



OPEN

## Genome wide association study and genomic risk prediction of age related macular degeneration in Israel

Michelle Grunin<sup>1,2</sup>, Daria Triffon<sup>1</sup>, Gala Beykin<sup>2</sup>, Elior Rahmani<sup>3</sup>, Regev Schweiger<sup>4,8</sup>, Liran Tiosano<sup>2</sup>, Samer Khateb<sup>2</sup>, Shira Hagbi-Levi<sup>2</sup>, Batya Rinsky<sup>2</sup>, Refael Munitz<sup>2</sup>, Thomas W. Winkler<sup>7</sup>, Iris M. Heid<sup>7</sup>, Eran Halperin<sup>4,5,6</sup>, Shai Carmi<sup>1,9</sup>✉ & Itay Chowers<sup>2,9</sup>✉

The risk of developing age-related macular degeneration (AMD) is influenced by genetic background. In 2016, the International AMD Genomics Consortium (IAMDGC) identified 52 risk variants in 34 loci, and a polygenic risk score (PRS) from these variants was associated with AMD. The Israeli population has a unique genetic composition: Ashkenazi Jewish (AJ), Jewish non-Ashkenazi, and Arab sub-populations. We aimed to perform a genome-wide association study (GWAS) for AMD in Israel, and to evaluate PRSs for AMD. Our discovery set recruited 403 AMD patients and 256 controls at Hadassah Medical Center. We genotyped individuals via custom exome chip. We imputed non-typed variants using cosmopolitan and AJ reference panels. We recruited additional 155 cases and 69 controls for validation. To evaluate predictive power of PRSs for AMD, we used IAMDGC summary-statistics excluding our study and developed PRSs via clumping/thresholding or LDpred2. In our discovery set, 31/34 loci reported by IAMDGC were AMD-associated ( $P < 0.05$ ). Of those, all effects were directionally consistent with IAMDGC and 11 loci had a P-value under Bonferroni-corrected threshold ( $0.05/34 = 0.0015$ ). At a  $5 \times 10^{-5}$  threshold, we discovered four suggestive associations in *FAM189A1*, *IGDCC4*, *C7orf50*, and *CNTNAP4*. Only the *FAM189A1* variant was AMD-associated in the replication cohort after Bonferroni-correction. A prediction model including LDpred2-based PRS + covariates had an AUC of 0.82 (95% CI 0.79–0.85) and performed better than covariates-only model ( $P = 5.1 \times 10^{-9}$ ). Therefore, previously reported AMD-associated loci were nominally associated with AMD in Israel. A PRS developed based on a large international study is predictive in Israeli populations.

Age-related macular degeneration (AMD) is the leading cause of blindness in the elderly population. The risk for developing AMD is strongly associated with the genetic background of the individual<sup>1,2</sup>. In 2005, AMD was the first disease for which genome-wide association studies (GWASs) have identified risk variants<sup>3,4</sup>. Via a seminal paper published in 2016, the International Age-Related Macular Degeneration Genomics Consortium (IAMDGC) has reported the genotyping of more than 30,000 AMD patients and controls of European ancestry and the discovery of 52 risk variants across 34 loci<sup>2</sup>.

Israel is home to a number of populations of distinct genetic ancestry, including Ashkenazi Jews, non-Ashkenazi Jews—predominantly North-African Jews and Middle-Eastern Jews, and Arabs—predominantly Palestinians, Bedouins, and Druze. These populations are genetically diverse, having genetic ancestry related to the Middle East, Africa, and Europe, with variable admixture proportions<sup>5–8</sup>. Some of the populations have experienced recent population-specific genetic drift due to founder events and endogamy<sup>7,9,10</sup>. The unique genetic

<sup>1</sup>Braun School of Public Health and Community Medicine, The Hebrew University of Jerusalem, POB 12271, 9112102 Jerusalem, Israel. <sup>2</sup>Department of Ophthalmology, Hadassah-Hebrew University Medical Center, POB 12000, 91120 Jerusalem, Israel. <sup>3</sup>Department of Computational Medicine, University of California, Los Angeles, Los Angeles, CA, USA. <sup>4</sup>Molecular Microbiology and Biotechnology, Tel Aviv University, Tel Aviv, Israel. <sup>5</sup>Department of Anesthesiology, University of California, Los Angeles, Los Angeles, CA, USA. <sup>6</sup>Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA, USA. <sup>7</sup>Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany. <sup>8</sup>Department of Genetics, University of Cambridge, CB21TN Cambridge, UK. <sup>9</sup>These authors contributed equally: Shai Carmi and Itay Chowers. ✉email: shai.carmi@huji.ac.il; chowers@hadassah.org.il

background of the Israeli populations suggests that the genetic architecture of AMD might be different in these populations compared to Europeans. In addition, the Israeli populations that have experienced strong genetic drift may harbor deleterious risk variants at a considerable frequency. This will increase power for discovering novel risk variants<sup>11</sup> as previously observed for other retinal diseases<sup>12–14</sup>.

Previous studies of the genetic basis of AMD in Israel found that the most prominent risk variants—the genes *CFH*<sup>15</sup> and *HTRA1/ARMS2*<sup>16</sup>—were associated with AMD. However, the *C2* locus, one of the top risk loci worldwide, was not associated with AMD in Israel<sup>17</sup>. The 2016 study of the IAMDGc included an Israeli cohort. However, it was analyzed jointly with the other studies, which was uninformative about Israeli-specific genetic architecture and risk variants. Searching for population-specific risk variants is important even beyond the population under study, as any discovered variants and biological pathways may provide insight into the pathogenesis of the disease.

Polygenic risk scores (PRSs) were recently developed for numerous diseases based on the results of large-scale GWASs<sup>18</sup>. Conceptually, a PRS is the count of risk alleles carried by an individual, where each allele is weighted by its effect size (usually the log odds-ratio), as estimated by GWAS. In practice, the list of single-nucleotide polymorphisms (SNPs) that are included in the score is optimized based on their strength of association (e.g., comparing multiple P-value cutoffs) and their correlation with other SNPs, in some methods accounting for prior assumptions on the genetic architecture of the disease<sup>19</sup>. While PRSs cannot unambiguously distinguish healthy and affected individuals (due to the small proportion of variance in disease liability they explain), individuals at the top PRS quantiles are at a particularly high risk<sup>20,21</sup>. These individuals can then be subjected to personalized screening or prevention.

