




ARTICLE

Positive predictive value highlights four novel candidates for actionable genetic screening from analysis of 220,000 clinicogenomic records

Kelly M. Schiabor Barrett¹, Alexandre Bolze¹, Yunyun Ni¹, Simon White¹, Magnus Isaksson¹, Lavania Sharma¹, Elissa Levin¹, William Lee¹, Joseph J. Grzymalski^{2,3}, James T. Lu¹, Nicole L. Washington^{1,4} and Elizabeth T. Cirulli^{1,4} 

PURPOSE: To identify conditions that are candidates for population genetic screening based on population prevalence, penetrance of rare variants, and actionability.

METHODS: We analyzed exome and medical record data from >220,000 participants across two large population health cohorts with different demographics. We performed a gene-based collapsing analysis of rare variants to identify genes significantly associated with disease status.

RESULTS: We identify 74 statistically significant gene–disease associations across 27 genes. Seven of these conditions have a positive predictive value (PPV) of at least 30% in both cohorts. Three are already used in population screening programs (*BRCA1*, *BRCA2*, *LDLR*), and we also identify four new candidates for population screening: *GCK* with diabetes mellitus, *HBB* with β -thalassemia minor and intermedia, *PKD1* with cystic kidney disease, and *MIP* with cataracts. Importantly, the associations are actionable in that early genetic screening of each of these conditions is expected to improve outcomes.

CONCLUSION: We identify seven genetic conditions where rare variation appears appropriate to assess in population screening, four of which are not yet used in screening programs. The addition of *GCK*, *HBB*, *PKD1*, and *MIP* rare variants into genetic screening programs would reach an additional 0.21% of participants with actionable disease risk, depending on the population.

Genetics in Medicine (2021) 23:2300–2308; <https://doi.org/10.1038/s41436-021-01293-9>

INTRODUCTION

Genetic conditions that are appropriate for population screening in US health programs are recommended to meet multiple criteria as proposed in guidelines by the CDC and/or American College of Medical Genetics and Genomics (ACMG) [1, 2]. Broadly, they must be conditions that affect a large number of people, have a genetic component with high penetrance in unselected populations, benefit from identifying at-risk individuals before they have fully developed the condition, have clear actionability for a change in clinical care upon genetic identification, and have the utility of screening confirmed by appropriate health economic analyses. An example of such conditions includes the CDC Tier 1 conditions—*BRCA*-related hereditary breast and ovarian cancer (HBOC), Lynch syndrome (LS), and familial hypercholesterolemia (FH) (Table 1)—which have highly penetrant and actionable genetic associations [1]. In contrast, the ACMG has identified 73 genes recommended for return of results of secondary findings, but most are not currently recommended for population screening because, although they have many of the same properties as CDC Tier 1, they are often too rare to be identified in population studies and have not undergone thorough analyses of their clinical and economic impact [3].

In health systems currently offering population genetic screening based on CDC Tier 1 conditions, roughly 1% of an unselected patient population harbors a pathogenic/likely pathogenic (P/LP) variant, and as many as 80% of these individuals are unaware of their elevated risk status [4, 5]. Leveraging available health-care

data, individuals with P/LP variants as a group display roughly 2–40 times higher risk of developing disease as compared to those without variants, and they also demonstrate penetrance averaging between 20% and 35% for personal history of relevant disease, increasing to 30–65% when family history is also considered, which helps contextualize lifetime risk of disease development (Table 1) [4–7]. This means that there is a high positive predictive value (PPV), generally >30%, when identifying individuals with P/LP variants.

Given the real world prevalence and penetrance seen thus far in genetic screening programs that detect and report P/LP variants, identifying additional common diseases where genetic variants confer a high PPV would expand the benefits of genomic medicine and population screening, as well as improve our understanding of disease biology.

In our opinion, the best candidates to expand genetic screening programs are those rare variants that predispose individuals to common diseases. Compared to common variants, rare variant associations are much more penetrant, resulting in direct and often more severe phenotypic effects that are also often relevant across ethnicities [8]. Significant rare variant associations at the population level not only distinguish differences in relative risk of disease between individuals with rare variants and control groups (often quantified as an odds ratio or OR), but also have high PPV, indicating a high probability for individuals with the variant to develop the disease in question. The high PPVs seen with many associated rare variants are similar to relationships established for

¹Helix, San Mateo, CA, USA. ²Renown Institute for Health Innovation, Reno, NV, USA. ³Desert Research Institute, Reno, NV, USA. ⁴These authors contributed equally Nicole L. Washington, Elizabeth T. Cirulli. ✉email: liz.cirulli@helix.com

Table 1. Positive predictive value (PPV) estimates from population-level genetic screening programs in health systems.

