Aggregated genomic data as cohort-specific allelic frequencies can boost variants and genes prioritization in non-solved cases of inherited retinal dystrophies

Ionut-Florin Iancu[1,2], Irene Perea-Romero[1,2], Gonzalo Núñez-Moreno[1,2,3], Lorena de la Fuente[1,3], Raquel Romero[1,2], Almudena Ávila-Fernandez[1,2], María José Trujillo-Tiebas[1,2], Rosa Riveiro-Álvarez[1,2], Berta Almoguera[1,2], Inmaculada Martín-Mérida[1,2], Marta Del Pozo-Valero[1,2], Alejandra Damián-Verde[1], Marta Cortón[1,2], Carmen Ayuso[1,2]# and Pablo Minguez[1,2,3]#

[1]Department of Genetics, Health Research Institute–Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain.

[2]Center for Biomedical Network Research on Rare Diseases (CIBERER), ISCIII, Madrid, Spain.

[3]Bioinformatics Unit, Health Research Institute–Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain.

Corresponding authors: cayuso@fjd.es and pablo.minguez@quironsalud.es

Abstract

Background

The introduction of next-generation sequencing in the diagnosis of genetic diseases has increased the known repertoire of causal variants and genes involved, as well as the amount of genomic information produced, that is not always shared or reused.

Methods

We built an allelic frequency database for a heterogeneous cohort of genetic diseases to explore the aggregated genomic information and boost the diagnosis in inherited retinal dystrophies (IRD). We retrospectively selected 5683 index-cases with clinical exome available, 1766 with IRD, and the rest with diverse genetic diseases. In our IRD cohort, 46% of the patients do not have conclusive diagnosis at the time of writing. We calculated the specific allele-frequencies of the solved and non-solved IRD subcohorts and compared them with suitable pseudocontrols that were used to prioritize variants. In addition, we developed a method to highlight genes with more frequent pathogenic

variants in non-solved IRD cases than in pseudocontrols weighted by the increment of benign variants in the same comparison. Our resource was also used to calculate the carrier frequency of deleterious variants in IRD genes.

Results

Prioritized variants were significantly enriched in deleterious variants in non-solved IRD cases but not in solved. Focusing on non-solved IRD cases, we prioritized variants with a significant increment of frequencies in cases compared to pseudocontrols. Among them, eight variants may contribute to explain the phenotype of 10 cases. Applied to variants of uncertain significance (VUS) monitored in our laboratory, we detected 11 more frequent in IRD than in pseudocontrols, and 10 of them were reclassified as likely-pathogenic according to ACMG guidelines. We also identified 18 genes with an accumulated pathogenicity in non-solved IRD samples for further studies that provided new insights in five cases. Most prevalent genes carrying pathogenic mutations are *ABCA4* (~7%) and *USH2A* (~3%).

Conclusions

A cohort-specific database of allele frequencies is able to diagnose cryptic non-solved IRD cases, reclassify VUS, propose candidate genes, and calculate CF on genes of interest. The database operates as an engine providing new hypotheses in non-solved cases as well as offering new resources for genetic counselling.

Background

Rare diseases are chronically debilitating or life-threatening, and have a prevalence in Europe of less than 1 of every 2000 people [1]. Inherited retinal dystrophies (IRD) are a group of rare diseases with a degenerative and progressive course and are caused by primary affection of photoreceptors and retinal pigmentary epithelial [2]. All together they affect 1 of every 3000-4000 people in the western world [3]. They are clinically heterogeneous, covering several syndromes (e.g., Usher, Bardet-Biedl -BBS-, or Joubert) [4–6], as well as non-syndromic forms as retinitis pigmentosa [2] and macular

dystrophies [7]. They have overlapped phenotypes and display any form of inherited patterns.

During the last two decades, next generation sequencing (NGS) techniques have transformed research on genetic rare diseases with a substantial increase in volume of available genomic data and knowledge generated [8,9]. Although several sequencing tests are available, a widespread approach in genetic diagnosis is to sequence the coding region of known clinically relevant genes (~4500), the so-called clinical exome (CE). A CE test detects several thousand variants, which need to be filtered and prioritized in order to highlight those responsible of the phenotype [10]. Regarding the task of filtering, in low prevalent diseases, apart from quality filters, a low population frequency is one of the first requirements to purge non-causal variants [11]. Several global genomic initiatives [12,13] provide allele frequencies on large populations, although frequencies from local cohorts provide a better estimation of real variant prevalence [14,15] and have proven to identify rare pathogenic variants [15,16].

In the absence of genomic information of *a priori* healthy people, Mendelian diseases can provide a good estimate of allele frequencies in the general population as pseudocontrols (PC) for other non-related diseases [17,18]. In the same terms, PC can also be applied for the calculation of carrier frequencies (CF) [16,19] of causal variants of non-related diseases in genes with a recessive inheritance pattern, as well as in the analysis of trios [18,19]. In more complex scenarios, where modifying and risk/protective variants may tune the effect of causal variants, the mutational landscape of a disease may help identifying: i) genetic pleiotropy together with causal variants in recessive forms [22], ii) digenic inheritance [23], or iii) disease-associated triallelic sites as in BBS [24].

With all these premises, we hypothesize that a database of variant allele frequencies calculated over a heterogeneous cohort of genetic diseases enriched in IRD cases can help to improve the detection of previously unnoticed, underrated or unknown causal variants and gene-disease associations. Additionally, it can assist to uncover disease

cases with overlapping phenotypes, as well as to describe the carrier frequencies of recessive variants. Thus, we built a database with the genomic data of a large cohort with various genetic diseases and developed methods to compare IRD-specific and PC frequencies. This tool is used as a global reanalysis platform to study frequent variants and over-mutated genes in IRD non-solved cases.

Methods

Cohort description

We retrospectively selected all index cases (N=5683) with a clinical exome test performed as a first tier approach at the Genetics and Genomics Department of the University Hospital Fundación Jiménez Díaz (UH-FJD, Madrid, Spain) from September 2015 to May 2021. The cohort included patients suffering from genetic diseases classified in 14 categories, with the largest disease group being IRD, with 1766 samples (Additional file 1:Table S1). The rest of the diseases were grouped in “other eye related diseases” (OERD) and “non-eye related diseases” (NRD). Based on the diagnostic status set by the molecular geneticists after CE inspection, all IRD cases were classified as: i) solved cases, when one or two pathogenic or likely pathogenic variants were found in dominant and recessive genes, respectively, ii) monoallelic cases, when only one pathogenic or likely pathogenic variant was found in a recessive gene in cases where this mode of inheritance was suspected; iii) cases with variants of uncertain significance (VUS) called here “VUS cases” with one VUS in dominant or 1-2 VUS in recessive genes; and iv) non-solved cases, when no potentially pathogenic or VUS variants were found. Numbers for monoallelic and VUS cases are provided together (Additional file 1:Table S2).

Sequencing tests

Samples were analyzed using targeted DNA sequencing with two different commercial sequencing panels: TruSightOne Sequencing Panel kit (TSO, Illumina, San Diego, CA),

and Clinical Exome Solution Sequencing Panel kit (CES, Sophia Genetics, Boston, MA). CES panel targets a total of 4828 genes and regulatory regions and TSO targets 4813 genes, with an overlap of 3567 genes between both panels (Additional file 1:Figure S1).

Bioinformatics reanalysis

In order to have a homogeneous variant calling and annotation of all sequencing tests, all sequenced data was reanalyzed using a custom bioinformatics pipeline for both single nucleotide variants (SNVs) and small insertions and deletions (indels) [25]. This pipeline included exonic, intronic, and UTR analysis. For variant calling, we included a 1000 base pairs padding for each target region, for both, TSO and CES clinical exome tests. The pipeline is based on GATK 4.1 variant caller [26], and uses BWA-MEM aligner [27] to the GRCh37/hg19 reference genome. The following databases were used for annotation: i) allele frequency: gnomAD [17], 1000genomes [28], and Kaviar [29]; ii) pathogenicity prediction: SIFT [30], PolyPhen [31], CADD [32], LRT [33], M-CAP [34], MetaLR [35], MetaSVM [35], MutationAssesor [36], MutationTaster [37], PROVEAN [38], and FATHMM [39]; iii) splicing prediction: ada_score [40] and rf_score [40]; iv) ClinVar [41]; v) conservation: phastCons20way [42] and phyloP20way [43]; vi) gene tolerance to loss of function (LoF) variants: LoFtool [44], and ExACpLI [12]; vii) constrained coding regions by means of gnomAD_CCR [45]; and viii) potential loss of heterozygosity regions, which were annotated with PLINK [46]. The pipeline is available at https://github.com/TBLabFJD/VariantCallingFJD.

Detection and removal of sample duplicates and cryptic relatedness

All known sample duplicates and relatives were removed prior to frequency calculation. In order to detect other possible sample duplicates and relatives, PLINK whole genome association analysis toolkit [46] was used to calculate inbreeding coefficients (identity-Identity-by-descent, IBD). First, single nucleotide polymorphisms (SNPs) pruning was

performed removing SNPs covered in less than 95% of the samples (PLINK parameter: geno 0.05), with less than 5% allelic frequency (PLINK parameter: maf 0.05), and in linkage disequilibrium (PLINK parameters: indep-pairwise 50 5 0.5). With the resulting SNPs, IBD was calculated for all sample combinations (PLINK parameter: genome). All samples with a PI_HAT score higher than 0.35 were removed.

Variant frequency calculation for IRD patients and pseudocontrols

After identifying and removing sample duplicates and relatedness (N=5683), variants (in vcf format) from the index-cases with a gnomAD MAF<0.1 were processed together and merged into a single multi-vcf file. Sequencing coverage was also calculated for each sample to distinguish between non-covered and non-mutated sites.

We developed in-house routines for allelic frequency calculation based on Hail python library for genomics data exploration and analysis (https://hail.is). The allele frequency (AF), allele number (AN), allele count (AC) and homozygotes count were obtained for the general cohort and for several subcohorts composed of IRD cases: 1) all IRD cases, 2) non-solved IRD cases, 3) solved IRD cases, 4) syndromic IRD cases. 5) non-syndromic IRD cases, and 6) macular dystrophies cases. To define the subcohort of IRD PC for all IRD subcohorts, we took samples from the subcohort NRD (N=3531). AF, AN and AC were calculated from this PC subcohort (PC-AF, PC-AN and PC-AC).

Definition of genes associated to IRD, OERD and NRD

Three disease specific gene panels were used in the inspection of variants and genes: i) IRD gene panel (244 genes; including 136 genes for syndromic IRD and 108 genes for non-syndromic IRD) as the virtual gene panel used in the diagnosis of IRD cases in the Genetics and Genomics Department of the UH-FJD, and extracted using RetNet, HGMD and literature searches (Additional file 1:Table S3), ii) OERD genes, including non-IRD genes with ocular phenotype (all genes linked with the HPO term “Eye Disease” – HP:0000478, N=1542 genes, Additional file 1:Table S4), and iii) NRD genes

(the rest of the genes included in TSO/CES panels, not related with eye diseases, N=3260, Additional file 1:Table S5). The genes included in each panel are listed in Tables S3-S5. Genes which are recommended by ACMG to report in case of secondary findings [47] (Additional file 1:Table S6) were excluded from the gene panels and analyses but the gene *RPE65* that belongs to the IRD panel.