A number of recent papers have developed or examined PRSs for AMD, showing that the PRS has considerable power to predict disease status and disease progression<sup>2,22–27</sup>. However, it is known that PRS accuracy can substantially decrease when evaluated in populations or ancestries other than the ones used for the original GWAS (usually European populations and ancestries)<sup>28,29</sup>. So far, no study has examined the accuracy of an AMD PRS in any of the Israeli sub-populations, which forms a barrier to the implementation of DNA-based risk stratification.

In this paper, we used data on 558 AMD cases and 325 controls to investigate the genetic basis of AMD in the Israeli populations. Our study had three main goals. (1) To determine whether previously identified risk variants (from the IAMDGc 2016 GWAS) are associated with AMD in Israel, either across all Israeli sub-populations or in a population-specific manner. (2) To discover putative new AMD risk variants by running a GWAS in the Israeli study, anticipating that despite the small sample size, we may be able to identify risk variants that have drifted to high frequencies in the Israeli founder populations. (3) To evaluate the accuracy in the Israeli population of a PRS generated based on the IAMDGc GWAS. We show that the vast majority of previously discovered risk variants are also associated with AMD in Israel. Accordingly, a PRS based on previously discovered variants has high predictive power. While our study was too small for discovering new risk variants at a genome-wide significance level, our study suggested a number of putative associations at an attenuated significance threshold.

## Results

### Replication of known AMD loci

A previous large-scale AMD GWAS by the IAMDGc ( $n = 33,976^2$ ) has discovered 34 associated loci. We examined the association of these loci with AMD status in our Israeli discovery set (403 AMD cases and 256 controls). Using the SNP with the lowest P-value in each locus, we found that most loci (31/34) were associated with AMD at a nominal significance level of  $P < 0.05$  with a direction of effect consistent with that of the IAMDGc (Supplementary Tables 2 and 3). The number of loci associated at the Bonferroni correction threshold ( $0.05/34 = 0.0015$ ) was 11/34 (Supplementary Table 2). The top ranked loci were *CFH* ( $P = 1.6 \times 10^{-9}$ ) and nearby loci on chr1, and *ARMS2/HTRA1* ( $P = 3.4 \times 10^{-9}$ ,  $5.1 \times 10^{-9}$ , respectively). The next significant locus was near *SYN3* ( $P = 5.7 \times 10^{-5}$ ). Association statistics for the known AMD risk loci for Ashkenazi Jews (AJ) (242 cases and 136 controls) and Arabs (36 cases and 30 controls) are reported in Supplementary Tables 4 and 5. We note that replication was to some extent expected, given that the majority of the Israeli cohort was included in the IAMDGc. However, the same 34 loci were associated with AMD at genome-wide significance even when all Israeli samples were excluded from the IAMDGc GWAS.

### Discovery GWAS

We next ran a GWAS in our discovery set (AMD cases:  $n = 403$ , controls:  $n = 256$ ). No novel variant was associated at the genome-wide significance threshold of  $5 \times 10^{-8}$ . Setting a more liberal threshold of  $5 \times 10^{-5}$ , and excluding variants in known risk loci, we identified four suggestive associations in the genes *C7orf50*, *IGDCC4*, *FAM189A1*, and *CNTNAP4* (Table 1; Fig. S2). None of these SNPs were associated with AMD in the IAMDGc data ( $P \geq 0.04$ , Table 1). The variant rs116928937 in *IGDCC4* is exonic. Its allele frequency in European Americans was 1.23% (in the Exome Variant Server (NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: <http://evs.gs.washington.edu/EVS/>))), compared to 2.66% here. It is a missense variant (c.3188G>T), and according to Polyphen<sup>30</sup> it is "probably-damaging".

We attempted to replicate the association of these four loci in a replication set of  $n = 155$  AMD cases and  $n = 69$  controls (Supplementary Table 1). We applied a Bonferroni corrected threshold of  $0.05/4 = 0.0125$ . Only the SNP in *FAM189A1* (rs1195500, chr15:29687047) replicated in the Israeli population ( $P < 0.0001$  in Fisher's exact test in both genotype and allele testing). The SNP rs12701455 (*C7orf50*) attained a P-value of 0.029 in the genotype-based test (Supplementary Table 1).

SNP ID	Genomic position (hg19)	Gene	P-value (entire cohort)	Odds ratio (95% confidence interval) Entire cohort, AJ subset	Minor allele frequency (cases/controls)	Minor allele frequency (gnomAD NFE/AJ)	P-value in IAMDGC—leave one out	Replication P-value Genotype (Fisher's exact)
rs12701455	chr7:1055409	C7orf50	$5.05 \times 10^{-6}$	0.5 (0.26–0.94)	0.147/0.251	0.309/0.212	0.78	0.028
rs116928937	chr15:65677446	IGDCC4	$3.73 \times 10^{-6}$	0.13 (0.02–0.83), 0.07	0.008/0.055	0.0185/0.0064	0.89	1
rs1506825	chr16:76483019	CNTNAP4	$6.91 \times 10^{-6}$	0.57 (0.34–0.96)	0.4373/0.5703	0.455/0.469	0.04	0.85
rs1195500	chr15:29,687,047	FAM189A1	$2.265 \times 10^{-5}$	0.51 (0.27–0.96), 0.37	0.129/0.227	0.26/0.15	0.27	0.00001
rs11689931 (AJ specific)	chr2:206440979	PARD3B	$3.19 \times 10^{-6}$ for AJ	0.07	0.20/0.37 in AJ	0.301/0.306		Was not tested

**Table 1.** SNP association with AMD in Israel Discovery Cohort. Statistics for the association of SNPs with AMD status in our Israeli discovery cohort. Genomic coordinates are in hg19. The gene is the nearest to the SNP. Allele frequencies were computed in gnomAD, for either non-Finnish Europeans (NFE) or Ashkenazi Jews (AJ). The first four rows provide details on four SNPs associated with AMD in our discovery cohort with  $P$ -value  $< 5 \times 10^{-5}$ . The last row presents details on the top-associated SNP in the AJ subset of our cohort.

