The Emirati Genome Program Enables Population-wide Penetrance Estimation and Novel Discovery for Inherited Retinal Disease

Budour Alkaf[1], Aashish Jha[2], Abdelrahman Saad[1], Ahmed Al Awadhi[1], Albarah El-Khani[1], Andreas Henschel[3], Antonio Milano[1], Asma Al Mannaei[4], Ayesha Al Ali[1], Azza Attia[1], Aziz Khan[5], Daniel Sanchez[1], Eduardo Beltrame[5], Fahed Al Marzooqi[1], Gurunath Katagi[1], Haiguo Wu[1], Hanil Sajad[1], Imam Chishty[1], Islam Eltantawy[1], Joseph Mafofo[1], Kristine Wong[1], Lawrence Petalidis[1], Maaz Shaikh[1], Marie Ibrahim[1], Mohamed Alameri[4], Mohamed El-Hadidi[6], Omar Soliman[1], Pierre Zalloua[3], Raony Cardenas[1], Sahar Al Marzooqi[4], Shalini Behl[1], Shobith Pejathaya[1], Shreyaa Chandrashekar[1], Thyago Cardoso[1], Val Zvereff[1], Vinay Kusuma[1], Vineet Kumar[1], Wadha Mohammed Abdulrahman[4], Youssef Idaghdour[2], Javier Quilez[1], Arif Khan[7], Tiago Magalhaes[1]

[1]M42, Abu Dhabi, United Arab Emirates
[2] New York University Abu Dhabi, Abu Dhabi, United Arab Emirates
[3] Khalifa University, Abu Dhabi, United Arab Emirates
[4]Department of Health of Abu Dhabi, Abu Dhabi, United Arab Emirates
[5]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, United Arab Emirates
[6] University of Birmingham, Dubai, United Arab Emirates
[7]Cleveland Clinic Abu Dhabi Hospital, Abu Dhabi, United Arab Emirates
*Corresponding author. Email: [balkaf@m42.ae]

Abstract

Interpreting rare variants remains a major challenge in genomic medicine, particularly for genetically and phenotypically heterogeneous disorders like inherited retinal diseases (IRDs). We introduce a population-scale framework that unites three unprecedented resources: whole genome data from 504,000 individuals in the Emirati Genome Program (EGP); electronic health record (EHR) data from 426,000 of these participants through phenoEGP, phenotype resource derived from Malaffi; and genetically reconstructed pedigrees, made possible by the rich familial structures within the EGP. Together, these datasets enable real-world penetrance estimation of rare variants at a population scale, stratification of individuals and families by clinical actionability, and the discovery of novel IRD-associated variants. Among 1,879 known variants, we found 96 high or complete penetrance, supporting clinical integration, while 36 showed null penetrance, warranting reclassification. We identified 2,131 asymptomatic individuals and 1,013 families as potential candidates for preventive care. Using phenotype-first familial prioritization, we identified 69 likely pathogenic variants among high-disease-burden families, including a novel biallelic *BBS12* variant confirmed through cross-cohort screening. We also identified rare, actionable conditions like undiagnosed *CYP27A1*-related cerebrotendinous xanthomatosis. Our integrative framework enables early detection of at-risk individuals, informs therapeutic decision-making, and uncovers novel disease loci, underscoring the power of population-scale genomic screening to advance precision medicine in rare diseases and provide a scalable model for other heterogeneous conditions.

Introduction

The EGP is among the world's largest population-scale sequencing initiatives, with ~30x whole-genome sequencing (WGS) data from over half a million individuals across the Emirati population. When linked through phenoEGP, a phenotype resource derived from Malaffi, the Abu Dhabi health information exchange, these genomes are connected to rich, 10-year longitudinal EHR encoded in ICD-10. Coupled with genetically reconstructed pedigrees, enabled by the rich distinctive familial structures, these resources converge to create a uniquely powerful setting for precision medicine. In this study, we bring these datasets together to establish a population-scale framework that quantifies real-world penetrance, refines variant interpretation, and stratifies individuals and families by clinical actionability, while uncovering novel disease-associated variants.

Here, we apply this framework to inherited retinal IRDs, a genetically and clinically heterogeneous group of disorders [1], whose complexity has long hindered accurate penetrance estimation and variant interpretation. Collectively, they represent the leading cause of vision loss worldwide, with an estimated incidence of 1 in 3,000 [2]. In the Arabian Gulf, the burden of IRDs is markedly elevated. Recent estimates from affected cohorts across 16 Arab countries, with nearly half of cases originating from Saudi Arabia, suggest that up to 5% of the population may be affected by a genetic disorder involving the eye and adnexa, significantly higher than in European populations [3] [4]. This disparity could be resulted from a combination of factors including high rate of consanguineous marriages[5], founder effects, and genetic drift making IRDs a leading cause of visual impairment in the region[6] [7].

More than 330 genes have been shown to be associated with IRDs (RetNet[8], accessed April 2025), encoding proteins involved in functions ranging from phototransduction to retinal structure. IRDs patterns of inheritance span autosomal dominant, autosomal recessive, X-linked, and mitochondrial modes. In addition to genetic heterogeneity, IRDs exhibit broad clinical variability, ranging in onset, severity, progression, and syndromic involvement, complicating accurate diagnosis and gene-based stratification[9] [10] [11].

Genomic screening has already transformed care in IRDs. Individuals with biallelic *RPE65* mutations may benefit from voretigene neparvovec, a gene therapy that delivers a functional copy of the *RPE65* gene to retinal cells. This treatment can preserve and improve vision, but only if administered early, before significant photoreceptor loss occurs[12]. Likewise, individuals homozygous for *CYP27A1* mutations may develop cerebrotendinous xanthomatosis (CTX), a metabolic disorder that causes progressive neurological and visual decline. Importantly, CTX is treatable with chenodeoxycholic acid (CDCA) replacement, which can halt or reverse symptoms if started early, but the condition is often misdiagnosed and left untreated [13] [14]. These examples highlight the value of genomic screening when variant pathogenicity is determined. Unfortunately, for most IRD associated genes, the correlation between pathogenicity (phenotype) and genetic penetrance (genotype) remains to be resolved.

IRD trait penetrance estimates have traditionally relied on small, clinically ascertained cohorts or family-based studies [15] [16] [17], approaches prone to ascertainment bias and risk overestimation. IRDs

also show marked variable expressivity and incomplete penetrance, both of which further complicate variant interpretation. Ellingford et al. [17] reported that within the same family, carriers of pathogenic *PRPF31* variants may show variable expressivity from severe symptoms to no phenotype at all. A recent IRD study in the United Arab Emirates (UAE) identified key regional variants (e.g., *ABCA4* c.5882G>A), but the cohort size (91 individuals from 74 families) limited generalizability [4].

Large-scale genomic studies have refined penetrance estimates for some cancer and cardiovascular allelic variants [18] [19]. Recent evidence have also revealed that a significant number of known pathogenic variants exhibit variable, and often low penetrance across ethnically diverse population cohorts [18] [20]. Comparable efforts on IRDs remain scarce, particularly in high-disease burden understudied populations, like the UAE. To address this gap, we analyzed data from 504,000 Emirati individuals, nearly half of the national population, making this the largest IRD penetrance study to date. By integrating genomic, clinical, and familial information, we systematically estimated variant-level penetrance, re-evaluated pathogenicity classifications, and identified both known and novel contributors to IRD risk. This work provides a scalable framework for ancestry-matched variant interpretation and establishes the foundation for precision genomic screening of vision disorders in the region and beyond.

Results

Study Population, Familial Structure, and IRD Burden

An overview of the study design, including genomic screening, phenotype prioritization, family-based mapping, and variant discovery is demonstrated in **Figure 1**. WGS data from 504,000 individuals enrolled in the EGP were analyzed, of whom 426,382 had complete phenoEGP data. Among them, 52.4% were women with mean age of 30.9 ± 17.4 years. The age distribution was broad, with 24.3% of the cohort aged 20–30 years and 9.6% under the age of 10 (**Table 1**).

A defining strength of the EGP lies in its unprecedented national scale and deep familial representation. The program covers nearly three-quarters of the Emirati population, providing a level of sampling density previously unachieved in population genomics. Combined with the high prevalence of consanguinity and large, multigenerational family networks characteristic of the Emirati population, this enables genome-wide reconstruction of relatedness across tens of thousands of families. Together, these features make the EGP uniquely suited to quantify population-level penetrance, validate variant pathogenicity through familial segregation, and uncover novel genetic contributors to inherited retinal diseases.

Using genome-wide kinship inference, 52,615 genetically inferred families were identified, of which 51,178 contained at least one individual with phenoEGP. These included 1,667 families with three or more generational lineages and 9,876 consanguineous families (defined by an inbreeding coefficient > 2× the population mean). The mean inbreeding coefficient across the cohort was $F = 0.037$, consistent with elevated endogamy in the region, over 10 times higher than what is typically observed in European populations ($F \approx 0.0026$) [21]. This deep population structure enables robust, consanguinity-aware modeling of variant segregation and penetrance at scale.