Variants discarded for analysis

Variants detected in the 5683 samples from our general cohort were further filtered out using two criteria: i) quality filtering, we removed 5% of variants with lowest AN, and ii) population filtering, in order to discard a population origin bias in our IRD subcohort compared to the rest of the cohort, we keep variants present in IRD solved or non-solved cases, and assuming no differences in population origin between IRD solved and non-solved cases, having a fold change between non-solved-AF and solved-AF >90% percentile, from them we rescue those having a non-solved-AF and solved-AF <0.1.

Determination of differentially frequent variants in IRD subcohorts compared to pseudocontrols

We define differentially frequent variants as those that have a higher frequency in a subcohort compared with a control subcohort. In order to extract variants differentially frequent in the IRD subcohorts (solved IRD, non-solved IRD, syndromic IRD, non-syndromic IRD and macular dystrophies) compared to the IRD PC subcohort, we calculated the FC of the AF in the IRD subcohort compared to the PC-AF for each of the variants. Based on the distribution of the log2 of FCs (log2(FC)) of all variants, we selected those above the 90% as the significant differential frequent variants in a subcohort, tagged as IRD-MFV (IRD most frequent variants) for any IRD subcohort. Variants annotated by ClinVar as “pathogenic” or “likely pathogenic” or with a CADD_PHRED ≥ 30 (top 0.1% most deleterious variants according to CADD) were

classified as deleterious, and variants annotated by ClinVar as “benign” or “likely benign” were classified as benign. We compared the proportion of deleterious variants in the IRD-MFV group with the rest of the variants (non-prioritized variants) for solved-IRD and non-solved-IRD cases. Furthermore, we performed this comparison grouping IRD cases as: syndromic forms, non-syndromic forms and macular dystrophies. A Fisher’s exact test was applied to compare the proportion of deleterious variants in these groups, p-value<0.05 was taken as significant.

VUS reclassification

We selected VUS whose reclassification can determine the diagnosis of an IRD case in our cohort. These VUS are reported in the diagnostic process at the Genetics Department of the UH-FJD if no pathogenic or likely pathogenic variant are found associated to the phenotype. Variants are classified using ACMG guides. In the IRD subcohort, there were reported 100 VUS fulfilling these criteria [10] (Additional file 1:Table S7). Of these, 63 fulfilled criteria to be within the database generated in this work, and were still classified as VUS according to ACMG (information taken from VarSome at the time of the analysis). For these VUS, we annotated IRD-AF (general IRD cohort) and PC-AF frequencies, calculated the FC for these two frequencies and selected two sets: 1) VUS with a log2(FC) >=1.5 (N=11), VUS with a log2(FC)>= 2.48 (value of the 90 percentile of the distributions of the log2(FC), N=6). For all selected VUS (N=11) we marked the specific ACMG criterion PS4 for which (“The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls”) and applied previous evidences to obtain a new ACMG classification.

Gene prioritization for IRD association

To prioritize genes in non-solved IRD cases, we selected for each gene included in the

database: i) deleterious variants annotated by ClinVar as “pathogenic” or “likely pathogenic” or with a CADD_PHRED ≥ 30 (top 0.1% most deleterious variants according to CADD), and ii) benign variants annotated by ClinVar as “benign” or “likely benign”. Genes with at least five deleterious and five benign variants were selected for further analysis. For each selected variant, a log2(FC) was calculated between non-solved IRD-AF and PC-AF. Finally, we applied Wilcoxon rank sum test to the distribution of log2(FC) for deleterious and benign variants in each gene. P-values were adjusted using FDR and genes with an adjusted p-value<0.05 were considered significant. This list of significant genes was classified into three different gene panels according to the relation degree with IRDs: i) IRD gene panel, ii) OERD gene panel, and iii) NRD gene panel (Additional file 1:Figure S2). This analysis was also performed for solved-IRD cases.

Carrier frequency calculation

Carrier frequency (CF) was calculated for genes in the non-syndromic IRD gene panel with at least three solved cases in our cohort. Genes were classified as with autosomal recessive or dominant inheritance patterns using the software DOMINO [48] and OMIM database [49]. In genes annotated as recessive, CF was calculated including variants classified: i) “pathogenic” or "likely pathogenic” in ClinVar; ii) or “pathogenic” or "likely pathogenic” in LOVD database; iii) or with a CADD_PHRED score ≥ 30; iv) or frameshift / stop-gain variants. The total AC of the variants selected was divided by the AN and the result multiplied by 2 (two alleles) and multiplied by 100 to represent the result as a percentage (0-100%), according to the equation 1.

$$CF = 2\left(\frac{\sum AC}{maxAN}\right)$$

Equation 1. Carrier frequency (CF) calculation.

For the gene *ABCA4,* the CF was also calculated excluding hypomorphic variants as

described in Hanany et al [19].

Results

A multi-disease cohort database of variant frequencies to study the aggregated signal in IRD genomic landscape

We compiled a heterogeneous cohort of 5683 patients with genetic diseases referred to the Genetics Department of the UH-FJD and with a clinical exome sequencing test available (see Methods). The cases were distributed into three groups of diseases as: inherited retinal dystrophies (IRDs), with 1766 cases, other eye-related diseases (OERDs) with 386 cases, and non-eye related diseases (NRD), 3531 cases (Figure 1A). Additionally, IRD cases were classified according to their diagnosis status as: solved (n=955 cases, 54%), non-solved (n=447 cases, 25%) and monoallelic-VUS (n=364, 21%) (Figure 1A). CEs were reanalyzed and variants detected and extracted for each case. After a quality filter and the removal of variants with a potential sub-population bias between IRD and other cases (Figure 1B and Methods) around 560K variants in 5046 genes from 5683 samples were left for further analyses.

Using this set of variants, an average IRD individual have approximately 4k non polymorphic variants (gnomAD MAF<0.1) within the specified sequenced region (Figure 1C), being mostly intronic (N=2486, 63%), missense (N=513, 13%), and synonymous (N=342, 9%). According to the inheritance pattern observed for the genes based on their associated diseases, 2203 variants (56%) have a recessive pattern, 1256 (32%) dominant, and 487 (12%) either without a clear pattern, undetermined, or X-linked. Regarding their pathogenicity, taken annotation from ClinVar and calculated over the total variants annotated, 379 (10%) are benign or likely benign, 105 (3%) VUS, and 13 (0,3%) pathogenic or likely pathogenic. However, there are still a large percentage of variants with missing or conflicting annotation (N=3449, 87%).

Variant frequencies, allele numbers and allele counts were calculated for IRD cases, solved-IRD cases and non-solved-IRD cases. From a cohort perspective, the

frequencies of pathogenic (including likely pathogenic) variants and VUS are unequally distributed over IRD associated genes when comparing solved-IRD and non-solved-IRD cases (Figure 1D). Top 5 genes with more pathogenic variants (including likely pathogenic) in solved-cases have a greater ratio pathogenic variants / VUS for IRD solved cases than for IRD non-solved cases.

IRD-specific highly frequent variants

For both, solved and non-solved IRD cases, we compared in every variant, its AF (solved-AF or non-solved-AF) with its AF in the PC subcohort (PC-AF), see Methods. We defined the "most frequent variants" in a IRD subcohort (IRD-MFVs) as those within the top 10% with the highest log2(FC) values in the comparison performed, that is, with log2(FC) of frequencies above 3.12 and 3.92 in solved and non-solved IRD cases, respectively (Figure 2A-B). The distribution of log2(FC) values in solved and non-solved IRD cases is shown in the additional file 1:Figure S3. Non-prioritized variants are defined as those with a higher frequency in IRD (FC>0) but below the significant threshold. Classifying IRD-MFVs according to their clinical relevance and removing those not informative (see Methods), we found IRD-MFVs enriched in deleterious variants in both solved and non-solved cases compared to non-prioritized variants (Fisher's exact test, p-values=4.77E-56 and 1.69E-32, respectively; Figure 2C-D and Additional file 1:Tables S9-10). Focusing on the type of genes where the IRD-MFVs are located, we divided IRD-MFVs as present in IRD-associated genes, OERD-related genes, and NRD-associated genes. In IRD-solved cases, regarding IRD-MFVs in IRD-genes only, the ratio deleterious / benign is 78% (177/256), that is significantly higher than the same proportion in non-prioritized variants (14%, 593/1926, Figure 2E, Additional file 1:Table S8). A different trend was observed in non-solved IRD cases where we found no significant differences in the percentages of deleterious/benign variants in IRD-genes between IRD-MFVs and non-prioritized variants (Figure 2F,

Additional file 1:Table S9). In IRD-MFVs in OERD-genes, we found also more deleterious variants in our prioritized set in both, solved and non-solved cases (Fisher's exact test, p-value=1.71E-06 and p-value=3.25E-06, respectively; Figure 2E-F, Additional file 1:Tables S8-S9). Finally, we also observed an enrichment of deleterious variants in the IRD-MFVs located in NRD-genes in solved and non-solved IRD cases (p-value=2.75E-08 and p-value=1.90E-28, respectively; Figure 2E-F, Additional file 1:Tables S8-S9). Furthermore, solved and non-solved cases were divided into disease sub-categories as syndromic, non-syndromic and macular dystrophy forms (Additional file 1:Figure S4), and the analysis was repeated for each sub-group. Thus, as for all solved cases considered as a whole; syndromic, macular dystrophies and non-syndromic forms behave very similarly, with more deleterious variants in the prioritized sets for IRD genes (Additional file 1:Figure S5A-C, Additional file 1:Tables S10-S11). For non-solved cases, we observed an increase in deleterious variants in the prioritized sets for macular dystrophy and non-syndromic forms (Additional file 1:Figure S5D-F and Additional file 1:Tables S10-S11).

In a reevaluation of non-solved IRD cases carrying an IRD-MFV, we found eight variants (three homozygous and five heterozygous) that helped to partially or fully characterize five cases and provided additional information that need further evaluation in other five cases (Table 1).

Another direct application of the IRD relative variant frequencies is the reevaluation of VUS. We extracted the FC of IRD-AF compared to PC-AF for a set of manually curated VUS whose reclassification could contribute to a conclusive diagnosis of an IRD case in our cohort, and present in the final dataset (N=63). Of them, six VUS are IRD-MFVs (log2(FC)>=2.48, see Methods), and 11 VUS are more frequent in IRD cases than in PCs with a log2(FC)FC>=1.5. ACMG classification was performed for the 11 VUS adding the ACMG criteria PS4 ("The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls") as true. Of them, 10 (91%) were reclassified as likely pathogenic/pathogenic. From the six VUS classified

as IRD-MFVs, we achieved a reclassification to likely pathogenic/pathogenic for all of them (Table 2).