Condition name	Primary diseases	Genes	PPV of heterozygous variants in unselected cohorts: personal history (plus family history)			
			Geisinger MyCode ⁴	UK Biobank ^{5c}	Mount Sinai BioMe ⁶	Healthy Nevada Project ^{7c}
Hereditary breast and ovarian cancer (HBOC)	Breast and ovarian cancers in females	<i>BRCA1, BRCA2^a</i>	15% ^b (48%)	28% (30%)	37% (57%)	32% (NA)
Familial hypercholesterolemia (FH)	Atherosclerotic cardiovascular disease	<i>LDLR, APOB, PCSK9</i>	NA ^d	21% (65%)	NA	18% (NA)
Lynch syndrome	Colorectal and uterine cancers	<i>MLH1, MSH2, MSH6, PMS2, EPCAM</i>	32% (60%)	22% (39%)	NA	29% (NA)

^aAdditional genes associated with hereditary breast cancer not examined in the studies shown here include *ATM, PALB2, CHEK2, TP53, PTEN, and CDH1*.
^bDoes not appear to be limited to females.
^cThe UK Biobank and Healthy Nevada Project cohorts referenced here overlap with the samples used in the present analysis, but the sample sizes are larger now, and the definition of likely disease-causing variants is different.
^dDoes not report an atherosclerotic cardiovascular disease phenotype, but does report 96% PPV for elevated LDL-C.

known P/LP variants. These results thus have both high clinical validity and high clinical utility when used prospectively to modify disease outcomes. When individuals with variants are identified prior to disease onset, proactive actions such as diagnostics, monitoring, and prophylactic risk reducing procedures, often beyond or different from the standard of care, can be employed to prevent or modify the disease for these individuals.

Because of this high PPV, the prospects of larger or additional cohorts for rare variant analyses are very different from potential benefits of larger sample sizes in common variant association analyses. While larger sample sizes in studies of common variants identify signals with smaller and smaller effect sizes, larger sample sizes in studies of rare variants allow for the identification of rare causal variants that can be used to very precisely inform an individual about their risk of disease. Here, we leverage exome and medical data from two large health-care cohorts to identify rare variant—common disease relationships that are statistically significant at the population level, with high PPV ($\geq 30\%$) and actionability relevant to the individuals with the variants, in line with the recommendations for population screening programs.

MATERIALS AND METHODS

Study design

As prior studies have shown, reducing the dimensionality of the genetic inputs can improve the power to detect associations with phenotypes when analyzing rare variants at the population level [9, 10]. Furthermore, differences in billing code (ICD) practices can artificially dampen diagnosis phenotype resolution both within and across cohorts and, like genetic signals from rare variants, they may also benefit from grouping methods [11]. Here, we performed genetic disease association analyses with two large exome-sequenced cohorts, the UK Biobank (UKB, $n = 189,495$) and Healthy Nevada Project (HNP, $n = 28,423$), using both gene and phenotype collapsing techniques.

Populations and genetic data

We utilized the OQFE version of the UKB PLINK-formatted exome files (field 23155) as well as the imputed genotypes from genome-wide association study (GWAS) genotyping (field 22801–22823). The HNP samples were sequenced and analyzed at Helix using the Exome+[®] assay as previously described [9]. The UKB participants range in age from 40 to 69 and are 55% female, while the HNP age range is from 18 to 89+ and is 68% female. The UKB is 83% British European ancestry, with another 10% of other European ancestry and 7% other ancestries, and the HNP is 77% general European ancestry, 14% Hispanic ancestry, and 9% other ancestries.

Phenotypes

HNP phenotypes were processed from Epic/Clarity Electronic Health Records (EHR) data as previously described [9]. UKB data were provided from the UKB resource (<http://www.ukbiobank.ac.uk/>, accessed August 2020). For HNP, International Classification of Diseases, Ninth and Tenth Revision ICD codes (ICD-9 and ICD-10-cm) were collected from available diagnosis tables (from problem lists, medical histories, admissions data, surgical case data, account data, claims, and invoices). For UKB, ICD codes (both ICD-9 and ICD-10) were collected from inpatient data, cancer registry table, and the first occurrences table (resource 593).

To map ICD to phecodes, ICD-9 (Phecode Map 1.2, used for both cohorts), ICD-10 (Phecode Map 1.2b to ICD-10 beta, used for UKB), and ICD-10-CM (Phecode Map 1.2b to ICD-10-CM beta, used for HNP) to phecode maps from the Phewas catalog were used to code individuals as a 1 if they had the phecode recorded at least once in their medical records, and otherwise 0 [12–14]. Analysis phenotypes were restricted to have cases in both cohorts, with at least 30 cases in the HNP data set ($n = 1,044$ phenotypes).

When identifying age at diagnosis, we required at least 5 years of medical history prior to the diagnosis, meaning the first diagnosis of any condition in the record must occur at least 5 years prior to the diagnosis in question, except for when diagnosis occurred in the first five years of life.