### Evaluating a polygenic risk score for AMD

We developed polygenic risk scores (PRSs) for AMD in Israel based on the results of the IAMDGC GWAS with the Israeli samples excluded (“leave-one-out”) and using two methods. The first method is clumping and thresholding (C + T), in which the most strongly associated SNP is retained from each LD block, as long as its  $P$ -value is under a threshold. The second method is LDpred2, which accounts for the influence of LD on effect sizes and incorporates a non-zero prior probability for having null effects. We generated nine C + T PRSs, corresponding to different  $P$ -value thresholds (exponentially increasing between  $5 \times 10^{-8}$  and 1), and four LDpred2 PRSs, corresponding to different values of the proportion of SNPs with non-zero effects and a sparsity parameter. For each PRS, we used logistic regression to predict AMD status based on age, sex, the first two principal components (a proxy of ancestry), and the PRS. We also fit a logistic regression model with covariates only. We used fivefold cross-validation to evaluate the accuracy of the various models, which we quantified using AUC (the area under the receiver operator characteristic curve (ROC curve)).

We compare the ROC curves of the most accurate (highest AUC) C + T model, the most accurate LDpred2 model, and the covariates-only model in Fig. 1. For C + T, the AUC was highest (0.79; 95% confidence interval (CI) 0.75–0.82) for the most stringent  $P$ -value threshold ( $5 \times 10^{-8}$ ), for a PRS that included 360 variants. Interestingly, the AUC showed a trend (not statistically significant) of decreasing with increasing  $P$ -value thresholds (Fig. S3). The best LDpred2 model (parameters  $P = 0.056$  and sparsity on) had a slightly higher AUC (0.82; 95% CI 0.79–0.85) than the best C + T model. The covariates-only model had a significantly lower AUC (0.72; 95% CI 0.69–0.76). This was also confirmed by DeLong’s test for two correlated ROC curves ( $P = 5.1 \times 10^{-9}$ ). Overall, our results suggest that including the PRS in the prediction model improves accuracy.

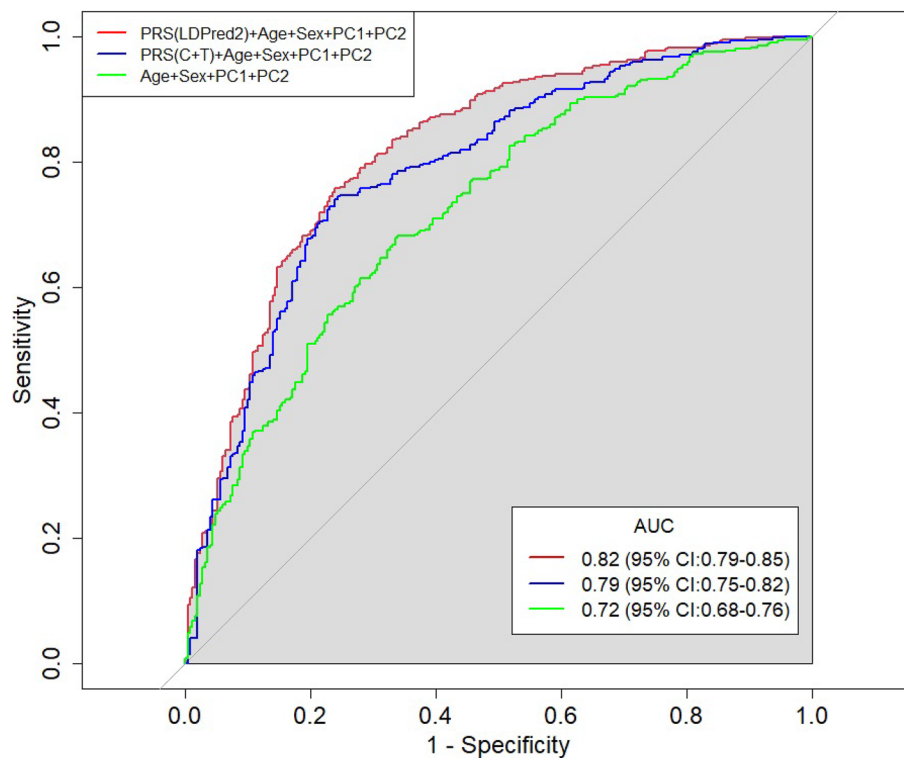
We show the distribution of the best LDpred2 PRS in cases and controls in Fig. 2A. The PRS distribution is different between cases and controls; however, considerable overlap exists. In Fig. 2B, we plot the proportion of cases in each quintile of the LDpred2 PRS, demonstrating that the proportion of cases steadily increases with increasing quintiles. Finally, we used Spearman’s correlation test to assess the correlation between age at diagnosis (measured here as age at blood draw) and the PRS among AMD cases. In Fig. S4, we show a modest, yet significant, negative correlation between the variables ( $\rho = -0.18$ ,  $P$ -value = 0.0003, using the best LDpred2 PRS), suggesting that the PRS may be associated not only with disease status but also with age of onset.

### Discussion

In this work, we studied the genetic basis of AMD in the Israeli populations. We confirmed that most of the known risk loci for AMD, as previously identified in a large international study, are also associated with AMD in Israel. This suggests that the genetic architecture of AMD is similar between the Israeli and other populations. We then performed a genome-wide association study in our cohort in an attempt to identify novel risk variants. As expected due to the small size of our cohort, no novel variants were detected at the genome-wide significance threshold. Setting a more relaxed threshold of  $5 \times 10^{-5}$ , we identified four suggestive variants.

One of the suggestive variants (rs1195500, in *FAM189A1*) replicated, after Bonferroni correction, in a small second set of cases and controls. *FAM189A1* is expressed in the brain (the Human Protein Atlas<sup>31</sup> (<https://www.proteinatlas.org/ENSG00000104059-FAM189A1/tissue>) and GTEx<sup>32</sup>) and in normal human retina (the Retinal Transcriptome data<sup>33</sup>), but its function is yet unknown. The *FAM189A1* gene is uncharacterized, and as such the function of its proteins is unknown. *FAM189A1* was recently discovered to have rare pathogenic variants in neurodevelopmental disorders, and is present in neuronal cells of the brain<sup>34</sup>. Additional studies will be required to replicate the association of this variant as well as validate the involvement of *FAM189A1* in AMD.