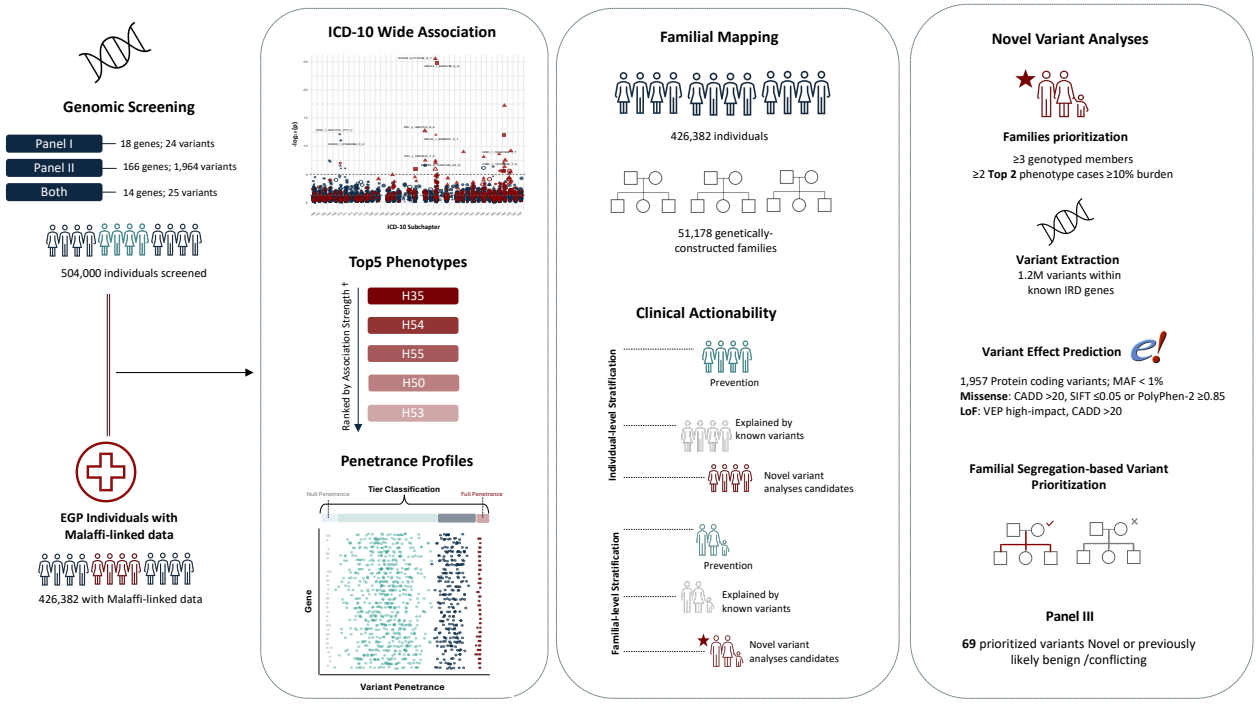


Figure 1. Study Design for Population-Scale Genomic and EHR-Based Analysis of Inherited Retinal Diseases (IRDs). Genomic screening of 504,000 individuals in the Emirati Genome Program (EGP) identified individuals harboring known and candidate IRD variants. 426,382 individuals with phenoEGP formed the analysis cohort used for all downstream analyses. ICD-10–wide association identified the Top5 phenotypes (top five IRD-related ICD-10 subchapters: H35, H50, H53, H54, H55) to define disease status and compute variant penetrance tiers. Familial mapping reconstructed 51,178 genetically defined families to assess segregation patterns and clinical actionability. Novel variant analyses, including functional prediction and family-based prioritization yielded 69 high-priority variants (novel or previously conflicting) in known IRD genes.

Table 1. Cohort characteristics and age-specific distribution of retinal dystrophies and severe visual dysfunction. This table summarizes the demographic profile of individuals from the Emirates Genome Program (EGP) with linked electronic health records (EHRs) from Malaffi (N = 426,382), including age distribution, sex, and mean age. The lower portion details the number and percentage of individuals diagnosed with retinal dystrophies and vision loss conditions, stratified by ICD-10 subchapter and age group.

Total EGP individuals with Malaffi Records	426,382
Sex [N(%)]	
Female	223,523 (52.4)
Male	202,491 (47.5)
Age [Mean±SD years]	30.9±17.4
Age distribution [N(%)]	
<10 years	40,780 (9.6)
10 to 20 years	80,078 (18.8)
20 to 30 years	103,760 (24.3)
30 to 40 years	71,487 (16.8)
40 to 50 years	66,761(15.7)
>50 years	63,512 (14.9)

		Age-specific Distribution of Retinal Dystrophies and Severe Visual Dysfunction [N(%)]						
Retinal and Vision-Loss Disease Clusters (ICD-10 H35/H54)	**ICD-10 Subchapter**	**All**	**<10**	**10 to 20**	**>20 to 30**	**>30 to 40**	**>40 to 50**	**>50**
Blindness (both eyes)	H54	190 (0.04)	9 (0.02)	23 (0.03)	19 (0.02)	27 (0.04)	21 (0.03)	91 (0.14)
Blindness (one eye)	H54	593 (0.14)	12 (0.03)	37 (0.05)	51 (0.05)	65 (0.09)	70 (0.10)	358 (0.56)
Legal Blindness (USA)	H54	103 (0.02)	2 (<0.01)	13 (0.02)	16 (0.02)	12 (0.02)	10 (0.01)	50 (0.08)
Low Vision (both/one eye)	H54	1974 (0.46)	76 (0.19)	538 (0.67)	366 (0.35)	174 (0.24)	185 (0.28)	635 (1.00)
Unqualified Visual Loss	H54	1500 (0.35)	68 (0.17)	301 (0.38)	304 (0.29)	122 (0.17)	174 (0.26)	531 (0.84)
Hereditary Dystrophies	H35	1304 (0.31)	61 (0.15)	184 (0.23)	170 (0.16)	154 (0.22)	218 (0.33)	517 (0.81)
Macular Degeneration	H35	7381 (1.73)	25 (0.06)	135 (0.17)	237 (0.23)	383 (0.54)	733 (1.10)	5868 (9.24)
Other Profilerative Retinopathies	H35	149 (0.03)	2 (<0.01)	6 (0.01)	9 (0.01)	8 (0.01)	20 (0.03)	104 (0.16)
Peripheral Degeneration	H35	3379 (0.79)	32 (0.08)	267 (0.33)	787 (0.76)	570 (0.80)	678 (1.02)	1045 (1.65)
Retinal Edema & Ischemia	H35	1148 (0.27)	10 (0.02)	42 (0.05)	63 (0.06)	107 (0.15)	163 (0.24)	763 (1.20)
Retinal Detachment/Serous	H35	856 (0.20)	5 (0.01)	13 (0.02)	31 (0.03)	90 (0.13)	262 (0.39)	455 (0.72)
Retinal Hemorrhage	H35	1128 (0.26)	28 (0.07)	16 (0.02)	44 (0.04)	71 (0.10)	120 (0.18)	849 (1.34)
ROP (Retinopathy of Prematurity)	H35	392 (0.09)	355 (0.87)	20 (0.03)	7 (0.01)	4 (0.01)	2 (<0.01)	4 (0.01)
Unspecified Retinopathy	H35	1493 (0.35)	24 (0.06)	95 (0.12)	184 (0.18)	163 (0.23)	254 (0.38)	773 (1.22)
Vascular Retinopathy	H35	4856 (1.14)	45 (0.11)	94 (0.12)	260 (0.25)	453 (0.63)	898 (1.35)	3106 (4.89)

To characterize the phenotypic spectrum of IRDs in the EGP, we applied a phenotype-first approach using EHR data from phenoEGP. Specifically, ICD-10 codes were filtered to include all diagnoses under subchapters: H35 (Retinal Disorders) and H54 (Visual Impairment including Blindness) (**Table**

1, Supplementary Table1). This yielded a curated cohort with clinically documented retinal dystrophies and vision loss, including both congenital and age-related causes. Notably, ROP (Retinopathy of Prematurity) and hereditary retinal dystrophies were enriched in younger age groups, while macular degeneration and vascular retinopathy increased sharply with age (**Table1**).

To assess clinical penetrance of IRD-associated genetic variants in a systematic and unbiased manner, the analysis was expanded to include all ICD-10 subchapters within Chapter VII (Diseases of the Eye and Adnexa, H00–H59). This broader set of diagnoses allowed the capture of ophthalmic phenotypes potentially linked to IRDs or overlapping endophenotypes. To enrich for rare, potentially Mendelian conditions and reduce noise from ubiquitous, nonspecific diagnoses, individual ICD-10 codes with ≥1% population prevalence (n = 50) were excluded, resulting in a refined set of 2,323 codes for downstream analysis. This filter removed high-frequency non-specific diagnoses, such as hordeolum and conjunctivitis, and retained clinically significant but lower-prevalence retinal and visual dysfunction codes (**Supplementary Table 2**). Age-stratified summaries of these clinically relevant low-frequency ICD-10 codes revealed clear enrichment of degenerative and vascular conditions in older individuals, while certain inflammatory and congenital disorders remained more evenly distributed or clustered in early age groups (**Supplementary Table 3**).

Prevalence of Known IRD-associated Genetic Variants

Two complementary panels of pathogenic genetic variants associated with IRD were defined: Panel I (50 variants in 33 genes), a population-specific set derived from clinically diagnosed Emirati IRD patients at Cleveland Clinic Abu Dhabi curated by Dr. Arif Khan (co-author), and Panel II (15,770 variants in 262 genes), a broader reference set comprising ClinVar [22]-designated pathogenic or likely pathogenic variants in RetNet [8] genes.

Genomic screening was performed in 504,000 individuals. For each variant, heterozygous individuals were defined as carrying one copy of the alternate allele, while homozygous individuals carried two copies. Across the cohort, a total of 92,476 heterozygous (18.3%) and 2,460 homozygous (0.49%) individuals, spanning 204 genes and 2,028 unique variants were identified. Specifically, Panel I included 26,827 heterozygous and 562 homozygous individuals, while Panel II encompassed 91,215 heterozygous and 2,440 homozygous individuals. **Supplementary Figure 1** illustrates the extent of overlap and divergence between the two panels in terms of carriers, variants, and genes. Notably, 14 genes and 25 variants were shared between both panels, accounting for 25,574 heterozygous and 570 homozygous individuals present in both sets. Overall, homozygous individuals exhibited higher inbreeding levels than heterozygotes, consistent with the expected impact of consanguinity on the likelihood of rare deleterious alleles being inherited in homozygous form (**Supplementary Figure 2**). Full genomic screening results for each variant are provided in **Supplementary Table 4**.

For known variants from Panel I or II detected in the EGP, population-specific enrichment was assessed by comparing minor allele frequencies (MAFs) against overall genome-based frequencies reported in gnomAD (version 4.1). Only 9 out of 50 variants in the Emirati-based Panel I were in gnomAD and the majority of these (n = 8, 88.9%) showed higher MAFs in EGP, underscoring strong regional enrichment likely due to founder effects or population-specific disease variants. In contrast,

among the 936 Panel II variants detected in the EGP, only 221 (23.6%) had higher MAFs in EGP than in gnomAD (**Supplementary Table 4**).

The most frequent IRD variants in the cohort were *ABCA4*:c.5882G>A (chr1:94008251:C>T) and *ABCA4*:c.2570T>C (chr1:94055128:A>G) (**Supplementary Table 4**). These two variants, when co-inherited either as compound heterozygotes or as homozygous double mutants, are known to cause severe Stargardt disease and were previously reported in affected Emirati individuals [23] [24]. In our dataset, both variants exhibited striking allele enrichment: *ABCA4*:c.5882G>A had a MAF of 1.95% in EGP (5.7-fold higher than in gnomAD, 0.34%), and *ABCA4*:c.2570T>C was over 4,000-fold more frequent in EGP (0.25%) compared to gnomAD (0.000062%). Notably, 867 compound heterozygotes and 35 homozygous double-mutant individuals were identified, underscoring the high population-level burden of this pathogenic haplotype in the Emirati population (**Supplementary Table 4**).

ICD-10-Wide Phenotype Associations and Selection of IRD-Relevant Diagnoses

Given the genetic and phenotypic heterogeneity of IRDs, as well as variability in ICD-10 coding practices across clinical settings, an ICD-10-wide association analysis was conducted to systematically identify the most relevant subchapters for downstream penetrance and burden estimation. The focus was on subchapters within the H00–H59 block (Diseases of the Eye and Adnexa) and calculated enrichment scores for each, defined as the sum of log(OR) × $-\log_{10}$(p-value) across all panel variants, separately for heterozygous and homozygous individuals (**Supplementary Table 5**).

