PSAP-genomic-regions: a method leveraging population data to prioritize coding and non-coding variants in whole genome sequencing for rare disease diagnosis

Marie-Sophie C. Ogloblinsky[1,*], Ozvan Bocher[1,2], Chaker Aloui[3], Anne-Louise Leutenegger[3], Ozan Ozisik[4], Anaïs Baudot[4], Elisabeth Tournier-Lasserve[3,5], Helen Castillo-Madeen[6], Daniel Lewinsohn[6], Donald F. Conrad[6], Emmanuelle Génin[1,7,¶], Gaëlle Marenne[1,*,¶]

[1]Univ Brest, Inserm, EFS, UMR 1078, GGB, Brest, France

[2]Institute of Translational Genomics, Helmholtz Zentrum München, Munich, Germany

[3]Université Paris Cité, Inserm, NeuroDiderot, Unité Mixte de Recherche 1141, Paris, France

[4]Aix Marseille Univ, INSERM, Marseille Medical Genetics (MMG), Marseille, France

[5]Assistance publique-Hôpitaux de Paris, Service de Génétique Moléculaire Neurovasculaire, Hôpital Saint-Louis, Paris, France

[6]Division of Genetics, Oregon National Primate Research Center, Oregon Health & Science University, Portland, Oregon, United States of America

[7]Centre Hospitalier Régional Universitaire de Brest, Brest, France

*Corresponding authors:

Email: marie-sophie.ogloblinsky@inserm.fr (M-S.O.); gaelle.marenne@inserm.fr (G.M.)

[¶]These authors contributed equally to this work

Abstract

The introduction of next generation sequencing technologies in the clinics has improved rare disease diagnosis. Nonetheless, for very heterogeneous or very rare diseases, more than half of cases still lack molecular diagnosis. Novel strategies are needed to prioritize variants within a single individual. The PSAP method was developed to meet this aim but only for coding variants in exome data. Here, we propose an extension of the PSAP method to the non-coding genome called PSAP-genomic-regions. In this extension, instead of considering genes as testing units (PSAP-genes strategy), we use genomic regions defined over the whole genome that pinpoint potential functional constraints.

We conceived an evaluation protocol for our method using artificially-generated disease exomes and genomes, by inserting coding and non-coding pathogenic ClinVar variants in large datasets of exomes and genomes from the general population.

PSAP-genomic-regions significantly improves the ranking of these variants compared to using a pathogenicity score alone. Using PSAP-genomic-regions, more than fifty percent of non-coding ClinVar variants were among the top 10 variants of the genome. On real sequencing data from 6 patients with Cerebral Small Vessel Disease and 9 patients with male infertility, all causal variants were ranked in the top 100 variants with PSAP-genomic-regions.

By revisiting the testing units used in the PSAP method to include non-coding variants, we have developed PSAP-genomic-regions, an efficient whole-genome prioritization tool which offers promising results for the diagnosis of unresolved rare diseases.

Keywords: rare diseases, non-coding variants, whole-genome sequencing, variant prioritization

Introduction

Each rare disease affects, by definition, a small number of individuals. However, as a whole, rare diseases affect about 350 million people world-wide (1). Approximately 80% of rare diseases have a genetic origin that mostly follows a Mendelian mode of inheritance (2–4). The advent of Next Generation Sequencing (NGS) and the development of variant pathogenicity prediction tools have allowed, in recent years, the identification of many genes involved in rare Mendelian diseases. Nonetheless, despite extensive efforts, the molecular diagnosis is still unknown for more than 50% of rare diseases cases (5–7). This can mainly be explained by the fact that many rare diseases are characterized by an extreme genetic heterogeneity, which results in only one individual carrying a specific pathogenic causal variant. This issue is referred to as the "n-of-one" problem (8).

With the advent of high throughput sequencing technologies in clinics, molecular diagnosis is now often sought through whole exome or whole genome sequencing (WES and WGS respectively). However, due to the large number of rare variants in each individual genome, causal variants are sought among very rare and highly pathogenic variants in genes relevant to the current known disease mechanism. The limited knowledge about gene functions and disease mechanisms can make this strategy unfruitful. To address the issue of variant prioritization at the level of an individual, the Population Sampling Method (PSAP) (8) was developed. PSAP computes, for each gene, a null distribution, which is the probability to observe in the general population a genotype with a CADD pathogenicity score (9) greater than or equal to the highest one to the highest one observed in the patient for this gene. This initial version of the PSAP method, which we will refer to as PSAP-genes, has been successfully applied to identify variants of interest in diverse phenotypes, including male infertility (10–12), recurrent pregnancy loss (13) and ciliary diskynesia (14).

A current hindrance to the application and generalization of PSAP-genes as a tool for diagnosis is its restriction to the coding parts of the genome. Indeed, the majority of variants reside in non-coding parts of the genome (15). Non-coding variants may contribute to explain part of the etiology of rare diseases (16), as suggested by the large number of GWAS hits located in non-coding regions of the genome (17). The involvement of non-coding pathogenic variants in rare diseases is further corroborated by the fact that non-coding regions are heavily involved in the regulation of gene expression. Several prediction tools have been developed to this end (18–20), but most of them lack a variant-based score for both coding and non-coding regions. In addition, to be performant, they often require multiple annotations like Human Phenotype Ontology (HPO) terms (21) to characterize the symptoms or disease of a patient . Thus, they rely on previous knowledge and rarely go beyond candidate genes.

To move beyond the gene as a natural unit of testing for the PSAP method, we need to use predetermined regions across the whole genome. These regions also need to be defined using functional information to be used as a cohesive unit for the construction of PSAP null distributions. This challenge of defining regions along the whole genome has been tackled by Bocher et al. in the context of rare-variant association testing (22): they describe CADD regions, which are characterized by a lack of observed variants with high functionally-Adjusted CADD Scores (ACS) in the gnomAD database (23). CADD regions are expected to reflect functional constraints. CADD regions present the key advantage of providing pre-defined and functionally-informed regions which can be used to construct PSAP null distributions.

We have made available a new implementation of the PSAP method using Snakemake (24) workflows, called Easy-PSAP (https://github.com/msogloblinsky/Easy-PSAP), which features null distributions constructed with up-to-date allele frequency data and pathogenicity scores. Here, we introduce PSAP-genomic-regions, an extension of the PSAP method to the non-coding genome by using the pre-defined CADD regions as testing unit instead of genes. This is an innovative strategy to prioritize variants at the scale of an individual genome. PSAP-genomic-regions is now available in Easy-PSAP. We

devised an evaluation protocol using artificially-generated disease exomes and genomes, obtained by inserting coding and non-coding ClinVar (25) variants in general population whole genomes from the 1000 Genomes Project (26) and exomes from the FrEnch EXome (FREX) project (27). We show the consistent improvement in prioritization by using PSAP-genomic-regions over pathogenicity scores alone for non-coding and then coding variants. For coding variants, we also demonstrate the good performance of PSAP-genomic-regions compared to PSAP-genes. On real-life data, we illustrate the power of PSAP-genomic-regions on WES data from six resolved cases of Cerebral Small Vessel Disease (CSVD) and WGS data from three families affected by male infertility. These two diseases are particularly relevant to test our method, monogenic forms of CSVD (28) and male infertility (29) being extremely heterogeneous.