*C7orf50* encodes for a newly discovered hormone, cholestin, that is secreted from the intestine in response to cholesterol absorption. This very recent discovery indicated that this gene, *C7orf50* is in charge of cholesterol synthesis in the liver, and is responsible for a reduction in circulating cholesterol levels<sup>35</sup>. This may be related to other cholesterol genes found in AMD<sup>2</sup>, as well as AMD’s connection to dietary cholesterol.



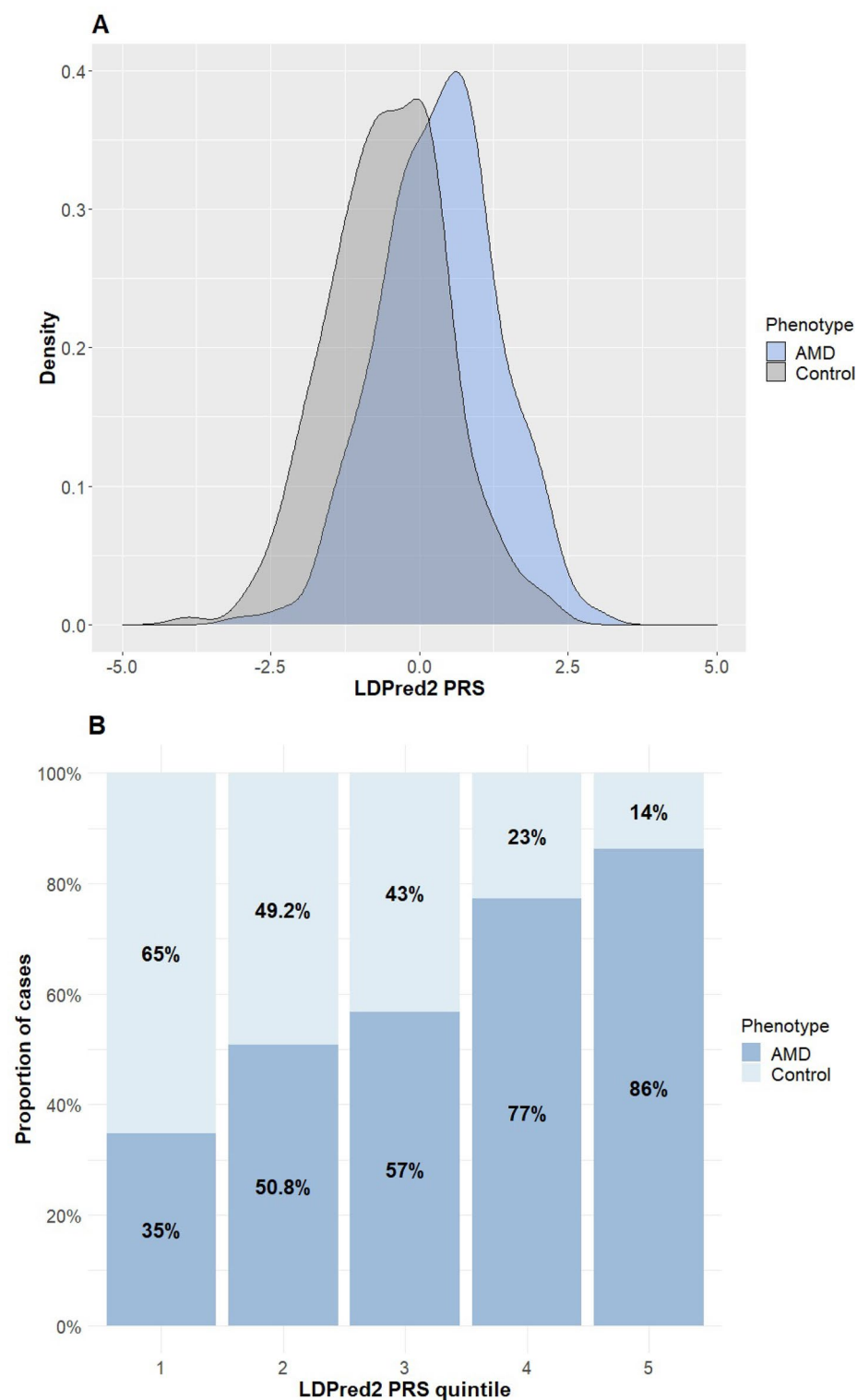
**Figure 1.** Prediction accuracy of selected PRS models. We show the ROC curve for the following AMD prediction models: the top C + T PRS (Fig. S3) + covariates (age, sex, PC1, and PC2); the top LDpred2 PRS + covariates; and a covariates-only model. The C + T PRS parameters were  $r^2 > 0.5$  and  $P < 5 \times 10^{-8}$ , and the LDpred2 PRS parameters were  $P = 0.056$  and sparse = TRUE. The AUC estimates (after cross-validation) are indicated on top of the plot.

The two other suggestive genes, *IDGCC4* and *CNTNAP4*, have been characterized. *IDGCC4* belongs to the immunoglobulin superfamily. This protein might have a role in regulating the immune system against "para-inflammatory" diseases like AMD<sup>36</sup>, but no studies clearly showed its immune system role. It was present in a study discussing methylation patterns in lung adenocarcinoma in smokers, and was methylated in advanced tumors (stage II–IV) as compared to early, stage I tumors<sup>37</sup>. Methylation patterns in neonates in this gene were associated with higher birth weight, as well as with higher maternal BMI and glucose levels, indicating that it also plays a role in metabolism, both at birth and at later life<sup>38</sup>. *CNTNAP4* is part of the contactin associated protein family (otherwise known as *Caspr4*), and is involved in both the development of myelinated axons and in stability of neuronal connections<sup>39,40</sup>. There is a known relationship between *CNTNAP4* and neurodegenerative diseases, and a CNV variant in this gene was found to be inversely associated with healthy aging. The CNV variant was found to be protective against cognitive impairment, Alzheimer's and Parkinson's, giving it a connection with normal neurological function<sup>41</sup>. Further investigation would be needed in order to ascertain their significance in the pathogenesis of AMD.