The top five subchapters with the strongest association signals were H35 (retinal disorders), H54 (visual impairment), H55 (nystagmus), H50 (strabismus), and H53 (visual disturbances) (**Supplementary Table 5** and **Supplementary Figure 3**). Importantly, these associations are not only statistically significant but also clinically meaningful. For example, H35 includes inherited and degenerative retinal conditions, H54 captures end-stage vision loss, and H53–H55 represent early symptoms such as photophobia, scotomas, and nystagmus, together encompassing the known phenotypic spectrum of IRDs. For all subsequent penetrance and burden analyses, IRD-affected individuals were defined as those having at least one diagnosis code within the top five enriched ICD-10 subchapters: H35, H54, H55, H50, and H53 (also referred to as "Top5 phenotypes" for short).

Population-scale, Real-world Penetrance Estimation

Clinical penetrance was defined at the variant level as the proportion of individuals harboring a given alternate allele who had at least one diagnosis from the Top5 phenotypes (H35, H54, H55, H50, H53) recorded in phenoEGP. Of the 2,028 unique variants identified through genomic screening, 1,879 (92.7%) could be evaluated for penetrance because they were observed in individuals with phenoEGP data, comprising 78,930 heterozygous and 2,150 homozygous carriers. For each variant, penetrance was estimated separately for heterozygous and homozygous carriers, stratified by mode of inheritance (AD, AR, X-linked, complex, or unknown).

For visualization in **Figure 2**, we applied two additional filters: (i) exclusion of heterozygous penetrance estimates for AR genes, and (ii) removal of variants observed only in singletons. After

these filters, 372 variants remained for display, providing a genome-wide overview of penetrance profiles stratified by zygosity. **Supplementary Figure 4** presents the full set of heterozygous penetrance estimates in AR genes, while detailed per-gene penetrance, odds ratio (OR), and risk difference (RD) results are shown for all 1,879 evaluable variants **in Supplementary Figure 5** (combined Top5 phenotypes) and **Supplementary Figures 6–10** (each individual subchapter).

To systematically interpret penetrance patterns in a population-scale context, a 7-tier classification schema was developed (**Supplementary Table 6**). This framework integrates observed penetrance proportions with a significance filter requiring at least two individuals having the relevant genotype (heterozygous or homozygous). Variants with 0% penetrance in ≥2 individuals were classified as significantly null (Tier 7), while those with ≥0.5 penetrance in ≥2 individuals were considered significantly high (Tiers 1–2). Intermediate tiers (Tiers 3–6) capture less confident or non-significant patterns, accounting for stochasticity in rare variant observations. While the schema was derived solely from penetrance values, estimates showed strong correlation with association metrics such as odds ratio and risk difference (**Supplementary Figure 11**).

Across the cohort, 37 variants (2.0%) were classified as Tier 7 (significantly null) (**Supplementary Tables 7-9**). These included: (i) 17 variants in AD genes showing no diagnoses among ≥2 heterozygous individuals, including five in *TREX1*, suggesting potential overestimation of pathogenicity or recessive inheritance (**Supplementary Tables 7 and 8**); (ii) 9 variants in AR genes with significantly null penetrance among homozygous individuals, including two variants in *USH2A*; (iii) 11 variants in five genes with other inheritance patterns, notably five in *ABCA4*, a gene known for complex allelic architecture and variable expressivity (**Supplementary Tables 7 and 9**).

Conversely, 96 variants (5.1%) in 52 genes met criteria for Tiers 1 or 2, indicating high or complete penetrance. These included 11 variants in genes *ABCA4* and *CNGB3* with complete penetrance (Tier 1) (**Supplementary Tables 7-9**). These variants represent robust candidates for clinical prioritization and integration into IRD diagnostic programs. Notably, several variants exhibited unexpected heterozygous effects in genes traditionally associated with AR inheritance, expanding the landscape of variant expressivity. This included five *USH2A* variants and six *CEP290* variants that reached Tier 2 penetrance among heterozygous carriers (**Supplementary Table 8**), warranting further investigation into their potential dominant or semi-dominant pathogenic roles.

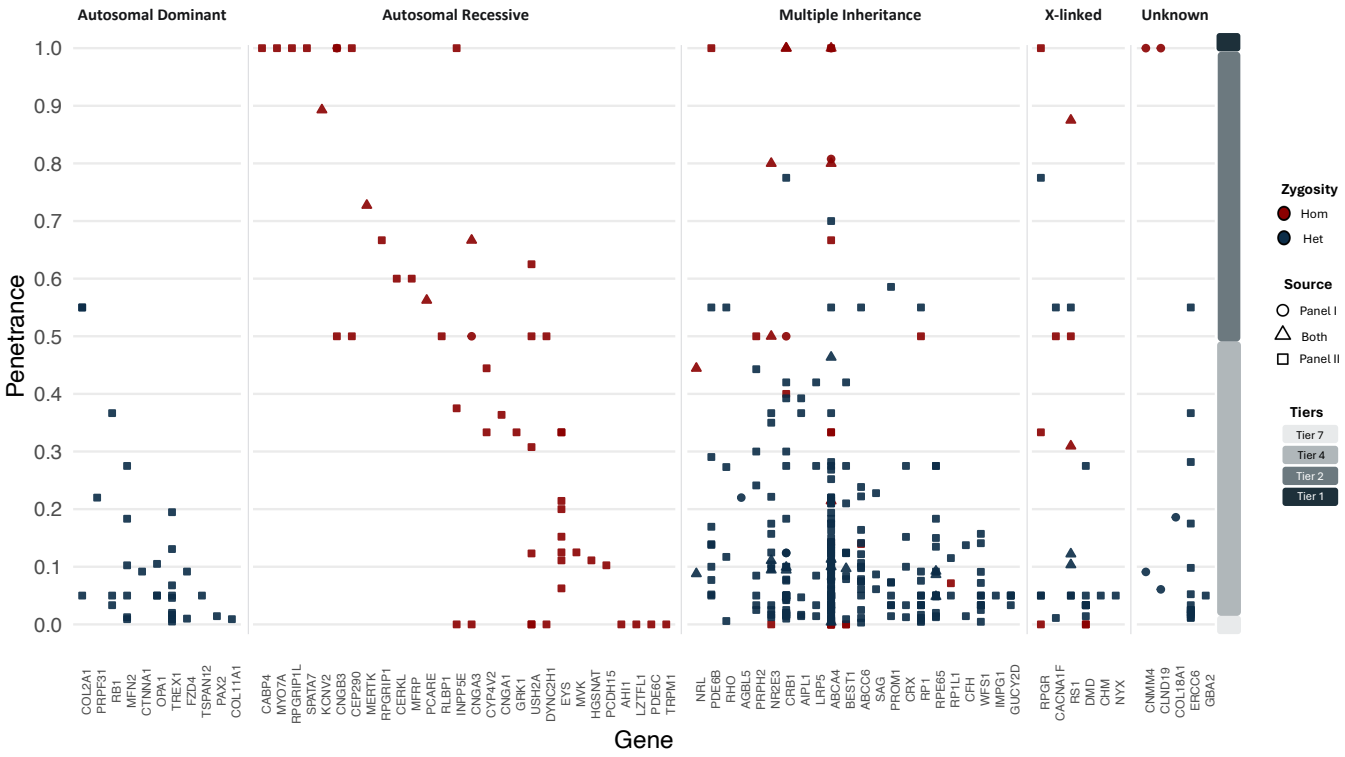


Figure 2. Penetrance of Pathogenic IRD Variants by Gene, Zygosity, and Mode of Inheritance. This plot summarizes variant-level clinical penetrance estimates for 372 pathogenic and likely pathogenic variants in RetNet genes, stratified by mode of inheritance. Each point represents a variant, with penetrance calculated as the proportion of individuals harboring variants presenting with any of the Top5 phenotypes (top five IRD-related ICD-10 subchapters: H35, H50, H53, H54, H55). Red symbols denote homozygous penetrance, and blue symbols denote heterozygous penetrance. Variant origin is indicated by shape (circle: Panel I, square: Panel II, triangle: both). Variants are annotated by penetrance tier: Tier 1 (fully penetrant, ≥2 harboring individuals), Tier 2 (high penetrance ≥0.5, ≥2 harboring individuals), Tier 4 (low penetrance <0.5, ≥2 harboring individuals), and Tier 7 (significantly null, penetrance = 0, ≥2 harboring individuals). Tiers requiring singleton harboring individuals (Tiers 3, 5, 6) are excluded, as penetrance estimates based on single harboring individuals were removed. Heterozygous penetrance for AR genes was removed for clarity (see Supplementary Figure 4 for the full dataset including AR heterozygotes). Within each inheritance mode, genes are ordered from highest to lowest mean penetrance.

To assess age-dependent expressivity, penetrance was stratified by age at analyses. Tier 1–2 variants were most prevalent in individuals aged 35 and above (5.5%), compared to 4.8% in 0–18 years and 3.6% in the 19–34 group (**Supplementary Tables 10–15**). Consistently, carriers with Top5 phenotypes were significantly older than carriers without Top5 phenotypes (median age 37.0 vs. 27.0 years, $p < 1\times10^{-4}$), further supporting the progressive nature of IRDs. Importantly, no age differences were observed between carriers with Tier 1–2 variants and Top5 phenotypes (median age 26.0 years) and carriers with Tier 7 variants without phenotypes (median age 28.0 years; $p = 0.11$), indicating that age alone does not fully explain penetrance tier differences.

To further validate our penetrance estimates, a sensitivity analysis restricted to individuals harboring only a single IRD-associated variant was performed, thereby minimizing potential confounding from multiple variants per individual. The overall proportions of significantly null and high-penetrance variants remained consistent in this subset, supporting the robustness of the tiering framework (**Supplementary Tables 16 and 17**). Notably, for the well-characterized *ABCA4* haplotype comprising c.5882G>A and c.2570T>C [24], we observed 7.8% penetrance among cis compound heterozygotes, 7.3% among trans compound heterozygotes, and 80.8% among homozygous double mutant, underscoring the markedly increased clinical impact when both pathogenic alleles are present in homozygosity (**Figure 3A**).

Finally, a subset of 27 Tier7 (significantly null penetrance) variants was revisited, comprising 14 in AD genes observed in the heterozygous state and 13 in any gene observed in the homozygous state. Among these, 15 variants had at least two submissions to ClinVar classified as pathogenic or likely pathogenic, suggesting prior clinical support for disease relevance. Notably, five Tier 7 variants in AD genes were associated with disorders that typically manifest later in adulthood (35–50 years) (**Supplementary Table 18**).