Prioritization of candidate genes based on weighted cohort-specific frequency of pathogenic and benign variants in non-solved IRD cases

In order to detect genes with an accumulated high pathogenicity in solved and non-solved IRD cases as good candidates to be involved in IRD phenotypes, for each IRD subcohort, we extracted deleterious and benign variants, and calculated the FC for their AF compared to the AF in PCs. For every subcohort and gene, the distributions of log2(FCs) in deleterious and benign variants were compared using a Wilcoxon rank sum test. Genes with a significant higher frequency in IRD cases of deleterious variants compared to benign variants (p-value<0.05) were selected. This revealed a number of genes with an accumulated pathogenicity in solved and non-solved IRD cases that we classified, as before, in three groups: IRD-genes, OERD-genes and NRD-genes. Actionable genes defined by ACMG were removed from the analysis (Additional file 1:Table S6). Thus, in IRD solved cases, we found 56 genes enriched in deleterious mutations. Of them, 22 (39%) were IRD genes, being the top 5 genes with more deleterious variants *ABCA4, USH2A, MYO7A, EYS and ADGRV1,* (Additional file 1:Table S12, and Figure 3A). In addition, 24 (43%) were OERD genes, highlighting here the top 5 with more deleterious variants *NEB, PAH, DNAH11, DNAH5* and *ATM* (Additional file 1:Table S12 and Figure 3A), and 10 (18%) were NRD genes, including top 5 genes *OBSCN, DYSF, SPTBN5, OTOF and SPTB*. Regarding non-solved cases we found 18 genes with an accumulated pathogenicity, with IRD-genes less represented (22%) and OERD-genes more present (55%) than in solved cases. Finally, 22% of the prioritized genes were NRD-genes (Figure 3B, and Table 3). Interestingly, there was a high overlap between IRD-genes and OERD-genes prioritized in solved and non-solved cases, 75% and 70% of the smallest group (non-solved cases), respectively. However, the overlap in NRD-genes was smaller, with only one gene

(25% of genes in non-solved cases) found in both IRD subcohorts (Additional file 1:Figure S6).

In a reevaluation of the non-solved cases with pathogenic, likely pathogenic or VUS variants in the prioritized genes, we found five cases carrying mutations possibly associated to the phenotype in two IRD associated genes, and one OERD gene (Table 4). Variants found in the gene *MYO7A* explained this case phenotype and helped to fully characterize it. Regarding OERD genes, the two pathogenic variants found in the gene *ADAMTSL4* helped to partially characterize this case.

Carrier's frequency of recessive diseases from a multi-disease cohort

We calculated the carrier frequency (CF) for 69 genes involved in recessive non syndromic IRDs using the frequency of pathogenic variants located in IRD genes in the pseudocontrol subcohort. We found three genes with a CF >= 1%, which represents a total of ~4% of the total analyzed (Table S13). We highlight *ABCA4* and *USH2A* with a CF of ~7% and ~3% respectively (Figure 4), the first being responsible of Stargardt disease, and the later causing Usher syndrome. In the case of *ABCA4* if hypomorphic variants are excluded as in a previous work performed by Hanany and collaborators [19], CF is reduced to ~4%. These genes are also the two most frequent in our IRD subcohort, found causal in 21% and 15% of the cases respectively (data not shown). The most frequent variant for *ABCA4* in the IRD subcohort and PC subcohort were NM_000350.3: c.3386G>T and NM_000350.3: c.3113C>T respectively, while *USH2A* had variant NM_007123.5: c.2276G>T as the most frequent in the two subcohorts.

Discussion

The recent application of NGS techniques has considerably increased our competence to study and diagnose rare diseases. Regarding IRDs, although new associated genes are still being discovered (see RetNet, https://sph.uth.edu/retnet for updates), the

diagnostic ratios need to be boosted, since up to ~46% remain unsolved (data from our own IRD cohort at the time of writing). At the same time, at the diagnostic setting, the genomic information from patients is being accumulated as databases or annotation systems, that can also be privative if only commercial solutions are used. Initiatives like gnomAD [17] or the Collaborative Spanish Variability Server [18] are acting as a crowdsource to recover this data from and to the community and offer it back as aggregated allele frequencies. Allelic frequency calculations have been proven to be useful in studying the prevalence of rare diseases [50], or integrated in the diagnostic analysis routines, as an additional annotation source to detect technical biases [51] or to interpret and classify variants [52]. These resources have also been used to study the CF of deleterious variants in recessive genes in a particular population [16]. In this study, we propose to stretch this approach out to develop a framework of methods able of highlighting variants and genes associated with a specific disease. Thus, the reuse of the genomic data will provide new discovery capabilities to a heterogeneous cohort of genetic diseases, an extra value to the resources invested previously, as well as allowing the patients to contribute to their own or others' future diagnosis. The main hypothesis behind this work is that, aside from a few causing mutations, the genome of patients with Mendelian diseases behaves similarly as those of healthy population, so patients with non-related diseases can act as pseudocontrols between each other. The fact of focusing on a heterogeneous but single-center cohort has two main advantages against using controls from larger genomic population databases [12,13,17]: first they can be technically more similar to the sequencing produced on the patients of interest; second, the geographical origin bias can be better controlled [15]; last, the phenotyping of patients can be fine-tuned by experts in order to create subcohorts of interest. A concept introduced here is the significantly different frequent variants, named as MFVs, that are variants more present in cases than in pseudocontrols, with statistical support. As a proof of concept, we found that considering the whole IRD cohort, the IRD-MFVs are enriched in deleterious mutations, suggesting that the prioritization is effective

there. Indeed, variants selected on solved IRD cases in IRD genes are mostly pathogenic (>78%) compared to benign, due to the detection of prevalent causal variants. In contrast, non-solved IRD cases show no differences in the proportion of deleterious / benign in prioritized and non-prioritized variants in IRD related genes, indicating that there are not many described pathogenic variants left out during diagnosis. In spite of this, our approach was able to solve or partially solve five cases with prioritized variants in IRD genes. We also highlight here the significant accumulation of deleterious variants in those prioritized located in OERD and NRD genes, in both IRD solved and non-solved cases. The rationale behind it might be different, though. While frequent pathogenic mutations in OERD and NRD genes in solved cases may suggest a more complex genotype scenario for Mendelian diseases [53,54], in non-solved cases we have to add the possibility of needing a disease re-evaluation, and the causing mutation being present in a not yet associated IRD gene. The exploration of syndromic and non-syndromic forms separately provides the same signal but with lower p-values for syndromic cases. Up to 10 non-solved cases gained new insights due to the reevaluation of our prioritized variants. Our variant prioritization approach was also applied for VUS reclassification, a major challenge to unlock the diagnosis of rare pathologies [10,55–57]. Indeed, in our IRD subcohort, top 5 genes with more deleterious variants in IRD solved cases, present a higher degree of uncertainty in variant annotation (proportion of VUS and deleterious variants) in unsolved IRD cases. For an initial list of 63 VUS whose reclassification may solve a pending case from the IRD cohort, we found 11 VUS more frequent in IRD cases than in pseudocontrols and 10 of them (~91%) changed their classification to likely pathogenic or pathogenic by the application of the ACMG PS4 criteria.

In parallel, we also aimed to highlight genes besides variants. Our method extracts genes having more frequent deleterious variants in IRD cases than in pseudocontrols, weighted by the relative frequency of benign variants in order to increase the disease association signal. Several discovery scenarios may fit into the results of this proposal.

First, finding underrated IRD genes with a role in IRD cases, such as *DYNC2H1* and *MYO7A* genes that solved or partially solved a total of four cases. Next, providing extra findings in complex cases, either syndromic cases, gene modifiers, or dual diagnosis. For instance, we found the gene *ADAMTSL4,* from the OERD panel, with two pathogenic variants in a non-solved syndromic IRD case with rod-cone dystrophy (HP:0000510) and lens luxation (HP:0012019) among other systemic findings. These mutations explained the lens luxation phenotype and helped to partially characterize this case. In non-classical Mendelian scenarios, the exploration of the mutational landscape of IRD may help for instance to identify more complex cases as: i) genetic pleiotropy together with causal variants in recessive forms [22], ii) digenic inheritance [23], or iii) triallelic sites associated to BBS [24]. Last, there is also the possibility of detecting genes not yet associated to IRD, that is candidate genes that may become IRD genes if further analyses are performed.

An additional interesting use of an internal database of allele frequencies is to have a cohort-specific CF estimation. This analysis can provide a better understanding on how deleterious variants are distributed in a general population and is relevant for their use in a public health strategy for genetic counselling. For instance, in the case of IRD, the gene with a higher carrier frequency is *ABCA4,* with carrier variants in ~7% of the population, which is in line with previous estimations [58]. Considering the curated set of pathogenic variants used by Hanany et al [19] for *ABCA4,* we obtained a similar CF (~6%). Nevertheless, excluding hypomorphic variants, as recommended in this study, CF drops to ~4%. The high CF obtained for this gene in our cohort can be explained partially by these variants.

It is reasonable to state that although pseudocontrols are suitable for providing a good estimation of general allele frequencies, the lack of healthy controls can be seen as a limitation. However availing of such control sample set is not always feasible for clinical setting and should be provided under the umbrella of national plans. The major constraint in the discovery capability of our database is that we are restricted to the

~5000 genes targeted in the clinical exome approaches, and thus, an implementation using data from whole exomes would be optimal. Our intention is to maintain and expand the database in number of cases but also in genomic regions. We should mention that IRD non-solved cases presented in the cohort can also have causing variants in non-coding regions, which are not covered with the clinical exome approach.

Although in this work we focus on IRD as the larger group of diseases in our cohort, the same methodology can be applied to other genetic rare diseases in our cohort as well as in other settings.

Conclusions

In conclusion, our cohort-specific database of allele frequencies has proven to be able to diagnose non-solved IRD cases, reclassify VUS, propose candidate genes, and calculate CF on genes of interest. We believe that the results shown here can highlight the importance of the reuse of genomic data produced in clinical settings, where the phenotyping is usually exhaustive and the patients waiting for a diagnosis or a genetic counselling can be directly benefited.