While our sample size is relatively small, it is, with 659 individuals in the base study and another 224 for replication, the largest AMD GWAS in Israel to date. While this was insufficient to discover any new associations at a genome-wide significance, our study was sufficiently powered to replicate most of the previously discovered associated loci. It was also sufficiently powered to detect an association between AMD status and a polygenic risk score, even when adjusting for standard covariates.

The evaluation of the AMD PRS in our Israeli cohort suggested multiple conclusions. First, the LDpred2 PRS had relatively high accuracy (AUC = 0.82), significantly better compared to not including the PRSs (Fig. 1). Second, with the simple clumping and thresholding approach, accuracy increased as more stringent P-value thresholds were used (Fig. S3). This could indicate that AMD is not as polygenic as other diseases. Third, as expected<sup>42,43</sup>, LDpred2 performed better than the C + T approach (Fig. 2A). Finally, high AMD PRS in our study associated not only with disease risk, but also with a lower age of onset (Fig. S4), as seen for other diseases such as colorectal cancer or cardiovascular disease<sup>44,45</sup>. Prospective studies will be required to further validate this finding.

Our results for high predictive power of the AMD PRS are in line with previous studies<sup>2,46</sup> in individuals of European ancestry. The transferability of the PRS to the Israeli population is perhaps expected given that the majority of our subjects had Ashkenazi Jewish ancestry, and given that PRSs for other diseases and traits were shown to have high accuracy in Ashkenazi Jews<sup>47–50</sup>. The transferability of PRSs into Ashkenazi Jews may be due to the high percentage of European ancestry in this population<sup>7</sup>. It is also consistent with our replication of the



**Figure 2.** Comparing the PRS between cases and controls. (A) The density of the top LDpred2 PRS (after regressing out the first two principal components) in AMD cases and controls in our study (n = 403 and 259, respectively). (B) The proportion of AMD cases in our study by PRS quintiles. We again used the top LDpred2 PRS.

known risk loci. However, a limitation of our study is that it does not allow a direct head-to-head comparison of the strength of the PRS association with AMD between our Israeli cohort and cohorts of individuals of European ancestry. Our sample size was also too small to evaluate the PRS accuracy in other sub-populations, which could be the goal of future studies. Further improvement of the PRS may be achieved via denser genotyping or larger and more diverse imputation reference panels. Additionally, multiple methods can leverage even small samples from a target non-European population to improve a PRS constructed using large GWASs in Europeans<sup>51,52</sup>. However, such efforts will require additional samples for evaluation of the resulting PRSs.

## Materials and methods

Our discovery set consisted of 403 AMD cases and 256 controls (659 total) recruited at Hadassah Medical Center, as previously reported<sup>53</sup>. Our cases included both atrophic and neovascular (a more advanced) AMD. The subjects' mean age was 75.4 years (SD: 2.76, range: 60–97) and 44.6% were female. The criteria for inclusion of AMD patients were: age > 60, AMD diagnosis according to AREDS (Age-Related Eye Disease Study)<sup>54</sup>, and choroidal neovascularization (CNV) and/or geographic atrophy. Diagnosis was also determined according to fluorescein angiogram and optical coherence tomography. Participants were included in all stages of AMD. We excluded individuals with other retinal diseases and individuals with other potential CNV causes such as myopia, trauma, or uveitis. Controls were over the age of 60 with a normal fundus examination and similar systemic exclusion criteria. The study was approved by the institutional ethics committee of Hadassah-Hebrew University Medical Center. All subjects signed informed consent forms that adhered to the tenets of the declaration of Helsinki.

We genotyped all subjects on the custom chip that was developed for the IAMDC. Genotyping on this chip was performed either via the IAMDC (at the Center for Inherited Disease Research (Johns Hopkins, USA)) or at the genomics core facility of the Technion (Israel), as previously described<sup>2</sup>. The custom chip, which was previously described, contains  $\approx 250,000$  tagging markers for imputation and  $\approx 250,000$  custom markers for AMD<sup>2</sup>. Samples genotyped at either center were merged into a single set and underwent joint imputation, quality control, and further downstream analysis.

We imputed the genomes of our subjects with the following reference panels: the 1000 Genomes Project (1 KG,  $n = 2504$ )<sup>55</sup> and the Ashkenazi Genome Consortium ( $n = 128$ )<sup>6</sup>. This strategy was shown to have the highest accuracy for imputing Ashkenazi genomes<sup>56</sup> and was applied here, given that 60% of our subjects have Ashkenazi ancestry. Unfortunately, a reference panel for non-Ashkenazi Jews or for the non-Jewish populations of Israel does not yet exist, and thus, all samples were imputed with both panels simultaneously. We phased our genomes using SHAPEIT<sup>57,58</sup> and performed imputation using a standard protocol<sup>58,59</sup>. We describe next the post-imputation quality control (QC) pipeline, as we previously developed<sup>53,60</sup>.

The chip was imputed to 37,126,112 variants. We performed QC according to standard protocols to remove low-quality variants and samples<sup>61</sup>. We excluded variants with imputation quality score  $R^2 < 0.6$ , variants with minor allele frequency  $< 0.01$  (removing 403,979 variants), and variants in Hardy–Weinberg disequilibrium (PLINK 1.9<sup>62,63</sup>). The sex of patients was confirmed using the sex-check option in PLINK. We excluded individuals who were related, having PIHAT  $> 0.3$  in PLINK. We performed principal components analysis (PCA) in PLINK and GCTA<sup>64</sup> to account for population stratification; the first two principal components were used as covariates in the association analysis (Fig. S1). The final variant count after filtering was 5,353,842 variants in 403 AMD patients and 256 controls.

