databases). These variants had high-quality metrics and 75% of them would not be captured by the rigorous application of available software, such as VQSR. We further validated this approach in three other independently processed and unrelated cohorts, demonstrating that our blacklist approach is generally, and perhaps universally, effective for filtering variants, and that the generation of blacklists specific to a given cohort significantly increases the number of variants filtered out. We provide a computational tool (ReFiNE) for automatically generating in-house cohort-specific blacklists. We show that our blacklist can be used in synergy with standard public database filtering, to remove variants displaying disproportionate enrichment in an internal cohort.

Public databases such as gnomAD, which represent major population groups (about half of individuals are of European ancestry and the others are a mixture of Admixed Americans, Africans/African Americans, South Asians, East Asians, and Others), are an invaluable resource for estimating the frequency of variants in the general population and in different genetic ancestry groups. However, cohort-specific exomes may contain common variants (e.g., >1%) that are absent from or rare in public databases, partly because they are population-specific variants less represented in gnomAD [as observed for African (31) and Middle Eastern individuals (27)]. Moreover, public databases, such as gnomAD, make considerable efforts to ensure the rigorous removal of FP variants to ensure that they provide high-quality, high-stringency information about variants. However, these public databases do not provide a list of filtered FP variants and their summary statistics for filtration purposes. We demonstrated this with 113 1KG genomes generated by our in-house pipeline, showing that 23% of the variants were absent from the public 1KG database, highlighting discrepancies between the analyzed and released data due to different bioinformatic procedures. Moreover, resources such as dbSNP are difficult to use for FP filtering because their FP variant rate is high (32). Therefore, even when using the latest versions of public databases and gene-level filtration (6, 7), ReFiNE is an effective tool for collecting data independently from external resources.

The technology associated with the NGS analyses (sequencing platform, targeting procedures, and software) is strongly associated with the calling of the variants. We and others have previously observed biases specific for WES and WGS (18) or variant-calling pipelines (33). Differences in technology can therefore lead to the misannotation of variants in a given cohort. The main sources of misannotation are as follows: (*i*) variants in gnomAD collected by different technologies (PCR for WES and PCR-free plus PCR for WGS) apply rigorous QC cutoffs based on high-quality technologies, resulting in higher proportions of variants from lower-quality technologies being removed; (*ii*) despite the presence of 15,496 genomes in gnomAD, some genomic regions remain poorly covered or not covered at all, whereas these regions are covered by our cohort and contain variants (2% of our BL-A); (*iii*) a recent comparative studied revealed strong discrepancies between the variant callers used in NGS analyses (34); these discrepancies have been highlighted by the differences between the gnomAD and ExAC databases (https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/); and (*iv*) the annotation of NGS variants in multiallelic positions is often problematic (35) because current annotation software [SNPeff (36), VEP (37), ANNOVAR (38)] cannot identify these variants efficiently. Indeed, 91.5% of our blacklist variants were located at multiallelic sites according to gnomAD's genome annotation. Each cohort is unique (in terms of technology, quality, ethnicities). Our blacklisting resource is intended to fill this gap, particularly for researchers without the large exome or genome databases required for filtering with computationally intensive methods, such as VQSR. ReFiNE can, thus, overcome anomalies in sequence alignment or variant-calling processes, such as large indel events (39).

We show here that analyses of variant frequency within internal cohorts constitute an additional method for filtering out variants too common to cause rare disease. The blacklists generated by ReFiNE are easy to use and rapidly identify NPVs that may confound the dissection of patient exomes. As WES and WGS are increasingly widely used for the investigation of genetic disorders in patients, it will be possible to apply the blacklisting approach described here and ReFiNE software to larger cohorts of patients, facilitating the effective identification of NPVs in these cohorts. However, caution is required when generating blacklists with ReFiNE from phenotypically homogeneous cohorts, particularly if of the same underrepresented ethnic origin, as this approach may remove TP variants in such conditions. Finally, such extensive, rapidly generated blacklists (1 h for 3,104 exomes) should increase the efficiency of NPV elimination from exomes and genomes, without the need for the large computer clusters required by current machine-learning algorithms, such as VQSR (a month for 3,104 exomes). As exome capture kits become increasingly efficient, and with the widespread adoption of WGS, the blacklists generated by ReFiNE will facilitate efficient noise reduction in NGS data, independently of the technology used, making it easy to find the needles in increasingly large haystacks of genetic variants in patients.

Materials and Methods

Website Resource. ReFiNE and precalculated blacklists are available on GitLab (40).

Patient Cohort. The 3,104 individuals studied here were selected from samples of diverse ancestral origins obtained by our laboratories and recruited with the help of clinicians. This sample was not random, but cohort-specific effects should not have biased the results, as the individuals included had a wide range of different infectious diseases and immune deficiency phenotypes. All study participants provided written informed consent for the use of their

DNA in studies aiming to identify genetic risk variants for disease. IRB approval was obtained from The Rockefeller University and Necker Hospital for Sick Children, along with a number of collaborating institutions. The exomes of 3,869 individuals suffering from neurological disease were obtained from the Greater Middle East (GME) Consortium, with recruitment according to a similar protocol (27). The exomes of 902 individuals suffering from severe infectious diseases (Infection cohort) were obtained from patients enrolled in studies coordinated by the laboratory of J.F. at École Polytechnique Fédérale de Lausanne (Lausanne, Switzerland). The exomes of 400 individuals in the Africa cohort were provided by the laboratory of L.Q.-M. at the Pasteur Institute (Paris, France).

WES. A summary of the technologies and pipelines used for the analysis of the different cohorts is provided in *SI Appendix*, Table S9.

Rockefeller PID exome sequences. Genomic DNA from peripheral blood mononuclear cells was extracted and sheared with a Covaris S2 Ultrasonicator. An adaptor-ligated library (Illumina) was generated, and exome capture was performed with SureSelect Human All Exon 37-, 50-, or 71-Mb kits (Agilent Technologies). Massively parallel WES was performed on a HiSeq 2000 or 2500 machine (Illumina), generating 72-, 100-, or 125-base reads. Quality controls were applied at the lane and fastq levels. Specifically, the cutoff used for a successful lane is Pass Filter > 90%, with over 250 M reads for the high-output mode. The fraction of reads in each lane assigned to each sample (no set value) and the fraction of bases with a quality score > Q30 for read 1 and read 2 (above 80% expected for each) were also checked. In addition, the FASTQC tool kit (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to review base quality distribution, representation of the four nucleotides of particular *k*-mer sequences (adaptor contamination). We used the Genome Analysis Software Kit (GATK) (version 3.4–46) best-practice pipeline to analyze our WES data (14). Reads were aligned with the human reference genome (hg19), using the maximum exact matches algorithm in Burrows–Wheeler Aligner (BWA) (41). PCR duplicates were removed with Picard tools (picard.sourceforge.net/). The GATK base quality score recalibrator was applied to correct sequencing artifacts. GATK HaplotypeCaller was used to identify variant calls. DP ≥ 5 and MQ ≥ 30 were used as standard hard filtering criteria (22). Variants were annotated with SnpEff (snpeff.sourceforge.net/). Exomes were annotated for PASS and non-PASS variants in gnomAD r2.0.2 (Exome Aggregation Consortium, Broad Institute) and the 1000 Genomes Project Phase 3 (www.internationalgenome.org/) databases. Joint genotyping followed by VQSR filtering was not used because there have been reports of fractions of variants unique to individual samples being missed (https://gatkforums.broadinstitute.org/gatk/discussion/4150/should-i-analyze-my-samples-alone-or-together), rendering this approach unsuitable for our studies. For the purpose of comparison between the blacklist and VQSR approaches, VQSR was calculated with VariantRecalibrator and ApplyRecalibration for both SNPs and indels, with ts_filter_level set to 99.0 and other settings as specified by GATK recommendations. We did not use the InbreedingCoeff as this is discouraged in situations in which the cohort includes members of the same family, as in our cohort. Similarly, we did not include DP among the parameters of the VQSR, as it is not suitable for targeted exome sequencing samples.

GME Consortium neurological exome sequences. WES for the GME Consortium was performed as previously described (27). Briefly, genomic DNA was extracted from peripheral blood mononuclear cells with Qiagen reagents and captured with the Agilent SureSelect Human All Exome 50-Mb kit. WES was performed on an Illumina HiSeq 2000. The GATK best-practice pipelines were used to analyze WES data (14). BWA was used to align reads with human reference genome NCBI Build 37 (41). The variant-call format files generated were annotated with the Rockefeller pipeline, as described above.

Africa exome sequences. Whole-exome sequences were obtained for 300 African samples (31), and these data were processed together with those for 100 European individuals (42). All samples were sequenced with the Nextera Rapid Capture Expanded Exome kit, which delivers 62 Mb of genomic content per individual, including exons, UTRs, and microRNAs. Using the GATK Best Practice recommendations (43), we first mapped read-pairs onto the human reference genome (GRCh37) with BWA, version 0.7.7 (41), and reads duplicating the start position of another read were marked as duplicates with Picard Tools, version 1.94 (picard.sourceforge.net/). GATK, version 3.5 (14), was then used for base quality score recalibration ("BaseRecalibrator"), insertion/deletion (indel) realignment ("IndelRealigner"), and SNP and indel discovery for each sample ("Haplotype Caller").

Infection exome sequences. WES for the Infection cohort was performed as previously described (44, 45). In brief, genomic DNA was extracted from whole blood with the QIAamp DNA blood kit and captured with the Agilent SureSelect Human All Exome 50-Mb kit (Agilent SureSelect Human all exon V4 or V5) or Illumina Truseq 65-Mb enrichment kit. WES was performed on an Illumina HiSeq 2000 or Illumina HiSeq 2500 machine. BWA-MEM was used to map reads onto the human reference genome hg19 decoy, and GATK, version 3.8 (or an earlier version of this software), was used for data processing and analysis, according to GATK best practice.

Blacklist Creation. The blacklists used in and provided with this manuscript were created by first collecting unique variants from 3,104 patient exomes and counting the occurrence of each variant (the number of patients reported to have the variant). The QC criteria used to collect these variants were equivalent to those used in gnomAD (MQ ≥ 30). However, we used a lower DP (DP ≥ 5), compatible with research approaches in which investigators want to retain as much information as possible. These criteria correspond to a high degree of QC despite low coverage, but may allow the discovery of true disease-causing variants, as illustrated by the example of the deletion of *ISG15*, which was initially identified by exome analysis despite a low DP of 4 (46). We did not use the QD value as a QC criterion due to the erroneous calls for some variants (https://gatkforums.broadinstitute.org/gatk/discussion/8912/most-variants-called). We explored the FN rate of the blacklists in the HGMD database and excluded variants that were present in the set of true disease-causing variants in HGMD according to further analyses (47). The measurement of variation at multiallelic sites was rendered more effective by separating variants into biallelic and multiallelic variant groups. Multiallelic variants represent a very specific challenge for the elimination of NPVs from exomes, as variants at multiallelic positions may occur individually in a small number of samples. Collectively, however, these variants may occur in a large proportion of the members of the cohort (i.e., many individuals may contain one of a number of variants at the position). The variants at multiallelic sites are often similar (e.g., G in the reference and an alternative of GA, GAA, GAAA, GAAAA, GAAAAA, etc.) but have remained resistant to removal from exomes by bioinformatic methods. For the capture of these variants, we collapsed all variants at multiallelic sites to a single value by calculating the total number of patients with any variant at the multiallelic position. When this number exceeded 1% of our cohort, all variants at the position concerned were included in the full blacklist. This procedure can thus identify variants present in only a few individuals but nevertheless occurring at positions with a high cumulative burden of variation in a cohort. We then considered biallelic variants. If the number of patients with any individual biallelic variant exceeded 1% of our cohort, the variant concerned was included in the full blacklist. For a schematic diagram of this pipeline, see *SI Appendix*, Fig. S1.

ReFiNE Generation and Usage. ReFiNE and subsequent analyses were performed in Python programming language (version 2.7.14; https://www.python.org/) and R, using both default and publicly available libraries. The Python Tkinter module was used to design and implement the graphical interface for ReFiNE. ReFiNE is available as a graphical interface program (including a command-line option) that can be run on a standard laptop and is compatible with comma-separated (CSV) files. ReFiNE can also generate blacklists from WGS data, although this application has yet to be extensively tested. ReFiNE includes an optional parameter for the exclusion of a list of variants from the blacklist regardless of their frequency in the in-house database. This option can be used to remove a small number of known true disease-causing HGMD variants, for example. We also provide precalculated blacklists generated from our cohort of 3,104 PID exomes with cutoffs of 1%, 3%, 5%, and 10%. These blacklists can be used for small cohorts for which it may not be possible to generate custom blacklists. We also provide the PID, Neuro, Infection, Africa, and combined blacklists used in this manuscript, annotated with gnomAD MAFs. Finally, we have constructed a public server (lab.rockefeller.edu/casanova/BL) containing all of the supplemental files, the ReFiNE program, and a user-friendly online tool that can be used to query whether a variant is included in our blacklist or to annotate lists of variants in a similar manner.

Statistics and Figures. The Scipy library (https://www.scipy.org/) was used for statistical analyses performed in Python. Seaborn (seaborn.pydata.org/) was used to generate figures in Python, together with matplotlib (https://matplotlib.org). Venn diagrams were generated with jvenn software (48). Wordclouds were generated with the WordCloud library (https://github.com/amueller/word_cloud). Prism (GraphPad) was also used for figure generation and statistical analysis.

Simulating minimum sample size and sample size saturation for blacklists. We determined the minimum number of samples required for the creation of safe blacklists by generating random blacklists based on 10, 50, 100, 250, 500, 1,000, 1,500, 2,000, 2,500, 3,000, 3,500, 4,000, 4,500, 5,000, 5,500, 6,000, or

6,500 individuals from the PID and Neuro cohorts. We weighted the random selection of individuals for the blacklists by project size (i.e., for a sample size of 10, we picked 4 individuals at random from the PID cohort and 6 at random from the Neuro cohort). The selection of individuals for each sample size was repeated 30 times, and full blacklists for each iteration were generated with ReFiNE. The median number of BL-A variants and a 99% CI based on a normal distribution were calculated for each sample size and plotted (*SI Appendix*, Fig. S18). The number of samples required to reach saturation for blacklist variants was predicted by fitting a logarithmic trendline to the blacklist dataset based on the coefficient of determination (R^2). The equation for this line was as follows:

$$y = 2{,}801.1 \times \ln(x) + 3{,}466.3,$$

where $R^2 = 0.7088$ (*SI Appendix*, Fig. S18). We defined saturation as the number of samples for which less than one cohort-specific variant was added to the blacklist per new exome. Based on the best-fit equation, we calculated the saturation point as 2,801 individuals.

Characterization of blacklisted variants by HW equilibrium/disequilibrium, occurrence in low-complexity regions, and allelic distribution across genetic ancestries. HW disequilibrium was calculated for the blacklisted variants found to be present in the European population ($n = 1{,}150$), which constituted the largest population of the PID cohort. χ^2 tests were used to assess HW equilibrium. Given the large number of tests performed and the heterogeneity of European origins in our European cohort, a stringent threshold of 10^{-8} for significance was used for significance. A total of 106 variants with a P value below 10^{-8} were considered to be in HW disequilibrium and were stratified by excess genotype as follows: excess of heterozygotes (observed no. of heterozygotes > expected no. of heterozygotes, 57 variants), excess wild-type homozygotes (observed no. of wild-type homozygotes > expected no. of wild-type homozygotes, and χ^2 for the wild-type homozygote > χ^2 for the alternative homozygote, 13 variants), excess alternative homozygotes (observed no. of alternative homozygotes > expected no. of alternative homozygotes, and χ^2 for alternative homozygotes > χ^2 for wild-type homozygotes, 36 variants).

The occurrence of the variants in low-complexity regions was assessed with the following tracks from the UCSC Genome Browser: RepeatMasker and Simple Repeats (group: Repeats), and GC percent (group: Mapping and Sequencing). RepeatMasker was created from the RepeatMasker program, which screens DNA sequences for interspersed repeats and low-complexity DNA sequences; Simple Repeats reports simple tandem repeats located by Tandem Repeats Finder (TRF), which was designed especially for this purpose. Variants were considered to occur in GC-rich regions in which the G+C content exceeded 80%.

The heterogeneity of ethnicity was assessed in the four largest genetic ancestry groups in our cohort (European, African, North African, and Middle Eastern), for the variants found to be in HW equilibrium in the European population. χ^2 tests were used to test the allelic distribution. In total, 203 variants with a P value below 10^{-8} were considered to be heterogeneous across ancestries. The ancestry driving heterogeneity was unequivocally determined for 67 variants, by testing the allelic distributions of four combinations of three populations from those mentioned above and determining the data for the missing population in the combination from the four that did not reach significance.

Sanger sequencing. DNA was extracted from 10 SV40-fibroblast cell lines from patients included in our cohort. PCR amplification was performed with Hot-Start Taq Blue DNA Polymerase (Denville Scientific), 85 ng of template genomic DNA, and the primers listed in *SI Appendix*, Table S10. Sanger sequencing was performed with the BigDye Terminator kit (Perkin-Elmer).

Analysis of variation in patient exomes. We identified the disease-causing mutation in patient D2 from a previous study (25), using a standard filtration pipeline. In brief, we removed variants with low-quality metrics (DP < 4, MQ < 40, QD < 2) that were common in public databases (variant frequency in gnomAD < 0.0001), variants of high-GDI genes (6), and variants with CADD scores below their gene-specific mutation significance cutoff (9). Gene burden was analyzed in our CMC cohort by first filtering each exome, as described above. We then compared the numbers of individuals with at least one variant for each mutated gene in the patient group between the patient ($n = 208$) and control ($n = 960$) groups in a one-tailed Fisher's exact test. The resulting P values were used to rank genes, to identify those with the highest levels of enrichment in patients.

ACKNOWLEDGMENTS. D.N.C. and P.D.S. gratefully acknowledge financial support from Qiagen, Inc., through a license agreement with Cardiff University. This work was supported by the National Institutes of Health (Grants P01AI061093, U24AI086037, R18AI048693, T32GM007280, R01AI088364, R01AI095983, and R01AI127564), the French National Research Agency (ANR 14-CE15-0009-01), the Jeffrey Modell Foundation, and the David S. Gottesman Immunology Chair and the Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai.