List of abbreviations

AC: Allele count

ACMG: American Colledge of Medical Genetics

AF: Allele frequency

AN: Allele number

BBS: Bardet Biedl syndrome

CE: Clinical exome

CES: Clinical exome solution

CF: Carrie frequency

FC: Fold change

HPO: Human phenotype ontology

UH-FJD: University Hospital Fundación Jiménez Díaz

IRD: Inherited retinal dystrophies

IRD-AF: Inherited retinal dystrophies allele frequency

IRD-MFV: Inherited retinal dystrophies most frequent variants

NGS: Next generation sequencing

NRD: Non-related diseases

OERD: Other eye related diseases

PC: Pseudocontrol

PC-AC: Pseucontrol's allele count

PC-AF: Pseucontrol's allele frequency

PC-AN: Pseucontrol's allele number

SNP: Single nucleotide polymorphism

SNV: Single nucleotide variant

TSO: TruSight One

UTR: Untranslated region

VUS: Variant of uncertain significance

Acknowledgments

We thank the patients for consenting to the use of their data for the study. We also thank all technical staff in the Genetics Department of the UH-FJD for conducting the sequencing and further analysis.

Funding

This work was supported by the Comunidad de Madrid (CAM, RAREGenomics Project, B2017/BMD-3721), Instituto de Salud Carlos III (ISCIII) of the Spanish Ministry of

Health (FIS; PI18/00579, PI19/00321, PI20/00851), Ramón Areces Foundation (4019/012), Centro de Investigación Biomédica en Red Enfermedades Raras (CIBERER, 06/07/0036), IIS-FJD BioBank (PT13/0010/0012), the Organización Nacional de Ciegos Españoles (ONCE), European Regional Development Fund (FEDER) and the University Chair UAM-IIS-FJD of Genomic Medicine. IFI is supported by a grant from the Comunidad de Madrid (CAM, PEJ-2017-AI/BMD7256) and ISCIII (IMPACT-Data; IMP/00019), IPR is supported by a PhD studentship from the predoctoral program from ISCIII (FI17/00192), GNM is supported by a grant from the Comunidad de Madrid (PEJ-2020-AI/BMD-18610), LdIF is supported by the platform technician contract of ISCIII (CA18/00017), RR is supported by a postdoctoral fellowship of the Comunidad de Madrid (2019-T2/BMD-13714), BA is supported by a Juan Rodes program from ISCIII (JR17/00020), AD is supported by a PhD studentship from the predoctoral program from ISCIII (FI18/00123) and PM is supported by a Miguel Servet program contract from ISCIII (CP16/00116, CPII21/00015).

Author information

Affiliations

Department of Genetics, Health Research Institute–Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain.

Ionut-Florin Iancu, Irene Perea-Romero, Gonzalo Núñez-Moreno, Lorena de la Fuente, Raquel Romero, Almudena Ávila-Fernandez, María José Trujillo-Tiebas, Rosa Riveiro-Álvarez, Berta Almoguera, Inmaculada Martín-Mérida, Marta Del Pozo-Valero, Alejandra Damián-Verde, Marta Cortón, Carmen Ayuso and Pablo Minguez

Center for Biomedical Network Research on Rare Diseases (CIBERER), ISCIII, Madrid, Spain.

Ionut-Florin Iancu, Irene Perea-Romero, Gonzalo Núñez-Moreno, Raquel Romero, Almudena Ávila-Fernandez, María José Trujillo-Tiebas, Rosa Riveiro-Álvarez, Berta Almoguera, Inmaculada Martín-Mérida, Marta Del Pozo-Valero, Marta Cortón, Carmen Ayuso and Pablo Minguez

Bioinformatics Unit, Health Research Institute–Fundación Jiménez Díaz University Hospital, Universidad Autónoma de Madrid (IIS-FJD, UAM), Madrid, Spain.
Gonzalo Núñez-Moreno, Lorena de la Fuente and Pablo Minguez

Contributions

Study concept and design: PM, CA and IFI. Bioinformatics methods: IFI, GNM, LdlF, RR and PM. Software implementation: IFI, GNM, LdlF and RR. Data analysis and interpretation: IFI, IPR, AAF, MJTT, RRA, BA, IMM, MDPV, MC, CA and PM. Drafting of the manuscript: IFI and PM. Manuscript reviewing and editing: IFI, IPR, GNM, LdlF, RR, AAF, MJTT, RRA, BA, IMM, MDPV, MC, CA and PM. Resources: CA and PM. The authors read and approved the final manuscript.

Corresponding authors
Carmen Ayuso and Pablo Minguez

Conflict of Interest statement. The authors have no conflict of interest to disclose.

Ethics declarations

Ethics Approval and consent to participate

The project was reviewed and approved by the Research Ethics Committee of UH-FJD (Ref. 2016/ 59) and fulfills the principles of the Declaration of Helsinki and subsequent reviews. All patients signed an informed consent before participating. All samples included in this work were pseudonymized and genomic data was only treated in aggregation.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Supplementary Information

Additional file 1:

Table S1. Diseases included in the allele frequency database, and number (N) of cases of each disease. Diseases are classified in three cathegories: i) Inherited Retinal

Dystrophies (IRD), other eye related diseases (OERD) and non related diseases (NRD).

Table S2. Diagnosis status for cases with inherited retinal dystrophies (IRD) at the Genetics Department of the HU-FJD. Cases classified in: solved, non-solved and a third group formed by VUS cases (cases with 1 or 2 VUS reported) and monoallelic cases (cases with a recessive inheritance pattern with 1 variant likely pathogenic / pathogenic reported).

Table S3. Gene panel used in the diagnosis of cases with inherited retinal dystrophies (IRD).

Table S4. Genes classified as involved in other eye related diseases (OERD). The list includes all genes linked with the HPO term "Eye Disease" – HP:0000478, but those included in the Table S3.

Table S5. Genes in the database of allelic frequencies not involved in an eye related disease (NRD).

Table S6. Genes that ACMG recommends to report secondary findings. These genes are filtered out from OERD and NRD gene panels.

Table S7. List with 63 VUS pending on reclassification at the genetics Department of the HU-FJD.

Table S8. Number of deleterious and benign within the prioritized (IRD-MFV) and non-prioritized sets in IRD solved cases, including those in all genes and in genes from the inherited retinal dystrophies (IRD) panel, the other eye related diseases (OERD) and non eye related diseases (NRD). P—values of the Fishers' exact test are shown.

Table S9. Number of deleterious and benign within the prioritized (IRD-MFV) and non-prioritized sets in IRD non-solved cases, including those in all genes and in genes from the inherited retinal dystrophies (IRD) panel, the other eye related diseases (OERD) and non eye related diseases (NRD). P—values of the Fishers' exact test are shown.

Table S10. Number of deleterious and benign within the prioritized (IRD-MFV) and non-prioritized sets in IRD solved syndromic (SY), IRD non-syndromic (NSY) and macular

dystrophies (MD) cases, including those in all genes and in genes from the inherited retinal dystrophies (IRD) panel, the other eye related diseases (OERD) and non eye related diseases (NRD). P-values of the Fishers' exact test are shown.

Table S11. Number of deleterious and benign within the prioritized (IRD-MFV) and non-prioritized sets in IRD non-solved syndromic (SY), IRD non-syndromic (NSY) and macular dystrophies (MD) cases, including those in all genes and in genes from the inherited retinal dystrophies (IRD) panel, the other eye related diseases (OERD) and non eye related diseases (NRD). P-values of the Fishers' exact test are shown.

Table S12. Genes prioritized in solved cases of inherited retinal dystrophies. They are classified in three gene panels: genes from the inherited retinal dystrophies (IRD) panel, the other eye related diseases (OERD) and non eye related diseases (NRD).

Table S13. Carrier frequency (in pseudocontrol cases) and frequency in cases with inherited retinal dystrophies for the top 10 genes with higher carrier frequency.

Figure S1. Intersection in genes included in the two clinical exomes used in the sequencing of the samples in the cohort: TruSightOne Sequencing Panel kit (TSO, Illumina, San Diego, CA), and Clinical Exome Solution Sequencing Panel kit (CES, Sophia Genetics, Boston, MA). Genes involved in inherited retinal dystrophies (IRD) are also highlighted.

Figure S2. Workflow to perform the gene prioritization in (A) solved and (B) non-solved with inherited retinal dystrophies (IRD).

Figure S3. Distribution of the values of the fold changes (log2(FC)) calculated between the allelic frequencies in two IRD subcohorts: (A) solved and (B) non-solved, and the allelic frequencies in the pseudocontrols. Percentiles 50%, 80%, 85%, 90%, 95% and 99% are shown in both groups.

Figure S4. Description of the cohort of cases with inherited retinal dystrophies (IRD). Number of cases grouped by diagnostic status (solved, non-solved and monoallelic-VUS) and IRD type (syndromic, non-syndromic, macular dystrophies and unknown).

Figure S5. Proportion of deleterious and benign variants in both solved (A, B and C)

and non-solved cases with inherited retinal dystrophies (D, E and F) for the different IRD subcohorts: syndromic, non-syndromic and macular dystrophies. The p-values representing the enrichment of deleterious variants in IRD-MFVs are shown. The genes in which the IRD-MFVs are located are grouped in: inherited retinal dystrophies (RD genes), other eye related diseases (OERD genes) and other non related diseases (NRD genes). Non-significant p-values are marked as “ns”.

Figure S6. Intersection of genes prioritized in solved and non-solved cases with inherited retinal dystrophies. The genes are grouped as involved in: inherited retinal dystrophies (IRD genes), other eye related diseases (OERD genes) and other non related diseases (NRD genes). We show only the names of the genes prioritized in non-solved IRD cases, in light green those unique to non-solved, and in dark green those in common with IRD solved cases.

References

1. Richter T, Nestler-Parr S, Babela R, Khan ZM, Tesoro T, Molsen E, et al. Rare Disease Terminology and Definitions—A Systematic Global Review: Report of the ISPOR Rare Disease Special Interest Group. Value in Health. 2015;18:906–14.

2. Ayuso C, Millan JM. Retinitis pigmentosa and allied conditions today: a paradigm of translational research. Genome medicine. 2010;2:34.

3. Wright AF, Chakarova CF, Abd El-Aziz MM, Bhattacharya SS. Photoreceptor degeneration: genetic and mechanistic dissection of a complex trait. Nature Reviews Genetics. 2010;11:273–84.

4. Sanchez-Navarro I, Da Silva LRJRJ, Blanco-Kelly F, Zurita O, Sanchez-Bolivar N, Villaverde C, et al. Combining targeted panel-based resequencing and copy-number variation analysis for the diagnosis of inherited syndromic retinopathies and associated ciliopathies. Scientific reports [Internet]. Sci Rep; 2018 [cited 2021 Dec 2];8. Available from: https://pubmed.ncbi.nlm.nih.gov/29588463/

5. Mockel A, Perdomo Y, Stutzmann F, Letsch J, Marion V, Dollfus H. Retinal dystrophy in Bardet–Biedl syndrome and related syndromic ciliopathies. Progress in Retinal and Eye Research. 2011;30:258–74.

6. Gana S, Serpieri V, Valente EM. Genotype-phenotype correlates in Joubert syndrome: A review. Am J Med Genet C Semin Med Genet. 2022.