Results

Construction of PSAP null distribution in coding and non-coding regions

The idea behind the original PSAP method, referred to as PSAP-genes, relies on the calculation of gene-specific null distributions of CADD pathogenicity scores. More precisely, for an individual exome or genome and in a given gene, PSAP-genes considers the genotype with the highest CADD score and evaluates the probability to observe such a high CADD score in this gene in the general population. PSAP-genes deals separately with heterozygote and homozygote variants in the autosomal dominant (AD) and the autosomal recessive (AR) models respectively. Here, we will focus on homozygote variants for the recessive model. As a result, PSAP-genes gives a p-value to the genotype with the highest CADD score in the gene for each gene, model, and individual. PSAP can also score compound heterozygote variants, i.e. two heterozygote variants in the same gene, thus also giving a PSAP p-value to the genotype with the second highest CADD score in the gene. This p-value allows the ranking of the genes for an individual exome or genome. The PSAP principle can be generalized to any genomic unit.

Here, with PSAP-genomic-regions, we extended the PSAP method to analyze whole-genome data using predefined CADD regions as testing units instead of genes (Fig 1). The same principle as before is employed, with the difference being that the genotype with the highest CADD score in the region can be coding or non-coding. We thus constructed PSAP-genomic-regions null distributions using the CADD pathogenicity score (PHRED scaled across the whole genome). Our novel strategy will be referred to as PSAP-genomic-regions-CADD. We also explored the use of another pathogenicity score, the ACS (22) (PHRED scaled CADD scores by “coding”, “regulatory” and “intergenic” regions) to mitigate the higher CADD scores of coding variants (PSAP-genomic-regions-ACS strategy). The PSAP-genomic-regions strategies were compared to the initial PSAP-genes strategy, also referred to as PSAP-genes-CADD.

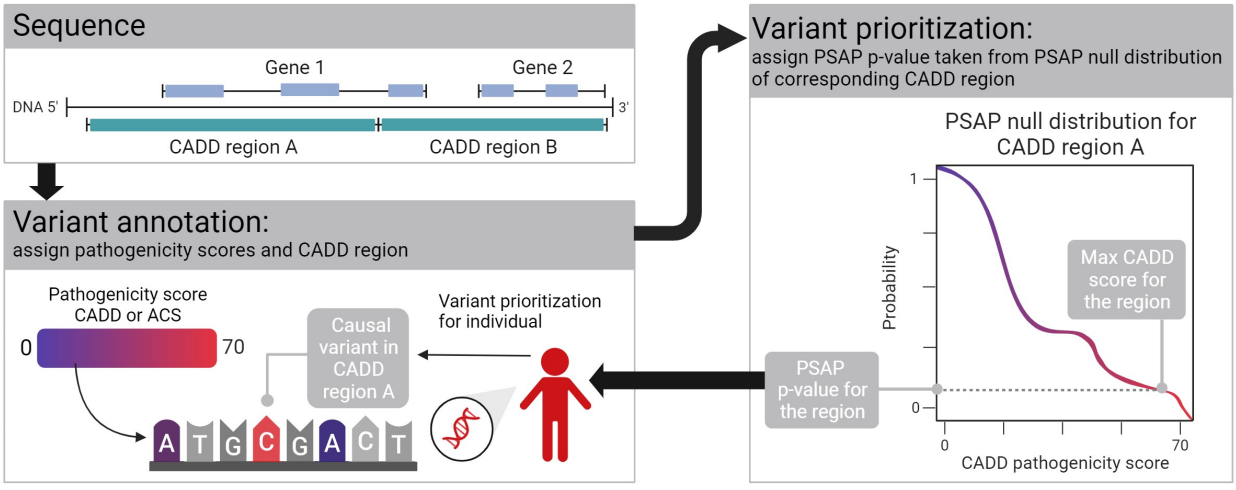

Fig 1. Description of the PSAP-genomic-regions strategy.

We calculated PSAP null distributions for SNVs in genes and CADD regions, in the hg19 and hg38 assemblies of the human genome. In hg19, PSAP null distributions were obtained for 19,283 genes and 119,695 CADD regions. In hg38 PSAP null distributions were obtained for 18,395 genes and 123,991 CADD regions. PSAP null distributions and their parameters (unit of testing, allele frequencies and pathogenicity score) can be found in S1 Table.

Evaluating the performance of PSAP-genomic-regions on artificially-generated disease exomes and genomes using ClinVar variants

Prioritization of non-coding pathogenic variants

First, to evaluate how PSAP-genomic-regions performed to prioritize non-coding pathogenic variants, we used artificially-generated disease genomes created by inserting non-coding ClinVar variants in the Non-Finnish Europeans (NFE) from the 1000 Genomes Project phase 4 (NFE genomes) (see Material & Methods and S1 File for the list of variants). Because the 1000 Genomes project is population-based, we

expect that some individuals might carry one or a few pathogenic variants in their genome. These pathogenic variants are characterized by a high CADD score and a low PSAP p-value. Thus, in order to summarize the rank of a ClinVar variant in an evaluation setting, we considered the best rank reached by the variant in at least 90% of the individuals.

Most of the NFE genomes carried a variant with a higher pathogenicity score or a lower PSAP p-value than most of the ClinVar variants (S1 Fig). We thus compared the percentage of the non-coding pathogenic variants ranked among the top N (N = 1, 10, 50 and 100) in at least 90% of the NFE genomes. The ranking at the individual level was done among all heterozygous variants for the ClinVar variants under the AD model, and across homozygous variants for the ClinVar variants under the AR model. Our main strategy PSAP-genomic-regions-CADD performed systematically better than using the CADD score alone (Fig 2A). The improvement was especially large for the top 10 ranking: 24.6% and 79.2% of ClinVar variants reached the top 10 with PSAP-genomic-regions-CADD for the AD and AR models, respectively, while no ClinVar variant reached the top 10 with CADD scores alone. For the prioritization of coding variants, the PSAP-genomic-regions-CADD strategy always outperformed PSAP-genomic-regions-ACS (S2 Fig).