1. Casanova JL, Conley ME, Seligman SJ, Abel L, Notarangelo LD (2014) Guidelines for genetic studies in single patients: Lessons from primary immunodeficiencies. *J Exp Med* 211:2137–2149.
2. Meyts I, et al. (2016) Exome and genome sequencing for inborn errors of immunity. *J Allergy Clin Immunol* 138:957–969.
3. Goldstein DB, et al. (2013) Sequencing studies in human genetics: Design and interpretation. *Nat Rev Genet* 14:460–470.
4. Stenson PD, et al. (2017) The Human Gene Mutation Database: Towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 136:665–677.
5. Lek M, et al.; Exome Aggregation Consortium (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291.
6. Itan Y, et al. (2015) The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci USA* 112:13615–13620.
7. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9: e1003709.
8. Kircher M, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46:310–315.
9. Itan Y, et al. (2016) The mutation significance cutoff: Gene-level thresholds for variant predictions. *Nat Methods* 13:109–110.
10. Itan Y, et al. (2013) The human gene connectome as a map of short cuts for morbid allele discovery. *Proc Natl Acad Sci USA* 110:5558–5563.
11. Bao R, et al. (2014) Review of current methods, applications, and data management for the bioinformatics analysis of whole exome sequencing. *Cancer Inform* 13:67–82.
12. MacArthur DG, et al. (2014) Guidelines for investigating causality of sequence variants in human disease. *Nature* 508:469–476.
13. Fuentes Fajardo KV, et al.; NISC Comparative Sequencing Program (2012) Detecting false-positive signals in exome sequencing. *Hum Mutat* 33:609–613.
14. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498.
15. Alcaïs A, et al. (2010) Life-threatening infectious diseases of childhood: Single-gene inborn errors of immunity? *Ann N Y Acad Sci* 1214:18–33.
16. Casanova JL (2015) Severe infectious diseases of childhood as monogenic inborn errors of immunity. *Proc Natl Acad Sci USA* 112:E7128–E7137.
17. Casanova JL (2015) Human genetic basis of interindividual variability in the course of infection. *Proc Natl Acad Sci USA* 112:E7118–E7127.
18. Belkadi A, et al. (2015) Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci USA* 112: 5473–5478.
19. Belkadi A, et al.; Exome/Array Consortium (2016) Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proc Natl Acad Sci USA* 113:6713–6718.
20. Jones E, Oliphant E, Peterson P (2001) SciPy: Open source scientific tools for Python, version 1.1.0. Available at https://www.scipy.org/. Accessed December 12, 2018.
21. Whiffin N, et al. (2017) Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med* 19:1151–1158.
22. Guo Y, Ye F, Sheng Q, Clark T, Samuels DC (2014) Three-stage quality control strategies for DNA re-sequencing data. *Brief Bioinform* 15:879–889.
23. Auton A, et al.; 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.
24. Buckley AR, et al. (2017) Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. *BMC Genomics* 18:458.
25. Kuehn HS, et al. (2016) Loss of B cells in patients with heterozygous mutations in IKAROS. *N Engl J Med* 374:1032–1043.
26. Toubiana J, et al.; International STAT1 Gain-of-Function Study Group (2016) Heterozygous STAT1 gain-of-function mutations underlie an unexpectedly broad clinical phenotype. *Blood* 127:3154–3164.
27. Scott EM, et al.; Greater Middle East Variome Consortium (2016) Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat Genet* 48:1071–1076.
28. Robinson JT, et al. (2011) Integrative genomics viewer. *Nat Biotechnol* 29:24–26.
29. Fazekas A, Steeves R, Newmaster S (2010) Improving sequencing quality from PCR products containing long mononucleotide repeats. *Biotechniques* 48:277–285.
30. Clarke LA, Rebelo CS, Gonçalves J, Boavida MG, Jordan P (2001) PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Mol Pathol* 54:351–353.
31. Lopez M, et al. (2018) The demographic history and mutational load of African hunter-gatherers and farmers. *Nat Ecol Evol* 2:721–730.
32. Mitchell AA, Zwick ME, Chakravarti A, Cutler DJ (2004) Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics* 20:1022–1032.
33. Hwang S, Kim E, Lee I, Marcotte EM (2015) Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci Rep* 5:17875.

34. Sandmann S, et al. (2017) Evaluating variant calling tools for non-matched next-generation sequencing data. *Sci Rep* 7:43169.
35. Campbell IM, et al. (2016) Multiallelic positions in the human genome: Challenges for genetic analyses. *Hum Mutat* 37:231–234.
36. Cingolani P, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; *iso-2; iso-3. Fly (Austin)* 6:80–92.
37. McLaren W, et al. (2016) The Ensembl variant effect predictor. *Genome Biol* 17:122.
38. Wang K, Li M, Hakonarson H (2010) ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164.
39. Ghoneim DH, Myers JR, Tuttle E, Paciorkowski AR (2014) Comparison of insertion/deletion calling algorithms on human next-generation sequencing data. *BMC Res Notes* 7:864.
40. Maffucci P, et al. (2018) Data from "Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis." GitLab. Available at https://gitlab.com/pmaffucci/refine. Deposited December 19, 2018.
41. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
42. Quach H, et al. (2016) Genetic adaptation and Neandertal admixture shaped the immune system of human populations. *Cell* 167:643–656.e17.
43. Van der Auwera GA, et al. (2013) From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43: 11.10.1–11.10.33.
44. Asgari S, et al. (2017) Severe viral respiratory infections in children with *IFIH1* loss-of-function mutations. *Proc Natl Acad Sci USA* 114:8342–8347.
45. Asgari S, et al.; Swiss Pediatric Sepsis Study (2016) Exome sequencing reveals primary immunodeficiencies in children with community-acquired *Pseudomonas aeruginosa* sepsis. *Front Immunol* 7:357.
46. Bogunovic D, et al. (2012) Mycobacterial disease and impaired IFN-γ immunity in humans with inherited ISG15 deficiency. *Science* 337:1684–1688.
47. Stenson PD, et al. (2003) Human Gene Mutation Database (HGMD): 2003 update. *Hum Mutat* 21:577–581.
48. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C (2014) jvenn: An interactive Venn diagram viewer. *BMC Bioinformatics* 15:293.

Supplementary Information for

Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis

Patrick Maffucci[#], Benedetta Bigio[#], Franck Rapaport, Aurélie Cobat, Alessandro Borghesi, Marie Lopez, Etienne Patin, Alexandre Bolze, Lei Shang, Matthieu Bendavid, Eric M Scott, Peter D Stenson, Charlotte Cunningham-Rundles, David N Cooper, Joseph G Gleeson, Jacques Fellay, Lluis Quintana-Murci, Jean-Laurent Casanova, Laurent Abel, Bertrand Boisson[‡], Yuval Itan[‡]

[#,‡]: Equal contributions

Corresponding authors:
Jean-Laurent Casanova – casanova@rockefeller.edu,
Yuval Itan – yuval.itan@mssm.edu

This PDF file includes:
Figs. S1 to S19
Tables S1 to S10

Figure S1. **Methodology for blacklist generation.** The blacklist was generated by first collecting unique high-quality variants (DP>=5, MQ>=30) from patient exomes and counting the occurrence of each variant. These variants were assembled into two classes: (1) biallelic, with a single alternative allele in our cohort; and (2) multiallelic, with two or more alternative alleles in the cohort, for which we collapsed all variants at a unique chromosomal position and summed the total number of patients containing these variants. We then collected the variants that had a frequency >=1% in the cohort (the Blacklist: "Common in-house variants"). Of these variants, 21.4% (167,144) were absent from gnomAD exome and genome databases. We considered these 167,144 variants to be "blacklist-annotated" (BL-A).

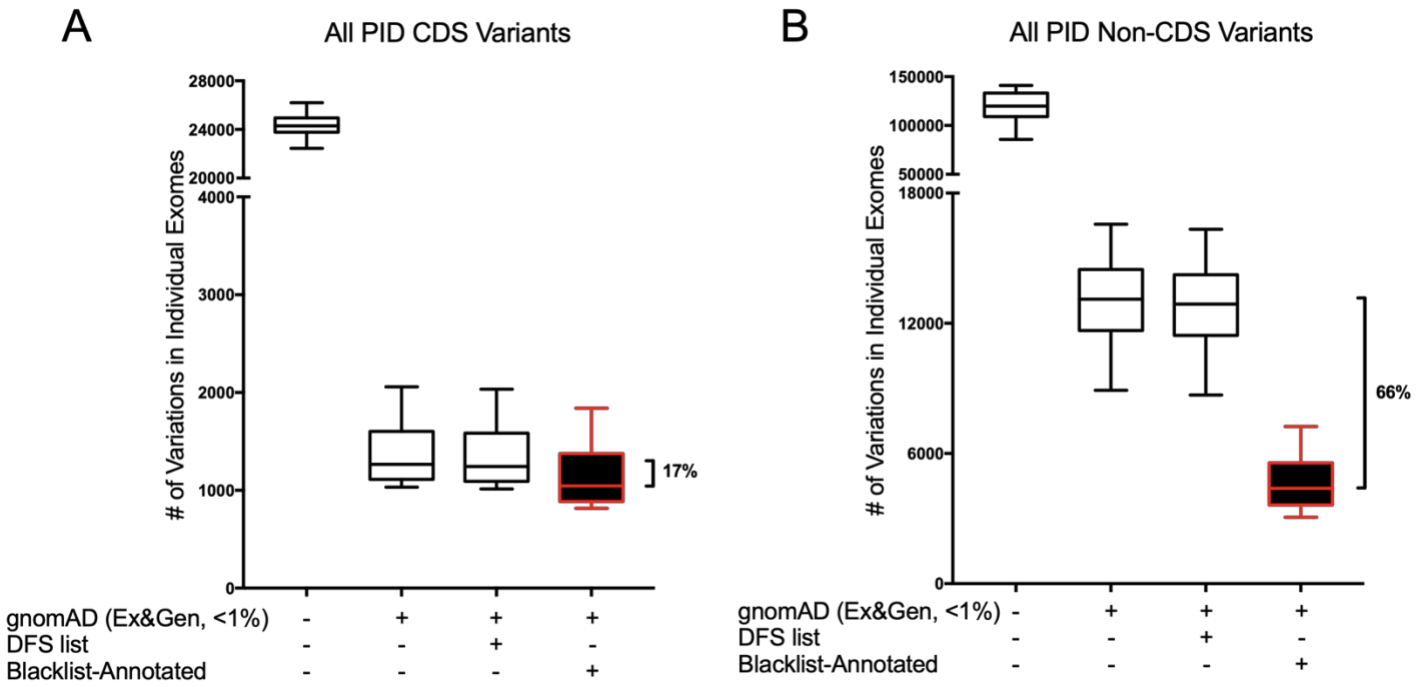


Figure S2. Filtering of coding sequence (CDS) or non-CDS variants in 3,104 PID exomes with the PID blacklist-annotated. Exomes were restricted to CDS (A) or non-CDS (B) variants and filtered by removing variants with a MAF greater than 0.01 in gnomAD. The remaining variants were filtered with the blacklist-annotated. Filtering with the DFS list is shown for comparison. Error bars represent the 10th-90th percentiles.

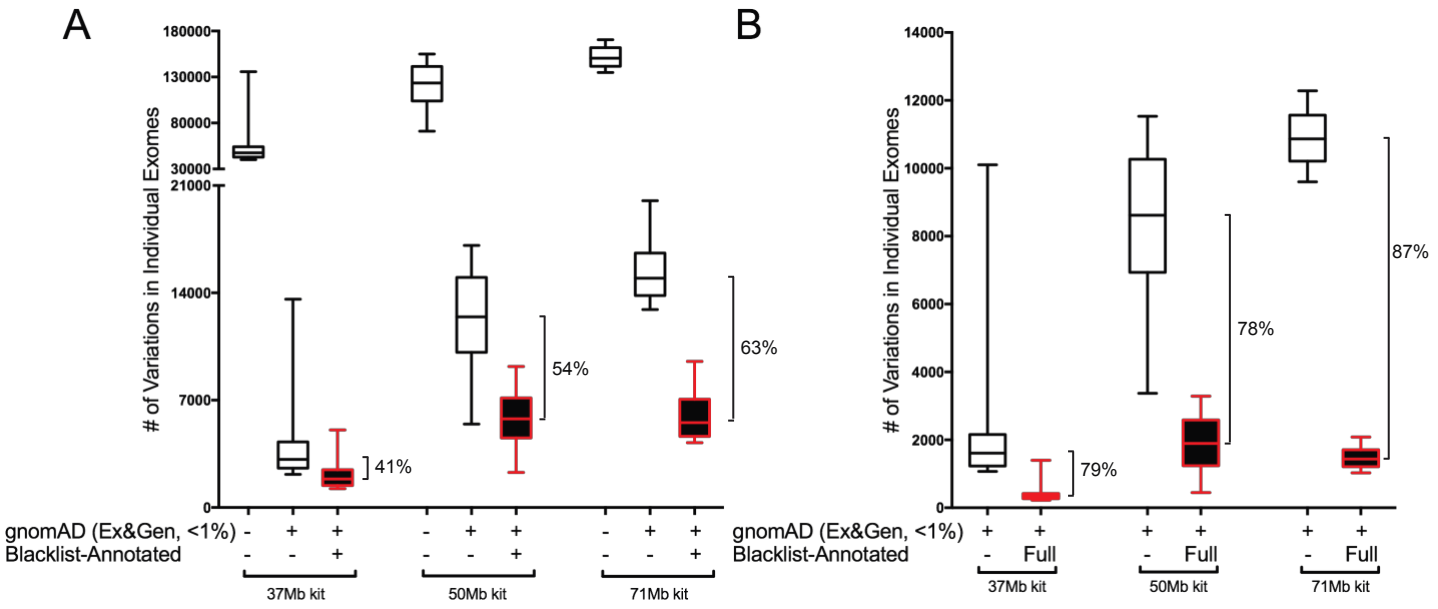


Figure S3. Filtering of 3,104 PID exomes broken down by the exome capture kit. PID exomes were captured with one of three SureSelect kits: 37 Mb (*n* = 96), 50 Mb (*n* = 727), or 71 Mb (*n* = 2,281). (A) Filtering of all variants in each exome, using gnomAD and the blacklist-annotated. gnomAD filtering performed by removing variants with a minor allele frequency greater than 0.01 in the databases. (B) Filtering of exomes restricted to cohort-specific variants with the blacklist-annotated. Error bars represent the 10th-90th percentiles.

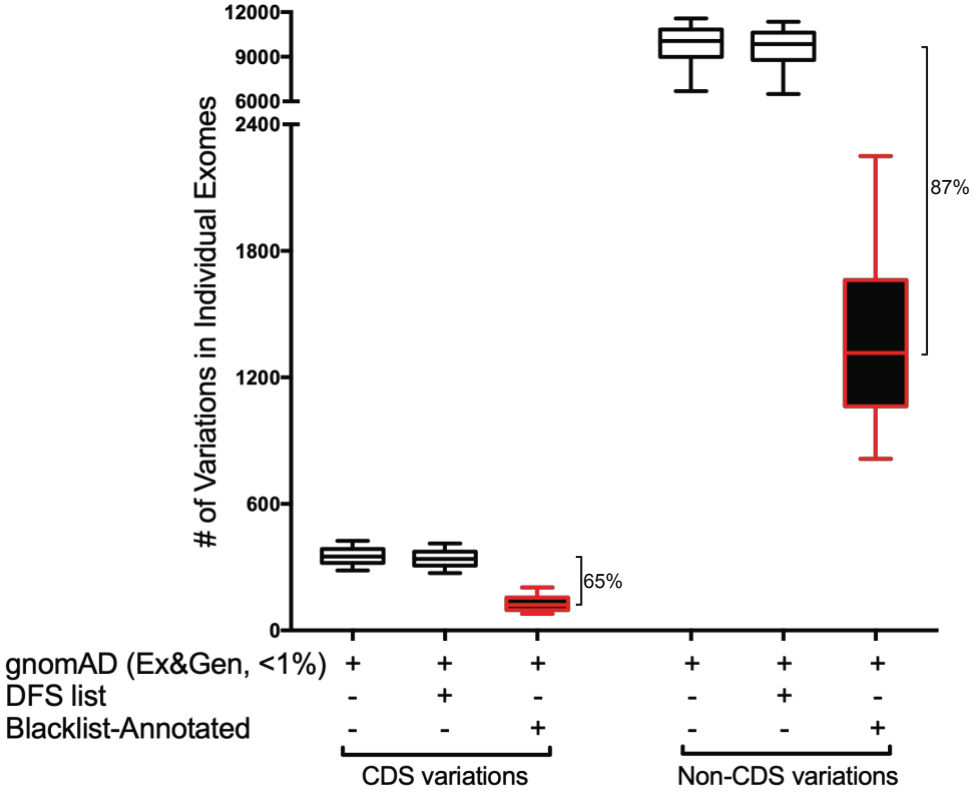


Figure S4. Filtering of coding sequence (CDS) and non-CDS variants in 3,104 PID exomes restricted to cohort-specific variations using the blacklist-annotated. DFS list shown for comparison. Error bars represent the 10th-90th percentiles.

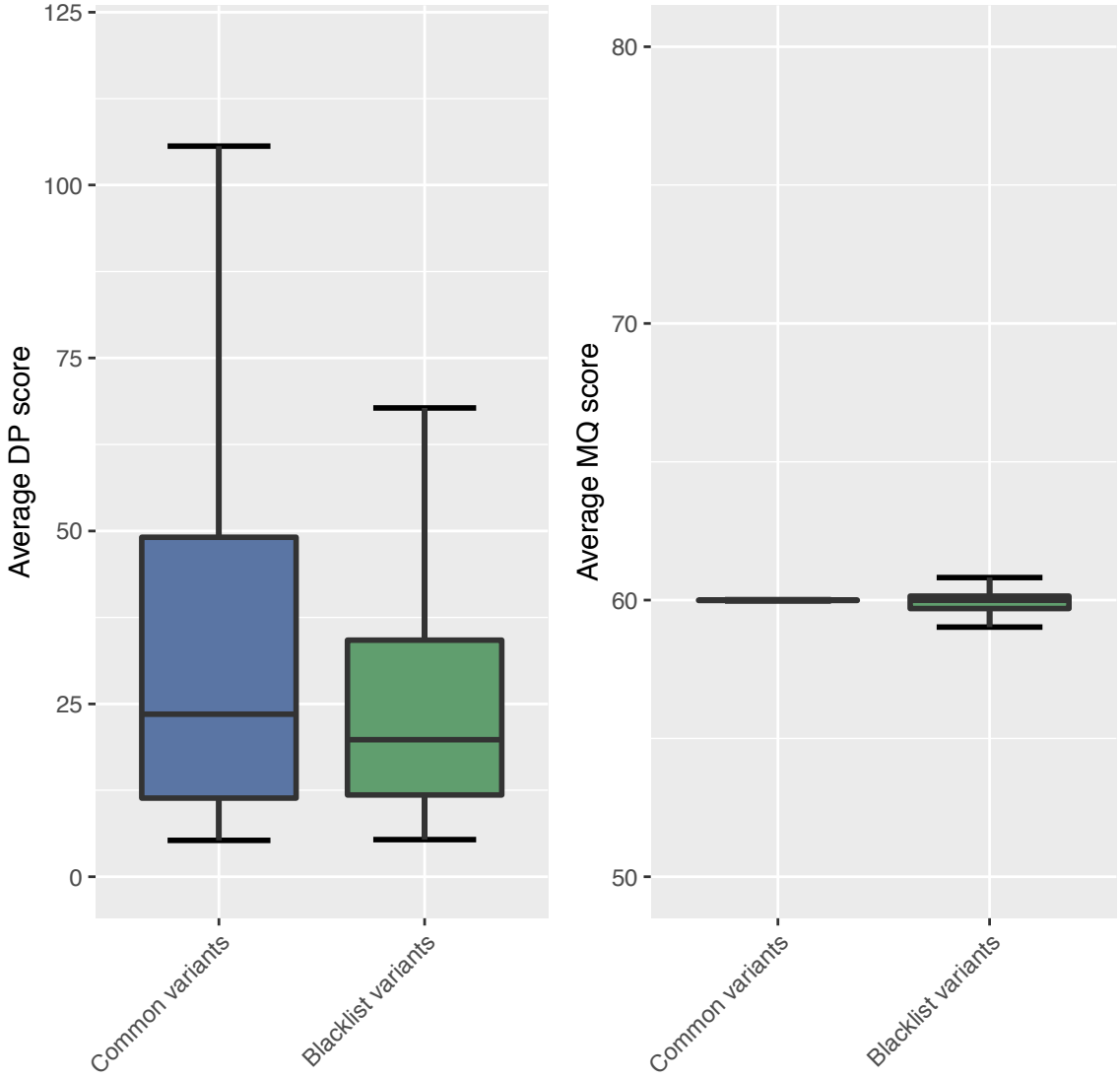


Figure S5. **Comparison of quality metrics for blacklisted and non-blacklisted variants.** Mean (A) read depth (DP) and (B) mapping quality (MQ) were calculated for common variants present in gnomAD with a MAF>1% (blue bar), and for blacklist-annotated variants (green bar). Error bars represent the upper and lower limits of 1.5 times the interquartile range.