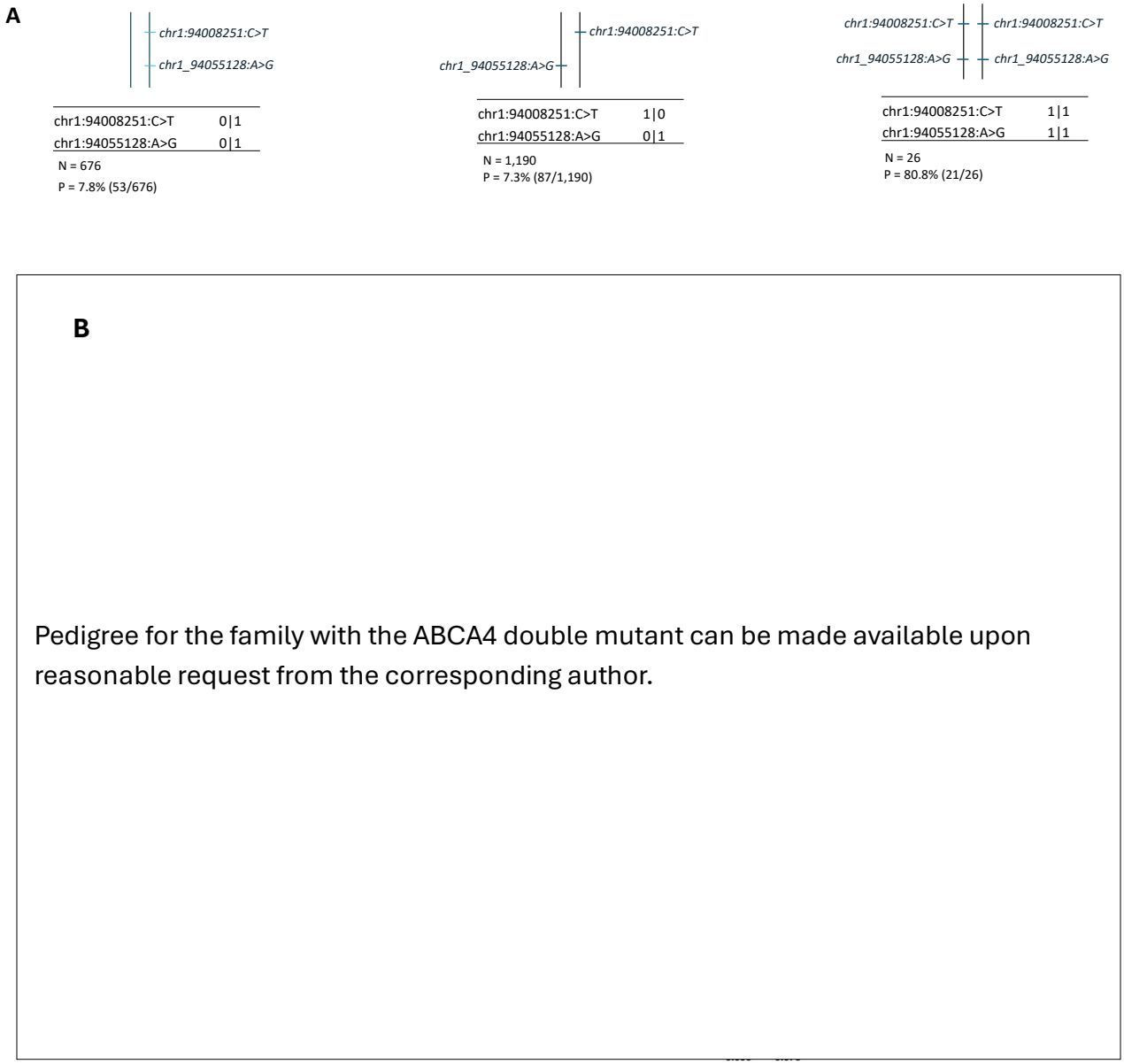


Figure 3. Pedigree and Genotypic Configuration of the ABCA4 Double Mutant (chr1:94055128:A>G; c.5882G>A and chr1:94008251:C>T; c.2570T>C) With Penetrance Estimates. (A) Schematic representation of the ABCA4 double mutant observed in cis, trans, or as double homozygous, annotated with the number of individuals identified in the cohort and the penetrance for each genotype, defined as the proportion of harboring individuals presenting with any of the Top5 phenotypes (top five IRD-related ICD-10 subchapters: H35, H50, H53, H54, H55). (B) Pedigree of a representative family segregating the ABCA4 double mutant, demonstrating full penetrance in double homozygotes and low penetrance in heterozygous harboring individuals (1/8 = 12.5%). Individual ages are shown inside symbols, with inbreeding coefficients (calculated using plink --het) displayed beneath. Affected individuals (cases) are annotated with their corresponding Top5 phenotypes. Color coding: white = non-harboring individual, healthy; dark red = homozygous, affected with Top5 phenotypes; grey = heterozygous, healthy; dark blue = heterozygous, affected with Top5 phenotypes. Question marks denote individuals not included in the EGP.

From Genomic Variants to Families: Translating Penetrance into Clinical Insights

To bridge variant-level penetrance with real-world familial and clinical outcomes, we leveraged the dense family representation within the EGP to examine how known IRD-associated variants aggregate and manifest across thousands of families.. Among 78,930 heterozygous and 2,150 homozygous individuals with phenoEGP, 8,220 had at least one diagnosis from the Top5 phenotypes (top five IRD-related ICD-10 subchapters: H35, H50, H53, H54, H55), mapping to 4,296 reconstructed families. To identify high-expressivity families, we filtered for those with ≥10 genotyped members, ≥2 variant carriers, and ≥50% penetrance among those harboring the variants, identifying 192 such families (**Supplementary Table 19**). These included a large family segregating a homozygous *ABCA4* double-mutant haplotype, *ABCA4*:c.5882G>A (chr1:94008251:C>T) and *ABCA4*:c.2570T>C with 80.8% penetrance (**Figure 3B**), demonstrating the marked phenotypic impact of carrying two pathogenic hits within this locus. We also identified two unrelated families harboring the same *KCNV2* variant (chr9:2718166:G>T; c.427G>T) with markedly different penetrance patterns (**Supplementary Figure 12**). One family exhibited 100% penetrance in homozygotes and 50% in heterozygotes individuals, while the other showed 50% penetrance in homozygotes and no expressivity in heterozygotes. This divergence in phenotypic outcome despite identical genotypes underscores the influence of modifying factors; genetic, epigenetic, or environmental, on disease expression within and across families.

To quantify the broader contribution of known variants to IRD burden, we focused on families with evidence of heritable disease, defined as having ≥3 genotyped individuals with phenoEGP, ≥2 Top5 phenotype cases, and a disease burden of ≥10%. Among 2,465 families meeting these criteria, known pathogenic variants explained 23.3% of H35 cases (with 77 families showing ≥50% of burden explained), 17.9% of H54 cases (2 families ≥50%), 16.7% of H55 cases (1 family ≥50%), and 20.9% of both H50 and H53 cases (32 and 135 families ≥50%, respectively). Within this group, 348 families (14.1%) had no individuals harboring any known variant and were prioritized as high-confidence candidates for novel discovery.

To translate our variant-level penetrance estimates into actionable insights, we stratified individuals and families based on clinical relevance. As illustrated in **Figure 4**, we categorized individuals and families across three tiers of clinical relevance. First, we identified 2,131 individuals harboring known IRD variants without any Top5 phenotype, representing candidates for (further investigation) preventive strategies. Within this group, 110 individuals had Tier 1 or Tier 2 variants, highlighting a subset with the highest likelihood of actionable outcomes. These individuals belonged to 1,013 families, including 36 families with high-tier individuals, offering further opportunities for early intervention in undiagnosed relatives. Second, a distinct set of individuals carried known IRD variants with clear phenotypic presentation, validating the penetrance-informed risk stratification framework. Finally, the remaining group comprised individuals and families with no known pathogenic variants but significant phenotypic burden, representing a valuable pool for novel variant discovery and gene-disease mapping efforts.