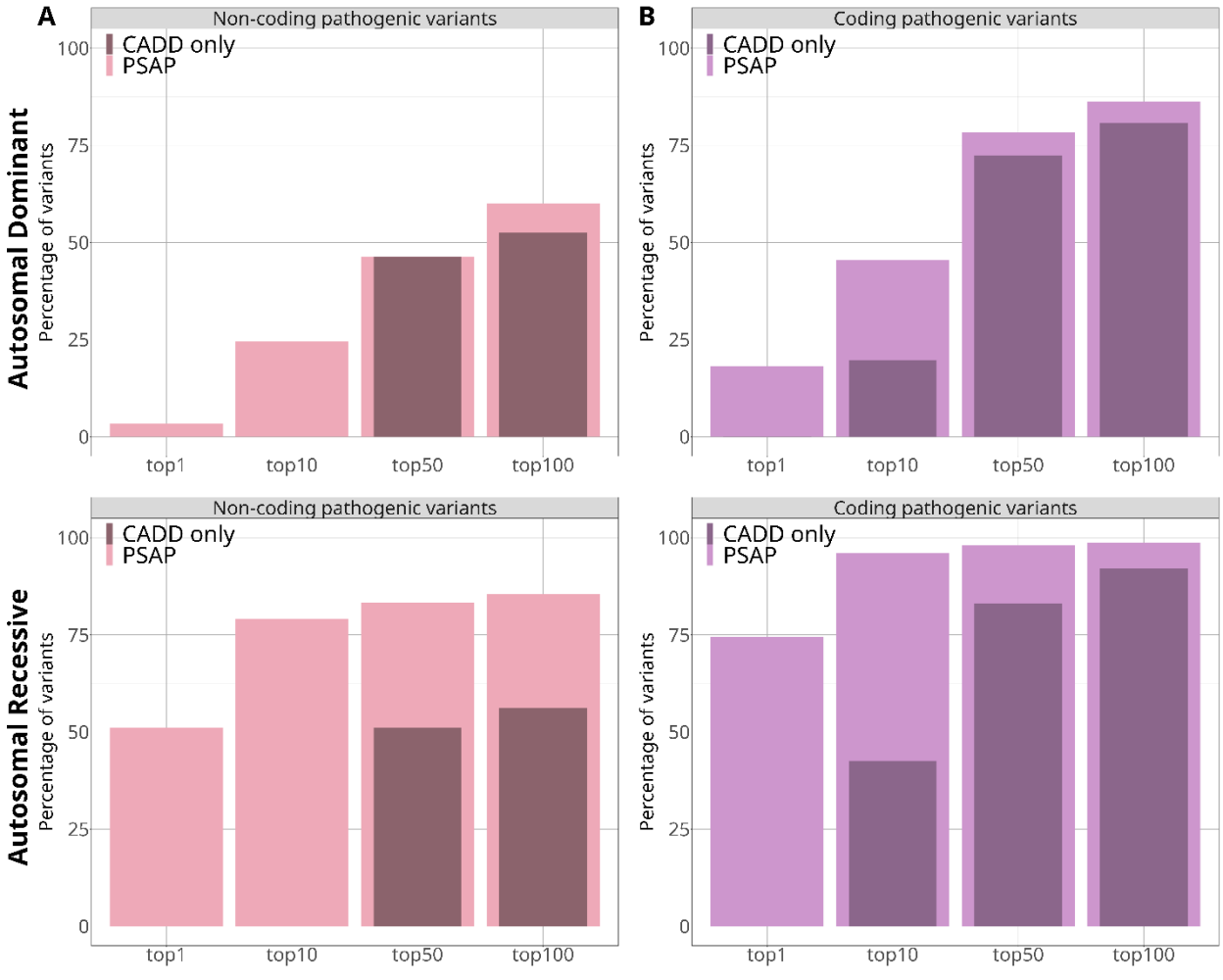

Fig 2. Comparison of the PSAP-genomic-regions-CADD strategy versus the CADD score alone in artificially-simulated disease genomes. Percentage of non-coding and coding pathogenic ClinVar variants reaching the top N of variants in at least 90% of NFE genomes, with PSAP-genomic-regions (darker shade of pink or purple) or the CADD score alone (lighter shade of pink or purple) (A) N = 175 non-coding AD variants and N = 96 non-coding AR variants (B) N = 4,965 coding AD variants and N = 2,680 coding AR variants.

Using the ACS scores improved the performance to detect non-coding-variants for the AD model (S2 Fig): 56.6% and 24.6% of variants reached the top 10 with PSAP-genomic-regions-ACS and PSAP-genomic-

regions-CADD, respectively. The gain in performance with PSAP-genomic-regions-ACS compared to PSAP-genomic-regions-CADD is not significant for the AR model for the top 10, top 50 and top 100. Nonetheless, we can note the pattern is different for the top 1 for the AR model: 51% with PSAP-genomic-regions-CADD to 5.5% with PSAP-genomic-regions-ACS. Indeed, switching from CADD score to ACS score has lowered the PSAP p-value of non-coding variants shared by more than 10% of NFE genomes. This led to a defect of the top rank reached by the ClinVar variants, as we considered the lowest rank reached in at least 90% of individuals. For instance, a variant in the CADD region R109138 shared by 70 of the NFE genomes went from a CADD score of 18.1 and a PSAP-genomic-regions-CADD p-value of 0.1 to an ACS of 22.2 and a PSAP-genomic-regions-ACS p-value of 5.18×10^{-10}. Thus, the ClinVar variants inserted in these individuals having a higher p-value than 5.18×10^{-10} do not rank first. Considering the overall more consistent performance of the PSAP-genomic-regions-CADD strategy, we chose to focus on this strategy, although we provide comparison with the PSAP-genomic-regions-ACS strategy which can have advantages for non-coding variants.

We further explored PSAP results for splicing ClinVar variants versus other type of non-coding ClinVar variants. Indeed, we observed that splicing variants are the major type of non-coding ClinVar variants. These splicing variants often had a very good ranking, especially with PSAP-genomic-regions-ACS (n=115 splicing variants among 175 non-coding AD variants and n=72 splicing variants among 96 non-coding AR variants; S3 Table; Panel A in S3 Fig). Splicing ClinVar variants have a much higher ACS than CADD scores (Panel B in S3 Fig) which results in better ranking than for other types of non-coding ClinVar variants using PSAP-genomic-regions-ACS p-values (Panel C in S3 Fig). As a consequence, the percentage of splicing ClinVar variants ranked in the top 10 was largely improved when using PSAP-genomic-regions-ACS, for the AD model especially which was less powerful with PSAP-genomic-regions-CADD to begin with (Panel D in S3 Fig).

The full results of ranking by PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS for the non-coding non-splicing pathogenic ClinVar variants can be found in S2 File. With PSAP-genomic-regions-ACS, around half of the non-coding non-splicing variants are ranked in the top 50 and top 10 of variants for more than 90% of NFE genomes for the AD and AR models, respectively (19 out of 45 variants for the AD model and 8 out of 21 variants for the AR model). The other half of variants present a less significant PSAP-genomic-regions-ACS p-value and a poorer ranking. PSAP-genomic-regions-CADD achieves a similar ranking of AR non-coding non-splicing variants (7 out of 21 variants) but a decreased prioritization for AD non-coding non-splicing variants (6 out of 45 variants). To confirm this pattern of ranking for non-coding non-splicing pathogenic variants on another set of variants, we evaluated with our artificially generated disease genomes protocol 320 non-coding SNVs used to train Genomiser (30). These variants were not associated with a mode of inheritance. Hence, we inserted them in the NFE genomes and scored them with both AD and AR PSAP-genomic-regions-ACS null distributions. Among the 320 non-coding variants, 169 reached the top 100 in at least 90% of NFE genomes, with either the AD or AR model (S3 File). This can be explained by the distributions of CADD scores compared to ACS scores for the ClinVar variants: the non-coding variants that do not reach the top 100 have a significantly lower CADD and ACS scores compared to all the other types of variants (S4 Fig). Overall, PSAP-genomic-regions prioritizes around half of non-coding ClinVar and Genomiser training variants in the top 100 of NFE genomes. The ones who have a higher ranking present much lower CADD and ACS scores and would never be well-ranked by any PSAP strategy.