7. Del Pozo-Valero M, Riveiro-Alvarez R, Martin-Merida I, Blanco-Kelly F, Swafiri S, Lorda-Sanchez I, et al. Impact of Next Generation Sequencing in Unraveling the Genetics of 1036 Spanish Families With Inherited Macular Dystrophies. Invest

Ophthalmol Vis Sci. 2022;63:11.

8. Martin-Merida I, Aguilera-Garcia D, Fernandez-San Jose P, Blanco-Kelly F, Zurita O, Almoguera B, et al. Toward the mutational landscape of autosomal dominant retinitis pigmentosa: A comprehensive analysis of 258 Spanish families. Investigative Ophthalmology and Visual Science. 2018;59.

9. Martin-Merida I, Avila-Fernandez A, Del Pozo-Valero M, Blanco-Kelly F, Zurita O, Perez-Carro R, et al. Genomic Landscape of Sporadic Retinitis Pigmentosa: Findings from 877 Spanish Cases. Ophthalmology. 2019;126:1181–8.

10. Iancu IFF, Avila-Fernandez A, Arteche A, Trujillo-Tiebas MJJ, Riveiro-Alvarez R, Almoguera B, et al. Prioritizing variants of uncertain significance for reclassification using a rule-based algorithm in inherited retinal dystrophies. NPJ genomic medicine. 2021;6(1):18.

11. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in medicine : official journal of the American College of Medical Genetics. NIH Public Access; 2015;17:405–24.

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

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

14. Mathieson I, Reich D. Differences in the rare variant spectrum among human populations. PLoS Genet. 2017;13:e1006581.

15. Dopazo J, Amadoz A, Bleda M, Garcia-Alonso L, Alemán A, García-García F, et al. 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. Molecular biology and evolution. 2016;33(5):1205-18

16. Fridman H, Yntema HGG, Mägi R, Andreson R, Metspalu A, Mezzavila M, et al. The landscape of autosomal-recessive pathogenic variants in European populations reveals phenotype-specific effects. American journal of human genetics. Am J Hum Genet; 2021;108:608–19.

17. 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;581:434–43.

18. Peña-Chilet M, Roldán G, Perez-Florido J, Ortuño FM, Carmona R, Aquino V, et al. CSVS, a crowdsourcing database of the Spanish population genetic variability. Nucleic acids research. 2021;49:D1130–7.

19. Hanany M, Rivolta C, Sharon D. Worldwide carrier frequency and genetic prevalence of autosomal recessive inherited retinal diseases. Proceedings of the National Academy of Sciences of the United States of America. Proc Natl Acad Sci U S A; 2020;117:2710–6.

20. Wang C, Sun L, Zheng H, Hu Y-Q. Detecting multiple variants associated with disease based on sequencing data of case-parent trios. J Hum Genet. 2016;61:851–60.

21. Cordell HJ, Barratt BJ, Clayton DG. Case/pseudocontrol analysis in genetic association studies: A unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. Genet Epidemiol. 2004;26:167–85.

22. Molinari E, Srivastava S, Sayer JA, Ramsbottom SA. From disease modelling to personalised therapy in patients with CEP290 mutations. F1000Res. 2017;6:669.

23. Gao F-J, Zhang S-H, Chen J-Y, Xu G-Z, Wu J-H. Digenic heterozygous mutations in EYS/LRP5 in a Chinese family with retinitis pigmentosa. Int J Ophthalmol. 2017;10:325–8.

24. Katsanis N, Ansley SJ, Badano JL, Eichers ER, Lewis RA, Hoskins BE, et al. Triallelic Inheritance in Bardet-Biedl Syndrome, a Mendelian Recessive Disorder. Science. 2001;293(5538):2256-9.

25. Romero R, de la Fuente L, Del Pozo-Valero M, Riveiro-Álvarez R, Trujillo-Tiebas MJ, Martín-Mérida I, Ávila-Fernández A, Iancu IF, Perea-Romero I, Núñez-Moreno G, Damián A, Rodilla C, Almoguera B, Cortón M, Ayuso C, Mínguez P. An evaluation of pipelines for DNA variant detection can guide a reanalysis protocol to increase the diagnostic ratio of genetic diseases. npj genomic medicine. 2022;7(1):7

26. 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;20:1297–303.

27. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

28. Durbin RM, Altshuler DL, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, et al. A map of human genome variation from population-scale sequencing. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2010;467:1061–73.

29. Glusman G, Caballero J, Mauldin DE, Hood L, Roach JC. Kaviar: an accessible system for testing SNV novelty. Bioinformatics. 2011;27:3216–7.

30. Ng PCC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. Nucleic Acids Res; 2003;31:3812–4.

31. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics. Curr Protoc Hum Genet. 2013;Chapter 7:Unit7.20.

32. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.

33. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009;19:1553–61.

34. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48:1581–6.

35. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and

integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet. 2015;24:2125–37.

36. Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol. 2007;8:R232.

37. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nature methods. 2010;7:575–6.

38. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One. 2012;7:e46688.

39. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34:57–65.

40. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42:13534–44.

41. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44:D862-868.

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

43. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15:1034–50.

44. Fadista J, Oskolkov N, Hansson O, Groop L. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics. 2017;33:471–4.

45. Havrilla JM, Pedersen BS, Layer RM, Quinlan AR. A map of constrained coding regions in the human genome. Nat Genet. 2019;51:88–95.

46. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.

47. Miller DT, Lee K, Chung WK, Gordon AS, Herman GE, Klein TE, et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2021;23:1381–90.

48. Quinodoz M, Royer-Bertrand B, Cisarova K, Di Gioia SA, Superti-Furga A, Rivolta C. DOMINO: Using Machine Learning to Predict Genes Associated with Dominant Disorders. Am J Hum Genet. 2017;101:623–9.

49. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic acids research. 2002;30:52–5.

50. Perea-Romero I, Gordo G, Iancu IFIFIF, Del Pozo-Valero M, Almoguera B, Blanco-Kelly F, et al. Genetic landscape of 6089 inherited retinal dystrophies affected cases in Spain and their therapeutic and extended epidemiological implications. Scientific Reports. 2021;11(1):1526.

51. Maffucci P, Bigio B, Rapaport F, Cobat A, Borghesi A, Lopez M, et al. Blacklisting variants common in private cohorts but not in public databases optimizes human exome analysis. Proc Natl Acad Sci U S A. 2019;116:950–9.

52. Musacchia F, Ciolfi A, Mutarelli M, Bruselles A, Castello R, Pinelli M, et al. VarGenius executes cohort-level DNA-seq variant calling and annotation and allows to manage the resulting data through a PostgreSQL database. BMC Bioinformatics. 2018;19:477.

53. Riordan JD, Nadeau JH. From Peas to Disease: Modifier Genes, Network Resilience, and the Genetics of Health. American Journal of Human Genetics. Cell Press; 2017;101:177–91.

54. Katsanis N. The continuum of causality in human genetic disorders. Genome Biology. BioMed Central Ltd.; 2016;17:233.

55. Mahecha D, Nuñez H, Lattig MC, Duitama J. Machine learning models for accurate prioritization of variants of uncertain significance. Hum Mutat. 2022;43:449–60.

56. Nicora G, Zucca S, Limongelli I, Bellazzi R, Magni P. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. Sci Rep. 2022;12:2517.

57. Castillo-Guardiola V, Rosado-Jiménez L, Sarabia-Meseguer MD, Marín-Vera M, Macías-Cerrolaza JA, García-Hernández R, et al. Next step in molecular genetics of hereditary breast/ovarian cancer: Multigene panel testing in clinical actionably genes and prioritization algorithms in the study of variants of uncertain significance. Eur J Med Genet. 2022;65:104468.

58. Riveiro-Alvarez R, Aguirre-Lamban J, Lopez-Martinez MA, Trujillo-Tiebas MJ, Cantalapiedra D, Vallespin E, et al. Frequency of ABCA4 mutations in 278 Spanish controls: an insight into the prevalence of autosomal recessive Stargardt disease. Br J Ophthalmol. 2009;93:1359–64.

Tables

Table 1. Prioritized variants in IRD non-solved cases adding knowledge to the phenotype in 10 cases, each row represents a different case. Previous and current diagnostic status of each case are shown. The gene panel column refers to the type of diseases that the gene has been associated with, inherited retinal dystrophies (IRD), other eye related diseases (OERD) and other non-eye related diseases (NRD).

Gene	Gene panel	Phenotype	Variant(s) (zygosity, AGMG classification)	Previous diagnostic status	Diagnostic status
CDH23	IRD	Several	NM_022124.6: c.6050-9G>A	Non-solved	Monoallelic

		phenotypes including rod-cone dystrophy (HP:0000510) and lens luxation (HP:0012019)	(heterozygous, pathogenic)		
MYO7A	IRD	Usher syndrome	NM_000260.4: c.1996C>T (heterozygous, pathogenic)	Non-solved	Solved
IFT88	ONRD	retinitis pigmentosa	NM_001318491.2: c.538G>T (heterozygous, pathogenic)	Non-solved	Pending
AP5Z1	OERD	macular dystrophy	NM_001364858.1: c.337C>T (homozygosis, likely pathogenic)	Non-solved	Pending
KIF1B	OERD	retinitis pigmentosa	NM_001365951.2: c.3340G>A (homozygosis, VUS)	Non-solved	Pending
KIAA2022	NRD	macular dystrophy	NM_001008537.3: c.4385del; hemyzygosis; likely pathogenic	Non-solved	Pending
TTPA	IRD	retinitis pigmentosa	NM_000370.3: c.227_235del (heterozygous, likely pathogenic)	Non-solved	Monoallelic
TTPA	IRD	macular dystrophy	NM_000370.3: c.227_235del (heterozygous, likely pathogenic)	Non-solved	Monoallelic
CDHR1	IRD	retinitis pigmentosa	NM_033100.4: c.2410_2485del (heterozygous, likely pathogenic)	Non-solved	Monoallelic
CDHR1	IRD	retinitis pigmentosa	NM_033100.4: c.2410_2485del (heterozygous, likely pathogenic)	Non-solved	Pending

Table 2. VUS selected with a Fold change (FC) higher than 2. In column "Status" reclassified variants are marked. Inheritance mode is annotated as autosomal recessive (AR), autosomal dominant (AD) and x-linked (XL). ACMG criteria are annotated.