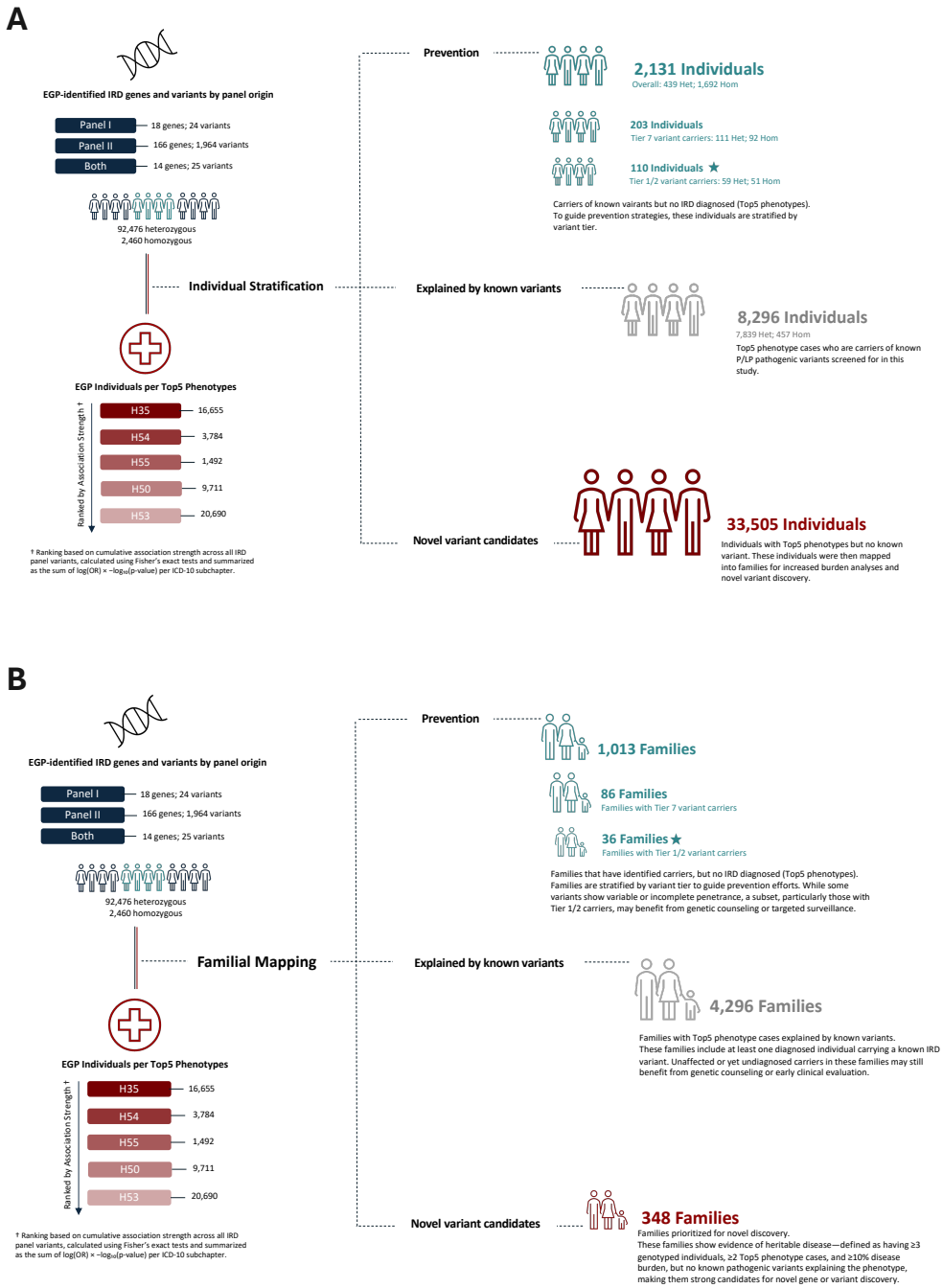


Figure 4. Integrated Genomic–Phenotypic Mapping to Stratify Individuals and Families by Clinical Actionability. This flowchart summarizes the integration of genomic screening data, phenoEGP, and family structure to stratify: (A) Individuals: carriers of IRD variants, classified into three tiers of actionability: those eligible for prevention strategies (n = 2,131), individuals with Top5 IRD-related phenotypes explained by known pathogenic variants (n = 8,296), and 33,505 individuals prioritized as novel variant candidates. (B) Families: 1,013 families identified for preventive care, 4,296 families explained by known variants, and 348 families prioritized for novel variant discovery, defined as having ≥3 carriers and ≥3 affected members. These analyses enable penetrance estimation, early clinical action, and identification of high-priority families for research and care. Genomic screening covered 204 genes, and 2,028 variants identified in the EGP cohort, and ICD-10-wide association analyses highlighted the Top5 phenotypes linked to IRD, used for downstream stratification.

Discovery and Prioritization of Novel Variants in High-Disease Burden IRD Families

We examined whether families showing high IRD burden but no known variants harbored novel alleles in established IRD genes. We focused on families with increased IRD burden specifically driven by diagnoses within H35 (retinal disorders) or H54 (visual impairment), the two most genetically enriched and clinically relevant ICD-10 subchapters identified in our association analysis (**Supplementary Table 5**). Of the 348 families previously prioritized for having increased IRD burden and 0% contribution from known variants across the Top5 phenotypes (top five IRD-related ICD-10 subchapters: H35, H50, H53, H54, H55), 219 met this refined criterion of concentrated disease burden in H35 or H54.

Across the 3,504 genotyped individuals in these 219 families (510 with H35 diagnoses, 111 with H54), we extracted all variants located within RetNet genes, yielding 1.19 million polymorphic sites. These variants were annotated using the Ensembl Variant Effect Predictor [25] (VEP), incorporating functional impact predictions from CADD [26], PolyPhen-2 [27], and SIFT [28].

Missense variants were prioritized if they occurred in protein-coding transcripts, had a CADD score >20, and were predicted to be damaging (PolyPhen score ≥0.85 or SIFT score ≤0.05). Additional variant types, including stop-gain, splice-site, and frameshift mutations, were prioritized if they were classified as high-impact by VEP, located in coding regions, and had CADD scores >20. After excluding variants with a cohort MAF >1%, we retained 1,957 rare high-priority variants. To further refine this list, we applied a segregation-based filtering strategy, designed to identify variants with strong genotype–phenotype concordance within families. Specifically, for each variant, we assessed its segregation pattern in families where multiple individuals had available genotypes and IRD-related diagnoses. Variants were excluded if they were found in families where both those with and without the variants exhibited IRD phenotypes, as this pattern suggests poor penetrance or lack of causal association. This family-based segregation filter is uniquely suited for large national cohorts with extended familial networks.

This prioritization process yielded 69 high-priority candidate variants across 34 families. Of these, 40 variants were not present in ClinVar, including 19 that were also absent from gnomAD, underscoring their potential novelty. Eight variants had been previously classified as benign or likely benign, while another 22 were reported as variants of uncertain significance (VUS) or with conflicting interpretations. Penetrance estimates for all 69 candidate variants were calculated across the full EGP cohort (N = 426,382 individuals), enabling population-scale assessment of their expressivity (**Supplementary Table 20).**

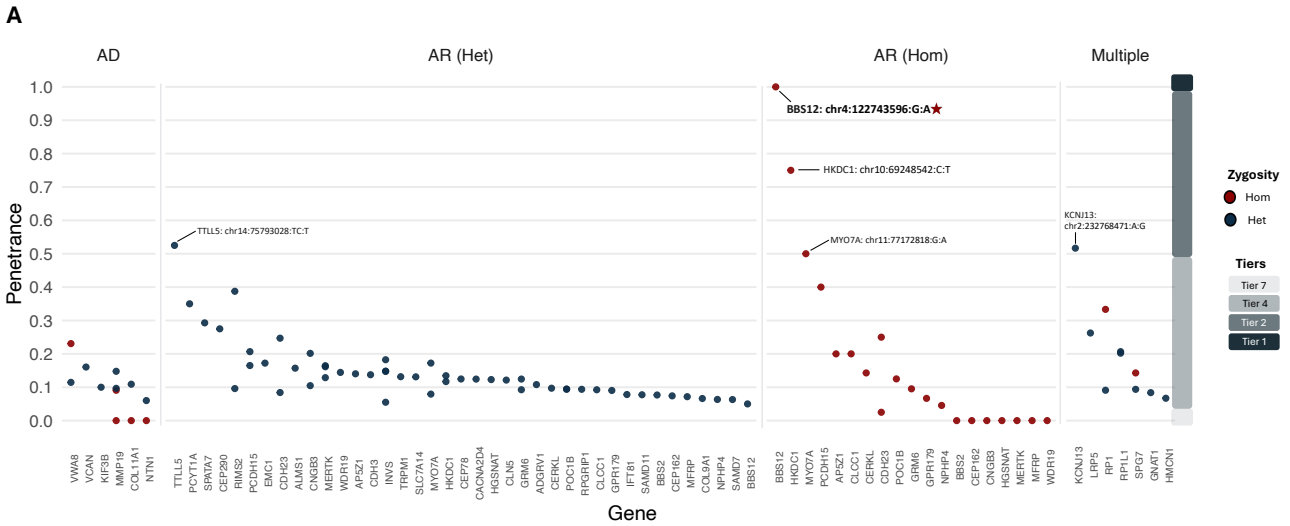
Based on penetrance estimates, 5 of the 69 variants met criteria for Tier 1 or Tier 2 (as defined in **Supplementary Table 6**). These included a biallelic stop-gain variant in *BBS12* (chr4:122743596:G>A), with a 100% penetrance among four homozygous individuals (**Figure 5A**, marked with a red star). The first two homozygous individuals were identified through our family-based prioritization strategy, which assessed variant segregation within IRD-enriched families. Both had concordant Top5 phenotypes. EGP-wide screening of the same *BBS12* variant then uncovered two additional affected siblings in an independent family. Their heterozygous sister and mother were

also carriers, but the mother phenoEGP data, which initially masked this family. Together, these findings demonstrate a consistent mode of inheritance across families and highlight the power of population-scale reverse phenotyping in complementing pedigree-based discovery (**Figure 5B**).

CYP27A1 and CTX: A Genomic Screening Success Case Beyond Classical IRDs

While our primary focus was on known and novel variants within established IRD genes, our genomic screening approach also uncovered clinically actionable findings in non-canonical genes associated with early visual dysfunction. A notable example is *CYP27A1*, implicated in CTX, a treatable autosomal recessive disorder in which early vision symptoms frequently occur. Given its clinical relevance and potential for early intervention, we screened 179 known pathogenic or likely pathogenic variants in *CYP27A1* across the cohort and identified 733 individuals harboring variants across 37 variants: 731 heterozygous and 2 homozygous individuals.

One homozygous individual presented with a constellation of vision-related and neurodevelopmental diagnoses, including autism, speech and language disorders, and unspecified eye disorders (**Supplementary Figure 13**). Notably, this individual also harbored a pathogenic *ABCA4* variant, and their family had been independently flagged in our IRD analysis. Several relatives heterozygous for both *CYP27A1* and *ABCA4* variants showed mild or subclinical visual phenotypes, suggesting possible modifier effects or partial expressivity. This case highlights a key strength of the EGP platform: its ability to detect clinically actionable, high-impact findings beyond traditional IRD classifications. While many variants identified in our IRD screen may require future therapeutic development, CTX has an existing, low-cost treatment, making early identification critical. This is a gap that population-scale genomic screening is uniquely positioned to fill.



B

Pedigrees for families with the novel *BBS12* variant can be made available upon reasonable request from the corresponding author.

Figure 5. Penetrance of Novel IRD Variants and Supporting Familial Evidence. (A) Variant-level clinical penetrance estimates for novel candidate IRD-associated variants identified through segregation-based prioritization. Variants are stratified by mode of inheritance (autosomal dominant, autosomal recessive heterozygotes, autosomal recessive homozygotes, and multiple inheritance). Penetrance is defined as the proportion of carriers with any of the Top5 phenotypes (top five IRD-related ICD-10 subchapters: H35, H50, H53, H54, H55). Variants are further categorized into penetrance tiers, with a Tier 1 fully penetrant BBS12 variant (chr4:122743596:G>A) annotated by a star. (B) Pedigrees for two families carrying the same novel variant: Family A (discovery), where the variant was initially identified, and Family B (independent), subsequently detected through EGP-wide screening. Circles represent females and squares males; ages are shown inside each symbol. Grey shapes denote heterozygous carriers without Top5 phenotypes (healthy), white shapes denote non-carriers, and dark red shapes denote homozygous carriers affected with one or more of the Top5 phenotypes (annotated below each symbol). Values beneath each symbol indicate the individual's inbreeding coefficient as estimated by plink --het.

Discussion

The EGP, one of the world's largest national sequencing initiatives, provides an unprecedented opportunity to advance rare disease genomics in a population that has been historically underrepresented in global reference datasets. Genomic interpretation for rare diseases is often limited by a lack of large-scale, population-specific data, particularly for genetically and phenotypically heterogeneous conditions like IRDs. In this study, we introduce a scalable population-level framework that combines WGS, 10-year longitudinal EHR data, and reconstructed familial structure to estimate penetrance, prioritize clinically relevant variants, and identify novel contributors to IRD risk. By leveraging data from over 500,000 Emiratis in the national genome program, we demonstrate the power of integrated, population-specific analyses to generate high-confidence variant classifications and guide precision medicine in historically understudied, consanguineous populations.