We performed the discovery GWAS on case–control status using logistic regression in PLINK. To account for population stratification, we used the first two principal components as covariates. The other covariates were age at blood draw and sex. We generated Manhattan and Q–Q plots with *qqman*<sup>65</sup>. For genome-wide significance, we used a P-value threshold of  $5 \times 10^{-8}$ . To detect suggestive associations, we used a threshold of  $5 \times 10^{-5}$ . We computed the frequency of risk alleles (either in Europeans or in Ashkenazi Jews) using gnomAD (<http://gnomad.broadinstitute.org/>) and, if exonic, also in the Exome Variant server (<http://evs.gs.washington.edu/EVS>). Variants that were outside gene boundaries were reported to nearest gene. Variants contained within a gene were reported with that gene.

To determine whether previously discovered associations replicate in our study, we considered variants within the 34 known loci that were identified in the IAMDC 2016 GWAS<sup>2</sup> (Table 5 in the IAMDC GWAS paper). For each locus (LD block) we retained the variant with the lowest P-value. We considered a nominal significance level of  $P = 0.05$  or a Bonferroni corrected level of  $P = 0.05/34 = 0.0015$ .

To test for population-specific replication, we separately studied Ashkenazi Jews (AJ; 242 cases, 136 controls) and Arabs (36 cases and 30 controls). All subjects self-reported their ancestry, and the self-reported ancestry was validated against the genetic ancestry, as determined by their clustering in PCA (Fig. S1). We identified Arab subjects based on self-report (36 cases and 30 controls). We considered all variants in linkage disequilibrium (LD;  $r^2 > 0.05$  in AJ, using hg19 linkage blocks as per the original Fritsche et al.<sup>2</sup> paper) to belong to the same locus.

We note that 549/649 of our subjects were part of the original IAMDC GWAS<sup>2</sup> (out of a total of 33,976 individuals). Therefore, some degree of replication is expected just by virtue of this sample overlap. However, given that the Israeli samples were less than 2% of the total IAMDC sample, the effect of the overlap is expected to be small. To validate that, we excluded the Israeli samples from the IAMDC dataset and re-ran the GWAS analysis (remaining  $n = 33,515$ ; the “base” study). This created a new, “leave-one-out” GWAS, which did not include any Israeli samples. Indeed, all 34 loci that associated with AMD at genome-wide significance in the original GWAS were also associated with AMD at genome-wide significance in the leave-one-out GWAS.

To replicate putative discoveries from the present study, we recruited additional 155 AMD cases and 69 controls (total 224) according to the same criteria as in the original discovery set. We used this case/control sample to validate the suggestively associated variants from the discovery set. Four variants passed the  $5 \times 10^{-5}$  genome-wide threshold in the discovery set, after excluding variants in known AMD risk loci. We genotyped

these four variants in our entire replication set using the KASP assay (LGC Group, Middlesex, UK) with custom primers. All heterozygotes were confirmed using Sanger sequencing (Macrogen, Seoul, Korea). We tested the association using EPACTS (<https://genome.sph.umich.edu/wiki/EPACTS>) and R using two tests. For each SNP, an allelic test compared the proportion of minor alleles between cases and controls. A genotypic test compared the proportion of homozygotes to the minor allele out of all homozygotes between cases and controls.

To generate a polygenic risk score for AMD, we first performed quality control according to standard protocols. We removed variants with strand-ambiguous variants from the base study's summary statistics. Duplicated variants were removed from both studies, separately. Variants with mismatching alleles were also removed. This has left 4,070,992 overlapping variants between the two studies (directly genotyped or imputed). We then extracted effect sizes from the leave-one-out IAMDGC GWAS, where the Israeli samples were excluded. We used these effect sizes to compute the PRS for individuals in the Israeli study ( $n = 659$ ; the "target" study).

We generated polygenic risk scores using two approaches for variant selection: clumping and thresholding (C + T) and LDpred2. Briefly, in C + T, index variants are sequentially selected based on having the lowest P-value, and nearby variants in LD with the index variants are removed. Index variants with P-value under a threshold are retained<sup>66,67</sup>. We used PLINK to implement clumping and computed LD ( $r^2$ ) using the target study. We set the clumping parameters to  $r^2 > 0.5$  and  $\pm 500$  kb and used nine P-value thresholds:  $5 \times 10^{-8}$ , and  $10^{-7}$ ,  $10^{-6}$ ,  $10^{-5}$ ,  $10^{-4}$ , 0.001, 0.01, 0.1, and 1. The minimum P-value cutoff was set to match the IAMDGC genome-wide significance thresholds.

LDpred2 is a Bayesian method for deriving polygenic scores based on summary statistics while explicitly accounting for LD. Briefly, causal effect sizes are assumed to be a mixture of a normal distribution and a point mass at zero. Posterior mean effects are computed using Gibbs sampling based on the LD matrix and an estimate of the heritability<sup>42,43</sup>. We set the SNP heritability ( $h^2$ ) to 0.47 based on a previous IAMDGC estimate<sup>2</sup> and used five proportions of causal variants ( $p$ ) spaced on a log scale:  $10^{-5}$ ,  $1.8 \times 10^{-4}$ , 0.0032, 0.056, and 1. We also included a third parameter of sparsity (true/false). The analysis was restricted to HapMap3 variants, as recommended by LDpred2 developers. A list of 1,054,330 HapMap3 variants were downloaded from the online repository provided<sup>43</sup> and 715,011 overlapping variants entered the analysis. The LD matrix was computed using the target study. We used the R package bigsnpr to compute the LD matrix and generate the grid of scores<sup>43</sup>. To avoid confounding by ancestry, in both methods we regressed the scores on the first two principal components and used the residuals as the scores in subsequent analyses.

We used PLINK to calculate PRSs for each of the 659 subjects. Overall, we obtained nine C + T PRSs (nine P-value cutoffs) and ten LDpred2 PRSs, of which four were reported as valid by LDpred2 ( $P = 0.056$ , sparse = FALSE;  $P = 0.056$ , sparse = TRUE;  $P = 1$ , sparse = FALSE;  $P = 1$ , sparse = TRUE). To evaluate the accuracy of each score, we used logistic regression of the disease status on the PRS and the following covariates: age, sex, PC1, and PC2. We also included a logistic regression model based on covariates only. We measured the accuracy of each model using the area under the curve (AUC) of the receiver operator characteristic curves (ROC curves), computed using fivefold cross validation. We used the R package *pROC* (<https://cran.r-project.org/web/packages/pROC/pROC.pdf>) to generate and analyze ROC curves, AUCs, and AUC confidence intervals ( $ci.auc()$ ), and the R package *caret* (<https://cran.r-project.org/web/packages/caret/vignettes/caret.html>) for cross-validation. Individuals with missing age data were excluded from the analysis (four cases and five controls).