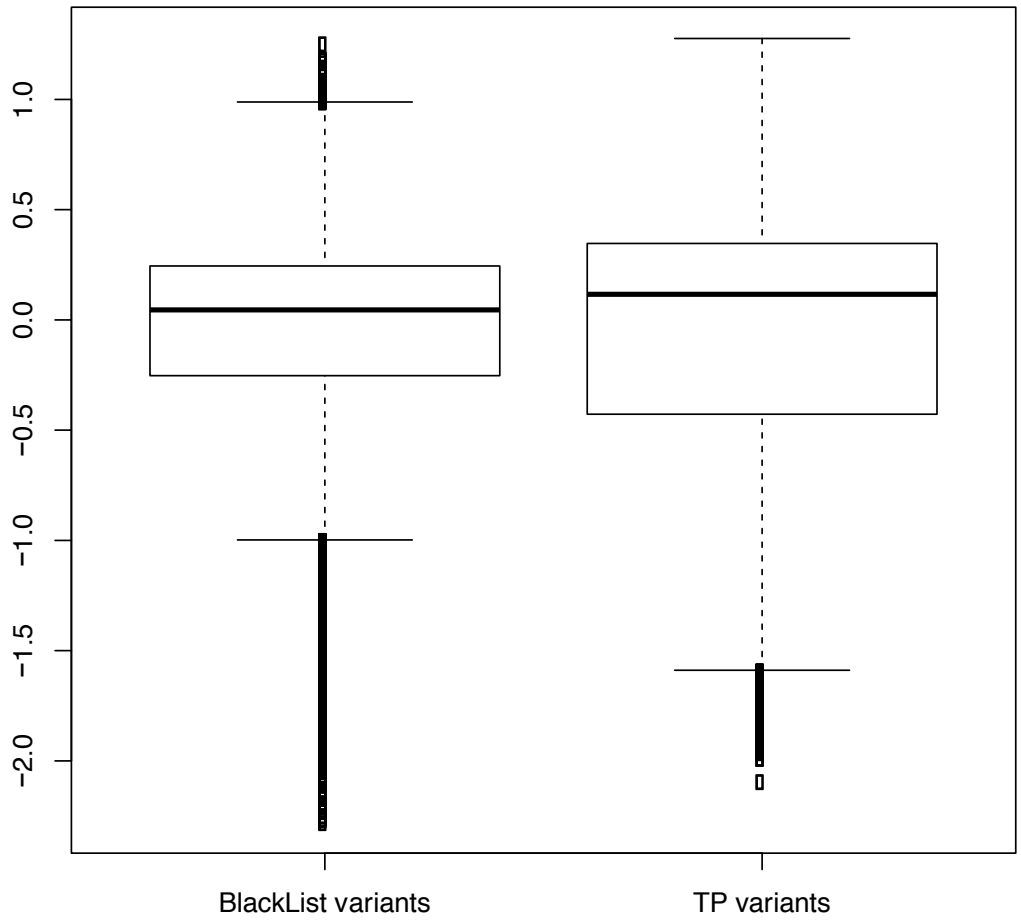


Figure S6. **Comparison with machine learning-based filtering methods.** We applied random forest scoring functions to blacklist-annotated variants and to a set of true-positive (TP) variants present in both the gnomAD dataset and our cohort with a MAF exceeding 1% in each dataset. The score distributions are almost identical, indicating that the blacklist-annotated variants are not distinguishable from TP variants according to this standard classification method.

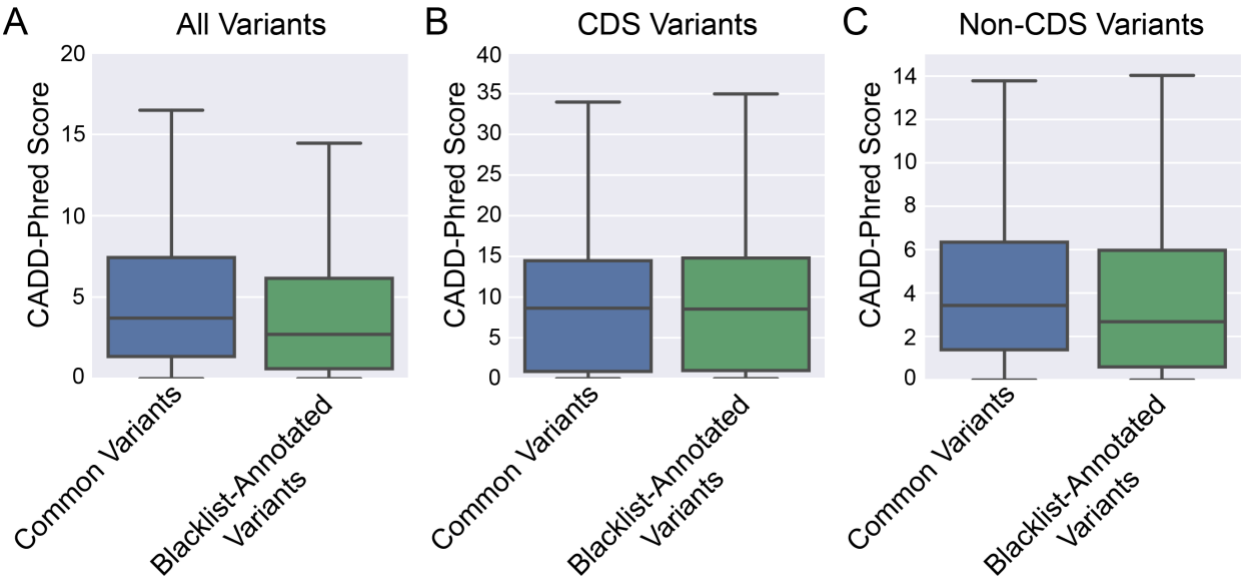


Figure S7. Comparison of CADD scores between blacklisted and non-blacklisted variants. Mean CADD scores were calculated for common variants present in gnomAD exome and genome databases with a MAF>1% (blue bar), or blacklist-annotated variants (green bar). Calculations were performed for all (A), CDS (B), and non-CDS (C) variants. Error bars represent the upper and lower limits of 1.5 times the interquartile range.

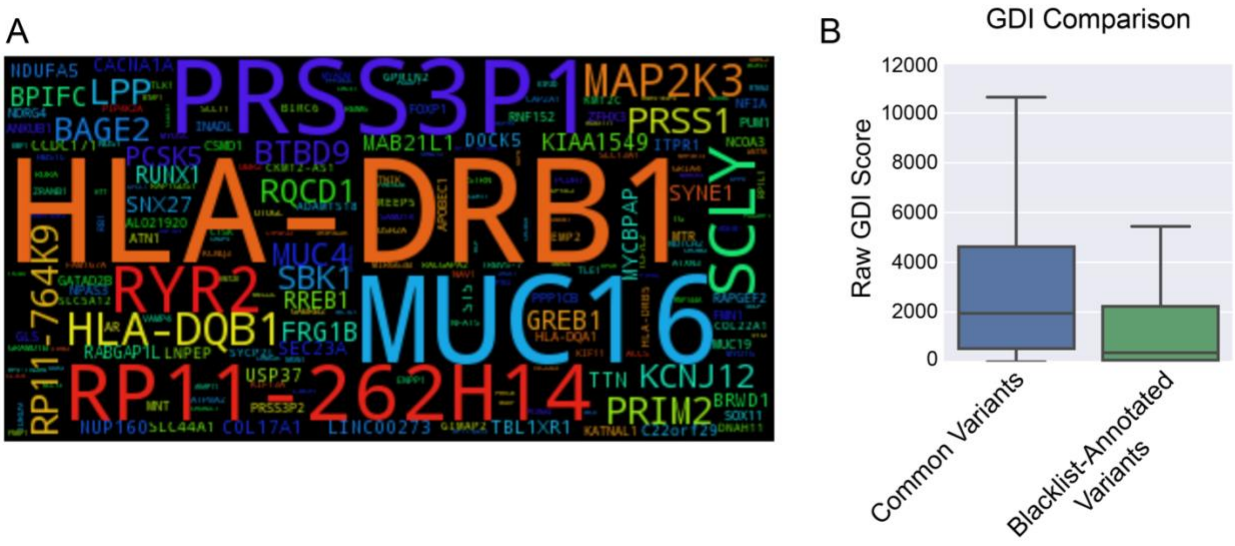


Figure S8. Characteristics of the most frequent genes in the blacklist-annotated. (A) Depiction of the top ranking genes in the blacklist-annotated according to the number of variants. The size of the text is proportional to the number of variants of the gene in the blacklist-annotated. (B) Comparison of GDI scores between the 1,000 most common genes in all the common in-house variants (gnomAD) and blacklist-annotated variants. Error bars represent the upper and lower limits of 1.5 times the interquartile range.

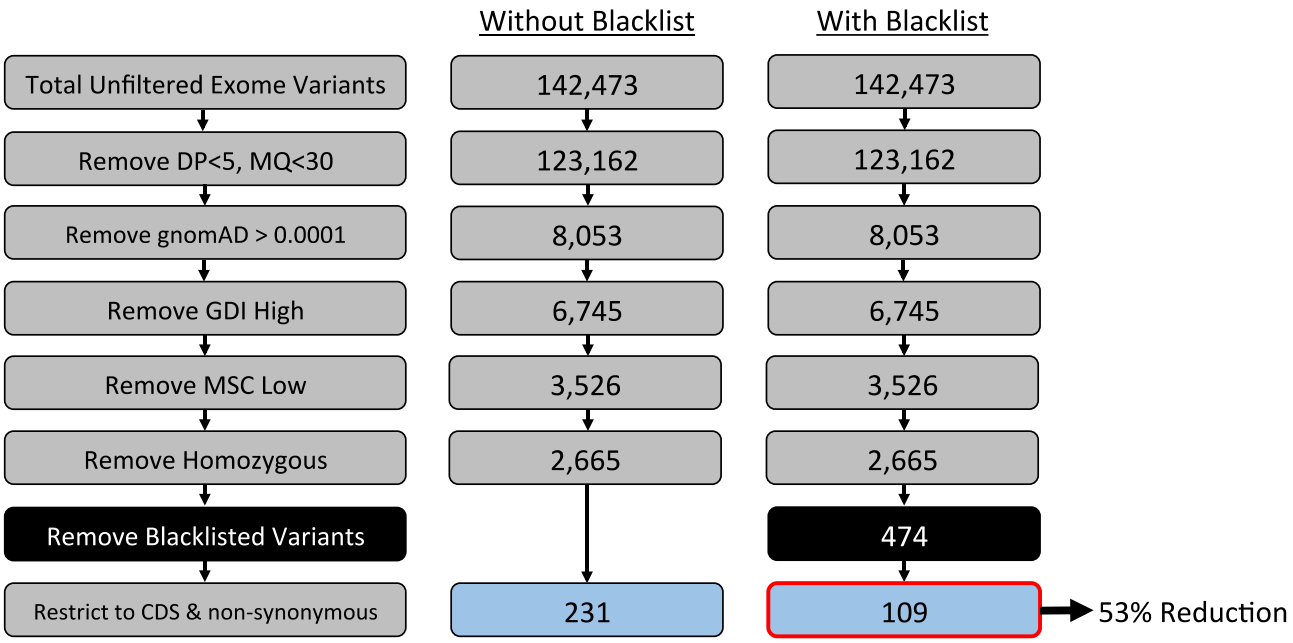


Figure S9. Practical analysis of a single patient exome by blacklisting. The practical utility of the blacklist approach was demonstrated with the exome of a patient with a published disease-causing mutation. The patient's exome was filtered with a standard pipeline with and without application of the blacklist-annotated. The numbers in each box represent the number of variants remaining in the exome after each filtering step. GDI: gene damage index; MSC: mutation significance cutoff.

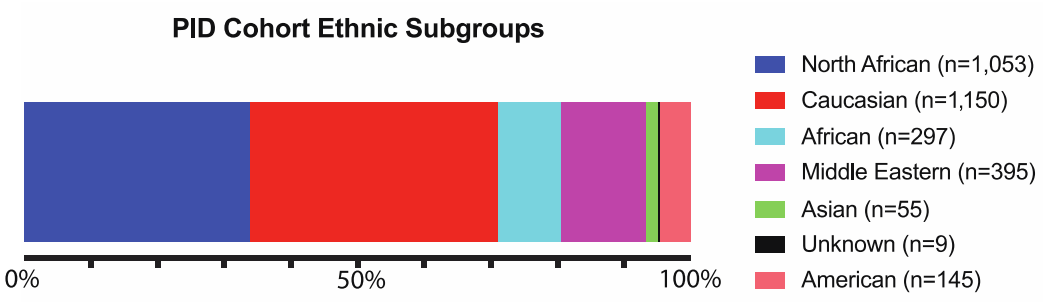


Figure S10. Representation of ethnic subgroups in 3,104 PID exomes. The distribution of the genetic ancestry groups in the PID cohort, as determined by PCA analysis.

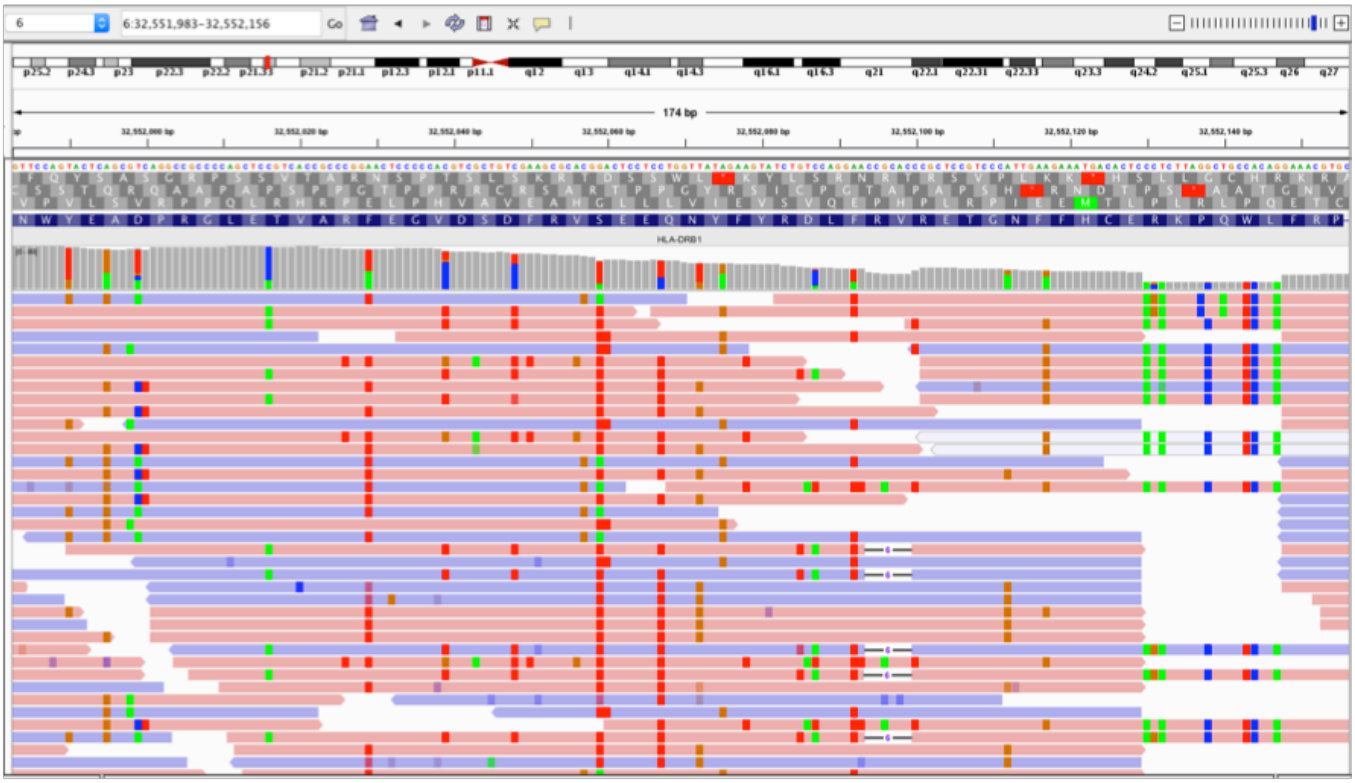


Figure S11. **Investigation of a biallelic *HLA-DRB1* variant: 6-32551960-T-TCC**
IGV screenshot of the WES alignment surrounding position 32,551,960 on chromosome 6.

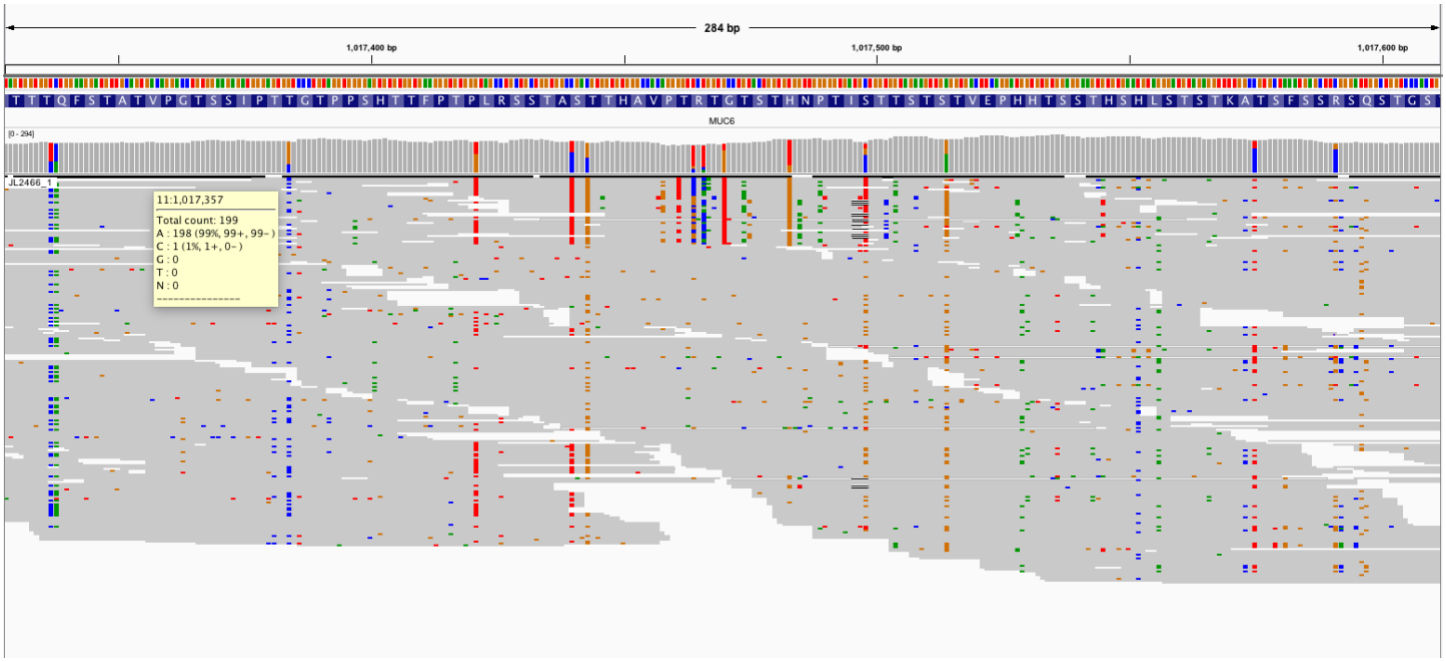


Figure S12. Investigation of a biallelic *MUC6* variant: 11-1017470-G-T
IGV screenshot of the WES alignment surrounding position 1,017,280 on chromosome 11.