Gene	HGVSc	Inheritance	log2(FC)	VarSome	Criteria	PS4	Status
COL11A1	NM_080629.2: c.4838C>A	AD	1.6	3	PM2,PP2	4	Reclassified
RP1	NM_006269.2: c.2497T>C	AR/AD	1.7	3	PM2,BP4	4	Reclassified
GDF6	NM_001001557.4: c.125G>T	AR/AD	2.0	3	PM1,PM2,PP2,PP5,BP4,BP6	3	Pending
MERTK	NM_006343.3: c.2435A>G	AR	2.0	3	PM1,PM2,PP3,BP1	4	Reclassified
CDH23	NM_022124.5: c.4231G>A	AR	2.0	3	PM2,PP2,PP3	4	Reclassified
CDHR1	NM_001171971.3: c.1589C>G	AR	2.6	3	PM1,PM2,BP1	4	Reclassified
CACNA1F	NM_001256789.3: c.4009-3C>G	XL	2.7	3	PM2,BP4	4	Reclassified
IMPG2	NM_016247.4: c.1460A>T	AR/AD	2.7	3	PM2,BP4	4	Reclassified
OFD1	NM_003611.2: c.87T>G	XL	2.7	3	PM2,PP3,BP1	4	Reclassified
NYX	NM_022567.2: c.505A>G	XL	2.9	3	PP2,PP3	4	Reclassified
WFS1	NM_001145853.1: c.1597C>T	AR/AD	3.7	3;4	PM2,PM5,PP3	4	Reclassified

Table 3. Gene list in non-solved cases ordered by Gene panel used (IRD, OERD and ONRD) and by number of deleterious variants. Inheritance mode associated for each gene is annotated in the "Inheritance" column, as recessive, dominant or R / D (either recessive or dominant). The gene panel column refers to the type of diseases that the gene has been associated with, inherited retinal dystrophies (IRD), other eye related diseases (OERD) and other non-eye related diseases (NRD).

Gene	Deleterious	Benign	FDR	Gene Panel	Inheritance
MYO7A	10	42	1.08E-03	IRD	Recessive
DYNC2H1	9	43	9.81E-04	IRD	Recessive
LAMA1	7	22	3.56E-02	IRD	Recessive
HMCN1	6	43	2.75E-02	IRD	Dominant
NEB	12	59	4.55E-03	OERD	Recessive
PAH	11	12	1.32E-02	OERD	Recessive
ALS2	8	11	1.87E-02	OERD	Recessive
DNAH9	7	13	1.32E-02	OERD	Recessive
HSPG2	6	44	1.87E-02	OERD	R / D
DNAH5	6	43	1.87E-02	OERD	Recessive
PLEC	6	80	1.87E-02	OERD	Dominant
ADAMTSL4	5	16	2.70E-02	OERD	Recessive
NDUFV1	5	8	1.32E-02	OERD	Recessive
COL4A3	5	15	1.87E-02	OERD	R / D
OBSCN	9	91	1.32E-02	NRD	Recessive
CAPN3	7	7	1.28E-02	NRD	Recessive

MYOM1	6	22	1.32E-02	NRD	Recessive
ABCB11	6	7	1.87E-02	NRD	Recessive

Table 4. Details about five IRD non-solved cases with variants adding knowledge to the phenotype. The gene panel column refers to the type of diseases that the gene has been associated with, inherited retinal dystrophies (IRD) and other eye related diseases (OERD).

Gene	Gene panel	Phenotype	Variant(s) and classification	Diagnostic status
DYNC2H1	IRD	macular dystrophy	NM_001080463.2: c.3793C>T (VUS); NM_001080463.2: c.1468C>T (VUS)	Pending on VUS reclassification
DYNC2H1	IRD	macular dystrophy	NM_001080463.2: c.988C>T (pathogenic)	Monoallelic
DYNC2H1	IRD	macular dystrophy	NM_001080463.2: c.7966C>T (likely pathogenic)	Monoallelic
MYO7A	IRD	Usher syndrome	NM_000260.4: c.1996C>T (pathogenic); NM_000260.4: c.3764del (pathogenic)	Solved
ADAMTSL4	OERD	Several phenotypes including rod-cone dystrophy (HP:0000510) and lens luxation (HP:0012019)	NM_001288607.2: c.2594G>A (pathogenic); M_001288607.2: c.767_786_del (pathogenic)	Variants explain Lens luxation phenotype

Figure Legends

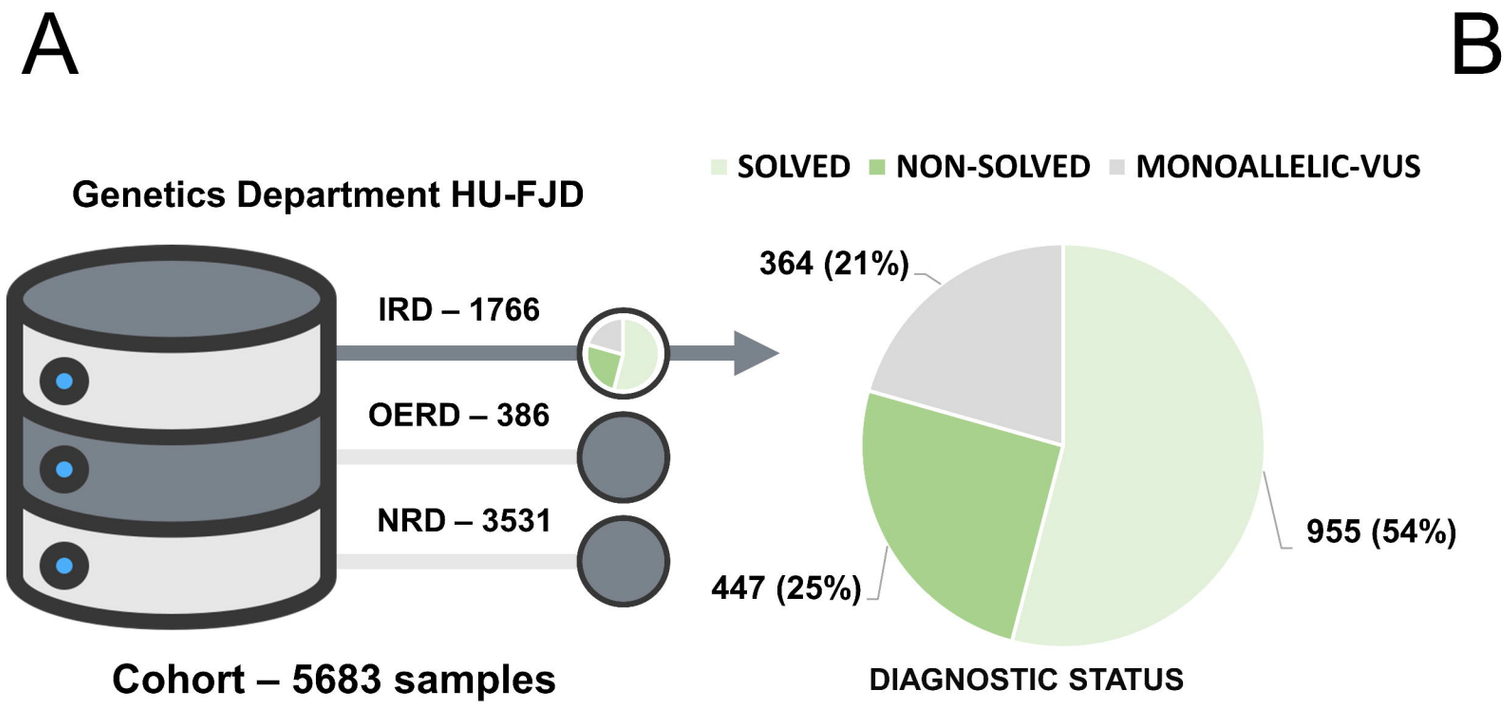
Figure 1. Database and cohort description. A) The cohort of patients with suspected rare genetic diseases at the Genetics Department of HU-FJD was divided into three subcohorts. An IRD subcohort of 1766 samples, an Other Eye Related Diseases (OERD) of 386 cases and a pseudocontrol subcohort of non-eye related diseases (NRD) with 3531 samples. IRD diagnostic status of the samples included in IRD subcohort was: solved, non-solved and a third group formed by VUS cases (cases with 1 or 2 VUS reported) and monoallelic cases (cases with a recessive inheritance pattern with 1 variant likely pathogenic / pathogenic reported). B) Flow chart of different filters applied to variants according to quality control (QC) and population (POP) filters. C) Summary of the variants included in the database in an average IRD case, values represent the average of all IRD samples. D) Proportion of pathogenic and VUS variants detected in IRD cases in the genes with more pathogenic variants in IRD solved cases, top 5 genes are shown.

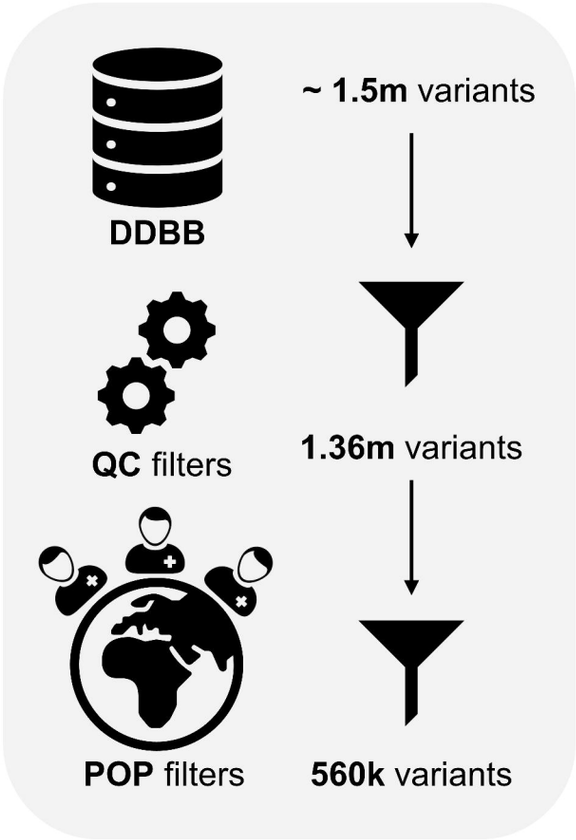
Figure 2. Allele frequency (AF) comparison between solved and non-solved IRD and pseudocontrol subcohorts. Variant AFs are compared in inherited retinal dystrophies (IRD) solved (A) and non-solved (B) subcohorts against pseudocontrols

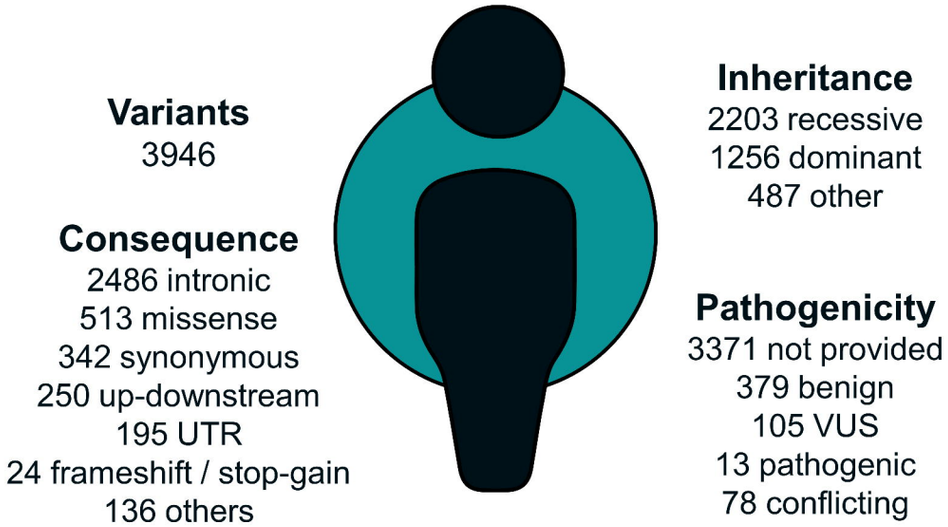
(PC) using fold changes (FC). IRD more frequent variants (IRD-MFVs), highlighted in dark, were defined using FC thresholds at .90 percentiles of all FC in each comparison (A-B). Proportion of deleterious and benign variants in both solved (C) and non-solved IRD cases (D) and the p-values representing the enrichment of deleterious variants in IRD-MFVs. Enrichment analyses are also performed dividing the IRD-MFVs according to the genes in which they are located, grouping them in: IRD genes, other eye related diseases (OERD genes) and other non-eye related diseases (NRD genes) (E-F). Total number of deleterious variants in each group is noted at the top of the red bars. Non-significant p-values are marked as “ns”.