PSAP-genomic-region is also relevant for the analysis of exome data. Indeed, exome sequencing captures variants outside of the bounds of coding regions (31), such as intronic variants. We explored the prioritization of non-coding ClinVar variants located within the WES-targeted regions of the FREX individuals using our artificially-generated disease exomes protocol (N=48 variants for the AD model and N=64 variants for the AR model, Panel A in S5 Fig). For both PSAP-genomic-regions-CADD and PSAP-

genomic-regions-ACS, there was a large increase in prioritization performance compared to using only the pathogenicity scores. Because there are fewer variants in an exome background than in a genome background, the rankings of these non-coding ClinVar variants were better in FREX than in NFE genomes. The best ranking was achieved using PSAP-genomic-regions-ACS, with 82% and 90.3% of variants reaching the top 10 for the AD and AR models, respectively, whilst PSAP-genomic-regions-CADD achieved a similar ranking for AR variants. Most of these non-coding pathogenic variants were splicing variants (40 out of 73 variants for the AD model and 56 out of 64 variants for the AR model), and half of them were considered as having a functional "HIGH IMPACT" (26 variants for the AD model and 22 variants for the AR model). Hence, prioritizing variants with PSAP-genomic-regions allows identifying more variants even in exome data, that are in addition functionally-relevant.

Prioritization of coding pathogenic variants

Similar evaluations were performed for ClinVar coding variants inserted in either WGS from 1000G NFE individuals or WES from FREX. As observed for non-coding pathogenic variants, PSAP-genomic-regions outperformed the pathogenicity scores alone (Fig 2B, Panel B in S5 Fig). However, in the context of coding pathogenic ClinVar variants, we observed that the strategy of PSAP-genomic-regions-CADD provided better prioritization compared with the PSAP-genomic-regions-ACS strategy. We observed that 18.2% and 74.6% of the coding variants reached the top 1 in at least 90% of genomes backgrounds with the PSAP-genomic-regions-CADD for the AD and AR model respectively, against no variants with the CADD score alone, and against 5.3% and 2.5% reaching the top 1 with PSAP-genomic-regions-ACS. In the exome background and with PSAP-genomic-regions-CADD, 38.7% and 89.8% of AD variants reached the top 1 and top 50, respectively; 80.3% and 97.9% of AR variants reached the top 1 and the top 50, respectively.

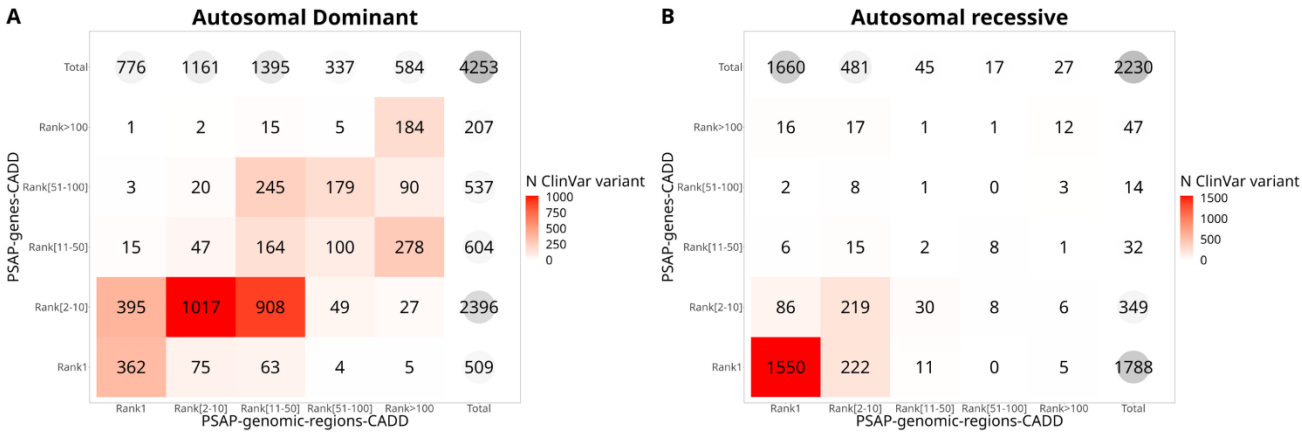

Fig.3. Comparison of PSAP-genomic-regions-CADD and PSAP-genes-CADD strategies in artificially-simulated disease genomes. Number of coding pathogenic ClinVar variants reaching rank [x-y] of variants in at least 90% of 1000 Genomes Project NFE individuals for each strategy.

We also compared the number of coding ClinVar variants reaching the tops in NFE genomes between PSAP-genomic-regions-CADD strategy and the initial PSAP-genes-CADD strategy (Fig 3). More differences were observed across the two PSAP strategies for the AD than for the AR model (Fig 3A). There were 362 variants ranked first and 1,017 variants ranked [2-10] in common between the two strategies. However, 908 variants that were ranked [2-10] with PSAP-genes-CADD were [11-50] with PSAP-genomic-regions-CADD, and 395 variants that were ranked [2-10] with PSAP-genes-CADD were ranked first with PSAP-genomic-regions-CADD. Regarding variants that are ranked more than a 100 with PSAP-genomic-regions-CADD, 278 of them are ranked [11-50] and 90 are ranked [51-100] by PSAP-genes-CADD. Regarding the AR model (Fig 3B), PSAP-genomic-regions-CADD performed similarly to PSAP-genes-CADD, and the majority of variants were ranked first with both strategies (1,550 variants). Even more promising results can be found when looking at the same comparison of ranks within the FREX exomes (S6 Fig). For instance, in the AD model, 592 variants that were ranked [2-10] with PSAP-genes-CADD are ranked first with PSAP-

246 genomic-regions-CADD, against 115 variants ranked [2-10] with PSAP-genomic-regions-CADD that
247 become first with PSAP-genes-CADD.

248

249 Application of PSAP-genomic-regions to real data with different modes
250 of inheritance

251 To illustrate our method in real-life settings, we analyzed two datasets (S4 Table), one with an AD
252 mode of inheritance and the other with an AR mode of inheritance. The first dataset consisted of WES
253 data for six individuals affected by monogenic forms of CSVD (32). Using PSAP-genomic-regions-CADD, all
254 of the causal variants were ranked at least in the top 100 in each patient (Fig 4). The contribution of CADD
255 regions as a unit of testing was especially visible for the variant in *COL4A2* and one variant in *HTRA1* which
256 were not well-ranked using genes as testing unit (rank 110 and 193 respectively with genes, and rank 3
257 and 69 with CADD regions). Using their maximal CADD score by gene or CADD region alone, these variants
258 would not have been prioritized in the top 100 for five out of six individuals.