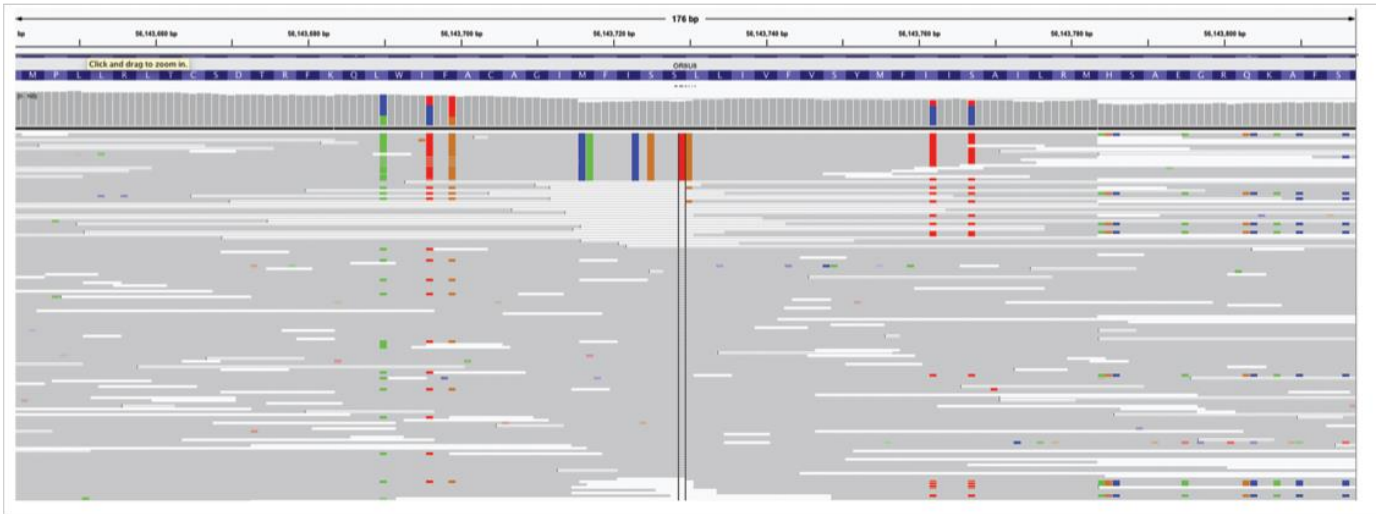


Figure S13. Investigation of biallelic *OR8U1* variants: 11,56143784,C,T and 11,56143803,A,G
IGV screenshot of the WES alignment surrounding position 11,56143784 on chromosome 11.

Figure S14. Investigation of a biallelic *HRNR* variant: 1-152195728-AT-A
IGV screenshot of the WES alignment at position 152195728 on chromosome 1.

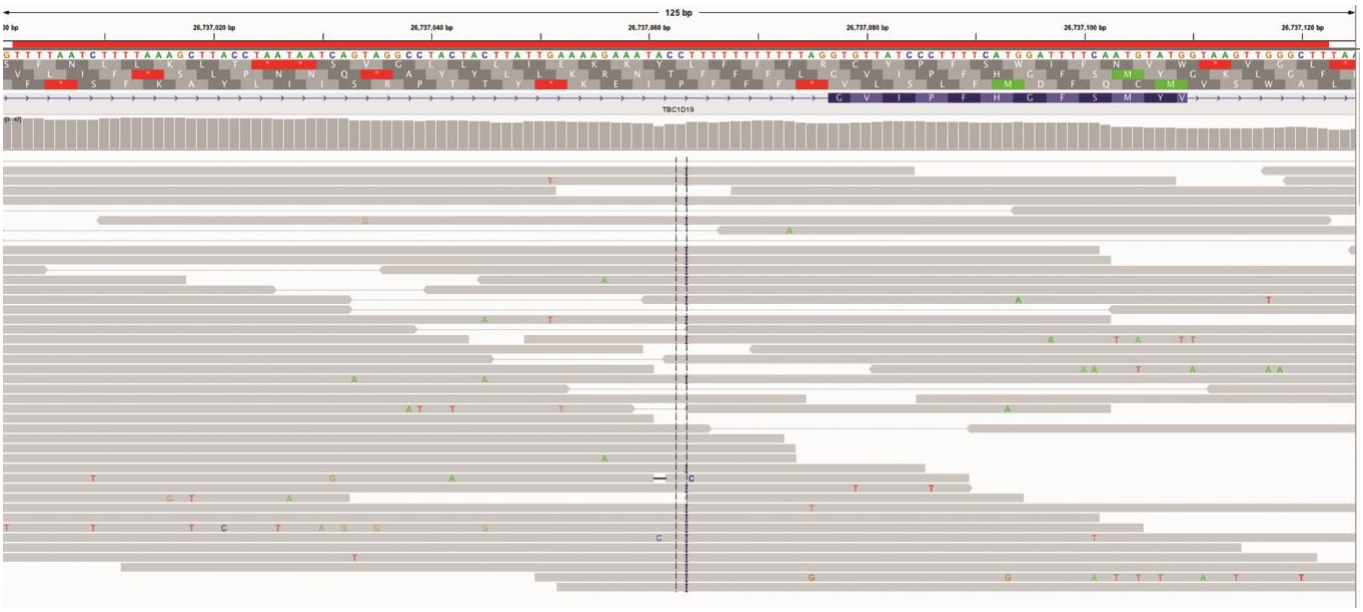


Figure S15. Investigation of a multiallelic *TBC1D19* variant: 4-26737063-C-CT
IGV screenshot of the WES alignment at position 26737063 on chromosome 4.

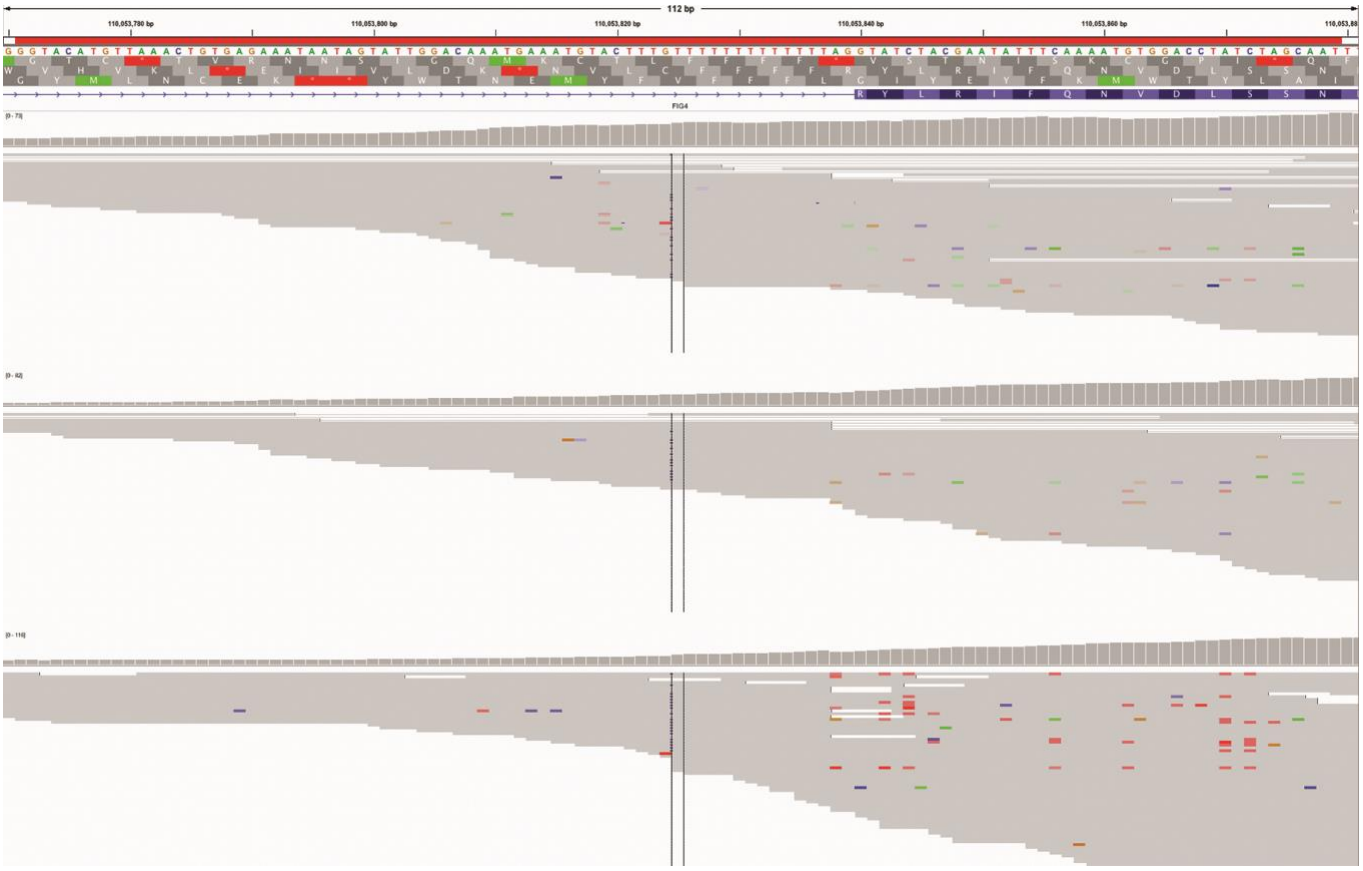


Figure S16. Investigation of a multiallelic *FIG4* variant: 6-110053824-G-GT
IGV screenshot of the WES alignment at position 110053824 on chromosome 6.

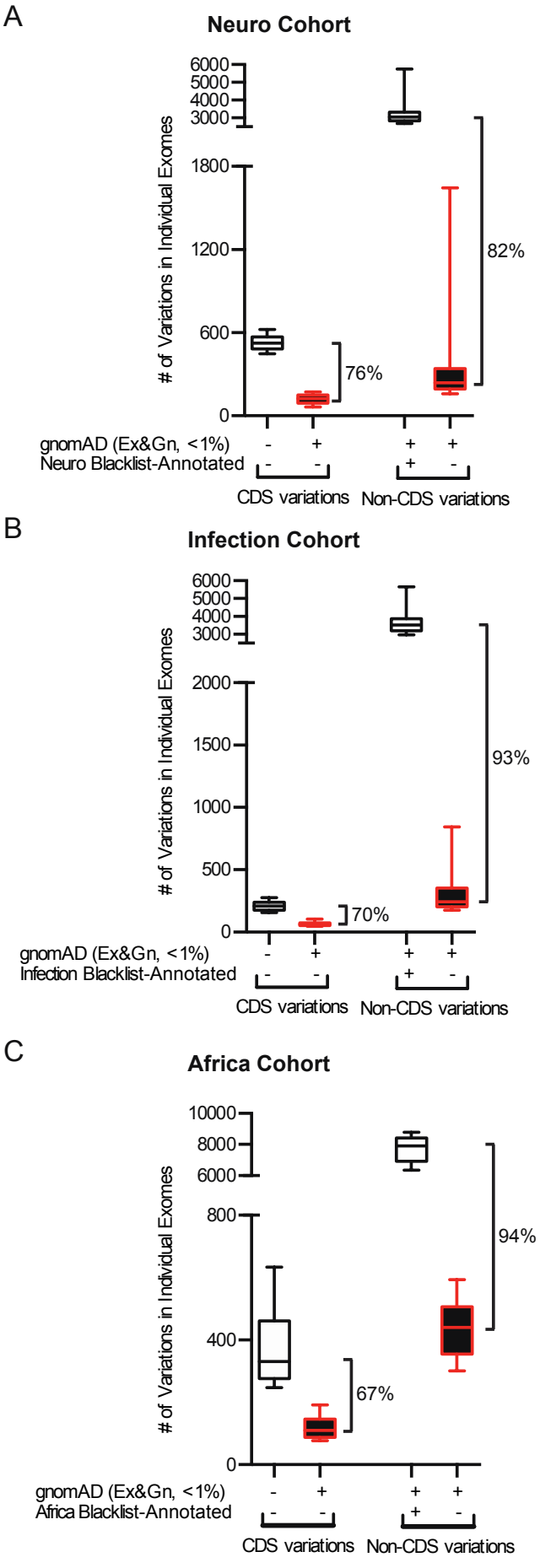


Figure S17. Filtering of coding and non-coding sequence variants in (A) 3,869 Neuro exomes restricted to cohort-specific variants with the Neuro blacklist-annotated, (B) 902 Infection exomes restricted to cohort-specific variants with the Infection blacklist-annotated, (C) 400 Africa exomes restricted to cohort-specific variants with the Africa blacklist-annotated. Error bars represent the 10th-90th percentiles.

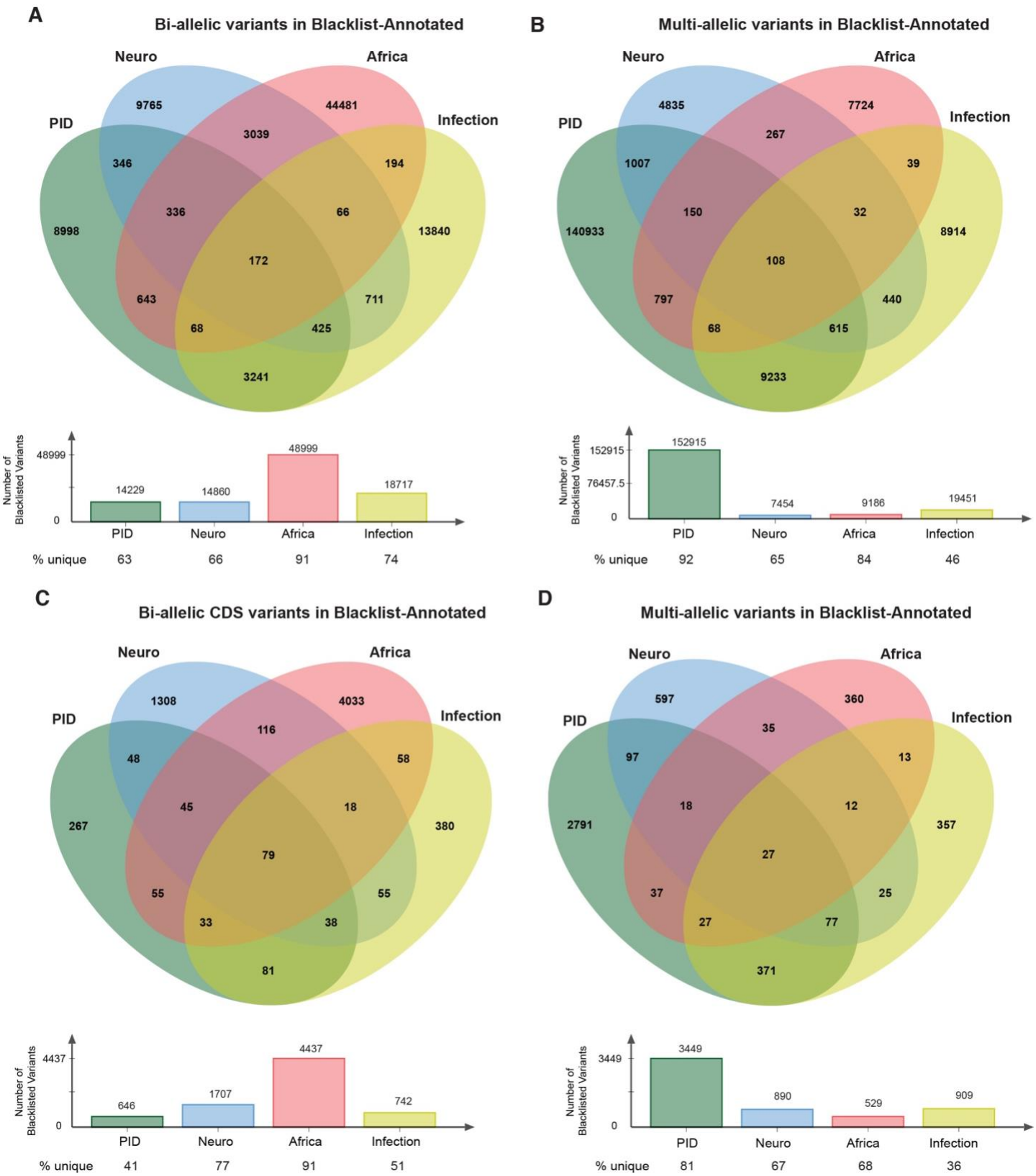


Figure S18. Relationship between the four blacklists. Common and unique biallelic (A), multiallelic (B), biallelic restricted to CDS (C), and multiallelic restricted to CDS (D) variants from the Blacklist-Annotated in the PID, Neuro, Africa and Infection cohorts.

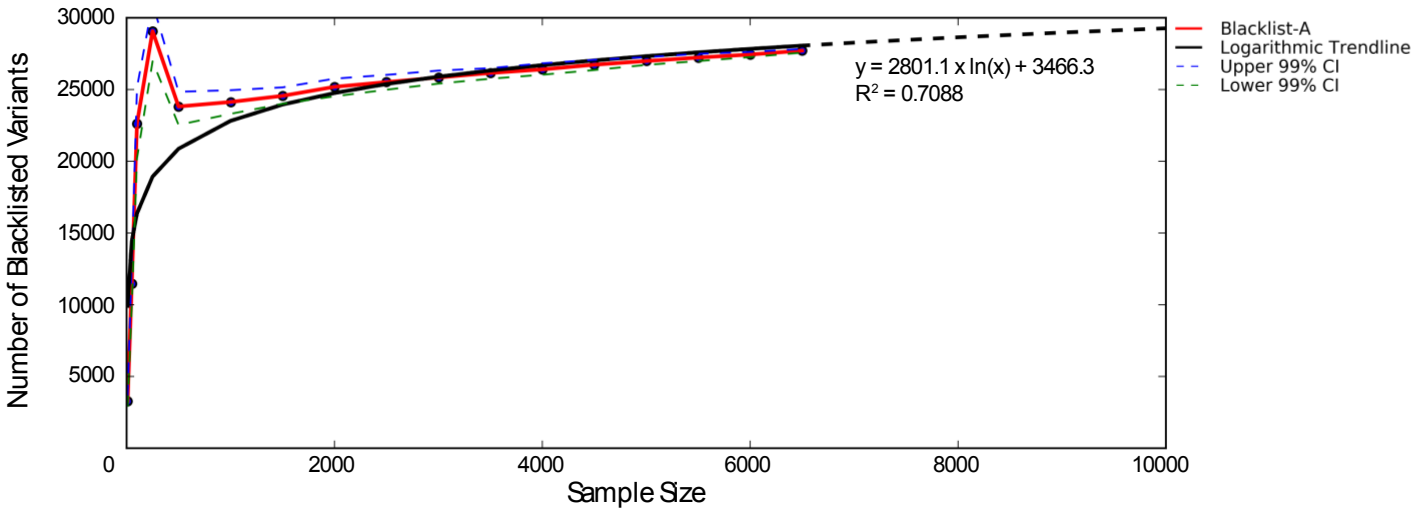


Figure S19. Relationship between sample size and number of blacklist variants. Estimation of the number of exomes required to create a saturated blacklist for CDS variants. Overlays in red, gray and green indicate that blacklist generation is unsafe, safe and optimal, respectively. The green vertical line indicates the suggested minimal sample size.