Our population-scale analysis yielded high-confidence penetrance estimates for over 1,800 known IRD variants, enabling both refinement of clinical variant classification and identification of actionable alleles. Our penetrance-based tiering framework provides nuanced lens for evaluating variant pathogenicity in the context of IRDs. We found 96 variants across 52 genes that met Tier1 and 2 criteria, which demonstrated high or complete penetrance in a population-matched cohort, represent strong candidates for clinical prioritization. These variants not only reinforce known gene–phenotype relationships but also offer a path forward for expanding diagnostic panels and refining risk stratification strategies, particularly in high-risk populations, like the UAE.

At the same time, our identification of a distinct subset of variants with significantly null penetrance (Tier 7) highlights critical limitations in current pathogenicity annotations. The presence of autosomal dominant variants with no documented disease among multiple carriers, and autosomal recessive variants lacking homozygous disease manifestation, raises concerns about potential overclassification in public databases such as ClinVar. This is especially notable in genes like *TREX1* and *ABCA4*, whose allelic complexity, incomplete penetrance, or misassigned inheritance models may contribute to erroneous clinical interpretations. These findings align with growing concerns raised in recent studies [29] [16] [30] regarding the reliability of existing variant classifications, and underscore the importance of large-scale, ancestry-matched penetrance data for the systematic re-evaluation of variant pathogenicity, improved risk prediction, and more accurate clinical translation.

We also observed a notable number of variants with potentially unexpected heterozygous effects in genes traditionally associated with AR inheritance, including *USH2A* and *CEP290*. While this may indicate dosage sensitivity, dominant-negative mechanisms, or partial loss-of-function, further functional validation is required. Importantly, our results highlight that penetrance may not be uniform across families, even for the same variant. This inter-family variability aligns with prior studies demonstrating the influence of ascertainment bias and genetic background on penetrance estimates[16] [31] [32]. While this underscores the limitations of relying solely on single-family or phenotypically enriched studies, our data also reaffirm the value of clinically ascertained variants when evaluated in broader contexts. For instance, variants identified through local clinics (Panel I,

derived from Cleveland Clinic Abu Dhabi and curated by Dr. Arif Khan) demonstrated both higher penetrance and significant population enrichment, confirming their diagnostic relevance. A striking example is the *ABCA4* double mutant allele (c.5882G>A & c.2570T>C), previously linked to Stargardt disease [24 23]. In our cohort, this allele showed >80% penetrance in homozygotes across the EGP, and 100% within an extended family carrying three homozygous double mutant individuals. This dual perspective, acknowledging both intra-family variability and the power of curated clinical observations, supports a hybrid framework that integrated population-scale genomic analyses with clinically guided variant discovery.

In addition to reclassifying known IRD-associated variants, our framework revealed clinically actionable findings in genes not traditionally considered IRD-related but relevant due to overlapping phenotypes. One such example is *CYP27A1*, associated with CTX, a rare, progressive metabolic disorder in which visual symptoms often precede systemic manifestations [13 14]. In most clinical settings, such cases are exceptionally difficult to diagnose due to their rarity and phenotypic complexity. However, the scale of the EGP, covering nearly half of the national population, enabled us to identify a homozygous *CYP27A1* individual likely representing an undiagnosed CTX case. This highlights the transformative potential of population-scale genomic screening to detect rare but treatable diseases early, ensuring timely diagnosis and targeted intervention for individuals who might otherwise remain undiagnosed for years.

Beyond diagnostic yield, the stratification framework developed in this study provides a clinically actionable roadmap for implementing genomic screening at scale. By combining variant pathogenicity, phenotypic expressivity, and penetrance within reconstructed family structures, we were able to identify individuals and families with high potential for early diagnosis, genetic counseling, or preventive care, even in the absence of current symptoms. Crucially, the extensive familial structures reconstructed within the EGP empowered us to move beyond individual-level associations and quantify how known pathogenic variants cluster within families, offering a window into heritable disease architecture. This enabled the identification of genetically enriched families for clinical follow-up, while also flagging families with unexplained yet significant IRD burden, high-confidence candidates for novel gene discovery.

Notably, one such family harbored a biallelic stop-gain variant in *BBS12*, a chaperonin-like protein essential for BBSome assembly and ciliary function in photoreceptors, whose disruption causes Bardet–Biedl syndrome (BBS), a ciliopathy characterized by retinal degeneration, obesity, and renal anomalies [33 34]. Although no cure currently exists for Bardet–Biedl syndrome, targeted therapies are emerging: for instance, setmelanotide (Imcivree) has FDA approval for chronic weight management in BBS, and GLP-1 receptor agonists show promising metabolic benefits in preclinical models [35]. Early genetic diagnosis in BBS12-mutant families could therefore facilitate timely multidisciplinary support, such as obesity interventions, vision monitoring, and renal care. Importantly, the pathogenicity of this variant was confirmed at the population level when genome-wide screening revealed an additional independent family carrying the same *BBS12* variant, in which all homozygous individuals were affected, demonstrating complete (100%) penetrance and reinforcing its clinical and biological relevance. While we were unable to assess broader phenotypic associations in the

discovery family due to a lack of metabolic trait data, this represents a compelling next step. Such findings demonstrate how integrating genomic data with family structure enables earlier detection and intervention, even in genetically complex conditions like Bardet–Biedl syndrome.

Conclusions, Limitations, and Future Directions

Conclusions

This study introduces a scalable framework for estimating variant penetrance and identifying both known and novel contributors to IRDs in a highly consanguineous, understudied Arab population. Leveraging WGS, ICD-10-coded health records, and reconstructed family structure across over 500,000 individuals, we identified 96 high-confidence variants with high or complete penetrance, uncovered novel likely pathogenic variants in unsolved families, and flagged 36 variants with null penetrance warranting re-evaluation. Our penetrance-informed stratification allowed us to identify individuals and families with clear opportunities for clinical action, ranging from those eligible for early screening and genetic counseling to those harboring treatable but previously undiagnosed conditions such as *CYP27A1*-related CTX. At the same time, we flagged hundreds of phenotypically burdened families without known pathogenic variants, enabling a focused search for novel disease contributors. Together, these findings highlight a scalable framework for variant prioritization, risk stratification, and discovery, one that bridges genotype, phenotype, and family context to guide precision screening and unlock clinical insights at the national level.

Limitations

A key limitation of our approach lies in the reliance on ICD-10-coded diagnoses from phenoEGP, the national electronic health record system. Diagnostic completeness may vary across clinics and hospitals, and individuals diagnosed outside the participating institutions may be misclassified as unaffected, potentially underestimating penetrance. Additionally, coding inconsistencies and differences in clinical practice may introduce variability in phenotype capture. While our stringent filtering and disease subchapter prioritization strategies help mitigate these issues, residual noise is likely. Genotyping errors, annotation inaccuracies, or incomplete understanding of variant biology may also confound results, especially for variants with complex mechanisms.

Future Work

Building on these findings, future efforts should prioritize the clinical translation of high-confidence, high-penetrance variants into actionable practice. Variants demonstrating complete or near-complete genotype–phenotype concordance are strong candidates for immediate inclusion in national diagnostic pipelines, premarital screening programs, and prospective genomic newborn screening. To maximize their impact, this should be coupled with structured clinical integration strategies, ranging from genetic counseling to phenotype surveillance protocols tailored by variant effect and inheritance. In parallel, individuals and families identified through this framework can inform genetics-based eligibility assessments for emerging therapies, particularly in cases with established treatment pathways or ongoing clinical trials.

References

1. Chen, T.-C. *et al*. Genetic characteristics and epidemiology of inherited retinal degeneration in Taiwan. *NPJ Genomic Med*. **6**, 16 (2021).
2. Karali, M. *et al*. Genetic epidemiology of inherited retinal diseases in a large patient cohort followed at a single center in Italy. *Sci. Rep*. **12**, 20815 (2022).
3. Berger, W., Kloeckener-Gruissem, B. & Neidhardt, J. The molecular basis of human retinal and vitreoretinal diseases. *Prog. Retin. Eye Res*. **29**, 335–375 (2010).
4. Jaffal, L. *et al*. The genetic landscape of inherited retinal dystrophies in Arabs. *BMC Med. Genomics* **16**, 89 (2023).
5. Al-Gazali, L. & Hamamy, H. Consanguinity and dysmorphology in Arabs. *Hum. Hered*. **77**, 93–107 (2014).
6. Chekroun, I. *et al*. Genomics of rare diseases in the Greater Middle East. *Nat. Genet*. **57**, 505–514 (2025).
7. Khan, A. O. Ocular genetic disease in the Middle East. *Curr. Opin. Ophthalmol*. **24**, 369–378 (2013).
8. Daiger, SP; Rossiter, BJF; Greenberg, J; Christoffels, A; Hide, W. Data services and software for identifying genes and mutations causing retinal degeneration. in vol. 39 S295 (Investigative Ophthalmology & Visual Science).
9. Galvin, O. *et al*. The Impact of Inherited Retinal Diseases in the Republic of Ireland (ROI) and the United Kingdom (UK) from a Cost-of-Illness Perspective. *Clin. Ophthalmol. Auckl. NZ* **14**, 707–719 (2020).
10. Gao, F.-J. *et al*. Genetic and Clinical Findings in a Large Cohort of Chinese Patients with Suspected Retinitis Pigmentosa. *Ophthalmology* **126**, 1549–1556 (2019).

11. Fenner, B. J. *et al*. Gene-Based Therapeutics for Inherited Retinal Diseases. *Front. Genet.* **12**, 794805 (2021).

12. Russell, S. *et al*. Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *Lancet Lond. Engl.* **390**, 849–860 (2017).

13. Koyama, S. *et al*. Cerebrotendinous Xanthomatosis: Molecular Pathogenesis, Clinical Spectrum, Diagnosis, and Disease-Modifying Treatments. *J. Atheroscler. Thromb.* **28**, 905–925 (2021).

14. van Hellenberg Hubar, J. L., Joosten, E. M. & Wevers, R. A. Cerebrotendinous xanthomatosis. *Clin. Neurol. Neurosurg.* **94 Suppl**, S165-167 (1992).

15. Ullah, M. *et al*. A comprehensive genetic landscape of inherited retinal diseases in a large Pakistani cohort. *NPJ Genomic Med.* **10**, 31 (2025).

16. Paterson, A. D., Seok, S.-C. & Vieland, V. J. The effect of ascertainment on penetrance estimates for rare variants: Implications for establishing pathogenicity and for genetic counselling. *PloS One* **18**, e0290336 (2023).