Gene-based collapsing

Variant annotation was performed with VEP 99 [15]. Coding regions were defined according to Gencode version GENCODE 33, and the Ensembl canonical transcript was used to determine variant consequence [16, 17]. Variants were restricted to CDS regions. Genotype processing was performed in Hail 0.2.54-8526838bf99f.

For the collapsing analysis, samples were coded as a 1 for each gene if they had a qualifying variant and a 0 otherwise [9]. We defined “qualifying” as coding (stop_lost, missense_variant, start_lost, splice_donor_variant, inframe_deletion, frameshift_variant, splice_acceptor_variant, stop_gained, or inframe_insertion) and not PolyPhen or SIFT benign (PolyPhen benign is <0.15, SIFT benign is >0.05). We also ran a loss-of-function (LoF) model that only included LoF variants (stop_lost, start_lost, splice_donor_variant, frameshift_variant, splice_acceptor_variant, or stop_gained). Variants were only included if their minor allele frequency (MAF) was below 0.1% in all gnomAD populations as well as locally within each population analyzed. Only variants that passed our MAF and predicted function thresholds were included, regardless of known P/LP status.

CNVs calls in HNP data

The Helix Exome+[®] platform includes a copy-number variant (CNV) caller, allowing us to incorporate rare CNVs at exon-level resolution into our gene-based collapsing analysis for the HNP samples [18]. Briefly, CNVs with the PASS QC filter were annotated with overlapping canonical transcripts (CT). For the collapsing analysis, rare CNV events were screened using both exon and event-level frequency information from within the cohort (<0.1% for each), as well as by relevant CNV type—deletions of at least one exon of the gene for LoF model, and deletions or duplications for damaging. Information on how many individuals carried CNVs in each significantly associated gene can be found in Table S1. Including CNVs increased the median frequency of individuals with variants in each gene by ~8%.

Genetic analysis

We used regenie for the genetic analysis [19]. Briefly, this method builds a whole-genome regression model using common variants to account for the effects of relatedness and population stratification, and it accounts for situations where there is an extreme case-control imbalance, which can lead to test statistic inflation with other analysis methods. The covariates we included were age, sex, age*sex, age*age, sex*age*age, and bioinformatics pipeline version as appropriate.

As previously described, a representative set of 184,445 coding and noncoding linkage disequilibrium (LD)-pruned, high-quality common variants were identified for both the creation of principal components and for building the whole-genome regression model [9].

We performed two main analyses: (1) all ancestries together and (2) only European ancestry, with 10 European ancestry-specific principal components included as additional covariates. When collapsing rare (MAF <0.1%) causal variants across a gene and analyzing with a linear mixed model or whole-genome regression, signals tend to be consistent whether restricting to one ancestry or analyzing across all ancestries [9]. This method works in this setting because analyses of collapsed rare variants are less influenced by ethnic background than are analyses of the common variants used in a typical GWAS, in large part because causal variants are being grouped together as opposed to tagging variants.

Meta-analysis was performed using the weighted Z-score *p* value in METAL [20] on the summary stats from each separate analysis. QQ plots showed no test statistic inflation. We required at least one individual to have the variant in both the UKB and the HNP groups, and the meta *p* value to be lower (better) than the *p* values for either individual cohort.

To identify significant associations, we used a conservative Bonferroni correction for multiple tests for all genes that had individuals with qualifying variants ($p < 1 \times 10^{-9}$).

PPV cutoff

To classify gene-disease relationships that would be strong candidates for population screening, we first calculated the PPV (percent of individuals with the variant who develop the condition) of each significant gene-based association by grouping individuals based on age, either all ages (ages 18–89+) or only 60+, to better estimate lifetime risk. Based on the PPV of genetic conditions typically reported in existing genetic screening programs (Table 1), we selected a PPV threshold of ≥ 0.3 to partition our association results. We applied this threshold to both the all ages and

lifetime risk groups, and we included those associations from the 60+ group even if the PPV was lower prior to age 60.

RESULTS

Population-level associations

Our gene-based collapsing analysis of rare variants included 15,857 genes in the coding model, 15,617 of which were also in the LoF model. For the phenotypes, we used phecodes to reduce the phenotype complexity from >20,000 ICD 9 and 10 codes to simply 1,044 medically relevant phenotypes based on available electronic health records (EHR) for both HNP and UKB cohorts. Our meta analysis across both data sets identified 74 statistically significant associations ($p < 1 \times 10^{-9}$) between 27 genes and 49 phecodes (Table 2 and Table S1). While most of the significant associations were obtained with a LoF model, 29 were associations found with coding models, including eight genes for which there was no significant LoF association (the association was only with the coding model).