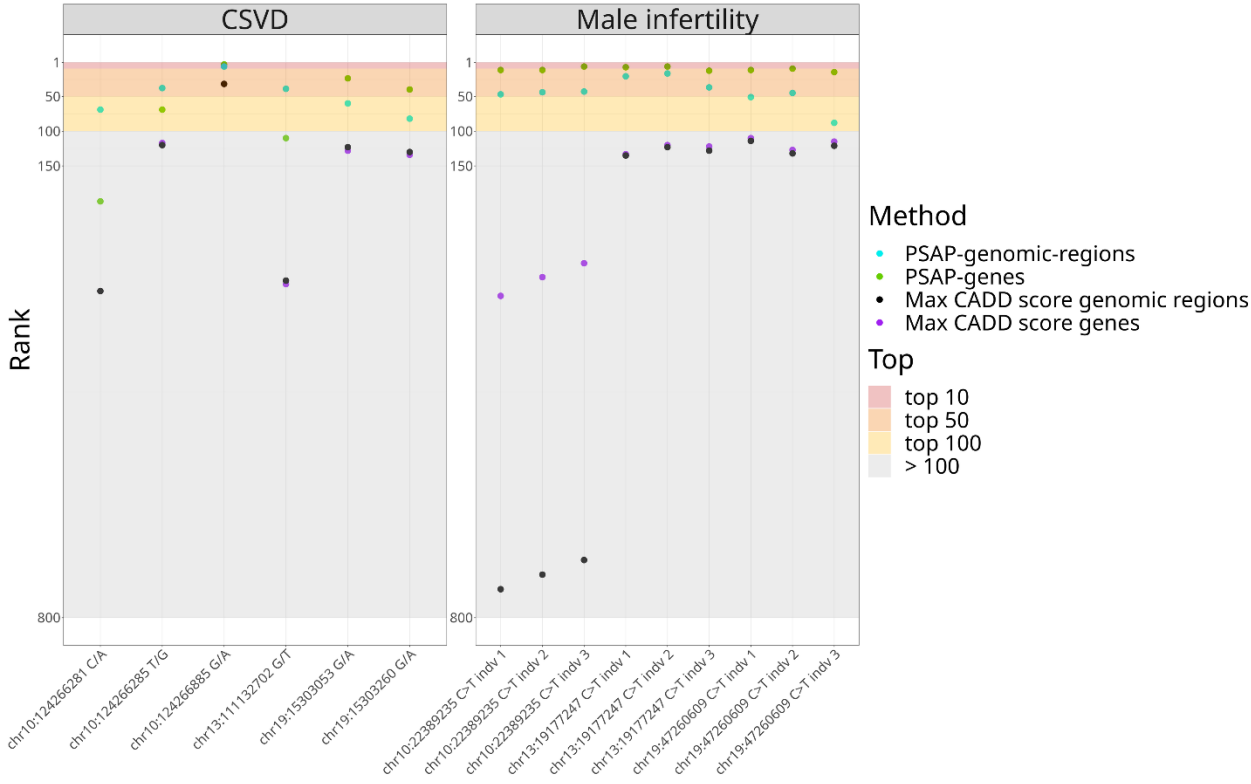

Fig. 4. Prioritization of 6 known CSVD mutations and 3 male infertility candidate variants with PSAP-genomic-regions-CADD, PSAP-genes-CADD and the maximal CADD score on genes or CADD regions.

The second dataset consisted of WGS data for 9 individuals from three families with clinically diagnosed male infertility (33). All causal variants fell within the top 20 of variants with prioritization by PSAP-genes-CADD, and within the top 50 for at least one case per family with PSAP-genomic-regions-CADD (within top 100 for all cases, Fig 4). PSAP-genomic-regions-CADD did not improve the ranking of these coding variants, which was expected considering the large number of variants in a WGS analysis (see S4 Table for the total number of variants in each analysis). The prioritization from PSAP-genomic-regions-CADD was still interesting to narrow the set of candidates for causal variants. In clinics when the CADD score alone is used, these variants would not have been prioritized (CADD score < 25, and rank > 100 with the maximal CADD score strategy). PSAP-genomic-regions-CADD thus allow a

relevant prioritization of coding pathogenic variants in WGS sequencing and an unbiased exploratory analysis at the scale of the whole genome.

Using PSAP-genomic-regions-ACS or the ACS score alone, almost all of the CSVD and male infertility coding pathogenic variants had a rank greatly exceeding the top 100 (S4 Table). The only exception is one variant in *HTRA1* (10:124266885 G/A) that was ranked 3 by PSAP-genomic-regions-ACS and 10 by the maximal ACS score alone. This *HTRA1* variant was a splicing variant, which confirms the good performance of the PSAP-genomic-regions-ACS strategy on this type of variant.

Discussion

Variant prioritization, especially in the case of very heterogeneous rare diseases, is a clinically-relevant methodological challenge for both clinicians and researchers. Mounting evidence suggests that current methods of analysis and their restriction to the coding genome are a hindrance to the discovery of new genetic variants implicated in rare diseases (16). We have developed PSAP-genomic-regions, an extension of the PSAP method to the whole genome using functionally-relevant genomic regions. PSAP-genomic-regions broadens the scope of variants evaluated by PSAP and addresses the issue of variant prioritization at an individual whole-genome scale.

PSAP-genomic-regions has been thoroughly tested and validated by using simulations emulating real-life scenarios of causal variant prioritization. PSAP-genomic-regions achieves a prioritization of coding pathogenic SNVs in the top 100 variants of an exome or genome which is a relevant number of variants to analyze for clinicians. Without use of prior knowledge on the disease, PSAP-genomic-regions achieves relevant variant prioritization within millions of variants to analyze, which is illustrated by the ranking of 6 variants involved in CSVD and 3 variants involved in familial cases of male infertility in the top 100 of WES and WGS data respectively. PSAP-genomic-regions thus helps with the diagnosis of such heterogeneous diseases in conjunction with other relevant information like the mode of transmission, prevalence or type of variant involved.

PSAP-genomic-regions also allows the scoring of variants otherwise discarded from the analysis, like splicing variants with a high predicted functional impact, and other non-coding variants of proven clinical significance. The only scenario for which PSAP-genomic-regions is not advantageous compared to the PSAP-genes strategy is for prioritizing coding variants in WGS data. In that case, using coding CADD regions, i.e. the coding parts of CADD regions for the analysis still yields better results compared to PSAP-genes (S7 Fig). Our simulations using known pathogenic variants have shown which PSAP strategy

performs the best depending on the type of data and variant expected to be involved in the disease mechanism. If there is no expected type of variant: we advise on the use of the PSAP-genomic-regions-CADD strategy, which gives the overall best results. For coding variants prioritization specifically, PSAP-genomic-regions-CADD gives the best results in WES, and PSAP-coding-genomic-regions-CADD performs best in WGS data. Finally, if no coding variant of interest for the disease is found with PSAP-genomic-regions-CADD or PSAP-coding-genomic-regions-CADD, PSAP-genomic-regions-ACS can be applied to look for non-coding variants of interest especially for an AD expected model of transmission.