Table S1. VQSR status of blacklist-annotated (BL-A) variants

	# of VQSR PASS (%)	# of VQSR non-PASS (%)
Blacklist	125,614 (75.2%)	41,530 (24.8%)

Table S2. Blacklist-annotated variants in HGMD or ClinVar database

Chr	Position	Ref.	Alt.	HGMD	ClinVar	gnomAD	Gene	Disease	Status	Consequence	cDNA	Protein	rs ID	Publication (PMID)
4	88929173	C	CGAG	x		PASS	PKD2			inframe insertion	c.307_308insAGG	p.Glu102dup	rs547253972	
8	100844596	G	T	x	x	-	VPS13B	Cohen syndrome		splice acceptor variant	c.9406-1G>T		rs386834119	23188044, 16917849, 15154116
10	89720633	C	CT	x	x	PASS	PTEN	Hereditary cancer-predisposing syndrome		intron	c.802-18C>T		rs376702513	25394175, 18951446
12	102796022	A	T	x	x	PASS	IGF1	Insulin-like growth factor I deficiency	begign/likely benign	3' UTR variant	c.*297T>A		rs70961704	
13	20763685	A	AC	x	x	PASS	GJB2	Deafness, autosomal recessive 1	2 alleles one closed to 1%	frameshit	c.35dupG	p.Val13CysfsTer35	rs398123814	9482292, 24503448
21	47545369	A	AC	x		PASS	COL6A2			frameshit	c.1817-10_1817-9insC	p.Asp163ArgfsTer3	rs149954350	
X	66765161	A	T	x	x	PASS	AR	Infertility, male	Not tested.	Missense	c.173A>T	Gln58Leu	rs200185441	12801573, 24737579, 23637914
X	153006092	C	T	x		RF;AC0	ABCD1			stop gained	c.1699C>T	p.Gln567Ter	rs201114595	

Table S3. Biallelic and multi-allelic blacklist-annotated variants in the PID, Neuro, Infection and Africa cohorts

Blacklists	Biallelic		Multiallelic		Total	
	Count	**% of Total**	**Count**	**% of Total**	**Count**	**% of Total**
PID	14,229	8.5	152,915	91.5	167,144	100
Neuro	14,860	66.6	7,454	33.4	22,314	100
Infection	18,717	49.0	19,451	51.0	38,168	100
Africa	48,999	84.2	9,186	15.8	58,185	100

Table S4. Bi-allelic and multi-allelic blacklist-annotated variants by repetitive regions
(STR: short tandem repeats, Alu, GC-rich regions, other repetitive regions**)**

Occurrence of blacklisted variants in complex regions						
	Multi-allelic		**Bi-allelic**		**Total**	
	Count	**% of Total**	**Count**	**% of Total**	**Count**	**% of Total**
In complex regions	118,154	77.3	6,711	47.2	124,865	74.7
Not in complex regions	34,761	22.7	7,518	52.8	42,279	25.3
Breakdown by complex regions						
	Multi-allelic		**Bi-allelic**		**Total**	
	Count	**% of Total**	**Count**	**% of Total**	**Count**	**% of Total**
STR	65,646	55.6	2,457	36.6	68,103	53.5
Alu elements	44,866	38.0	1,713	25.5	46,579	36.7
GC-rich regions	4,314	3.7	1,742	26.0	6,056	6.2
Other repeat regions	3,328	2.8	799	11.9	4,127	3.6

Table S5. Hardy-Weinberg of bi-allelic CDS blacklist-annotated (BL-A) variants in Caucasian individuals

CDS bi-allelic variants in Caucasian Individuals (n = 1150)			
Total	**$<10^{-8}$**	**$>=10^{-8}$**	**% Disequilibrium**
622	74	548	12
CDS bi-allelic variants in disequilibrium by excess genotype			
	excess het	**excess hom alt**	**excess hom WT**
Counts	35	28	11
%	47.3	37.8	14.9
DP	163.0	20.5	15.6

Table S6. Ethnicity distribution of bi-allelic CDS blacklist-annotated (BL-A) variants in Hardy-Weinberg equilibrium

Ethnicity Distribution of CDS bi-allelic variants in HW equilibrium				
	Total	**$<10^{-8}$**	**$>10^{-8}$**	**Ethnical Disequilibrium (%)**
Counts	548	200	348	36.5
Causal Ethnicity for Disequilibrium				
	Middle Eastern	**African**	**Caucasian**	
Counts	20	20	6	
%	43.5	43.5	13.0	

Table S7: Biallelic blacklist annotated CDS variants in Hardy-Weinberg disequilibrium

Var	Gene	Unique	Exome_gnomAD	Genome_gnomAD	Obs het	Obs hom	Obs wt	HW_ Disequilibrium	DP Avg	Figure
4,88536886,CAGTGACAGCAGCAACAGCAGTGACAGCAGCGAT,C	DSPP	unique	PASS	PASS	352	75	150	7.01E-09	50	
6,136599910,T,TGTATCGCTTCTTTCTAGAATGAGATCTTGATCTTGATCA	BCLAF1	unique	PASS	AC0;RF	348	0	797	1.33E-09	210	
6,31324025,G,GT	HLA-B	unique	PASS	PASS	689	43	401	3.1E-32	23	
6,31324603,C,T	HLA-B	unique	PASS	PASS	717	253	173	1.18E-18	61	
6,32489852,A,ACGG	HLA-DRB1	unique	PASS	RF	608	102	357	9.33E-12	49	
6,32551960,T,TCC	HLA-DRB1	multi-01	PASS	PASS	631	113	394	1.03E-09	90	Sup. Figure 11
6,32552056,A,G	HLA-DRB1	multi-01	RF	InbreedingCoeff;RF	720	0	425	2.62E-54	152	Sup. Figure 11
6,32552085,G,GC	HLA-DRB1	multi-01	PASS	InbreedingCoeff	950	47	148	1.35E-114	124	Sup. Figure 11
6,32552093,A,T	HLA-DRB1	multi-01	RF	RF	528	0	610	2.2E-24	109	Sup. Figure 11
6,32552140,T,A	HLA-DRB1	multi-01	PASS	PASS	846	16	253	3.37E-86	64	Sup. Figure 11
6,32552144,A,C	HLA-DRB1	multi-01	PASS	PASS	953	28	119	1.13E-134	58	Sup. Figure 11
6,32557610,T,C	HLA-DRB1	multi-01	.	.	451	0	693	1.01E-16	55	Sup. Figure 11
7,100550245,G,T	MUC3A	unique	InbreedingCoeff	InbreedingCoeff	533	0	192	3.3E-55	547	
7,100551331,G,T	MUC3A	unique	PASS	PASS	850	0	177	2.49E-113	508	
7,142470773,A,G	PRSS3P1	unique	.	.	992	0	153	1.85E-147	213	
7,142231826,T,C	TRBV10-1	unique	PASS	PASS	1046	0	99	4.59E-178	236	
10,94018,T,G	TUBB8	unique	RF;AC0	AC0;InbreedingCoeff;RF	404	0	729	2.81E-13	51	
11,1093430,C,CCACCACGGTGACCCCAACCCCAACACCCACCGGCACACAG ACCCCAACAACGACACCCATCAGCACCAA	MUC2	unique	PASS	PASS	740	0	404	8.39E-59	171	
11,1016961,G,T	MUC6	multi-02	RF;AC0	AC0;RF	444	0	700	3.83E-16	306	Sup. Figure 12
11,1016972,G,A	MUC6	multi-02	.	InbreedingCoeff;RF	733	0	411	3.16E-57	280	Sup. Figure 12
11,1017040,G,GA	MUC6	multi-02	RF;InbreedingCoeff	InbreedingCoeff;RF	863	0	281	3.01E-93	237	Sup. Figure 12
11,1017458,A,G	MUC6	multi-02	RF;AC0	InbreedingCoeff;RF	1055	0	89	3.72E-184	231	Sup. Figure 12
11,1017470,G,T	MUC6	multi-02	.	.	908	0	70	1.12E-161	253	Sup. Figure 12
11,1018483,C,G	MUC6	multi-02	InbreedingCoeff	InbreedingCoeff	1015	0	129	3.5E-160	110	Sup. Figure 12
11,48387118,G,A	OR4C5	unique	InbreedingCoeff	InbreedingCoeff	1144	0	0	9.03E-251	125	
11,56143784,C,T	OR8U1	multi-03	InbreedingCoeff	InbreedingCoeff	1102	0	42	8.52E-217	122	Sup. Figure 13
11,56143803,A,G	OR8U1	multi-03	InbreedingCoeff	InbreedingCoeff	1071	0	73	1.1E-194	117	Sup. Figure 13
12,11244067,A,ATT	TAS2R43	multi-04	PASS	AC0;RF	660	203	266	6.36E-09	60	
12,11244070,T,C	TAS2R43	multi-04	PASS	PASS	665	210	251	8.68E-10	60	
15,23685604,TC,T	GOLGA6L2	unique	InbreedingCoeff	InbreedingCoeff	970	1	159	1.23E-140	308	
15,23686113,C,CTGCTCTTACATCTTCTCG	GOLGA6L2	unique	PASS	RF	342	0	785	1.92E-09	401	
15,90294306,C,A	MESP1	unique	PASS	PASS	649	172	270	4.5E-11	23	
19,8999561,G,C	MUC16	unique	RF	InbreedingCoeff;RF	619	0	525	4.27E-36	69	
19,4511350,T,A	PLIN4	unique	.	InbreedingCoeff	713	417	6	1.05E-50	140	
19,50463670,T,G	SIGLEC11	unique	PASS	PASS	404	9	730	3.97E-09	42	

Table S8. Sanger sequencing of 3 variants from blacklist annotated in patient exomes.

Variant					Characterization						Databases			Quality			WES Total			WES Genotype of 10 individuals			Sanger sequence of 10 individuals			Variant Status		
Gene	Chr	Pos	Ref	Alt	BL category	Diseq.	HW Eq. p-value[a]	Repeat region	CCDS	Ethnic Heterogenity	% of cohort with variant	ExAC 0.3.1	GnomAD r2.0.2	Mean DP	Mean MQ	Mean QD	WT[c]	Het	Hom	WT[c]	Het	Hom	WT	Het	Hom	Variant	Call problem	Suspected reason
HRNR	1	152,195,728	AT	A	Multi allelic	nd	nd	No	No	nd	98.3	-	Yes	42.3	60.2	18.3	44	170	2890	0	0	10	nc	nc	nc	nc	Yes	Short stretch of T
TBC1D19	4	26,737,063	C	CT	Multi allelic	nd	nd	No	No	nd	91.8	-	Yes	24.1	60.2	15.1	210	877	2017	0	5	5	nc	nc	nc	nc	Yes	Short stretch of T
FIG4	6	110,053,824	G	GT	Multi allelic	nd	nd	No	No	nd	88.6	-	Yes	28.4	60.0	13.9	349	1231	1524	0	6	4	nc	nc	nc	nc	Yes	Short stretch of T

nc : Not confirmed by Sanger sequencing due to poor quality.

Table S9. Summary of the technology employed for each cohort

Cohort							
	Size	**Kit**	**Sequencer**	**Aligner**	**Reference Genome**	**Caller**	**Annotator**
PID	3,104	Agilent 37, 50, 71 Mb	Hiseq 2000, 2500	bwa(v0.7.12)	hg19	GATK (v3.4-46)	snpEff
Neuro	3,869	Agilent 50 Mb	Hiseq 2000	bwa (v0.7.5)	GRCh37	GATK (v.3.1-1)	snpEff
Africa	400	Nextera Rapid Capture Expanded Exome 61 Mb	Hiseq 2500	bwa (v0.7.7)	GRCh37	GATK (v.3.5)	snpEff
Infection	902	Agilent 50 Mb, Illumina 65Mb	Hiseq 2000, 2500	bwa (v0.7.10)	hg19 decoy	GATK (v3.8)	snpEff

Table S10. Primers for PCR and sanger sequencing

Gene	Forward primer (5’ → 3’)	Reverse primer (5’ → 3’)
FIG4	CTGTCTTGCCCAAAGTCTGC	TTCTCATTCTGCTTTTACCCGC
HRNR	GGCGTGGAGTTCTTACCTTC	CACTCTCTTGCTACATGGCTTG
TBC1D19	CTTTCTGACATTTATGAACAGAG	GTGATTAGAAATAAAGTGGTG

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Detection of homozygous and hemizygous partial exon deletions by whole-exome sequencing

Benedetta Bigio[1,2,3], Yoann Seeleuthner[2,3], Gaspard Kerner[2,3], Melanie Migaud[2,3], Jérémie Rosain[2,3], Bertrand Boisson[1,2,3], Carla Nasca[4], Anne Puel[2,3], Jacinta Bustamante[1,2,3,5], Jean-Laurent Casanova[1,2,3,6,7], Laurent Abel[1,2,3,#,*], Aurelie Cobat[2,3,#,*]

[1]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065, USA
[2]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, 75015 Paris, France
[3]University of Paris, Imagine Institute, 75015 Paris, France
[4]Laboratory of Neuroendocrinology, The Rockefeller University, New York, NY 10065, USA
[5]Study Center of Immunodeficiencies, Necker Hospital for Sick Children, 75015 Paris, France
[6]Pediatric Hematology-Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France
[7]Howard Hughes Medical Institute, New York, NY 10065, USA

[#] indicates equal contributions

* To whom correspondence should be addressed.
Tel: +33 1 42 75 43 14;
Fax: +33 1 42 75 42 24;
E-mail: aurelie.cobat@inserm.fr, laurent.abel@inserm.fr

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

ABSTRACT

The detection of copy number variations (CNVs) in whole-exome sequencing (WES) data is important, as CNVs may underlie a number of human genetic disorders. The recently developed HMZDelFinder algorithm can detect rare homozygous and hemizygous (HMZ) deletions in WES data more effectively than other widely used tools. Here, we present HMZDelFinder_opt, an approach that outperforms HMZDelFinder for the detection of HMZ deletions, including partial exon deletions in particular, in typical laboratory cohorts that are generated over time under different experimental conditions. We show that using an optimized reference control set of WES data, based on a PCA-derived Euclidean distance for coverage, strongly improves the detection of HMZ deletions both in real patients carrying validated disease-causing deletions and in simulated data. Furthermore, we develop a sliding window approach enabling HMZDelFinder-opt to identify HMZ partial deletions of exons that are otherwise undiscovered by HMZDelFinder. HMZDelFinder_opt is a timely and powerful approach for detecting HMZ deletions, particularly partial exon deletions, in laboratory cohorts, which are typically heterogeneous.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

INTRODUCTION

Copy number variations (CNVs) are unbalanced rearrangements, classically covering more than 50 base pairs (bp), that increase or decrease the number of copies of specific DNA regions (1,2). There is growing evidence to implicate CNVs in common and rare diseases (1,3-5). CNVs have also been linked to adaptive traits, in environmental contexts for example (3). It has been recently estimated that CNVs affect ~5–10% of the genome, suggesting that a number of potentially disease-causing CNVs have yet to be discovered (1,6). Next-generation sequencing (NGS) techniques, such as whole-genome and whole-exome sequencing (WGS and WES), provide unprecedent opportunities for studying CNVs. Computational tools using data from WGS have been successfully used to detect CNVs (7-10), but WES-based methods have met with more limited success, mostly due to the nature of targeted enrichment protocols (11-13). Common WGS-based methods use breakpoints, the regions in which the rearrangements occur, to detect CNVs. By contrast, WES focuses on noncontiguous genomic targets (the exons), and most breakpoints are not sequenced. Hence, current WES-based approaches for detecting CNVs use the read depth (or coverage information) as a proxy for copy number information.

The HMZDelFinder algorithm is a recently developed coverage-based method for detecting rare homozygous and hemizygous (HMZ) deletions (14). This subset of CNVs may result in null alleles and a complete loss of gene function. Their identification may, therefore, lead to the discovery of novel genes or variations underlying Mendelian diseases. HMZDelFinder jointly evaluates the normalized per-interval coverage of all the samples of the entire dataset, making it possible to detect rare exonic HMZ deletions while minimizing the number of false-positive calls due to low-coverage regions. HMZDelFinder outperformed other CNV-calling tools, such as CONIFER (15), CoNVex (16), XHMM (17), ExonDel (18), CANOES (19), CLAMMS (20) and CODEX (21), particularly for the detection of single-exon deletions (i.e. deletions spanning only one exon) (14). However, two major limitations remain to be addressed. First, HMZDelFinder has been optimized to detect HMZ deletions from an entire dataset (>500) of homogeneous exome data. Its performance for typical laboratory cohort, which include exome data generated over time, often under different conditions, is, therefore, not optimal. Second, HMZDelFinder was not designed for the systematic detection of partial exon deletions (i.e. deletions spanning less than one exon). Here, we provide HMZDelFinder_opt, a method that extends the scope of HMZDelFinder

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

by improving the performance of the algorithm for the calling of HMZ deletions in typical laboratory cohorts, which are generated over time, and by allowing the systematic detection of partial exon deletions.

MATERIALS AND METHODS

Patient Cohort.

The 3,954 individuals used in this study were recruited in collaborations with clinicians, and most of them present different severe infectious diseases. Probands' family members account for the rest. Although these individuals do not form a random sample, they were ascertained through a number of distinct phenotypes and in different countries. Cohort-specific effects are, therefore, not expected to bias patterns of variation. All study participants provided written informed consent for the use of their DNA in studies aiming to identify genetic risk variants for disease. IRB approval was obtained from The Rockefeller University and Necker Hospital for Sick Children, along with a number of collaborating institutions.