Importantly, the ethnic makeup of the two cohorts was quite different despite each being predominantly of European ancestry, and our analysis results were similar whether restricting to European ancestry or analyzing across ethnicities (Table S1), consistent with our previous study showing that collapsed rare variant signals tend to be consistent across ancestries [9].

Applying PPV to highlight associations for population genetic screening

We identified seven genes that passed our PPV cutoff of 0.3 (meaning at least 30% of individuals who carried qualifying variants developed the condition). It is important to note that we required the PPV to be above this threshold for *both* cohorts, indicating that the predictive power of the genetic association is applicable across different health systems, population demographics, and countries. Additionally, the ORs for these associations were all >4 in both cohorts, indicating a substantial increase in risk. As expected, some of the statistically significant associations that meet or exceed this threshold cover gene-disease relationships that are already tested in existing population screening programs: *BRCA1* and *BRCA2* with breast cancer (*BRCA1* $p = 8.77 \times 10^{-28}$, OR = 14.2; *BRCA2* $p = 3.96 \times 10^{-45}$, OR = 8.5), and *LDLR* with coronary atherosclerosis ($p = 1.46 \times 10^{-12}$, OR = 17.5). Additionally, we observed several statistically significant associations that have just as strong or stronger PPVs than these conditions, including LoF variants in *HBB* with hemoglobinopathies ($p = 1.91 \times 10^{-29}$, OR = 197.2), LoF variants in *PKD1* and with cystic kidney disease ($p = 4.54 \times 10^{-48}$, OR = 78.5), coding variants in *GCK* with diabetes mellitus ($p = 1.46 \times 10^{-33}$, OR = 11.3), and coding variants in *MIP* with cataracts ($p = 1.56 \times 10^{-10}$, OR = 4.6) (Table 2 and Fig. 1). The remaining significant associations have PPV <0.3 and would have more limited utility if communicated to patients under this paradigm (Table 2 and Table S1).

Importantly, each high-PPV gene-disease association identified here is actionable at some level, further supporting their suitability for inclusion in population screening programs. While some of the conditions have clearly established preventive guidelines based on genetics, all would benefit from earlier diagnosis. Since genetic screening for highly penetrant conditions can lead to a more accurate diagnosis, the resulting medical management guidelines for the patients are likely to be improved. For example, treatment recommendations for maturity onset diabetes of the young (MODY) vary depending on the genetic status of the patient. Individuals who have a *GCK* variant generally do not need treatment and can benefit from a reduced need for surveillance so long as any hyperglycemia remains the mild fasting hyperglycemia typically seen with *GCK*. Clinical actionability, medical management, surveillance methods, and genetics-dependent care

Table 2. Population-level significant rare variant gene–disease ($p < 1 \times 10^{-9}$) associations.

Gene	Model	Phenotype	P value	OR	PPV ≥ 0.3 in both cohorts	
					Age 60+	All ages
<i>HBB</i>	LoF	Other hemoglobinopathies	1.91E-129	197.2	+	+
<i>PKD1</i>	LoF	Cystic kidney disease	4.54E-48	78.5	^b	+
<i>GCK</i>	Coding	Type 2 diabetes	1.46E-33	11.3	+	+
<i>LDLR</i>	LoF	Coronary atherosclerosis ^a	1.46E-12	17.5	+	+
<i>BRCA2</i>	LoF	Malignant neoplasm of female breast ^a	3.96E-45	8.5	+	-
<i>BRCA1</i>	LoF	Malignant neoplasm of female breast ^a	8.77E-28	14.2	+	-
<i>MIP</i>	Coding	Cataract	1.56E-10	4.6	+	-
<i>JAK2</i>	Coding	Myeloproliferative disease ^a	6.41E-62	7.6	-	-
<i>COL4A4</i>	LoF	Hematuria	8.96E-23	4.6	-	-
<i>TTN</i>	LoF	Atrial fibrillation and flutter ^a	1.91E-17	1.8	-	-
<i>MSH6</i>	LoF	Malignant neoplasm of uterus	2.11E-17	19.6	-	-
<i>MYBPC3</i>	LoF	Other hypertrophic cardiomyopathy	5.07E-17	70.2	-	-
<i>IFT140</i>	LoF	Cyst of kidney, acquired	3.81E-16	10.2	-	-
<i>NF1</i>	LoF	Other benign neoplasm of connective and other soft tissue	1.25E-15	14.9	-	-
<i>PKD2</i>	Coding	Cystic kidney disease	1.85E-15	3.9	-	-
<i>TET2</i>	LoF	Neutropenia ^a	2.34E-15	4.8	-	-
<i>VWF</i>	Coding	Von Willebrand disease	2.77E-15	6.7	-	-
<i>SF3B1</i>	Coding	Myeloproliferative disease	4.21E-13	13	-	-
<i>CDKN2A</i>	Coding	Melanomas of skin	5.57E-13	10.2	-	-
<i>TSHR</i>	Coding	Hypothyroidism not otherwise specified	1.52E-12	1.9	-	-
<i>PALB2</i>	LoF	Malignant neoplasm of female breast	8.22E-12	5.0	-	-
<i>ASXL1</i>	LoF	Myeloproliferative disease	9.75E-12	13.0	-	-
<i>PROC</i>	Coding	Phlebitis and thrombophlebitis ^a	1.35E-11	4.9	-	-
<i>ATM</i>	LoF	Malignant neoplasm of female breast	2.49E-11	4.9	-	-
<i>SLC22A12</i>	Coding	Gout	4.86E-11	0.1	-	-
<i>MLH1</i>	LoF	Colon cancer	1.54E-10	240.7	-	-
<i>SLC4A1</i>	Coding	Other hereditary hemolytic anemias	1.99E-10	19.8	-	-