To the best of our knowledge, there is no other score of predicted pathogenicity for all possible SNVs comparable to CADD. The main pathogenicity prediction scores developed to date were described and compared in a recent review (34). Multiple benchmarks on the subject show conflicting conclusions depending on the variant testing set (35,36). A significant limitation of some of the most popular tools, such as SIFT (37), PolyPhen-2 (38), VEST (39), and REVEL (40), is their restriction to analyzing only missense variants. In contrast, CADD stands out as it is a meta-predictor, integrating scores from SIFT, PolyPhen-2, phyloP (41), and GERP (42), and enabling the scoring of any SNV with pre-computed scores, and any InDel in the genome with on-request scores. Additionally, CADD is trained on a much larger number of variants compared to other machine-learning methods, while using a relatively modest number of features. Similar types of methods aim at prioritizing more constrained regions in the non-coding genome (18,20) or distinguishing deleterious non-coding variants from neutral ones (18,43). However, most of these prediction methods either do not provide a variant-specific score, or are not defined in both coding and non-coding parts of the genome. Other well-known methods for identification of pathogenic variants in exome and genome data rely on the use of HPO terms to make a prediction, like Exomiser (44) or Genomiser (30), making in comparison PSAP-genomic-regions an unmatched prioritization tool. As any other bioinformatics variant prioritization method, it has to be used in conjunction with other lines of evidence like the expression of the associated gene in a tissue of interest or segregation of variants if

familial data is available to ultimately lead to any genetic diagnosis of a patient. PSAP-genomic-regions does not make assumption on the type of variants and does explore the whole genome. The ranking by p-values coming from the application of PSAP-genomic-regions to an individual's variants is a useful way to narrow-down the list of variants to further investigate for both researchers and clinicians in different scenarios. Other criteria for variant filtering will depend heavily on the type of disease studied. The clinical interpretation of pathogenicity for non-coding variants is more challenging than for coding variants and can be improved by applying modified ACMG guidelines which can help pick out potentially candidate regulatory elements to explain the phenotype of the patient (45).

The method most comparable to the strategy followed by PSAP-genomic-regions is the recently-developed machine-learning algorithm FINSURF (46). FINSURF aims to predict the functional impact of non-coding variants in regulatory regions and has been applied to known pathogenic variants inserted in WGS data like we did. Nonetheless it has been difficult to compare properly the two methods considering FINSURF only scores non-coding variants in predefined regulatory regions, and the set of variants used to train the method is not available.

The main limitation of PSAP-genomic-regions comes from the score used to calibrate null distributions, namely the CADD score. We have observed that known pathogenic non-coding ClinVar variants that were not well-ranked by PSAP-genomic-regions had significantly lower CADD and ACS scores compared to splicing and better-ranked non-coding variants. Because such CADD score is likely to be seen in the general population, PSAP-genomic-regions will not be able to prioritize such a variant with at a low rank. We also observed that some CADD regions were badly-calibrated and resulted in the assignment of very low PSAP-genomic-regions p-values to putatively neutral variants in the 1000 Genomes Project. As allele frequencies from larger databases and more accurate pathogenicity scores become available, this will lead to an improvement of the PSAP method as well. The most recent release of the CADD score v1.7

(47) notably integrates regulatory annotations and may further improve the prioritization of non-coding pathogenic variants when integrated in PSAP-genomic-regions.

Many avenues of further development and improvement are open for PSAP-genomic-regions, including the inclusion and scoring of InDel variations and structural variants. Exploring the combination of the PSAP-genomic-regions p-values with other metrics or information coming from omics analysis could also improve prediction. Finally, the flexibility of the PSAP method makes it potentially adaptable to other more complex models like digenic and oligogenic models of inheritance, considering the increasing availability of information coming from gene networks and biological pathways.

Materials and Methods

Construction of PSAP null distributions

The first parameter is the units in which to construct the PSAP null distribution. Here we considered two unit strategies: the genes and the CADD regions (S1 Table). For the genes, the coding regions of genes were defined based on the biomaRt R package: the gene coding sequences were retrieved from Ensembl (48) by requesting the "genomic_coding_start" and "genomic_coding_end", on both the hg19 and hg38 builds. To account for splicing regions, the coding regions were extended by two bases on both sides of the gene coding regions. In total, 19,780 genes were retrieved in hg19 and 23,163 in the hg38 build. For the CADD regions, their coordinates were downloaded from https://lysine.univ-brest.fr/RAVA-FIRST/ for the hg19 build and were lifted over to hg38 using the Ensembl Assembly Converter. CADD regions coordinates in hg38 are available on Easy-PSAP GitHub (https://github.com/msogloblinsky/Easy-PSAP). There were 135,224 CADD regions in hg19 and 131,970 in hg38. For the coding CADD regions, i.e. the

371 coding parts of CADD regions, we considered the intersection of the CADD regions and the gene coding
372 regions for each build, which yielded 37,978 coding CADD regions in hg19 and 52,340 in hg38.

373 The second parameter is the allele frequencies database. Here we considered the global allele
374 frequencies from the gnomAD database to calibrate the PSAP null distributions: gnomAD genome r2.0.1
375 for hg19 and gnomAD V3 (49) for hg38. For our purpose, we considered only single nucleotide variants
376 (SNVs) annotated as PASS by the Variant Quality Score Recalibration (VQSR) of GATK (50) and located in
377 well-covered regions. Well-covered regions in gnomAD genome were defined as regions for which 90% of
378 individuals have coverage at depth 10. Variants not seen in gnomAD genome, not annotated as PASS or
379 not located in well-covered regions (gnomAD genome version according to the build) have a frequency of
380 0 and thus did not contribute to the construction of the null distributions.

381 To ensure reliability of PSAP null distribution, it is crucial that the units are well covered in the
382 database from which the allele frequencies are taken. Thus, we only considered units for which at least
383 half of the unit was well-covered (as defined previously) in gnomAD genome (version according to the
384 build). Coding regions of genes and well-covered regions in gnomAD genome were intersected to get the
385 percentage of each gene's coding regions that were well-covered in the database. The same steps were
386 carried out with CADD regions as genomic units for PSAP, for hg19 and hg38 builds. PSAP null distributions
387 were thus constructed for 19,283 and 18,395 genes in hg19 and hg38 respectively, 119,695 and 123,991
388 CADD regions, and 34,397 and 35,226 coding CADD regions in hg19 and hg38 respectively.

389 The third parameter is the pathogenicity score. Here, for the evaluation of PSAP on coding variants,
390 we used the version 1.6 of CADD (51) for each build, accessible on the CADD website
391 (https://cadd.gs.washington.edu/). For the evaluation on non-coding variants, which tend to have lower
392 CADD scores than coding variants (52), we followed the strategy described in Bocher et al.(22) to adjust

393 the RAW CADD score v1.6 of all possible SNVs on a PHRED scale stratifying by type of genomic regions:
394 “coding”, “regulatory” and “intergenic”, resulting in “adjusted CADD scores”, referred to as “ACS”.