WES and bioinformatic analysis

WES and bioinformatics analysis were performed as previously described (22). Briefly, genomic DNA was extracted and sheared with a Covaris S2 Ultra-sonicator. An adaptor-ligated library (Illumina) was generated, and exome capture was performed with either SureSelect Human All Exon kits (V5-50Mb, V4-50Mb, V4-71Mb, or V6-60Mb) from Agilent Technologies, or xGen Exome Research 39Mb Panel from Integrated DNA Technologies (IDT xGen). Massively parallel WES was performed on a HiSeq 2000 or 2500 machine (Illumina), generating 100- or 125-base reads. Quality controls were applied at the lane and fastq levels. Specifically, the cutoff used for a successful lane is Pass Filter > 90%, with over 250 M reads for the high-output mode. The fraction of reads in each lane assigned to each sample (no set value) and the fraction of bases with a quality score > Q30 for read 1 and read 2 (above 80% expected for each) were also checked. In addition, the FASTQC tool kit (www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to review base quality distribution, representation of the four nucleotides of particular k-mer sequences (adaptor contamination). We used the Genome Analysis Software Kit (GATK) (version 3.2.2 or 3.4-46) best-practice pipeline to analyze our WES data(23). Reads were aligned with the human reference genome (hg19), using the maximum exact matches algorithm in Burrows–Wheeler Aligner (BWA)(24). PCR duplicates were removed with Picard tools (picard.sourceforge.net/). The GATK base quality score recalibrator was applied to correct sequencing artifacts.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Positive controls

The five WES samples used as positive controls carry rare HMZ disease-causing deletions that were confirmed with state-of-the-art molecular approaches (25-27). Specifically, these HMZ deletions comprise one or more exons and have different lengths as follows (SI Table 1). P1 carries a deletion of exons 21 to 23 in *DOCK8* (10,800 bp) that was validated by multiplex ligation-dependent probe amplification (MLPA). The deletion in *DOCK8* was functionally linked to staphylococcus infection (25). P2 had a deletion of exon 5 in *NCF2* (134 bp) that was also validated by MLPA and found to be causal in chronic granulomatous disease (manuscript in preparation). P3's deletion spanned exons 2 to 8 in *IL12RB1* (13,000 bp) and was validated by sanger sequencing. This deletion was demonstrated to be causal for a Mendelian susceptibility to mycobacterial disease (26). P4 has a deletion of the entire *CYBB* (3,400,000 bp) validated by MLPA and CGH array that resulted in chronic granulomatous disease (27). Finally, P5 is a patient with hyper IgE syndrome carrying a deletion of exons 7 to 15 in entire *DOCK8* (28,000 bp) that was validated by Sanger sequencing. *CYBB* is on the X chromosome while all other genes are autosomal.

HMZDelFinder-opt

The general workflow used in HMZDelFinder-opt is depicted in SI Figure 1. First, HMZDelFinder_opt computes coverage profiles from the BAM files of the entire dataset. Second, the Principal component analysis (PCA) is calculated from a covariance matrix based on standardized coverage profiles and a k nearest neighbors algorithm is used to select the reference control set. Third, the BAM file of a given sample and the BAM files of the reference control set are used as input of HMZDelFinder to detect HMZ deletions. Fourth, when HMZDelFinder_opt is provided with the parameter -sliding_window_size and the related size, it will employ a sliding window approach for identification of partial deletions of exons. Each of these steps is described in the following paragraphs.

Principal component analysis (PCA) and k nearest neighbors algorithm

The PCA was performed on the coverage profile of the 3,954 WES using per-exon coverage. Specifically, for each sample, the coverage profile was calculated using the mean depth of coverage of the 194,528 exons from the consensus coding sequences (CCDS) annotation of GRCh37 obtained using biomaRt (28). The PCA was then performed using the 'prcomp' function from R 3.5.1 on the scaled coverage profiles. To select the reference

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

control set for a given sample, we computed pairwise weighted Euclidean distances between individuals i and j based on the first 10 principal components from the PCA using the 'dist' function of R 3.5.1, using the formula:

$$dist(i,j) = \sqrt{\sum_{k=1}^{10} \lambda_k \left(PC_{ki} - PC_{kj}\right)^2}$$

where PC is the matrix of principal components (PCs) calculated on common variants and λ_k the eigenvalue corresponding to the k-th principal component PC_k.

HMZDelFinder

We used the HMZDelFinder algorithm as described (14). In brief, HMZDelFinder calculates per-exon read depth (reads per thousand base pairs per million reads; RPKM) to detect HMZ deletions. For our purpose of covering all the coding regions, we employed an interval file containing all coding sequences from Gencode. For a given interval, the criteria to call a deletion are as follows: 1) RPKM < 0.65 and 2) frequency of the deletion within the dataset ≤ 0.5%. Filtering criteria at the interval and sample levels include removal of low quality intervals (RPKM median < 7 across all samples) and removal of low quality samples (2% with highest number of calls). When using the optional absence of heterozygosity (AOH) step, HMZDelFinder uses VCF files to filter out deletions not falling in AOH regions, assuming that rare and pathogenic homozygous deletions are likely to be located within larger AOH regions due to the inheritance of a shared haplotype block from both parents. Finally, to prioritize deletions, z-scores are computed. The z-score of a deletion measures the number of standard deviations between the coverage of the deleted interval in a given sample compared to the mean coverage of the same interval in the rest of the dataset. A very low z-score indicates high mean coverage with low variance in the dataset and very low (or no coverage at all) in a given sample. Hence, lower z-scores denote higher confidence in a given deletion.

Sliding window approach and simulated data

We simulated deletions of variable size in 200 randomly selected individuals among our in-house cohort but excluding the oldest samples (V4-50Mbp capture kit), due to a lower quality than present standards. Two different exons were selected to undergo simulated deletions: a favorable case, exon 11 from LIMCH1 gene (409bp) with a mean coverage of approximately 85X in our samples, and an unfavorable case, exon 4 from RPL15 gene (406 bp) with a mean coverage of 15X in our samples. For both exons, we deleted a segment of 25%, 50%, 75% or

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

100% of the exon size, using the '-v' argument of the 'bedtools intersect' command (bedtools v1.9) on the BAM file to remove all reads overlapping the segment. We then ran HMZDelFinder and HMZDelFinder_opt (with and without the --sliding_windows parameter) on the whole BAM files. Specifically, we applied a sliding window approach, in which each exon was divided into 100 bp windows, with 50 bp overlaps, and BAM files for individual exomes were transformed into per-window read depths. In a separate analysis, we used 50 bp windows, with 25 bp overlaps.

Analysis of common deletions

To determine whether some of the called deletions were previously reported as common deletions, we utilized the CNVs from the Gold Standard track (hg19 version dated 2016-05-15) of the Database of Genomic Variants (DGV), a highly curated resource that collects CNVs in the human genome (29). We retained only entries with field 'variant_sub_type' equal to 'Loss' and frequency greater than 1%. We then crossed the retained entries with the deletions called by HMZDelFinder and HMZDelFinder_opt in the positive controls. Deletions were considered common in the DGV database when they overlapped at least 50% with the retained entries from the DGV database.

RESULTS

Optimization of the reference control set in HMZDelFinder_opt

We first aimed to improve the performance of HMZDelFinder for detecting HMZ deletions in typical heterogeneous laboratory cohorts, which were generated over time and in different experimental settings (e.g. capture kit). We reasoned that comparing a given sample with an optimized reference control set would limit the impact of the background variability intrinsic to exome data, thereby improving the performance of HMZDelFinder. We designed the optimized reference control set as a selection of samples with similar coverage profiles (SI Figure 1). We did this by first performing a principal component analysis (PCA) of the depth of coverage for consensus coding sequences (CCDS) for 3,954 exomes from our in-house cohort, including mostly patients with severe infectious diseases. As expected, given the different sequencing conditions used for whole-exome sequencing (SI Table 2), the coverage profiles of the samples were highly variable (Figure 1). The first two principal components (PCs) of the PCA identified six distinct clusters, mostly reflecting the capture kit used (Figure 1). Interestingly, two different clusters (clusters 1 and 2 on Figure 1) corresponded to the V4-71Mb

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

capture kit, the difference between these clusters being associated mostly with a minor change in the sequencing chemistry of the kit, leading to a significant improvement in coverage profile for the more recently generated exome data (SI Figure 2). We then used the first 10 PCs to calculate the pairwise weighted Euclidean distances between all samples (30) (see methods). We used this metric to determine, for each sample of interest, the closest neighbors, for use as the reference control set in HMZDelFinder_opt.

We then compared the performances of HMZDelFinder_opt and HMZDelFinder, using five WES samples carrying validated rare HMZ disease-causing deletions of different lengths as positive controls (SI Table 1, methods). Specifically, we tested the ability of HMZDelFinder_opt and HMZDelFinder to detect the validated deletions, and we also compared the total numbers of deletions called and their *z*-scores (see Methods). In HMZDelFinder_opt, we compared reference control sets of different size (ranging from 50 to 500, SI Figure 3), selected for each sample as described above. In HMZDelFinder, we used the entire dataset, consisting of 3,954 WES samples. For both approaches, the final set of called deletions for each sample was narrowed down to the capture kit corresponding to the patient WES data. We chose to benchmark HMZDelFinder because it has been shown to perform at least as well as, and sometimes better than several widely used and actively maintained detection tools (14).

Both HMZDelFinder and HMZDelFinder_opt successfully detected all five confirmed HMZ deletions in the positive controls, regardless of the size of the reference control set (Table 1). However, HMZDelFinder_opt detected a smaller total number of deletions than HMZDelFinder (Table 1). Specifically, the total number of deletions ranged from one to 21 deletions for HMZDelFinder_opt, and from 11 to 2,586 for HMZDelFinder, suggesting that a smaller number of false-positive calls were obtained with HMZDelFinder_opt. Using the optional filtering step based on the absence of heterozygosity (AOH) information for HMZDelFinder (see methods) decreased the number of deletions detected, but this number nevertheless remained much higher than that for HMZDelFinder_opt (Table 1). We hypothesized that the large difference between the two methods for P1 reflected the low quality of exome data for this patient. Indeed, the mean coverage and the proportion of bases with coverage above 10x were much lower for P1 than for the other four patients (e.g. only 68.9% of bases had a coverage above 10x for P1, versus >99% for the other patients) (SI Table 1), leading to a large number of likely false positive deletions detected when not using an appropriate reference control set with similar coverage. Consistently, the number of deletions detected for P1 with HMZDelFinder_opt was larger with the largest

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

reference sample size (500) (Table 1). We therefore performed subsequent HMZDelFinder_opt analyses with a reference sample size of 100, which provided a good compromise between the algorithm performance and computation time.

We then compared the rankings of the confirmed deletions between the two algorithms, using the *z*-score provided by HMZDelFinder (see method). While the two approaches ranked the confirmed disease-causing deletions for P1 and P5 first, HMZDelFinder_opt ranked higher the confirmed disease-causing deletions for P2, P3 and P4 than HMZDelFinder (Table 1; Figure 2). Moreover, *z*-scores were consistently better with HMZDelFinder_opt (Figure 2) than with HMZDelFinder, leading to a more specific discovery of true HMZ deletions. Again, using the AOH option for HMZDelFinder slightly improved the ranking (Table 1), but did not change the *z*-score ranking. Together, these results suggest that HMZDelFinder_opt gives better *z*-scores for deletions than HMZDelFinder, which should lead to higher sensitivity in the general case.

Finally, we studied the HMZ deletions called by both approaches, in addition to the validated ones, to determine whether some of the deletions identified were reported as common deletions. We used the CNVs from the gold standard track of the Database of Genomic Variants (DGV), a highly curated resource containing CNVs from the human genome (29). We focused on the positive controls with high data quality (P2, P3, P4 and P5), and found that the HMZ deletions called by HMZDelFinder_opt were more enriched in common deletions (frequency > 1%) than those called by HMZDelFinder (SI Table 3). Among the 6 and 303 additional HMZ deletions called by HMZDelFinder–opt (with the reference control set of 100 exomes) and HMZDelFinder, 50% and 1%, respectively, were present in the DGV database (SI Table 3), suggesting that the deletions called by HMZDelFinder_opt were enriched in true deletions. Overall, these findings demonstrate that the use of an appropriate reference control set of WES data based on a PCA-derived coverage distance improves the performance of HMZDelFinder. These results also provided a first validation of HMZDelFinder_opt for five confirmed disease-causing HMZ deletions.

Detection of HMZ partial exon deletions by HMZDelFinder_opt

In HMZDelFinder, individual exome BAM files are transformed into per-exon read depths, facilitating a more efficient detection of single-exon HMZ deletions than can be achieved with other classical CNV-calling algorithms (14). Here, we aimed to address the need for the identification of even smaller HMZ deletions, spanning less

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

than an exon (partial exon deletions). To this end, we used HMZDelFinder_opt with a sliding window approach, in which each exon was divided into 100 bp windows, with 50 bp overlaps, and BAM files for individual exomes were transformed into per-window read depths. We tested this approach by simulating deletions in two exons of similar size (~400 bp) but with different mean coverages in a randomly selected dataset of 200 WES samples from our in-house cohort. The deletions spanned 100%, 75%, 50% or 25% of either exon 11 of *LIMCH1* (409 bp, ~85x mean coverage) or exon 4 of *RPL15* (406 bp, ~15x mean coverage). We used these datasets to compare the performances of HMZDelFinder_opt with sliding windows of 100 bp (HMZDelFinder_opt+sw100), HMZDelFinder_opt without sliding windows (HMZDelFinder_opt), and the original HMZDelFinder. For HMZDelFinder_opt+sw100 and HMZDelFinder_opt ,we used reference control sets of size 100.

For deletions spanning the full exon (100%), we confirmed that HMZDelFinder_opt had a detection rate (98% and 93% for exons with higher and lower coverage, respectively; Figure 3) similar to that of HMZDelFinder (98% and 93% for exons with higher and lower coverage, respectively). However, the total number of HMZ deletions called by HMZDelFinder_opt was only one eighth the total number of HMZ deletions called by HMZDelFinder (median number of HMZ deletions: 2 vs. 13 SI Figure 4). The detection rate was slightly higher when sliding windows were used (detection rate for HMZDelFinder_opt+sw100 of 99% and 94% for exons with a higher and lower coverage, respectively), but at the cost of a slightly larger total number of HMZ deletions called than for HMZDelFinder_opt (median number of deletions: 5 vs. 2). Nevertheless, the total number of HMZ deletions called by HMZDelFinder_opt+sw100 remained lower than the total number of HMZ deletions called by HMZDelFinder.

For partial exon deletions, the detection rates of HMZDelFinder and HMZDelFinder_opt were much lower, at less than 10% for deletions spanning 75% of the exon and 0% for deletions spanning 25% or 50% of the exon. Conversely, HMZDelFinder_opt+sw100 succeeded in detecting simulated deletions spanning 50% or 75% (200 bp or ~300 bp) of both exon 11 of *LIMCH1* and exon 4 of *RPL15* in 99% of the samples, with a median number of called HMZ deletions of 5 (Figure 3, SI Figure 4). For deletions spanning 25% of the exon (~100 bp), HMZDelFinder_opt+sw100 had a detection rate of 74% for the exon with the highest coverage in *LIMCH1,* but it failed to detect the deletions in the exon with the lowest coverage in *RPL15*. We assessed the performance of this method further, using a smaller sliding window of 50 bp in size, and a step size of 25 bp, to improve granularity. We found that the use of smaller sliding windows with HMZDelFinder_opt+sw50 greatly increased

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

the detection rate for deletions spanning 25% of the exon with the lowest coverage, exon 4 of *RPL15* (93% for sw50 vs. 1% for sw100) and of the exon with the highest coverage in *LIMCH1* (98% for sw50 vs. 74% for sw100) (Figure 3). Thus, the use of a sliding window makes it possible to detect HMZ partial exon deletions that would otherwise be missed, and the use of simulated data further validated HMZDelFinder_opt.

DISCUSSION

WES offers unprecedent opportunities for identifying HMZ deletions as novel causal determinants of human diseases, but it poses a number of computational challenges. Most current methods for detecting HMZ deletions compare the depth of coverage between a given exome and the rest of the exomes in the dataset. However, coverage depth is heavily dependent on sequencing conditions, which are continually evolving in typical laboratory settings. Thus, the exome data generated over time are inevitably heterogeneous, complicating the discovery of deletions. Using HMZDelFinder_opt with both validated disease-causing deletions and simulated data, we demonstrated that the *a priori* selection of a reference control set with a coverage profile similar to that of the WES sample studied reduced the number of deletions detected, while improving the ranking of the true HMZ deletion. These results are consistent with a recent report showing that the selection of an appropriate reference control set with multidimensional scaling significantly improves the sensitivity of various CNV callers (31). In further support for our findings, the ranking of the known deletion and the number of additional deletions detected by HMZDelFinder_opt start worsening with increasing numbers of controls in the reference set, including neighbors with a less similar coverage profile, as illustrated, for P1, in SI Fig. 3A.

In addition to providing an optimized tool for detecting deletions in typical laboratory cohorts, HMZDelFinder_opt also fills the gap in the study of deletions spanning less than an exon, by providing the first tool for the systematic identification of partial exon deletions. Existing CNV callers are optimized for the detection of either large deletions (usually spanning more than three exons), or deletions of full single exons (14,32). Other established callers, such as GATK, are not designed to detect CNVs and can therefore identify deletions of only a few dozen base pairs (typically up to 50 bp, https://gatkforums.broadinstitute.org/gatk/discussion/5938/using-gatk-tool-how-long-insertion-deletion-could-be-detected and (33)). The human genome contains ~235,000 exons, about 20% of which are larger than 200 bp (34). HMZDelFinder_opt therefore makes possible the systematic discovery of currently unknown HMZ deletions in ~47,000 exons that are not detectable with other

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

tools. In sum, we describe HMZDelFinder_opt, a method for improving the detection of HMZ deletions in heterogeneous exome data that can be used to identify partial exon deletions that would otherwise be missed, through an extension of the scope of HMZDelFinder.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

DATA AVAILABILITY

The code for the PCA-based selection and sliding window is available in the GitHub repository (https://github.com/casanova-lab/HMZDelFinder_opt/).

ACKNOWLEDGEMENT

We thank the members of the Human Genetics of Infectious Diseases Laboratory for helpful discussions. We also thank Yelena Nemiroskaya, Dominick Papandrea, Mark Woollett, Dana Liu (St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, New York, USA), and Cécile Patissier, Lazaro Lorenzo-Diaz, Christine Rivalain (Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, Paris, France) for their assistance.

FUNDING

This research was supported in part by the National Institutes of Health (NIH) (grants R01AI088364, R37AI095983, U19AI111143, R01AI127564, P01AI061093 to J.-L.C.), the National Center for Research Resources and the National Center for Advancing Sciences of the NIH (grant 8UL1TR001866), the Yale Center for Mendelian Genomics and the GSP Coordinating Center funded by the National Human Genome Research Institute (NHGRI) (UM1HG006504 and U24HG008956), the Rockefeller University, the St. Giles Foundation, Howard Hughes Medical Institute, Institut National de la Santé et de la Recherche Médicale (INSERM), University of Paris, the Integrative Biology of Emerging Infectious Diseases Laboratory of Excellence (ANR-10-LABX-62-IBEID), the French Foundation for Medical Research (FRM) (EQU201903007798), the SCOR Corporate Foundation for Science, and the French National Research Agency (ANR) under the “Investments for the future” (grand number ANR-10-IAHU-01), GENMSMD (ANR-16-CE17.0005-01, to JB), ANR-LTh-MSMD-CMCD (ANR-18-CE93-0008-01 to A.P), Fonds de Recherche en Santé Respiratoire (SRC2017 to J.B.), ProgLegio project (ANR-15-CE17-0014). and ECOS Nord (C19S01-63407 to J.B.).