We visually inspected the discriminatory power of the PRS using plots of the density of the PRS in cases and controls (using kernel density estimation), and the proportion of AMD cases across quintiles (fifths) of the PRS distribution. Both plots were generated with the R package *ggplot2* (<https://cran.r-project.org/web/packages/ggplot2/index.html>). Finally, we computed Spearman's rank correlation coefficient to examine the association between the PRS and age at blood draw (as a proxy of the age at diagnosis) among AMD cases. Plots were generated with R package *ggpubr* (<https://cran.r-project.org/web/packages/ggpubr/index.html>).

### Ethical approval

The study was approved by the institutional ethics committee. All subjects signed informed consent forms that adhered to the tenets of the declaration of Helsinki.

### Data availability

The full IAMDGC dataset values can be accessed at: <http://amdgenetics.org/> including the entire IAMDGC and the Jerusalem dataset specifically. In addition, the GWAS summary statistics and code utilized in this manuscript can be found by contacting the corresponding author via reasonable request. The Related Manuscript Variant supplementary file contains all nomenclature for HGVS for all variants.

Received: 22 December 2023; Accepted: 24 May 2024

Published online: 06 June 2024

### References

- DeAngelis, M. M. *et al.* Genetics of age-related macular degeneration (AMD). *Hum. Mol. Genet.* **26**, R45–R50 (2017).
- Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48**, 134–143 (2016).
- Haines, J. L. *et al.* Complement factor H variant increases the risk of age-related macular degeneration. *Science (80-)*. **308**, 419–421 (2005).
- Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science (80-)*. **308**, 385–389 (2005).
- Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466**, 238–242 (2010).
- Carmi, S. *et al.* Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun.* **5**, 4835 (2014).

7. Waldman, S. *et al.* Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14(th) century. *Cell* **185**, 4703–4716.e16 (2022).
8. Agranat-Tamir, L. *et al.* The genomic history of the Bronze Age Southern Levant. *Cell* **181**, 1146–1157.e11 (2020).
9. Granot-HersHKovitz, E. *et al.* A study of Kibbutzim in Israel reveals risk factors for cardiometabolic traits and subtle population structure. *Eur. J. Hum. Genet.* **26**, 1848–1858 (2018).
10. Zidan, J. *et al.* Genotyping of geographically diverse Druze trios reveals substructure and a recent bottleneck. *Eur. J. Hum. Genet.* **8**, 1093–1099 (2014).
11. Zeggini, E. Using genetically isolated populations to understand the genomic basis of disease. *Genome Med.* **6**, 83 (2014).
12. Zelinger, L. *et al.* A missense mutation in DHDDS, encoding dehydrodolichyl diphosphate synthase, is associated with autosomal-recessive retinitis pigmentosa in ashkenazi jews. *Am. J. Hum. Genet.* **88**, 207–215 (2011).
13. Zlotogora, J. & Chemke, J. Medical genetics in Israel. *Eur. J. Hum. Genet.* **3**, 147–154 (1995).
14. Beryozkin, A. *et al.* Whole exome sequencing reveals mutations in known retinal disease genes in 33 out of 68 Israeli families with inherited retinopathies. *Sci. Rep.* **5**, 131187 (2015).
15. Chowers, I. *et al.* Association of complement factor H Y402H polymorphism with phenotype of neovascular age related macular degeneration in Israel. *Mol. Vis.* **14**, 1829–1834 (2008).
16. Chowers, I. *et al.* Sequence variants in HTRA1 and LOC387715/ARMS2 and phenotype and response to photodynamic therapy in neovascular age-related macular degeneration in populations from Israel. *Mol. Vis.* **14**, 2263–2271 (2008).
17. Asleh, S. A. A. *et al.* Lack of association between the C2 allele of transferrin and age-related macular degeneration in the Israeli population. *Ophthalmic Genet.* **30**, 161–164 (2009).
18. Babb de Villiers, C., Kroese, M. & Moorthe, S. Understanding polygenic models, their development and the potential application of polygenic scores in healthcare. *J. Med. Genet.* **57**, 725–732 (2020).
19. Wang, Y., Tsuo, K., Kanai, M., Neale, B. M. & Martin, A. R. Challenges and opportunities for developing more generalizable polygenic risk scores. *Annu. Rev. Biomed. Data Sci.* **5**, 293–320 (2022).
20. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* <https://doi.org/10.1038/s41576-018-0018-x> (2018).
21. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* <https://doi.org/10.1038/s41588-018-0183-z> (2018).
22. Heesterbeek, T. J. *et al.* Genetic risk score has added value over initial clinical grading stage in predicting disease progression in age-related macular degeneration. *Sci. Rep.* **9**, 6611 (2019).
23. Colijn, J. M. *et al.* Genetic risk, lifestyle, and age-related macular degeneration in Europe: The EYE-RISK Consortium. *Ophthalmology* **128**, 1039–1049 (2021).
24. Yan, Q. *et al.* Genome-wide association studies-based machine learning for prediction of age-related macular degeneration risk. *Transl. Vis. Sci. Technol.* **10**, 29 (2021).
25. de Breuk, A. *et al.* Genetic risk in families with age-related macular degeneration. *Ophthalmol. Sci.* **1**, 100087 (2021).
26. Wąsowska, A. *et al.* Polygenic risk score impact on susceptibility to age-related macular degeneration in Polish patients. *J. Clin. Med.* **12**, 295 (2022).
27. Yu, C. *et al.* Predictive performance of an updated polygenic risk score for age-related macular degeneration. *Ophthalmology*. <https://doi.org/10.1016/j.ophtha.2024.01.033> (2024).
28. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
29. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
30. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* <https://doi.org/10.1002/0471142905.hg0720s76> (2013).
31. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (80-)*. **347**, 1260419–1260419 (2015).
32. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
33. Farkas, M. H. *et al.* Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of novel transcribed sequence via significant alternative splicing and novel genes. *BMC Genom.* **14**, 486 (2013).
34. Safizadeh Shabestari, S. A. *et al.* Overlapping pathogenic de novo CNVs in neurodevelopmental disorders and congenital anomalies impacting constraint genes regulating early development. *Hum. Genet.* **142**, 1201–1213 (2023).
35. Hu, X. *et al.* A gut-derived hormone regulates cholesterol metabolism. *Cell* **187**, 1685–1700.e18 (2024).
36. Parmeggiani, F. *et al.* Mechanism of inflammation in age-related macular degeneration. *Mediat. Inflamm.* **2012**, 546786 (2012).
37. Tessema, M. *et al.* Genome-wide unmasking of epigenetically silenced genes in lung adenocarcinoma from smokers and never smokers. *Carcinogenesis* **35**, 1248–1257 (2014).
38. Lin, X. *et al.* Developmental pathways to adiposity begin before birth and are influenced by genotype, prenatal environment and epigenome. *BMC Med.* **15**, 50 (2017).
39. Zou, Y. *et al.* Structure and function of the contactin-associated protein family in myelinated axons and their relationship with nerve diseases. *Neural Regen. Res.* **12**, 1551–1558 (2017).
40. Spiegel, I., Salomon, D., Erne, B., Schaeren-Wiemers, N. & Peles, E. Caspr3 and caspr4, two novel members of the caspr family are expressed in the nervous system and interact with PDZ domains. *Mol. Cell. Neurosci.* **20**, 283–297 (2002).
41. Iakoubov, L. *et al.* A common copy number variation (CNV) polymorphism in the CNTNAP4 gene: Association with aging in females. *PLoS One* **8**, e79790 (2013).
42. Vilhjálmsón, B. J. *et al.* Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).
43. Privé, F., Arbel, J. & Vilhjálmsón, B. J. LDpred2: Better, faster, stronger. *Bioinformatics* **36**, 5424–5431 (2020).
44. Thomas, M. *et al.* Genome-wide modeling of polygenic risk score in colorectal cancer risk. *Am. J. Hum. Genet.* **107**, 432–444 (2020).
45. Vaura, F. *et al.* Polygenic risk scores predict hypertension onset and cardiovascular risk. *Hypertens. (Dallas, Tex. 1979)*. **77**, 1119–1127 (2021).
46. Qassim, A. *et al.* Risk stratification and clinical utility of polygenic risk scores in ophthalmology. *Transl. Vis. Sci. Technol.* **10**, 14 (2021).
47. Gettler, K. *et al.* Common and rare variant prediction and penetrance of IBD in a large, multi-ethnic, Health System-based Biobank Cohort. *Gastroenterology* **160**, 1546–1557 (2021).
48. Belbin, G. M. *et al.* Toward a fine-scale population health monitoring system. *Cell* **184**, 2068–2083.e11 (2021).
49. Fahed, A. C. *et al.* Transethnic transferability of a genome-wide polygenic score for coronary artery disease. *Circulat. Genom. Precis. Med.* **14**, e003092 (2021).
50. Privé, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* **109**, 12–23 (2022).
51. Cai, M. *et al.* A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *Am. J. Hum. Genet.* **108**, 632–655 (2021).
52. Kachuri, L. *et al.* Principles and methods for transferring polygenic risk scores across global populations. *Nat. Rev. Genet.* <https://doi.org/10.1038/s41576-023-00637-2> (2023).