17. Ellingford, J. M., Hufnagel, R. B. & Arno, G. Phenotype and Genotype Correlations in Inherited Retinal Diseases: Population-Guided Variant Interpretation, Variable Expressivity and Incomplete Penetrance. *Genes* **11**, (2020).

18. Forrest, I. S. *et al*. Population-Based Penetrance of Deleterious Clinical Variants. *JAMA* **327**, 350–359 (2022).

19. Fahed, A. C. *et al*. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat. Commun.* **11**, 3635 (2020).

20. Ciesielski, T. H., Sirugo, G., Iyengar, S. K. & Williams, S. M. Characterizing the pathogenicity of genetic variants: the consequences of context. *Npj Genomic Med.* **9**, 3 (2024).

21. Schraiber, J. G., Shih, S. & Slatkin, M. Genomic tests of variation in inbreeding among individuals and among chromosomes. *Genetics* **192**, 1477–1482 (2012).

22. National Center for Biotechnology Information (NCBI). ClinVar [VCF data].

23. Khan, A. O. PHENOTYPE-GUIDED GENETIC TESTING OF PEDIATRIC INHERITED RETINAL DISEASE IN THE UNITED ARAB EMIRATES. *Retina Phila. Pa* **40**, 1829–1837 (2020).

24. Khan, A. O. HOMOZYGOSITY FOR A NOVEL DOUBLE MUTANT ALLELE (G1961E/L857P) UNDERLIES CHILDHOOD-ONSET ABCA4-RELATED RETINOPATHY IN THE UNITED ARAB EMIRATES. *Retina Phila. Pa* **40**, 1429–1433 (2020).

25. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).

26. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

27. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

28. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).

29. Richards, S. *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. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17**, 405–424 (2015).

30. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

31. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* **15**, 901–913 (2005).

32. Manrai, A. K. *et al*. Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med*. **375**, 655–665 (2016).

33. Stoetzel, C. *et al*. Identification of a novel BBS gene (BBS12) highlights the major role of a vertebrate-specific branch of chaperonin-related proteins in Bardet-Biedl syndrome. *Am. J. Hum. Genet*. **80**, 1–11 (2007).

34. Hildebrandt, F. & Zhou, W. Nephronophthisis-associated ciliopathies. *J. Am. Soc. Nephrol. JASN* **18**, 1855–1871 (2007).

35. Nowak-Ciołek, M. *et al*. Collaborative effort: managing Bardet-Biedl syndrome in pediatric patients. Case series and a literature review. *Front. Endocrinol*. **15**, 1424819 (2024).

36. Illumina. DRAGEN secondary analysis.

37. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* **1303**, (2013).

38. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl*. **34**, 3094–3100 (2018).

39. Sentieon. (2025).

40. Zheng, Z. *et al*. Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat. Comput. Sci*. **2**, 797–803 (2022).

41. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*. **47**, D1038–D1043 (2019).

42. Illumina Inc. Infinium Global Screening Array-24 v3.0 BeadChip Data Sheet. (2019).

43. Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

44. Manichaikul, A. *et al*. Robust relationship inference in genome-wide association studies. *Bioinforma. Oxf. Engl*. **26**, 2867–2873 (2010).

45. Csardi, G. & Nepusz, T. The Igraph Software Package for Complex Network Research. *InterJournal* **Complex Systems**, 1695 (2005).

46. Sinnwell, J. P., Therneau, T. M. & Schaid, D. J. The kinship2 R package for pedigree data. *Hum. Hered.* **78**, 91–93 (2014).

Methods

Ethical Compliance

All genomic sequencing and participant consent procedures were conducted under the governance of the Emirati Genome Program (EGP), which operates under national regulatory and ethical oversight. This study was approved by the Department of Health Abu Dhabi Human Research Ethics Committee for the use of EGP data in genomic screening, EHR linkage via Malaffi (resulting in the integrated dataset referred to as phenoEGP), and downstream analyses (IRB protocol number [DOH/ADHRTC/2025/1032]). All participants provided informed consent for research use of their de-identified genomic and phenotype data. All analyses adhered to relevant national guidelines on human subject research and data protection.

Study Design

An overview of the study design is shown in **Figure 1**. We performed population-scale genomic screening of 504,000 Emirati individuals enrolled in the EGP. Genomic screening targeted pathogenic and likely pathogenic variants from two curated gene panels: Panel I, consisting of variants previously identified in local IRD diagnostic clinics, and Panel II, derived from ClinVar[22]-annotated variants in known IRD-associated genes[8]. The curation strategy and inclusion criteria for both sets are described in detail in the *Panel Curation* section below.

Of the total cohort, 426,382 individuals (84.6%) had linked EHR data available through phenoEGP, a phenotype resource integrating EGP genomes with Malaffi-derived records and were included in downstream phenotype-based analyses. We conducted an ICD-10-wide association analysis to identify vision-related subchapters most enriched among variant carriers, from which the Top5 phenotypes (top five IRD-related ICD-10 subchapters: H35, H50, H53, H54, H55) were selected for downstream analyses. Individuals with at least one diagnosis in these Top5 subchapters were considered to have IRD-related disease status.

Clinical penetrance was estimated for each variant separately in heterozygous carriers (individuals with one copy of the alternate allele) and homozygous carriers (individuals with two copies), defined as the proportion with a diagnosis in one or more of the Top5 phenotypes (the IRD-related ICD-10 subchapters H35, H50, H53, H54, and H55). We reconstructed familial relationships through kinship inference and mapped both disease status and carrier status into family structures. Based on these data, families were stratified into three categories: (1) families with carriers but no IRD diagnosis (potential prevention opportunities for further investigation), (2) families with IRD explained by known variants from Panel I or II, and (3) high-disease burden families with IRD but without explanatory variants in Panel I or II, prioritized for novel discovery.. Functional annotation and prioritization of candidate variants in these unsolved families was performed using variant effect predictors, as described in the *Variant Annotation* section.

Study Population

Participants were enrolled in the EGP; a nationwide initiative launched in 2019 by the Department of Health of Abu Dhabi (DOH) to sequence the genomes of Emirati citizens for research and precision medicine. Our current analysis included whole-genome sequencing that had been completed for 504,000 individuals recruited between January 2020 and October 2024. Additional details on sequencing platforms and variant curation are provided in the following section.

Among these, 426,382 individuals had linked phenoEGP available through phenoEGP. Individuals without phenoEGP data or missing essential demographic information were excluded from phenotype-based analyses. All age information reported in this study reflects each individual's age at the time of analysis and was used consistently in age-stratified and comparative analyses.

Electronic Health Record Integration and Phenotype Extraction

EHRs were accessed through Malaffi (Arabic for "my file"), Abu Dhabi's health information exchange (HIE) operated by Abu Dhabi Health Data Services (ADHDS), an M42 company established in partnership with the Department of Health Abu Dhabi (DOH). Malaffi provides centralized, longitudinal clinical records for patients across public and private healthcare facilities in the Emirate. For this study, we used the subset of Malaffi data specifically linked to participants in the EGP hereafter referred to as phenoEGP. PhenoEGP comprises 10 years of longitudinal clinical data (2013–2024), including diagnoses, procedures, prescriptions, and laboratory results, integrated with genomic data from the EGP to enable large-scale genotype–phenotype analyses.

A total of 426,382 individuals from the 504,000 sequenced in the EGP had phenoEGP data and were included in phenotype-based analyses. Diagnoses were encoded using the International Classification of Diseases, Tenth Revision (ICD-10). To define vision-related phenotypes, we focused on subchapters within ICD-10 Chapter VII: Diseases of the Eye and Adnexa (H00–H59). To enrich for rare, potentially Mendelian conditions and reduce noise from ubiquitous, non-specific diagnoses, we excluded individual ICD-10 codes with ≥1% prevalence in the EGP population (n= 50). This filtering resulted in a refined set of 2,323 ICD-10 codes used for downstream association and penetrance analyses.

We then conducted an ICD-10-wide association analysis using this filtered set to identify the ICD-10 subchapters most significantly enriched among variant carriers, as detailed in the Statistical Analysis section. Based on the strength of association (ranked by cumulative $-\log_{10}P$ values across all variants), the five most enriched subchapters were selected for downstream phenotypic stratification. These included: H35 (retinal disorders), H54 (visual impairment), H55 (nystagmus), H50 (strabismus), and H53 (visual disturbances), were selected for downstream analyses. Together, these five subchapters, referred to as the Top5 phenotypes, served as proxies for IRD manifestations across the phenoEGP.

In the case of *CYP27A1*, a gene associated with CTX, a manual review of phenoEGP was conducted for homozygous individuals identified, as CTX-related features may extend beyond vision-specific ICD-10 subchapters.

Whole Genome Sequencing, Sample and Variant QC

Whole genome sequencing was performed to a target depth of 30x in one of three platforms: Illumina, MGI, or Oxford Nanopore Technologies (ONT). Reads were aligned to the human genome assembly GRCh38 with the use of the DRAGEN Bio-IT Platform 3.9 (Illumina)[36], BWA-MEM algorithm (MGI)[37] and Sentieon's acceleration of Minimap2 (ONT)[38]. Variant calling was conducted using DRAGEN Bio-IT Platform 3.9 (Illumina)[36], Sentieon's HaplotypeCaller algorithm (MGI)[39], and Clair2 (ONT)[40].

Sample-level quality control was applied to exclude individuals with low sequencing coverage (<26x), sex discordance, low mapping rate (<70%), or signs of contamination (contamination >1% or Het/Hom ratio ≥2.5). Variant-level QC excluded variants with low genotype quality or depth, using platform-specific thresholds as outlines in **Supplementary Table 21**. Only high-confidence variants and samples passing all filters were retained for downstream analyses.

As part of the genomic screening workflow, additional cross-platform enrichment analyses computed using Fisher's exact tests were performed to detect variants disproportionately over- or under-called in specific platforms. These were used to exclude variants likely affected by platform-specific technical bias from downstream analyses.

Panels Curation and Genomic Screening

To evaluate the genomic landscape of IRDs, we curated two variant sets from complementary sources: (1) Panel I consisted of pathogenic or likely pathogenic variants previously identified in local IRD diagnostic clinics, including confirmed clinical cases from Cleveland Clinic Abu Dhabi. These variants represent known pathogenic alleles within the UAE population and were curated by Dr. Arif Khan, one of the co-authors of this study, (2) Panel II included all ClinVar[22]-annotated pathogenic or likely pathogenic variants (as of November 2024) with overall star ratings of ≥1corresponding to single submitter/conflicting interpretations, multiple submitters with no conflict, or expert panel review, located in genes previously associated with IRD (as of April 2025), based on expert curation and literature support[8]. The mode of inheritance for each gene was annotated based on OMIM[41] to inform zygosity-aware interpretation during screening.