395 Easy-PSAP (https://github.com/msogloblinsky/Easy-PSAP) was used to generate null distributions
396 according to the previously described input files and parameters. This resulted in 4 sets of null
397 distributions for the AD and AR models for both hg19 and hg38 assemblies (S1 Table).

398

399 Evaluating the performance of PSAP-genomic-regions using artificially-
400 generated disease exomes and genomes

401 To evaluate the ability of PSAP-genomic-regions to prioritize known pathogenic variants in an
402 individual, we leveraged artificially-generated disease exomes and genomes using available general
403 population cohorts. These different PSAP strategies (see Table 1) were compared in terms of their
404 performances to prioritize the known pathogenic variants.

405 The pathogenic ClinVar (25) SNVs with coordinates in hg19 and hg38 were downloaded from the NCBI
406 website (https://www.ncbi.nlm.nih.gov/clinvar/, accessed on the 3rd of June 2022). Some of these ClinVar
407 variants had an annotated mode of inheritance ("moi autosomal recessive" and "moi autosomal
408 dominant"). From ClinVar, there were 14,056 variants annotated as AD and 12,758 variants annotated as
409 AR. Variants were filtered out to keep only autosomal pathogenic SNVs having as review status either
410 “reviewed by expert panel” or “criteria provided, multiple submitters, no conflicts”, which are the two
411 best review status in ClinVar. There were 1,518 AD and 1,118 AR variants meeting these criteria.

412 For variants which did not have an annotated mode of inheritance, we used a curated version of the
413 database OMIM, hOMIM (53) to retrieve a mode of inheritance, and kept variants that were always
414 associated with an AD or AR mode of inheritance in hOMIM. The same filtering was applied, which left

3,641 additional variants for the AD and 1,706 for the AR model. In total, we had a set of 5,159 variants for the AD model and 2,824 variants for the AR model. Among these ClinVar variants, 4,965 and 2,680 variants were coding SNVs respectively for the AD and AR models. Similarly, 175 and 96 variants were non-coding variants for the AD model and AR models, among which 48 variants for the AD model and 64 for AR model fell within the boundaries covered by FREX exomes. The list of pathogenic ClinVar variants and their mode of inheritance can be found in S1 File.

We inserted each variant from our curated list of pathogenic ClinVar variants successively in each of the 533 high coverage NFE genomes and each of the 574 exomes from the FREX project. An individual-focused QC was applied on both datasets using the RAVAQ R package (54): we performed a genotype and variant QC with default parameters corresponding to standard GATK hard filtering criteria, mean allele balance computed across heterozygous genotypes and call rates, except for MAX_AB_GENO_DEV = 0.25, MAX_ABHET_DEV, MIN_CALLRATE and MIN_FISHER_CALLRATE "disabled".

We conducted the artificially-generated disease genome and exome evaluation with PSAP null distributions in hg19 and hg38 respectively, to match with the build of the data. We then applied the 3 PSAP strategies mentioned previously (PSAP-genes-CADD, PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS). For each strategy, we kept the maximal pathogenicity score (CADD or ACS) for each unit (gene or CADD regions) and then ranked the units according to their PSAP p-value or to their pathogenicity score alone within each genome or exome. We compared the PSAP-genes-CADD and PSAP-genomic-regions-CADD strategies to using the maximal CADD score alone by gene or CADD regions, respectively; and the PSAP-genomic-regions-ACS strategy to using the maximal ACS score by CADD region. For each ClinVar variant, we retrieved its rank within each genome or exome. Coding ClinVar variants were evaluated with the 3 PSAP strategies whereas non-coding ClinVar variants were evaluated with the novel PSAP-genomic-regions-CADD and PSAP-genomic-regions-ACS strategies (see S2 Table for more details).

Patient data analysis

The PSAP strategies were applied to real WES data from six unrelated patients affected by a CSVD for which the causal variant is known, which allowed a comparison of performance between the different strategies. The full description of the dataset can be found in [Aloui et al. 2021] (32), with the exception of the QC process. For this analysis, the same QC as for the FREX and 1000 Genomes Project datasets was performed. We applied PSAP-genes-CADD and PSAP-genomic-regions-CADD in hg19 to the six resolved CSVD patients' exome data. The other PSAP parameters were the ones by default as described previously. Two of the individuals had a causal pathogenic variant in the gene *NOTCH3* (19:15303053 G/A and 19:15303260 G/A), one individual in the gene *COL4A2* (13:111132702 G/T) and three individuals in the gene *HTRA1* (10:124266285 T/G, 10:124266281 C/A and 10:124266885 G/A). The rank of the known CSVD variants among other heterozygote variants in the patient's exome according to its PSAP p-value for the 2 strategies was then retrieved.

The PSAP strategies were also applied to WGS data of three families with clinically diagnosed forms of male infertility (33) and for which a pathogenic recessive variant was prioritized using a computational pipeline featuring the initial PSAP-genes implementation. Three affected individuals were analyzed for each family. The description of the whole dataset and candidate variant filtering process can be found in [Khan and Akbari et al. 2023] (33), except for the QC that was performed in the same way as for the CSVD data. Two other families were resolved from the same dataset, but considering that the causal variants were deletions we did not include them in the current analysis. The prioritized pathogenic variants were in the genes: *SPAG6* (chr10:22389235 C/T) for family 3, *TUBA3C* (chr13:19177247 C/T) for family 7 and *CCDC9* (chr19:47260609 C/T) for family 4. We applied PSAP–genes-CADD and PSAP-genomic-regions-

460 CADD in hg38 to the 9 cases and retrieved the rank of the known male infertility variants among other

461 homozygote variants in the patient's genomes according to its PSAP p-value for the 2 strategies.

References

1. Sequeira AR, Mentzakis E, Archangelidi O, Paolucci F. The economic and health impact of rare diseases: A meta-analysis. Health Policy and Technology. 2021 Mar 1;10(1):32–44.

2. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM®). Nucleic Acids Research. 2009 Jan 1;37(suppl_1):D793–6.

3. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research. 2015 Jan 28;43(D1):D789–98.

4. Ehrhart F, Willighagen EL, Kutmon M, van Hoften M, Curfs LMG, Evelo CT. A resource to explore the discovery of rare diseases and their causative genes. Sci Data. 2021 May 4;8(1):124.

5. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nat Rev Genet. 2018 May;19(5):253–68.

6. Boycott KM, Rath A, Chong JX, Hartley T, Alkuraya FS, Baynam G, et al. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. The American Journal of Human Genetics. 2017 May 4;100(5):695–705.

7. Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. The American Journal of Human Genetics. 2015 Aug 6;97(2):199–215.

8. Wilfert AB, Chao KR, Kaushal M, Jain S, Zöllner S, Adams DR, et al. Genomewide significance testing of variation from single case exomes. Nat Genet. 2016 Dec;48(12):1455–61.

9. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Mar;46(3):310–5.

10. Wyrwoll MJ, Temel ŞG, Nagirnaja L, Oud MS, Lopes AM, van der Heijden GW, et al. Bi-allelic Mutations in M1AP Are a Frequent Cause of Meiotic Arrest and Severely Impaired Spermatogenesis Leading to Male Infertility. The American Journal of Human Genetics. 2020 Aug 6;107(2):342–51.

11. Kasak L, Punab M, Nagirnaja L, Grigorova M, Minajeva A, Lopes AM, et al. Bi-allelic Recessive Loss-of-Function Variants in FANCM Cause Non-obstructive Azoospermia. The American Journal of Human Genetics. 2018 Aug 2;103(2):200–12.

12. Salas-Huetos A, Tüttelmann F, Wyrwoll MJ, Kliesch S, Lopes AM, Conçalves J, et al. Disruption of human meiotic telomere complex genes TERB1, TERB2 and MAJIN in men with non-obstructive azoospermia. Hum Genet. 2021 Jan;140(1):217–27.

13. Kasak L, Rull K, Yang T, Roden DM, Laan M. Recurrent Pregnancy Loss and Concealed Long-QT Syndrome. J Am Heart Assoc. 2021 Aug 16;10(17):e021236.

14. Bustamante-Marin XM, Horani A, Stoyanova M, Charng WL, Bottier M, Sears PR, et al. Mutation of CFAP57, a protein required for the asymmetric targeting of a subset of inner dynein arms in Chlamydomonas, causes primary ciliary dyskinesia. PLoS Genet. 2020 Aug 7;16(8):e1008691.

15. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences. 2009 Jun 9;106(23):9362–7.

16. Posey JE. Genome sequencing and implications for rare disorders. Orphanet Journal of Rare Diseases. 2019 Jun 24;14(1):153.

17. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019 Jan 8;47(Database issue):D1005–12.

18. Gussow AB, Copeland BR, Dhindsa RS, Wang Q, Petrovski S, Majoros WH, et al. Orion: Detecting regions of the human non-coding genome that are intolerant to variation using population genetics. PLOS ONE. 2017 Aug 10;12(8):e0181604.

19. Huang YF, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. 2017 Apr;49(4):618–24.

20. Vitsios D, Dhindsa RS, Middleton L, Gussow AB, Petrovski S. Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning. Nat Commun. 2021 Mar 8;12(1):1504.

21. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. Am J Hum Genet. 2008 Nov 17;83(5):610–5.

22. Bocher O, Ludwig TE, Oglobinsky MS, Marenne G, Deleuze JF, Suryakant S, et al. Testing for association with rare variants in the coding and non-coding genome: RAVA-FIRST, a new approach based on CADD deleteriousness score. PLOS Genetics. 2022 Sep 16;18(9):e1009923.

23. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

24. Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. Bioinformatics. 2012 Oct 1;28(19):2520–2.

25. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D1062–7.

26. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

27. Génin E, Redon R, Deleuze J, Campion D, Lambert J, Dartigues J, et al. The French Exome (FREX) Project: A Population-based panel of exomes to help filter out common local variants. Genetic Epidemiology. 2017;41(7):691–691.

28. Rannikmäe K, Henshall DE, Thrippleton S, Ginj Kong Q, Chong M, Grami N, et al. Beyond the Brain. Stroke. 2020 Oct;51(10):3007–17.

29. Houston BJ, Riera-Escamilla A, Wyrwoll MJ, Salas-Huetos A, Xavier MJ, Nagirnaja L, et al. A systematic review of the validated monogenic causes of human male infertility: 2020 update and a discussion of emerging gene–disease relationships. Human Reproduction Update. 2022 Feb 1;28(1):15–29.

30. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet. 2016 Sep 1;99(3):595–606.

31. Guo Y, Long J, He J, Li CI, Cai Q, Shu XO, et al. Exome sequencing generates high quality data in non-target regions. BMC Genomics. 2012 May 20;13:194.

32. Aloui C, Hervé D, Marenne G, Savenier F, Le Guennec K, Bergametti F, et al. End-Truncated LAMB1 Causes a Hippocampal Memory Defect and a Leukoencephalopathy. Annals of Neurology. 2021;90(6):962–75.

33. Khan MR, Akbari A, Nicholas TJ, Castillo-Madeen H, Ajmal M, Haq TU, et al. Genome sequencing of Pakistani families with male infertility identifies deleterious genotypes in SPAG6, CCDC9, TKTL1, TUBA3C, and M1AP. Andrology. 2023 Dec 10;

34. Garcia FA de O, Andrade ES de, Palmero EI. Insights on variant analysis in silico tools for pathogenicity prediction. Frontiers in Genetics [Internet]. 2022 [cited 2024 Feb 14];13. Available from: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.1010327

35. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, et al. Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. 2018 Sep 6;46(15):7793–804.

36. Anderson D, Lassmann T. An expanded phenotype centric benchmark of variant prioritisation tools. Hum Mutat. 2022 May;43(5):539–46.

37. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1;31(13):3812–4.

38. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Current Protocols in Human Genetics. 2013;76(1):7.20.1-7.20.41.

39. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics. 2013;14 Suppl 3(Suppl 3):S3.

40. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. The American Journal of Human Genetics. 2016 Oct 6;99(4):877–85.

41. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010 Jan;20(1):110–21.

42. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol. 2010 Dec 2;6(12):e1001025.

43. Caron B, Luo Y, Rausell A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. Genome Biology. 2019 Feb 11;20(1):32.

44. Smedley D, Jacobsen JOB, Jager M, Köhler S, Holtgrewe M, Schubach M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015 Dec;10(12):2004–15.

45. Ellingford JM, Ahn JW, Bagnall RD, Baralle D, Barton S, Campbell C, et al. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. Genome Medicine. 2022 Jul 19;14(1):73.

46. Moyon L, Berthelot C, Louis A, Nguyen NTT, Crollius HR. Classification of non-coding variants with high pathogenic impact. PLOS Genetics. 2022 Apr 29;18(4):e1010191.

47. Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. Nucleic Acids Research. 2024 Jan 5;52(D1):D1143–54.

48. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic Acids Research. 2022 Jan 7;50(D1):D988–95.

49. Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes [Internet]. bioRxiv; 2022 [cited 2023 Aug 30]. p. 2022.03.20.485034. Available from: https://www.biorxiv.org/content/10.1101/2022.03.20.485034v2

50. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

51. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine. 2021 Feb 22;13(1):31.

52. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research. 2019 Jan 8;47(D1):D886–94.

53. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, et al. Natural selection on genes that underlie human disease susceptibility. Curr Biol. 2008 Jun 24;18(12):883–9.

54. Marenne G, Ludwig TE, Bocher O, Herzig AF, Aloui C, Tournier-Lasserve E, et al. RAVAQ: An integrative pipeline from quality control to region-based rare variant association analysis. Genetic Epidemiology. 2022;46(5–6):256–65.