LoF loss of function, OR odds ratio, PPV positive predictive value.

^aA significant association was also found with another phenotype with a PPV that was higher than that for the main phenotype, but it was not a clinical endpoint of main interest (for example, acquired absence of breast for *BRCA1/2*, or hypercholesterolemia for *LDLR*). For full details, see Table S1. ^bThe PPV was >0.3 at age 60+ in UKB, but all 6 HNP *PKD1* LoF heterozygotes with cystic kidney disease were aged <60 .

pathways are summarized for these associations in Table 3 and discussed further below.

Overall, we find seven associations with high PPV, four of which would be novel for population screening and warrant examination in additional cohorts to quantify suitability of screening in more genetically diverse populations, how well population screening can catch the conditions early and change disease course, and the resulting economic impact.

DISCUSSION

Genetic screening programs that prospectively identify individuals who are likely to develop conditions that are treatable or preventable through medical interventions, especially when detected before disease onset or early in the disease course, could make substantial improvements to individual and public health. Rare variants that can be identified as causing common diseases in population-level analyses are the natural candidates for population screening programs due to their relatively high

penetrance and prevalence. Here, we find that when conditions identified from gene-based collapsing analyses of rare variants consistently have a penetrance of at least 30% (PPV ≥ 0.3), they have properties that make them excellent candidates for population screening programs (Table 3). Our analysis identified seven such conditions. Four of these—coding variants in *GCK* with diabetes mellitus, LoF variants in *HBB* with hemoglobinopathies, LoF variants in *PKD1* with cystic kidney disease, and coding variants in *MIP* with cataracts—are novel conditions for population screening. It is notable that these four associations have a PPV as high or higher than the other three associations we identified, which are already used in population screening programs: LoF variants in *BRCA1* and *BRCA2* with HBOC and LoF variants in *LDLR* with atherosclerosis. These associations all represent genetically driven subsets of common, complex diseases that are in line with recommended guidelines for population screening and present opportunities for precision medicine at scale (Table 3) [2]. We briefly discuss each association below and the potential benefits