In total, 15,795 variants spanning 251 genes were screened across both panels. As illustrated in **Supplementary Figure 1**, a subset of variants (and genes) was unique to each panel, while others were shared between both. The overlap in variant- and gene-level representation highlights the complementarity of clinically derived and database-curated variant sets. These curated variants were used to screen 504,000 individuals from the EGP cohort. Variants were extracted from whole-genome variant call files using bcftools v1.21. For everyone, carrier status (heterozygotes: individuals with one copy of the alternate allele, homozygotes: individuals with two copies of the alternate allele) was identified and linked to phenotypes through phenoEGP for downstream enrichment, penetrance, and disease burden analyses.

Haplotype Phasing of Biallelic Variant Carriers

To resolve the phase of co-occurring rare variants within *ABCA4*, we performed local genotype phasing using Beagle (version 22Jul22.46e). The analysis focused specifically on individuals heterozygous for both *ABCA4*:c.5882G>A (chr1:94008251:C>T) and *ABCA4*:c.2570T>C

(chr1:94055128:A>G). Phasing was used to determine whether the two *ABCA4* variants occurred in cis (on the same chromosome) or trans (on opposite chromosomes), a distinction critical for evaluating biallelic pathogenicity. The resulting phase information was incorporated into downstream penetrance estimation and familial segregation analyses to refine haplotype interpretation.

Kinship Inference and Family Structure Reconstruction

We first generated a genome-wide dataset in PLINK format simulating Illumina Global Screening Array-24 v3.0 (GSA)[42], which includes 608,896 variants. This array-based dataset was constructed for 504,000 individuals in the EGP and served as input for kinship analyses. To assess levels of autozygosity and consanguinity, we computed individual-level inbreeding coefficients (F-statistics) using PLINK, version 1.9[43], with the --het flag.

We defined a family as a set of individuals connected through inferred biological relationships based on genome-wide identity-by-descent (IBD) estimates. Kinship inference was performed using KING, version 2.2.7[44], which calculates pairwise kinship coefficients and classifies relationship types based on autosomal genotype data. Relationship categories included: Parent-offspring (PO): kinship coefficient ≈ 0.25 and IBD2 ≈ 0, Full siblings (FS): kinship coefficient ≈ 0.25 and IBD2 > 0, Second-degree relatives (e.g., grandparent–grandchild, avuncular, half-siblings): kinship coefficient ≈ 0.125, Third-degree relatives: kinship coefficient ≈ 0.0625.

To identify multigenerational family structures, we focused exclusively on PO relationships inferred using KING. Pairwise PO links were converted into a graph using the igraph package in R[45], where nodes represent individuals and edges represent directed PO relationships. Families were defined as connected components within this graph, i.e., groups of individuals connected through one or more PO links. This approach captures both nuclear and extended family structures, and in large families, may include multiple interconnected subunits. In cases where components were especially large, subfamilies were extracted for targeted analysis while retaining the overall pedigree context.

PO pairs were used to assign pedigree directionality using available age and sex information, assuming the older individual to be the parent. Pedigrees were constructed using the kinship2 package in R[46], supporting visualization and inheritance-based filtering. These reconstructed families formed the basis for mapping IRD diagnoses, assigning carrier status, and stratifying families by clinical relevance.

Family and Individual Classification Based on Genomic and Clinical Findings

Following family structure reconstruction, we performed clinical classification at both the individual and family levels, integrating genomic carrier status, phenoEGP-derived phenotypes, and pedigree structure to assess clinical actionability and prioritize follow-up strategies.

At the individual level, each participant was evaluated based on:

- Carrier status for curated pathogenic or likely pathogenic IRD variants, stratified by zygosity (heterozygous/homozygous) and variant origin (Panel I, Panel II, or both),

- Clinical status, based on diagnoses in phenoEGP, specifically the Top5 phenotypes (top five IRD-related ICD-10 subchapters: H35, H50, H53, H54, H55) previously identified as most enriched ICD-10 subchapters among IRD variant carriers.

Individuals were then assigned to one of three clinical actionability categories:

1. Prevention – heterozygotes or homozygotes with no documented diagnosis in the Top5 phenotypes. Further stratified by variant penetrance tier.
2. Explained by Known Variants – heterozygotes or homozygotes with a diagnosis in the Top5 phenotypes, supporting known variant–disease associations.
3. Novel Variant Candidates – individuals diagnosed with a Top5 phenotype but without any known pathogenic variants from Panel I or II.

At the family level, individuals were mapped into pedigrees based on inferred PO relationships, and families were similarly classified:

1. Prevention – families with variant carriers (heterozygotes or homozygotes) but no members diagnosed with a Top5 phenotype, further stratified by variant penetrance tier.
2. Explained by Known Variants – families with at least one carrier (heterozygotes or homozygotes) diagnosed with a Top5 phenotype.
3. Discovery – families with no known pathogenic variants from Panel I or II, but evidence of heritable disease, defined as ≥3 genotyped members, ≥2 Top5 diagnoses, and ≥10% family-level IRD burden.

This dual-level framework enabled downstream penetrance estimation, identification of clinically actionable cases, and prioritization of high-risk families for research and care

Variant Effect Prediction and Prioritization

To identify potentially novel IRD-associated variants in unsolved high-disease burden families, we extracted all polymorphic sites located within RetNet genes across individuals in families with elevated IRD burden and no contribution from known pathogenic variants. Variant annotation and consequence prediction were performed using the Ensembl Variant Effect Predictor (VEP)[25], version 112.0, which includes standardized functional annotations and impact classifications. Variants were prioritized based on molecular consequence and predicted deleteriousness: (1) Missense variants were retained if they occurred in protein-coding transcripts, had a CADD score >20, and were predicted to be damaging by either SIFT (≤0.05) or PolyPhen-2 (≥0.85); (2) Loss-of-function (LoF) variants, including stop gained/lost, splice acceptor/donor, frameshift, and start lost variants, were prioritized if annotated as high-impact by VEP and had CADD scores >20. Variants with minor allele frequency (MAF) >1% in the EGP cohort were excluded. ClinVar annotations were used to flag variants previously classified as benign, or of uncertain significance; variants absent from ClinVar and gnomAD (version 4.1) were considered potentially novel. To evaluate genotype–phenotype concordance, we applied a segregation filter, excluding variants observed in families where IRD diagnoses were present in both carriers and non-carriers.

Statistical Analyses

All statistical analyses were performed using R (version 2024.04.0+735) and Python (version 3.12.9) unless otherwise specified. Detailed descriptions of variant filtering and annotation are provided in the corresponding Methods sections.

(I) *ICD-10-wide association analyses*

To identify ICD-10 subchapters most relevant to IRD phenotypes, we conducted an association analysis across the 2,323 retained ICD-10 codes (filtered to exclude those with ≥1% population prevalence). For each code, we tested for enrichment among variant carriers (heterozygous and homozygous) versus non-carriers using Fisher's exact test, separately for each zygosity group. Association p-values were $-\log_{10}$ transformed and visualized by ICD-10 subchapter to prioritize high-confidence vision-related phenotype groupings. To quantitatively prioritize subchapters for downstream analyses, we calculated an enrichment score for each subchapter within the H00–H59 block (Diseases of the Eye and Adnexa), defined as:

$$\{Enrichment\ Score\} = \sum\left[\ log(\{OR\}) \times -log_{\{10\}(p)}\right]$$

The five subchapters with the strongest associations: H35, H54, H55, H50, and H53, were selected for downstream analyses. For clarity, these are collectively referred to as the "Top5 phenotypes" throughout the manuscript.

(II) *Penetrance estimation and tier classification*

Variant-level clinical penetrance was calculated as the proportion of carriers, separately for heterozygous and homozygous individuals, with at least one diagnosis among the Top5 phenotypes (H35, H54, H55, H50, H53). All variants, including singletons, were assigned penetrance values and included in descriptive analyses. However, for the purpose of interpretation, significance was defined based on carrier count: variants observed in two or more individuals with the same zygosity were considered significant, while singletons were treated as non-significant. This significance threshold was chosen to reflect both the limited statistical power available for rare variant analysis, particularly among homozygous carriers, and the need for clinically meaningful thresholds in the context of population screening. Observing consistent phenotype expression across multiple carriers provides stronger evidence of clinical relevance than p-values alone, which may be unstable at low counts. To guide variant interpretation and prioritization, we implemented a 7-tier penetrance schema (see **Supplementary Table 6**), based on penetrance magnitude and significance. To assess age-dependent expressivity, penetrance estimates were stratified by age group (0–18 years, 18–35 years, and 35+ years) based on each individual's age at analysis.. The proportion of high-penetrance variants (Tiers 1–2) was compared across age groups to evaluate whether additional variants reach full phenotypic expression in older individuals, consistent with the

progressive nature of IRDs. In parallel, we performed direct statistical comparisons of age between carriers with Top5 phenotypes and carriers without Top5 phenotypes, as well as between Tier-specific subgroups (e.g., Tier 1–2 with Top5 phenotypes vs. Tier 7 without phenotypes), to further evaluate the contribution of age to penetrance differences. In a separate sensitivity analysis, we reassessed tier classification after restricting to individuals carrying only a single IRD-associated variant, excluding potential compound heterozygotes or multi-variant genotypes. This allowed us to assess whether penetrance signals were confounded by additional variants acting independently or additively.

(III) *Risk difference and odds ratio calculation*

To further assess the association between individual variants and diagnoses within the Top5 phenotypes, we computed risk difference (RD) and odds ratio (OR) using Fisher's exact test. For each variant, separate 2×2 contingency tables were constructed for heterozygous and homozygous genotypes, comparing the proportion of affected individuals among carriers versus non-carriers. These association metrics complemented penetrance estimates by capturing population-level risk enrichment.

(IV) *Family-Level Estimation of IRD Burden and Variant Explainability*

To quantify the contribution of known pathogenic variants to IRD burden, we identified families with evidence of heritable disease, defined as those meeting all the following criteria: ≥3 genotyped members, ≥2 individuals with IRD-related ICD-10 code from the Top5 phenotypes, and ≥10% family-level IRD burden, calculated as:

$$Burden_{ICD} = \frac{Total\ number\ of\ individuals\ with\ ICD\ diagnosis}{Total\ number\ of\ individuals\ with\ ICD}$$

Among these families, we calculated the proportion of IRD burden that could be attributed to known IRD variants from Panel I or II, defined as the number of affected individuals who also carried a curated pathogenic or likely pathogenic variant:

$$Explained\ Burden_{ICD} = \frac{\{Het_{ICD} + Hom_{ICD}\}}{Total\ number\ of\ individuals\ with\ ICD}$$

These metrics were calculated separately for each of the Top5 phenotypes, where "ICD" refers to one of the five most enriched IRD-related ICD-10 subchapters. The resulting stratification was used to classify families for downstream discovery or clinical prioritization.