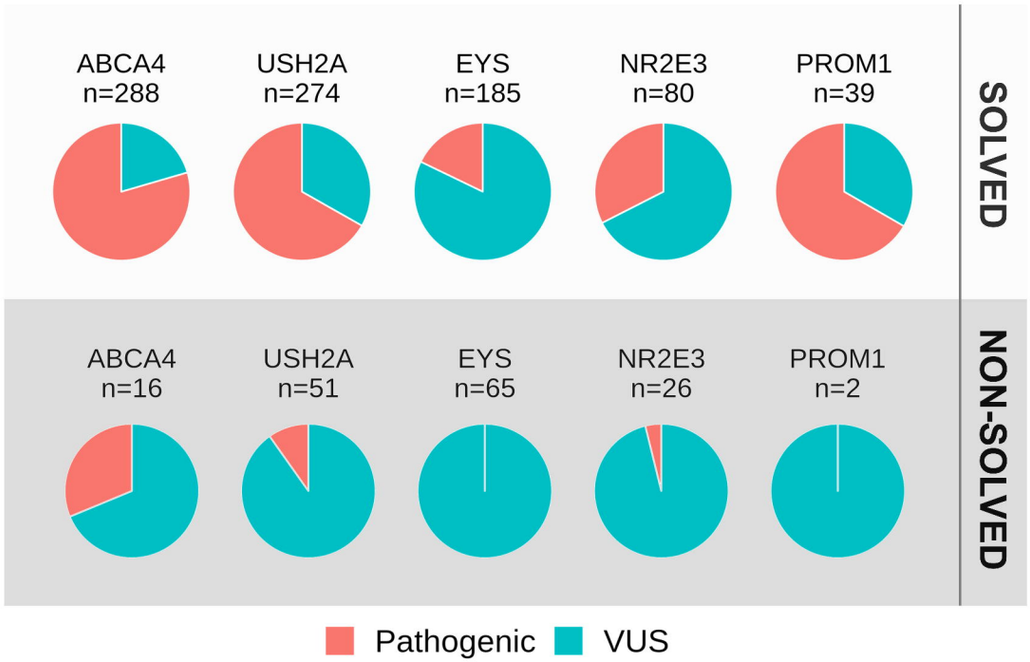
Figure 3. **Genes with a higher accumulated pathogenicity in solved and non-solved IRD cases compared to pseudocontrols.** Mean fold changes (FCs) in log2 scale for deleterious (Y-axis) and benign variants (X-axis) are shown for each gene. For IRD solved cases, significative genes with at least 8 deleterious variants are shown in the plot (A). For IRD non-solved cases all significative genes are shown (B).

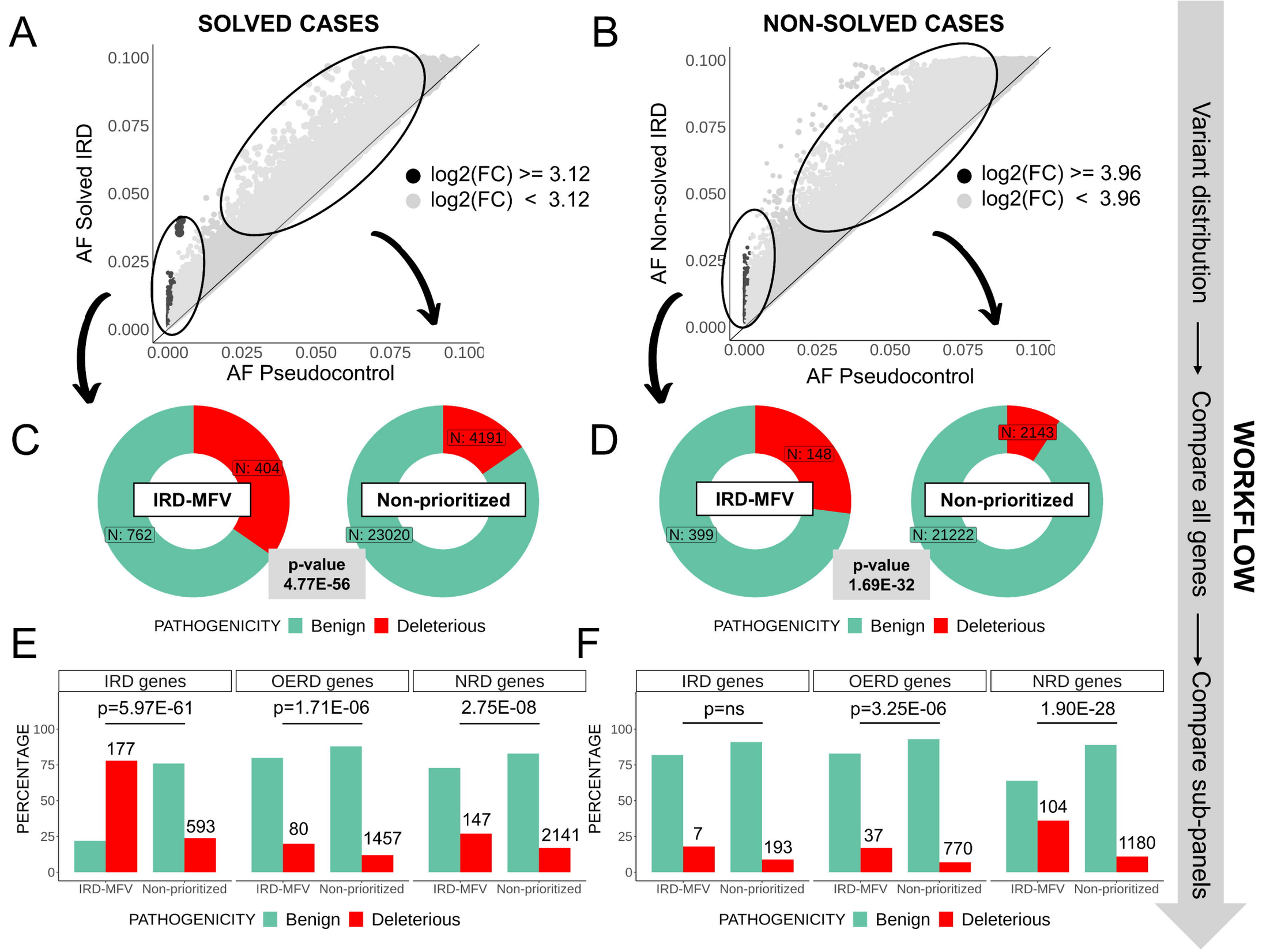
Figure 4. Carrier frequency. The carrier frequency (CF) was calculated for all IRD genes with recessive inheritance pattern, and at least one solved case in our cohort. In green is represented the CF of the genes and in orange the frequency in our IRD subcohort. The pseudocontrol subcohort was composed of 3531 cases, and the IRD subcohort had 1766 cases. Top 10 genes with higher CF are shown ordered decreasingly.

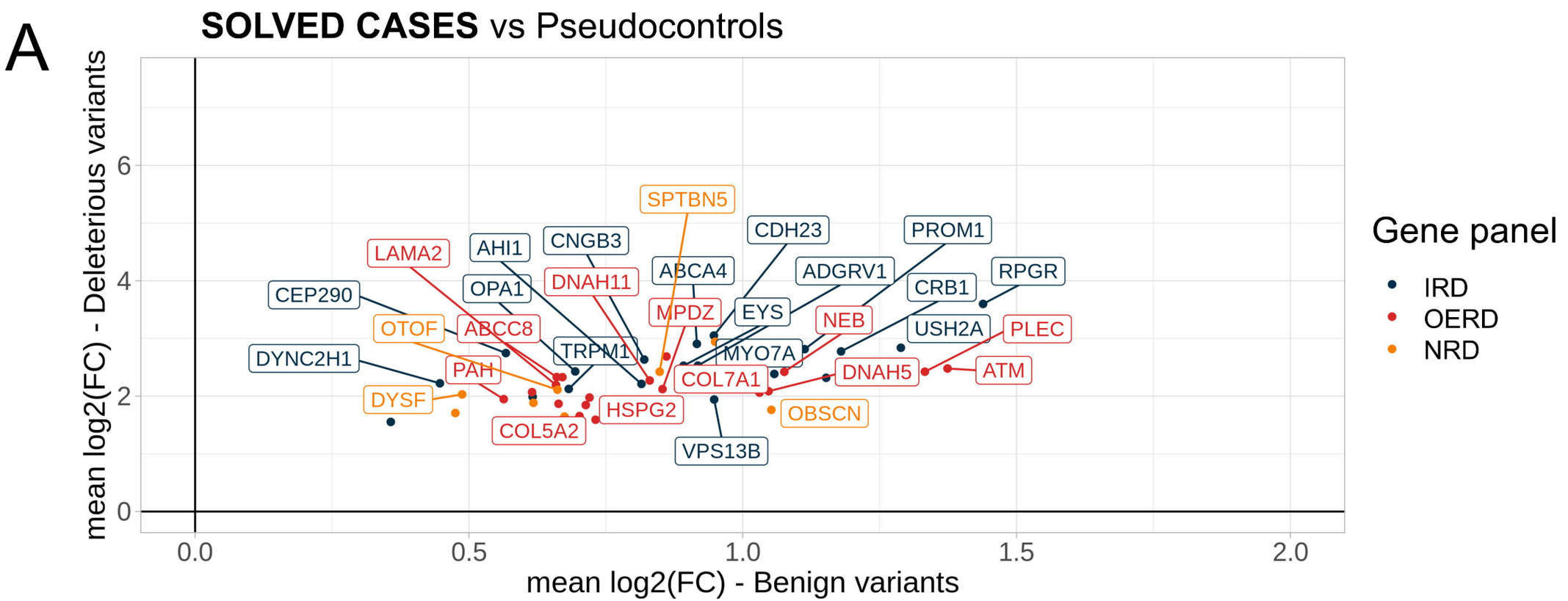

A
Genetics Department HU-FJD
IRD – 1766
OERD – 386
NRD – 3531
Cohort – 5683 samples
SOLVED
NON-SOLVED
MONOALLELIC-VUS
364 (21%)
955 (54%)
447 (25%)
DIAGNOSTIC STATUS