53. Lorés-Motta, L. *et al.* Association of genetic variants with response to anti-vascular endothelial growth factor therapy in age-related macular degeneration. *JAMA Ophthalmol.* <https://doi.org/10.1001/jamaophthalmol.2018.2019> (2018).
54. Age-Related Eye Disease Study Research, G. The Age-Related Eye Disease Study (AREDS): Design implications. AREDS report no. 1. *Control Clin. Trials.* **20**, 573–600 (1999).
55. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
56. Lencz, T. *et al.* High-depth whole genome sequencing of an Ashkenazi Jewish reference panel: Enhancing sensitivity, accuracy, and imputation. *Hum. Genet.* **137**, 343–355 (2018).
57. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
58. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).
59. van Leeuwen, E. M. *et al.* Population-specific genotype imputations using minimac or IMPUTE2. *Nat. Protoc.* **10**, 1285–1296 (2015).
60. Grunin, M. *et al.* Association of a variant in VWA3A with response to anti-vascular endothelial growth factor treatment in neovascular AMD. *Investig. Ophthalmol. Vis. Sci.* <https://doi.org/10.1167/iovs.61.2.48> (2020).
61. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* **5**, 1564–1573 (2010).
62. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. 1–22 (2014).
63. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
64. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
65. Turner, S. D. qqman: An R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* **3**, 731 (2018).
66. Privé, F., Vilhjálmsson, B. J., Aschard, H. & Blum, M. G. B. Making the most of clumping and thresholding for polygenic scores. *Am. J. Hum. Genet.* **105**, 1213–1221 (2019).
67. Choi, S. W., Mak, T.S.-H. & O'Reilly, P. F. Tutorial: A guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

### Author contributions

MG, DT, SC, IC, ER conceived and designed the work, MG, DT, ER, RS, RM designed and performed the experiments and data analysis, MG, DT, ER, RS, GB, LT, SK, SH-L, BR, RM, IC and SC provided data availability and data extraction and analysis, MG, DT, SC, IC wrote the manuscript, MG, DT, SC, IC, ER, RS, EH provided manuscript feedback, revision, and drafts.

### Funding

This work was supported by a Grants from the Israel Science Foundation: 3485/19 to I.C. and S.C. The contribution of the International AMD Genomics Consortium (IAMGDC) was supported by a Grant from NIH (R01 EY022310) and RES516564 to IMH. Genotyping was supported by a contract (HHSN2682012000081) to the Center for Inherited Disease Research. MG is supported by a grant from the Bright Focus Foundation (M2021006F). We would also like to acknowledge the Israeli Ministry of Science (3-17354 Grant Number).

### Competing interests

S.C. is a paid consultant to MyHeritage. All other authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-63065-0>.

**Correspondence** and requests for materials should be addressed to S.C. or I.C.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024