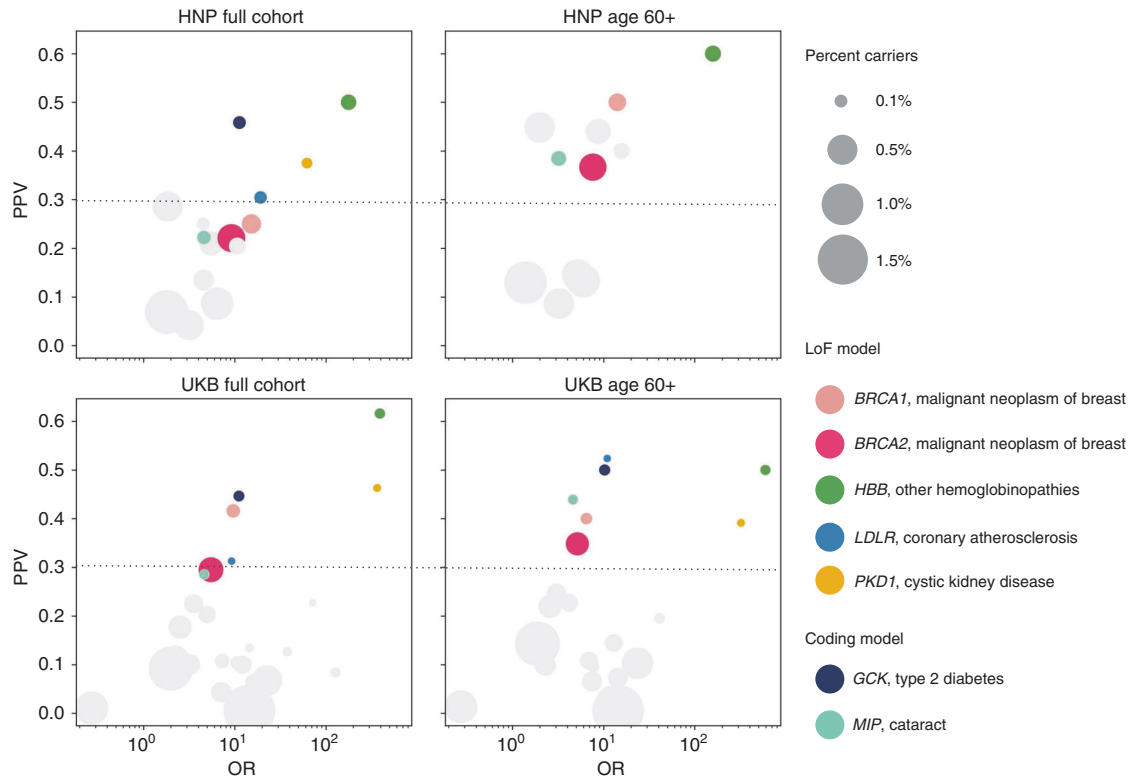


Fig. 1 Positive predictive value (PPV) vs. odds ratio (OR) for statistically significant associations. Shown is the significant association with the best PPV for each gene, as in Tables 2 and 3. The horizontal line indicates our PPV cutoff of 0.3 for high impact genes. The percent of the cohort with variants of interest for each gene is shown by the size of the circle. The seven genes with PPVs ≥ 0.3 in both cohorts are shown in colors as indicated in the legend, and the remaining genes are in gray. Because low sample sizes can produce unreliable results, only data points with at least five cases with variants are shown (this excludes *PKD1*, *GCK*, and *LDLR* in the HNP age 60+ subset). The three gene associations that are above the 0.3 cutoff in HNP age 60+ but had lower PPVs in UKB are detailed in Tables 2 and S1 and include *IFT140* with acquired cyst of kidney, *TSHR* with hypothyroidism, and *ATM* with malignant neoplasm of breast.

of returning rare variant screening results to relevant individuals given current clinical knowledge and practice.

GCK and type 2 diabetes

While often misclassified as type 2 diabetes (T2D), individuals with *GCK* variants typically have mild but stable fasting hyperglycemia and do not develop the microvascular complications typical of T2D [21]. The significant association ($p = 1.46 \times 10^{-33}$) and high PPV (0.5) we observe between *GCK* rare coding variants and T2D corroborates the misclassification of these cases seen in other studies, including ours [22]. Returning *GCK* results to relevant heterozygotes is actionable as it can help their health-care provider tailor the care they receive and set realistic goals for their glucose levels, which are unlikely to fall into the normal range regardless of lifestyle changes. With building evidence for no effect of oral or insulin treatment on glucose levels in *GCK* heterozygotes with mild hyperglycemia, identifying and terminating pharmaceutical treatments in these patients could lead to substantial lifestyle improvements and cost savings [23].

While *GCK* heterozygotes generally do not have problematic clinical outcomes for T2D, they are known to be at increased risk for developing gestational diabetes and are advised to be closely monitored during pregnancy [24]. Our analysis also identified a significant association between rare coding variants in *GCK* and gestational diabetes (Table S1), but the PPV did not pass our 0.3 cutoff (0.17 in HNP and 0.09 in UKB) because our main analysis for this trait included all females and was not restricted to pregnant females. However, when we limit our association analysis to include only females with pregnancy phenotypes in their medical records, we see the PPV for gestational diabetes rise to 1.0 for HNP

and 0.75 for UKB (respectively, 0 of 2,363 and 2 of 10,555 pregnant females without gestational diabetes were heterozygous for qualifying *GCK* variants), suggesting this may indeed be a genetic condition worthy of pre-pregnancy population screening. In particular, identifying whether the fetus has inherited a *GCK* variant from either the mother or father can be important for tailoring care during pregnancy: in a pregnancy where the fetus has a *GCK* variant, hyperglycemia in the mother should usually not be treated as it can lead to dangerously low birthweight, while treatment with insulin is more likely to be indicated if the fetus did not inherit the *GCK* variant [24].

PKD1 and chronic kidney disease

Autosomal dominant polycystic kidney disease (ADPKD, caused by variants in *PKD1* and *PKD2*) is the most common inherited kidney disorder, is the fourth leading cause of chronic kidney disease, and is often not diagnosed until later stages of the disease [25]. While there is currently no cure for ADPKD, early detection of ADPKD can provide the opportunity to treat comorbidities such as early onset hypertension, cardiovascular complications, and kidney disease progression can potentially be slowed with pharmaceutical intervention [26]. Genetic screening programs that include *PKD1* could help detect cases earlier and prioritize these patients for total kidney volume (TKV) measurements in addition to the more typical estimated glomerular filtration rate (eGFR) surveillance for better monitoring of disease progression.