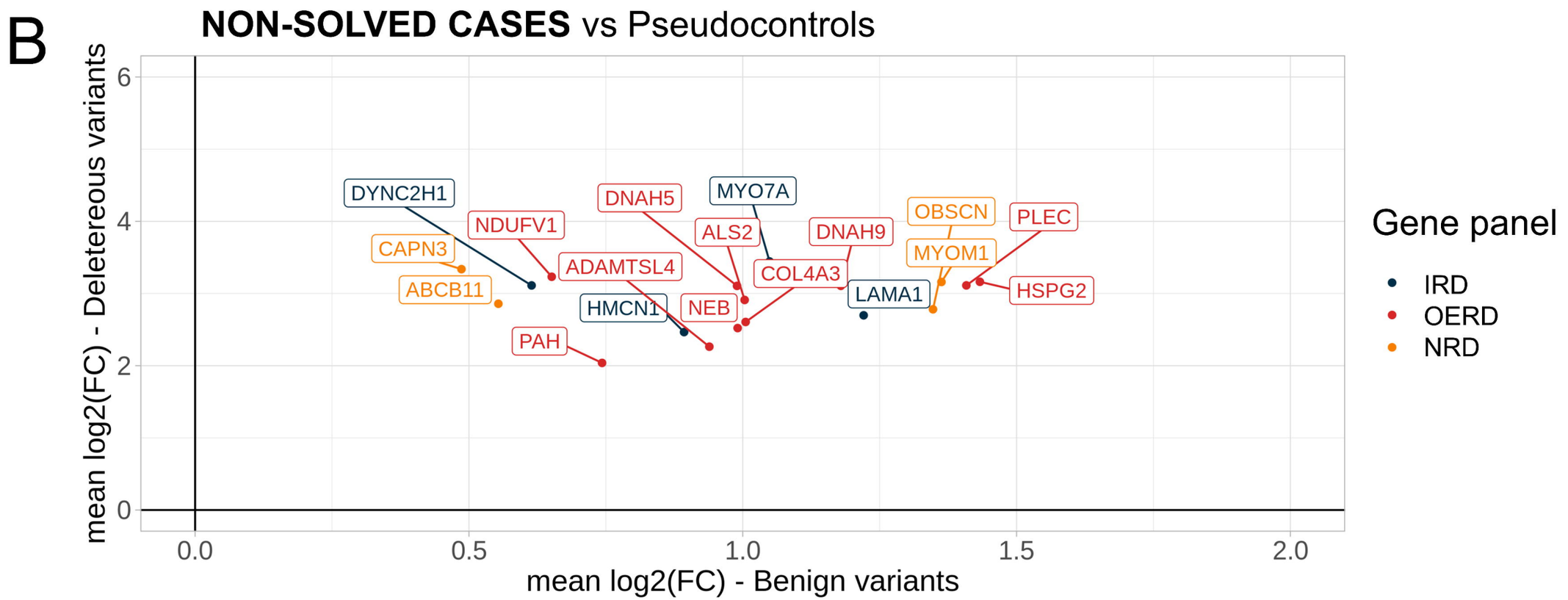

B
DDBB
~ 1.5m variants
QC filters
1.36m variants
POP filters
560k variants

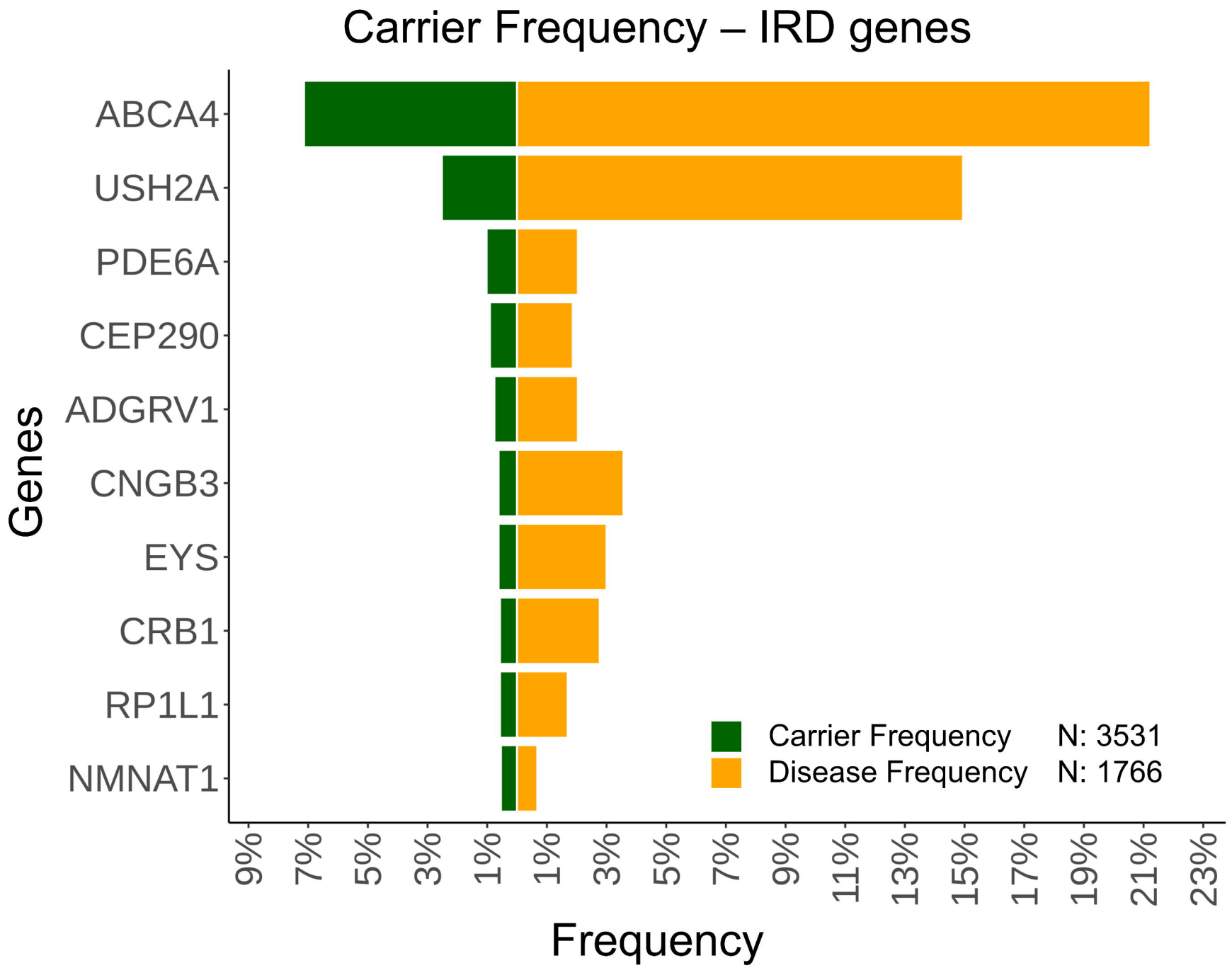

C
Variants
3946
Consequence
2486 intronic
513 missense
342 synonymous
250 up-downstream
195 UTR
24 frameshift / stop-gain
136 others
Inheritance
2203 recessive
1256 dominant
487 other
Pathogenicity
3371 not provided
379 benign
105 VUS
13 pathogenic
78 conflicting


D
ABCA4 n=288
USH2A n=274
EYS n=185
NR2E3 n=80
PROM1 n=39
SOLVED
ABCA4 n=16
USH2A n=51
EYS n=65
NR2E3 n=26
PROM1 n=2
NON-SOLVED
Pathogenic
VUS


A
SOLVED CASES
AF Solved IRD
AF Pseudocontrol
0.000
0.025
0.050
0.075
0.100
log2(FC) >= 3.12
log2(FC) < 3.12
B
NON-SOLVED CASES
AF Non-solved IRD
AF Pseudocontrol
log2(FC) >= 3.96
log2(FC) < 3.96
C
IRD-MFV
N: 404
N: 762
Non-prioritized
N: 4191
N: 23020
p-value
4.77E-56
D
IRD-MFV
N: 148
N: 399
Non-prioritized
N: 2143
N: 21222
p-value
1.69E-32
PATHOGENICITY Benign Deleterious
E
IRD genes
OERD genes
NRD genes
p=5.97E-61
p=1.71E-06
2.75E-08
PERCENTAGE
177
593
80
1457
147
2141
IRD-MFV
Non-prioritized
PATHOGENICITY Benign Deleterious
F
IRD genes
OERD genes
NRD genes
p=ns
p=3.25E-06
1.90E-28
7
193
37
770
104
1180
PATHOGENICITY Benign Deleterious
WORKFLOW
Variant distribution
Compare all genes
Compare sub-panels


A
SOLVED CASES vs Pseudocontrols
mean log2(FC) - Deleterious variants
mean log2(FC) - Benign variants
0
2
4
6
0.0
0.5
1.0
1.5
2.0
SPTBN5
CDH23
PROM1
LAMA2
AHI1
CNGB3
ABCA4
ADGRV1
RPGR
CEP290
OPA1
DNAH11
CRB1
MPDZ
EYS
NEB
OTOF
ABCC8
USH2A
PLEC
DYNC2H1
TRPM1
MYO7A
PAH
COL7A1
DNAH5
ATM
DYSF
HSPG2
OBSCN
COL5A2
VPS13B
Gene panel
IRD
OERD
NRD


B
NON-SOLVED CASES vs Pseudocontrols
mean log2(FC) - Deletereous variants
mean log2(FC) - Benign variants
0
2
4
6
0.0
0.5
1.0
1.5
2.0
DYNC2H1
DNAH5
MYO7A
OBSCN
PLEC
NDUFV1
ALS2
DNAH9
CAPN3
MYOM1
ADAMTSL4
COL4A3
ABCB11
LAMA1
HSPG2
HMCN1
NEB
PAH
Gene panel
IRD
OERD
NRD


Carrier Frequency – IRD genes
Genes
ABCA4
USH2A
PDE6A
CEP290
ADGRV1
CNGB3
EYS
CRB1
RP1L1
NMNAT1
Carrier Frequency N: 3531
Disease Frequency N: 1766
9%
7%
5%
3%
1%
1%
3%
5%
7%
9%
11%
13%
15%
17%
19%
21%
23%
Frequency