CONFLICT OF INTEREST

We declare no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

REFERENCES

1. Zarrei, M., MacDonald, J.R., Merico, D. and Scherer, S.W. (2015) A copy number variation map of the human genome. *Nature reviews. Genetics*, **16**, 172-183.
2. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alföldi, J., Khera, A.V., Francioli, L.C., Gauthier, L.D., Wang, H., Watts, N.A. *et al.* (2019) An open resource of structural variation for medical and population genetics. *bioRxiv*, 578674.
3. Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R. *et al.* (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet*, **39**, 1256-1260.
4. Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. (2009) Copy number variation in human health, disease, and evolution. *Annual review of genomics and human genetics*, **10**, 451-481.
5. Lee, C. and Scherer, S.W. (2010) The clinical context of copy number variation in the human genome. *Expert Rev Mol Med*, **12**, e8-e8.
6. Sharp, A.J., Cheng, Z. and Eichler, E.E. (2006) Structural variation of the human genome. *Annual review of genomics and human genetics*, **7**, 407-442.
7. Handsaker, R.E., Korn, J.M., Nemesh, J. and McCarroll, S.A. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet*, **43**, 269-276.
8. Zhou, B., Ho, S.S., Zhang, X., Pattni, R., Haraksingh, R.R. and Urban, A.E. (2018) Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *J Med Genet*, **55**, 735-743.
9. Gross, A.M., Ajay, S.S., Rajan, V., Brown, C., Bluske, K., Burns, N.J., Chawla, A., Coffey, A.J., Malhotra, A., Scocchia, A. *et al.* (2019) Copy-number variants in clinical genome sequencing: deployment and interpretation for rare and undiagnosed disease. *Genetics in Medicine*, **21**, 1121-1130.
10. Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q.B., Antipenko, A., Shang, L., Boisson, B., Casanova, J.-L. and Abel, L. (2015), *Proc. Natl. Acad. Sci. U.S.A.*, Vol. 112, pp. 5473-5478.
11. Kadalayil, L., Rafiq, S., Rose-Zerilli, M.J.J., Pengelly, R.J., Parker, H., Oscier, D., Strefford, J.C., Tapper, W.J., Gibson, J., Ennis, S. *et al.* (2015) Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform*, **16**, 380-392.
12. Fromer, M., Moran, J.L., Chambert, K., Banks, E., Bergen, S.E., Ruderfer, D.M., Handsaker, R.E., McCarroll, S.A., O'Donovan, M.C., Owen, M.J. *et al.* (2012) Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet*, **91**, 597-607.
13. Tan, R., Wang, Y., Kleinstein, S.E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A.S. and Zhu, M. (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat*, **35**, 899-907.
14. Gambin, T., Akdemir, Z.C., Yuan, B., Gu, S., Chiang, T., Carvalho, C.M.B., Shaw, C., Jhangiani, S., Boone, P.M., Eldomery, M.K. *et al.* (2017) Homozygous and hemizygous CNV detection from exome sequencing data in a Mendelian disease cohort. *Nucleic Acids Res*, **45**, 1633-1648.
15. Krumm, N., Sudmant, P.H., Ko, A., O'Roak, B.J., Malig, M., Coe, B.P., Quinlan, A.R., Nickerson, D.A. and Eichler, E.E. (2012) Copy number variation detection and genotyping from exome sequence data. *Genome research*, **22**, 1525-1532.
16. Amarasinghe, K.C., Li, J. and Halgamuge, S.K. (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics*, **14**, S2.
17. Fromer, M. and Purcell, S.M. (2014) Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Current protocols in human genetics*, **81**, 7.23.21-21.
18. Guo, Y., Zhao, S., Lehmann, B.D., Sheng, Q., Shaver, T.M., Stricker, T.P., Pietenpol, J.A. and Shyr, Y. (2014) Detection of internal exon deletion with exon Del. *BMC Bioinformatics*, **15**, 332.
19. Backenroth, D., Homsy, J., Murillo, L.R., Glessner, J., Lin, E., Brueckner, M., Lifton, R., Goldmuntz, E., Chung, W.K. and Shen, Y. (2014) CANOES: detecting rare copy number variants from whole exome sequencing data. *Nucleic Acids Res*, **42**, e97.
20. Packer, J.S., Maxwell, E.K., O'Dushlaine, C., Lopez, A.E., Dewey, F.E., Chernomorsky, R., Baras, A., Overton, J.D., Habegger, L. and Reid, J.G. (2016) CLAMMS: a scalable algorithm for calling common and rare copy number variants from exome sequencing data. *Bioinformatics (Oxford, England)*, **32**, 133-135.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

21. Jiang, Y., Oldridge, D.A., Diskin, S.J. and Zhang, N.R. (2015) CODEX: a normalization and copy number variation detection method for whole exome sequencing. *Nucleic Acids Res*, **43**, e39.
22. Maffucci, P., Bigio, B., Rapaport, F., Cobat, A., Borghesi, A., Lopez, M., Patin, E., Bolze, A., Shang, L., Bendavid, M. *et al.* (2019) Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis. *Proc Natl Acad Sci U S A*, **116**, 950-959.
23. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M. *et al.* (2011), *Nat. Genet.*, Vol. 43, pp. 491-498.
24. Li, H. and Durbin, R. (2009), *Bioinformatics (Oxford, England)*. Oxford University Press, Vol. 25, pp. 1754-1760.
25. Aydin, S.E., Kilic, S.S., Aytekin, C., Kumar, A., Porras, O., Kainulainen, L., Kostyuchenko, L., Genel, F., Kütükcüler, N., Karaca, N. *et al.* (2015) DOCK8 deficiency: clinical and immunological phenotype and treatment options - a review of 136 patients. *Journal of clinical immunology*, **35**, 189-198.
26. Rosain, J., Oleaga-Quintas, C., Deswarte, C., Verdin, H., Marot, S., Syridou, G., Mansouri, M., Mahdaviani, S.A., Venegas-Montoya, E., Tsolia, M. *et al.* (2018) A Variety of Alu-Mediated Copy Number Variations Can Underlie IL-12Rβ1 Deficiency. *Journal of clinical immunology*, **38**, 617-627.
27. Blancas-Galicia, L., Santos-Chávez, E., Deswarte, C., Mignac, Q., Medina-Vera, I., León-Lara, X., Roynard, M., Scheffler-Mendoza, S.C., Rioja-Valencia, R., Alvirde-Ayala, A. *et al.* (2020) Genetic, Immunological, and Clinical Features of the First Mexican Cohort of Patients with Chronic Granulomatous Disease. *Journal of clinical immunology*, **40**, 475-493.
28. Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., Arnaiz, O., Awedh, M.H., Baldock, R., Barbiera, G. *et al.* (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res*, **43**, W589-W598.
29. MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L. and Scherer, S.W. (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*, **42**, D986-D992.
30. Belkadi, A., Pedergnana, V., Cobat, A., Itan, Y., Vincent, Q.B., Abhyankar, A., Shang, L., El Baghdadi, J., Bousfiha, A., Alcais, A. *et al.* (2016) Whole-exome sequencing to analyze population structure, parental inbreeding, and familial linkage. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 6713-6718.
31. Kuśmirek, W., Szmurło, A., Wiewiórka, M., Nowak, R. and Gambin, T. (2019) Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers performance. *BMC Bioinformatics*, **20**, 266-266.
32. de Ligt, J., Boone, P.M., Pfundt, R., Vissers, L.E.L.M., Richmond, T., Geoghegan, J., O'Moore, K., de Leeuw, N., Shaw, C., Brunner, H.G. *et al.* (2013) Detection of clinically relevant copy number variants with whole-exome sequencing. *Hum Mutat*, **34**, 1439-1448.
33. Shigemizu, D., Miya, F., Akiyama, S., Okuda, S., Boroevich, K.A., Fujimoto, A., Nakagawa, H., Ozaki, K., Niida, S., Kanemura, Y. *et al.* (2018) IMSindel: An accurate intermediate-size indel detection tool incorporating de novo assembly and gapped global-local alignment with split read analysis. *Scientific Reports*, **8**, 5608.
34. Sakharkar, M.K., Chow, V.T.K. and Kangueane, P. (2004) Distributions of exons and introns in the human genome. *In Silico Biol*, **4**, 387-393.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

TABLES AND FIGURES

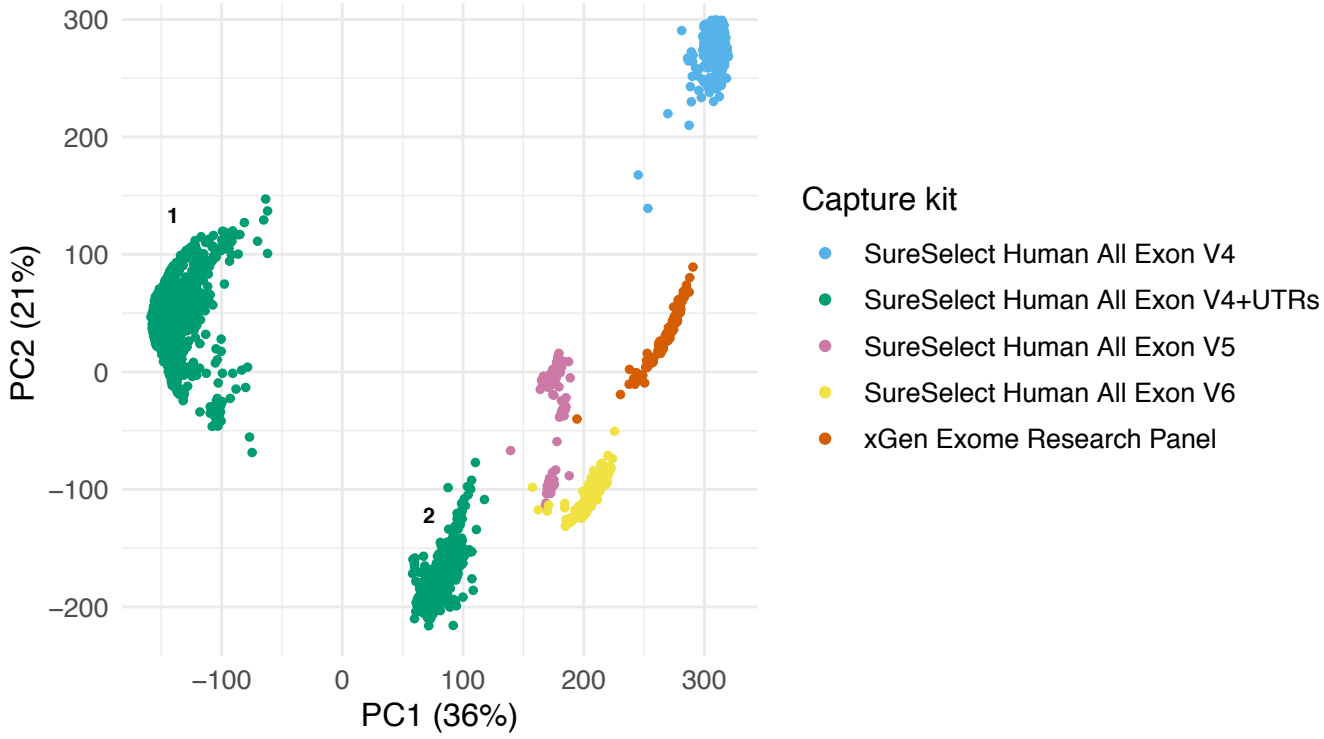


Figure 1: Principal Component Analysis (PCA) of the WES coverage. The PCA was computed from the coverage profiles of consensus coding sequences (CCDS) from 3,954 individuals. Dots are color-coded by the type of the capture kit used for sequencing.). Two different clusters (clusters 1 and 2) corresponded to the V4-71Mb capture kit. See also SI Figure 2.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

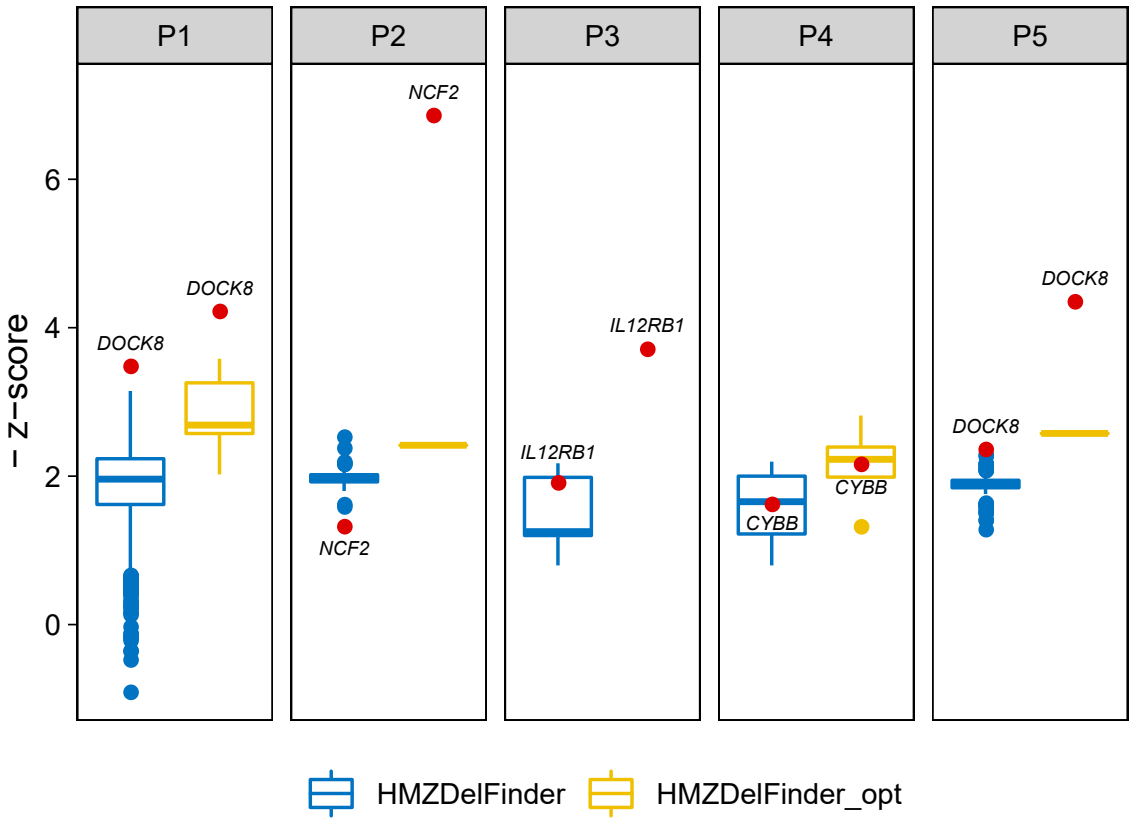


Figure 2: Comparison of the ranking of the deletions called by HMZDelFinder_opt and HMZDelFinder in five positive controls carrying validated rare HMZ disease-causing deletions. The ranking is expressed as - z-score. Lower z-scores (and higher ranking) indicate more confidence in a given deletion. The confirmed deletions ranked 1st in P1, P2, P3, P5 with HMZDelFinder_opt while they ranked 1st only in P1 and P5 with HMZDelFinder as shown by the red dots in the blue (HMZDelFinder) and yellow (HMZDelFinder_opt) distributions. The ranking was consistently higher with HMZDelFinder_opt than with HMZDelFinder. Results are shown for HMZDelFinder_opt using 100 as size of the reference control set.

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

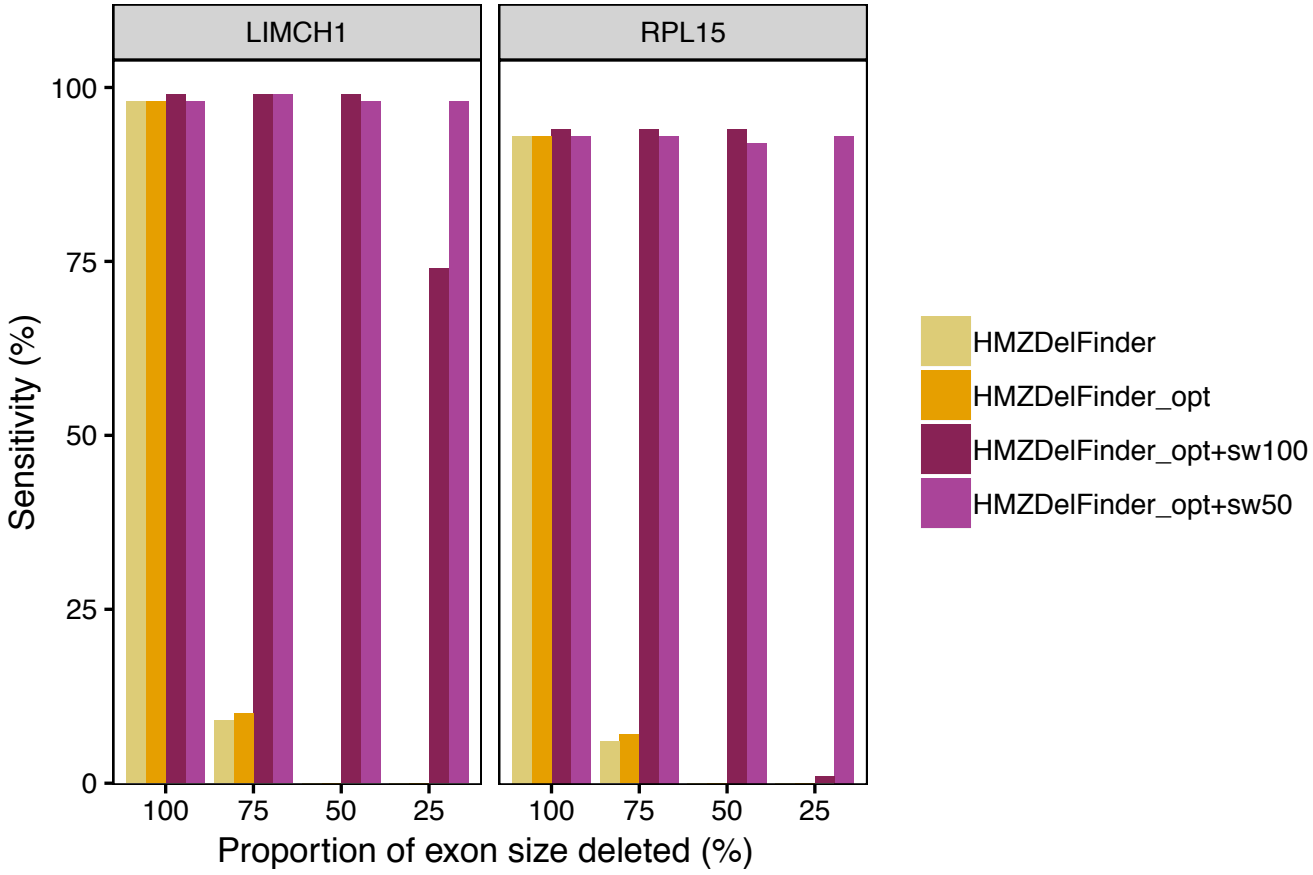


Figure 3: Comparison of HMZDelFinder_opt with or without sliding windows and HMZDelFinder by using simulated data. Proportions of deletions detected in simulated data in the higher (LIMCH1) or lower (RPL15) covered exons by using HMZDelfinder (yellow), HMZDelFinder_opt (orange), HMZDelFinder_opt+sw100 (red), HMZDelFinder_opt+sw50 (pink).

bioRxiv preprint doi: https://doi.org/10.1101/2020.07.23.217976. this version posted July 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

		P1	P2	P3	P4	P5
	KIT	V4-50MB	V6-60MB	V5-50MB	V5-50MB	V6-60MB
METHOD	**N NEIGHBORS**	**Confirmed deletion (Rank/Total number of deletions)**				
HMZDelFinder_opt	50	DOCK8 (1/11)	NCF2 (1/2)	IL12RB1 (1/1)	CYBB (3/5)	DOCK8 (1/3)
	100	DOCK8 (1/11)	NCF2 (1/2)	IL12RB1 (1/1)	CYBB (4/5)	DOCK8 (1/2)
	200	DOCK8 (1/11)	NCF2 (1/3)	IL12RB1 (1/1)	CYBB (4/5)	DOCK8 (1/3)
	500	DOCK8 (4/21)	NCF2 (1/2)	IL12RB1 (1/3)	CYBB (3/5)	DOCK8 (1/2)
HMZDelFinder	All	DOCK8 (1/2586)	NCF2 (120/120)	IL12RB1 (4/11)	CYBB (7/13)	DOCK8 (1/163)
HMZDelFinder AOH	All	DOCK8 (1/457)	NCF2 (37/37)	IL12RB1 (2/5)	CYBB (4/7)	DOCK8 (1/46)

Table 1: Comparison of the results between HMZDelFinder_opt and HMZDelFinder by using five positive controls carrying validated rare HMZ disease-causing deletions. Both HMZDelFinder_opt and HMZDelFinder (with or without AOH filtering step) detect the confirmed deletions. HMZDelFinder_opt detects a lower number of other deletions and ranks higher the confirmed deletion as compared to HMZDelFinder with or without AOH filtering step.