In addition to the association seen with *PKD1*, we also saw a significant association between the related gene *PKD2* and cystic kidney disease (CKD) (Table 2). This coding model association had a lower PPV (OR = 12.5; PPV = 0.03), compared to that of the *PKD1*

Table 3. Summary of PPV and clinical actionability for genes with significant associations and PPV ≥ 0.3 in our study.

Condition	Hereditary breast and ovarian cancer ^a	Type 2 diabetes ^a	Gestational diabetes	β -thalassaemia minor and intermedia ^b	Familial hypercholesterolemia ^a	Cataracts	Chronic kidney disease
Associated complications	Breast, ovarian, prostate, and pancreatic cancers and melanoma	Microvascular disorders	Macrosomia	Anemia, osteoporosis, iron overload	Atherosclerotic cardiovascular disease	Blurry vision up to blindness	Early onset hypertension, cyst infections
Nongenetic surveillance available to confirm or monitor disease?	Mammograms, MRIs, transvaginal ultrasound, regular CA-125 surveillance, PSA surveillance	Glucose and HbA1C surveillance	OGTT	CBC with smear, hemoglobin analysis	Lipid blood tests	Slit lamp eye exam	TKV, serum creatinine, eGFR
Treatment options (to mitigate, prevent, or reverse)	Risk-reducing surgeries	Diet, exercise, metformin, other oral hypoglycemic agents, insulin	Diet change and intense monitoring; insulin needed if diet and exercise do not help	Supplementation, blood transfusion, iron chelation	Statins. May also consider ezetimibe, bile acid sequestrants, niacin, PCSK9 inhibitors, LDL apheresis	Cataract surgery	Various drug options to slow progression
Associated gene	<i>BRCA1</i>	<i>BRCA2</i>	<i>GCK</i>	<i>HBB</i>	<i>LDLR</i>	<i>MIP</i>	<i>PKD1</i>
% with variant ^c	0.11%	0.36%	0.06%	0.06% ^c	0.03%	0.06%	0.03%
PPV ^c	0.43	0.35	0.82 ^e	0.59	0.5	0.43	0.44
Diagnose earlier with genetic screening, and cascade test?	Yes	Yes	Yes	Yes ^f	Yes	Yes	Yes
How are genetic cases treated differently?	Earlier monitoring/treatment	Earlier monitoring/treatment	Tailor treatment based on genotype of fetus	Usually do not give iron for anemia; reproductive counseling	Earlier monitoring/treatment	Earlier monitoring/treatment	Screen TKV to track disease progression [26]

CBC complete blood count, eGFR estimated glomerular filtration rate, MRI magnetic resonance image, OGTT oral glucose tolerance test, PPV positive predictive value, TKV total kidney volume.

^aConditions already part of existing population screening programs and part of CDC Tier 1.

^bComplications and treatments differ depending on the exact type of β -thalassaemia, with β -thalassaemia minor having no complications beyond mild anemia.

^cFrequency varies substantially by population.

^dCalculated for the significant phenotype and model with the highest PPV for this gene in this study.

^eWhen controls are restricted to pregnant females.

^fWhile symptomatic thalassaemia is generally diagnosed in childhood, 71% of the heterozygotes in the present study were diagnosed as adults.

LoF model (OR = 292; PPV = 0.44). Further investigation of the data sets revealed that LoF variants in *PKD2* had a PPV of 0.5 in UKB (OR = 490; p value 2.5×10^{-42}) but had not been included in the analysis because there were only 4 individuals with variants in total in HNP (OR ~61; PPV = 0.5). Despite the similar effect sizes between LoF variants in *PKD1* and *PKD2*, LoF variants in *PKD2* occurred in only 0.02% and 0.01% of the UKB and HNP populations, respectively, compared to 0.03% and 0.06% for *PKD1*. With the HNP study continuing to enroll more participants, we will likely see additional individuals with a *PKD2* variant and CKD, which would likely revise this screening recommendation to include both *PKD1* and *PKD2* for CKD.

HBB and hemoglobinopathies

Rare variants in *HBB* cause the recessive hemoglobinopathy β -thalassemia major, which is quite severe and presents early in life [27]. The statistically significant, dominant association between *HBB* rare variants and hemoglobinopathies and the high PPV (0.55, Table 3) found in our cohorts are driven by a mixture of some individuals who may have β -thalassemia intermedia, a less severe form of the disease that is sometimes inherited in a dominant fashion, and many individuals with β -thalassemia minor, who are generally asymptomatic but often have mild anemia [28, 29].