Detection of homozygous and hemizygous partial exon deletions by whole-exome sequencing

SUPPLEMENTARY INFORMATION

Benedetta Bigio[1,2,3], Yoann Seeleuthner[2,3], Gaspard Kerner[2,3], Melanie Migaud[2,3], Jérémie Rosain[2,3], Bertrand Boisson[1,2,3], Carla Nasca[4], Anne Puel[2,3], Jacinta Bustamante[1,2,3,5], Jean-Laurent Casanova[1,2,3,6,7], Laurent Abel[1,2,3,#,*], Aurelie Cobat[2,3,#,*]

[1]St. Giles Laboratory of Human Genetics of Infectious Diseases, Rockefeller Branch, The Rockefeller University, New York, NY 10065, USA

[2]Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Necker Hospital for Sick Children, 75015 Paris, France

[3]University of Paris, Imagine Institute, 75015 Paris, France

[4]Laboratory of Neuroendocrinology, The Rockefeller University, New York, NY 10065, USA

[5]Study Center of Immunodeficiencies, Necker Hospital for Sick Children, 75015 Paris, France

[6]Pediatric Hematology-Immunology Unit, Necker Hospital for Sick Children, 75015 Paris, France

[7]Howard Hughes Medical Institute, New York, NY 10065, USA

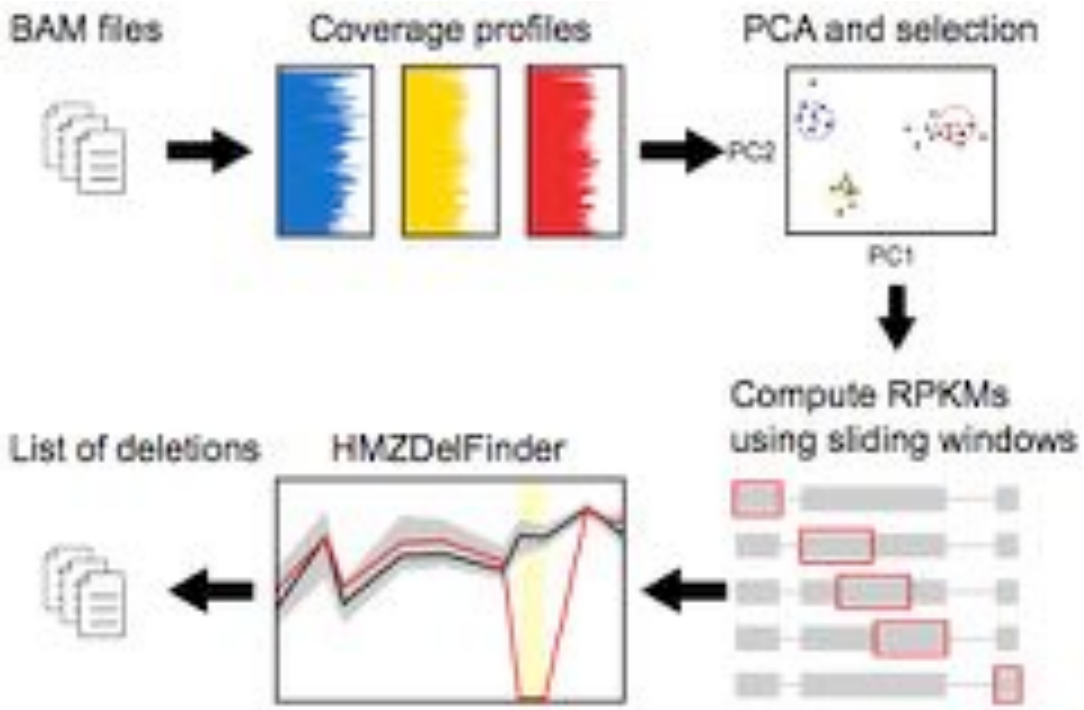
[#] indicates equal contributions

* To whom correspondence should be addressed.

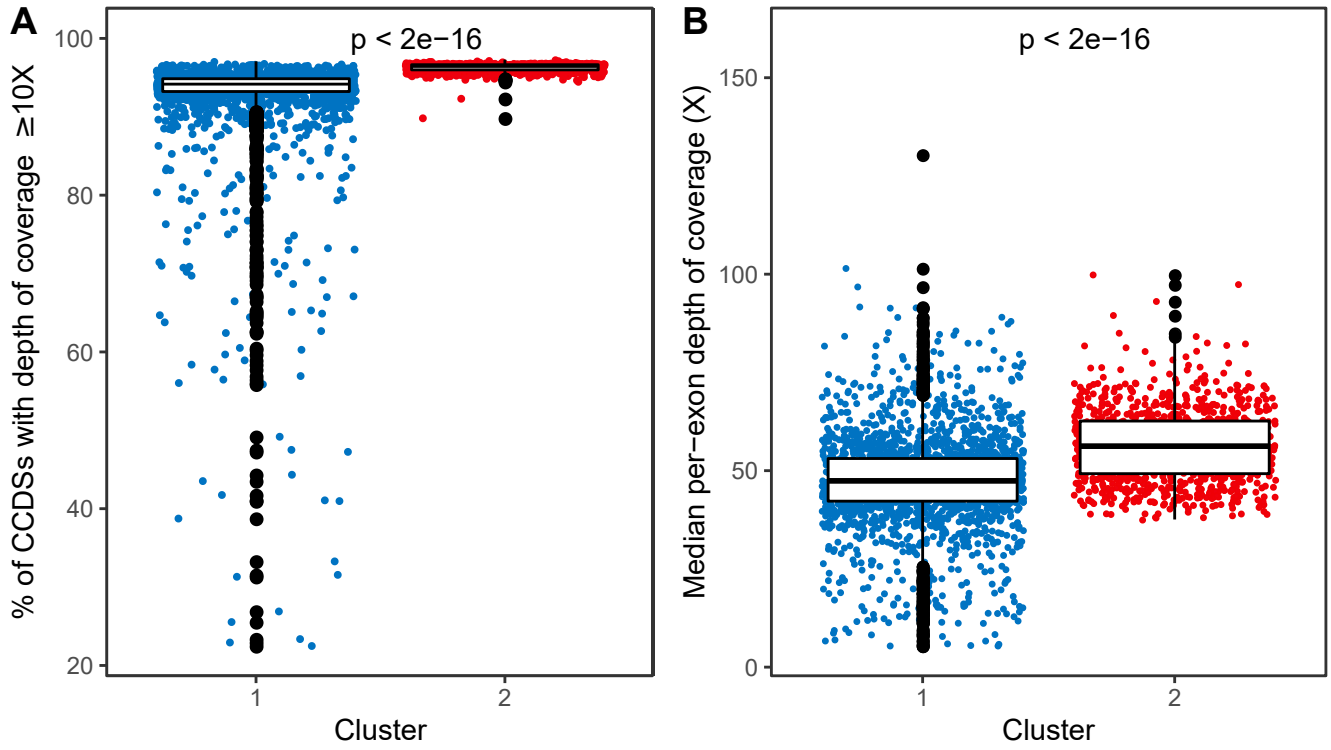
Tel: +33 1 42 75 43 14;

Fax: +33 1 42 75 42 24;

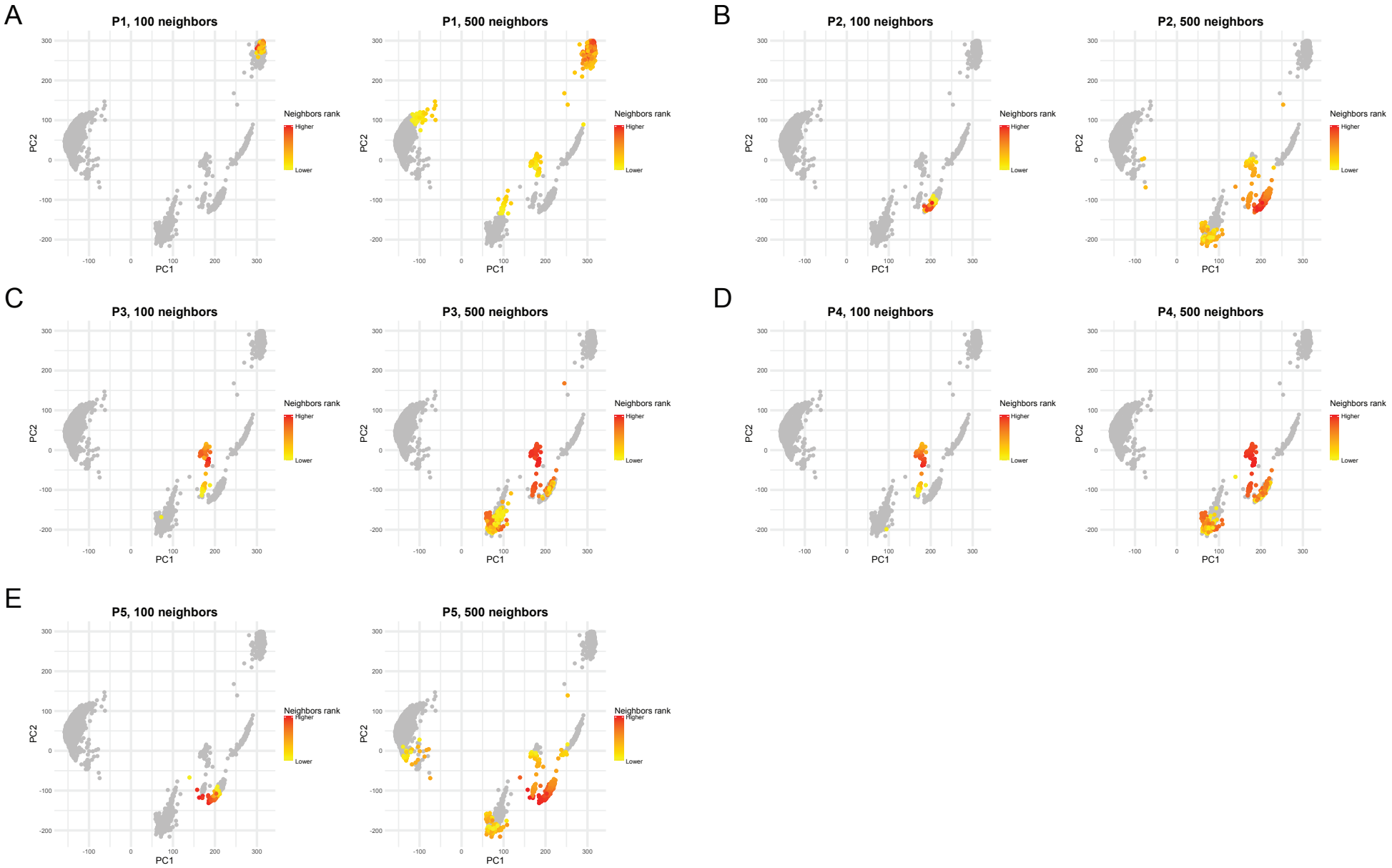
E-mail: aurelie.cobat@inserm.fr, laurent.abel@inserm.fr



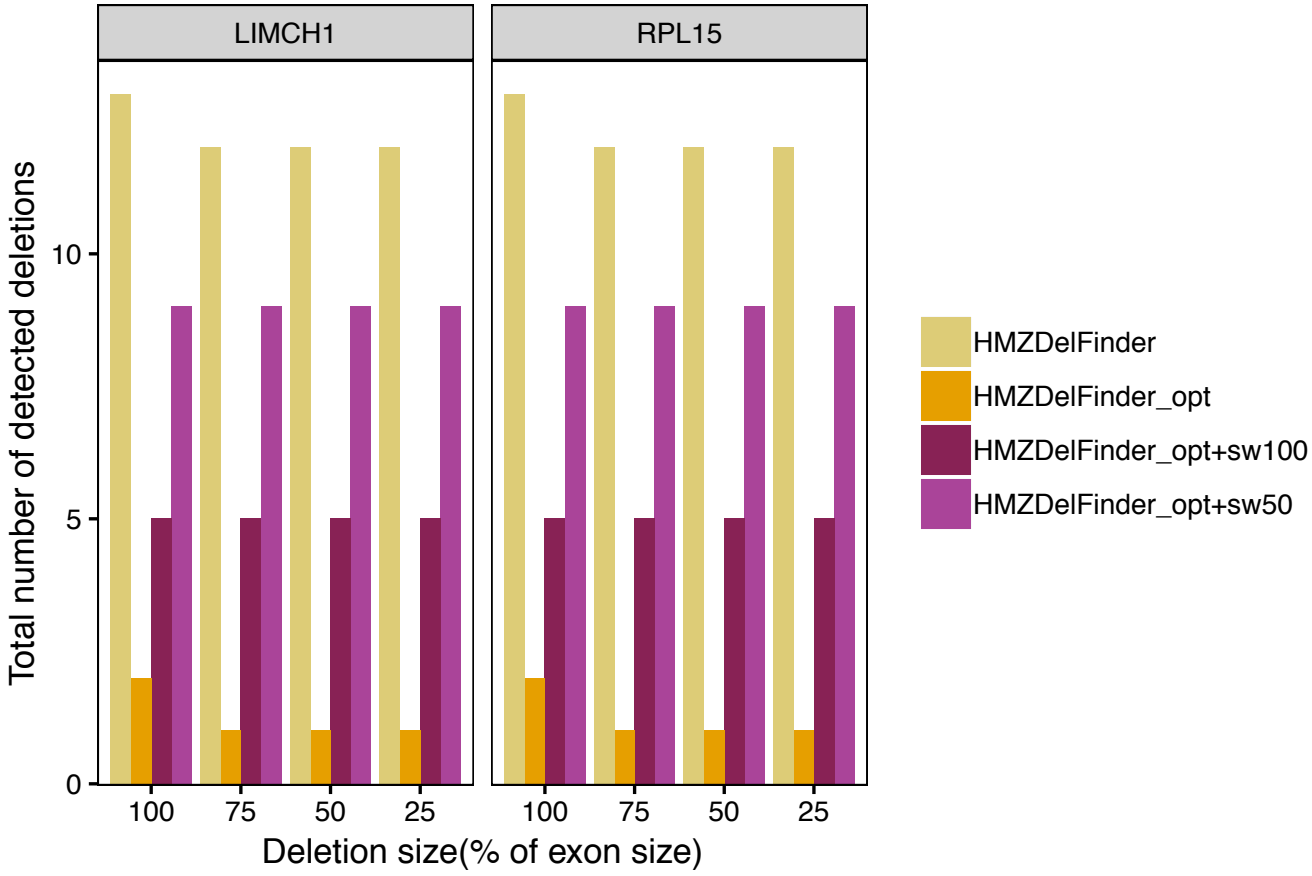
SI Figure 1: Schematic representation of the method employed by HMZDelFinder_opt to detect partial-exon homozygous and hemizygous deletions. First, HMZDelFinder_opt computes coverage profiles from the BAM files. The PCA is then calculated from a covariance matrix based on standardized coverage profiles and a k nearest neighbors algorithm is used to select the reference control set. The BAM file of a given sample and the BAM files of the reference control set are used as input of HMZDelFinder to detect homozygous and hemizygous deletions. In addition, HMZDelFinder_opt accepts a parameter (-sliding_window_size) to employ a sliding window approach for identification of partial-exon deletions.



SI Figure 2: Coverage in the two exome clusters revealed by PCA within the exomes generated by the V4-71Mb capture kit. Both the number of CCDSs with at least 10X (A) and the depth of coverage per exon (B) are significantly higher ($p<2^{-16}$) in the most recent V4-71Mb exomes (Cluster 2) than in the oldest V4-71Mb exomes (Cluster 1). Effect size: A, Cohen's d=0.45; B, Cohen's d=0.8.



SI Figure 3: Closest neighbors of the positive controls as function of the size of the reference control set. A total of 100 and 500 neighbors are showed for P1 (A), P2 (B), P3 (C), P4 (D), and P5 (E).



SI Figure 4: Median number of detected deletions in the simulated data in the higher (LIMCH1) or lower (RPL15) covered exons by using HMZDelfinder (yellow), HMZDelFinder_opt (orange), HMZDelFinder_opt+sw100 (red), HMZDelFinder_opt+sw50 (pink).

Patient	Confirmed Homozygous Deletion				Exome	
	Location	**Gene**	**Size (kbp)**	**Validation method**	**Mean Coverage**	**% Bases above 10%**
P1	Chr 9, Exons 21 to 23	DOCK8	10.8	MLPA	23	68.9
P2	Chr 1, Exon 5	NCF2	0.13	MLPA	115	99.5
P3	Chr 19, Exons 2 to 8	IL12RB1	13	Sanger sequencing	206	99.5
P4	Chr X, Whole gene	CYBB	3,400	MLPA and CGH array	156	99.2
P5	Chr 9, Exons 7 to 15	DOCK8	28	Sanger sequencing	66	99.5

SI Table 1: Validated rare HMZ disease-causing deletions and exome coverage in the five exomes used as positive controls.

Kit	Kit (full name)	Number (Percentage) of Exomes	Median Coverage (SD)	% bases above 10X
IDT-xGen	xGen Exome Research Panel v2 from Integrated DNA Technologies	188 (4.8%)	41.7 (9.5)	91.4
V4-50Mb	Agilent SureSelect Human All Exon V4	354 (9.0%)	50.0 (15.5)	83.2
V4-71Mb	Agilent SureSelect Human All Exon V4+UTRs	3095 (78.3%)	47.4 (10.2)	81.0
V5-50Mb	Agilent SureSelect Human All Exon V5	101 (2.6%)	72.4 (43.7)	70.3
V6-60Mb	Agilent SureSelect Human All Exon V6	216 (5.5%)	125.9 (38.6)	99.0

SI Table 2: Distribution of the capture kit in the 3,954 exomes and corresponding coverage metrics.

		P2	P3	P4	P5	TOTAL
	KIT	V6-60MB	V5-50MB	V5-50MB	V4-71MB	
METHOD	N NEIGHBORS	(COMMON DELETIONS/NUMBER OF OTHER DETECTED DELETIONS)				
HMZDelFinder_opt	50	0/1 (0%)	0/0 (-)	2/4 (50%)	2/2 (100%)	4/7 (60%)
	100	0/1 (0%)	0/0 (-)	2/4 (50%)	1/1 (100%)	3/6 (50%)
	200	0/2 (0%)	0/0 (-)	2/4 (50%)	2/3 (67%)	4/9 (44%)
	500	0/1 (0%)	0/2 (0%)	2/4 (50%)	1/1 (100%)	3/8 (38%)
HMZDelFinder	all	0/119 (0%)	0/10 (0%)	1/12 (8%)	2/162 (1%)	3/303 (1%)

SI Table 3: Number and percentage of common deletions (>1% frequency) among the detected deletions (other than the confirmed deletion)