Individuals with β -thalassemia minor are often misdiagnosed as having iron deficiency anemia. In our study, 30% of *HBB* LoF heterozygotes with a thalassemia diagnosis and 16% of heterozygotes without a thalassemia diagnosis had a diagnosis of iron deficiency anemia, driving a statistically significant association with this trait (Table S1; compared to only 6% of those without a *HBB* LoF variant). Furthermore, 12% of *HBB* LoF heterozygotes reported taking iron supplements, compared to 3% of those without *HBB* LoF variants. Medical records indicated hemochromatosis in 1.6% of *HBB* LoF heterozygotes vs. 0.4% of those without *HBB* LoF variants, 2.4% vs. 0.007% had hepatic fibrosis, and 2.2% vs. 0.3% had nonalcoholic cirrhosis, indicating that complications of iron overload can be a concern for *HBB* LoF heterozygotes. Additionally, the bloodwork available for members of these cohorts showed that 100% of the *HBB* LoF heterozygotes, regardless of thalassemia diagnosis status, had red blood cell (RBC) microcytosis (mean corpuscular volume [MCV] $<80 \mu\text{m}$ [3]; compared to 6% of those without LoF variants), indicating that many individuals with β -thalassemia minor may remain undiagnosed in these cohorts. For individuals with β -thalassemia intermedia, common complications include extensive iron overload in many tissues through increased intestinal absorption, as well as marked and progressive osteoporosis [27]. Not only can the diagnosis of thalassemia be directly confirmed via blood tests, but many screenings and treatments also exist to avoid or mitigate the phenotypic complications, including bone density scans, blood tests to assess iron overload, blood transfusions, splenectomy, folic acid supplementation, and iron chelation therapies [30, 31]. Early detection of *HBB* LoF heterozygotes is useful for reproductive planning and for helping physicians tailor treatment when considering the cause of the patient's anemia. In our study, only 29% of cases with *HBB* LoF variants with age of diagnosis available had been diagnosed as children, indicating that genetic screening of adults for this condition may be warranted.

MIP and cataract

While previous studies have implicated *MIP* variants in rare, familial, congenital cataracts, our results provide evidence for a more general role of *MIP* in cataracts [32–34]. The median age of cataract diagnosis in our study of adults was 61. Returning these genetic results at an earlier age provides an opportunity for health-care providers to encourage or even facilitate underutilized cataract screening and promote possible prevention strategies

such as limiting UV exposure. The added risk may encourage yearly eye exams, as well as safe and effective routine surgery, for those at higher than average risk based on their genetics [35]. Cataract screening is typically performed as part of a routine eye exam, but relatively few Americans keep up with this practice. In a survey of the eye care usage trends of nearly 300,000 adults from 1997 to 2005, eye care utilization rates in the 12 months prior to survey for those older than 65, a group who not only receive coverage for an annual eye exam through Medicare but are also the most likely to harbor an eye condition like cataracts, ranged from 50% to 65% [35]. In addition to personal utility, the timely treatment of cataracts can also have societal benefits. Cataract surgery was recently associated with a 61% reduction in car crash frequency in a cohort of nearly 3,000 drivers aged 60 and above who underwent cataract surgery over the course of the study period [36]. On a broader scale, a deeper understanding of this genetic association has the potential to guide the development of pharmaceuticals that may slow or even reverse cataract disease progression [37, 38].

Population-level clinical impact and future directions

When combining together the variant frequencies for all associations above our 0.3 PPV threshold, we find that population screening for these conditions could impact up to 1% of program participants (Table 3). Reassuringly, we identify genes (*BRCA1*, *BRCA2*, and *LDLR*) that are typically included in existing population health programs, which themselves account for more than half of the potential impact (0.47–0.73% of individuals have relevant variants in UKB and HNP, respectively). However, the inclusion of *HBB*, *GCK*, *PDK1*, and *MIP* in the same programs would reach an additional 0.19–0.36% of participants in each population (for UKB and HNP, respectively; this value will also differ by population, especially for *HBB*).

Recent economic evaluations have revealed that, in addition to personal utility, genetic screening programs are cost effective for payers, especially when performed earlier in life [39, 40]. Because all of the conditions identified here have evidence of improved outcomes when early actions are taken (Table 3), and given that there is a net increase in findings with the same amount of work at the population level (a single assay can just as easily screen one or all human genes), it is likely that the addition of these four conditions with the same or better PPV as existing population screening genes would only improve the cost effectiveness and overall economic benefit of a genetic screening program. However, additional work is still required by official clinical bodies to both evaluate the health economics of early intervention for these conditions and to translate these findings from research into clinical practice through official guidelines. In particular, guidelines will be needed to determine the type and frequency of screening modalities that will be needed for individuals who harbor risk alleles for these conditions. It is also important to include genetic counselors as a part of the return of results process and provide educational materials for all health-care providers involved in the communication of results. Therefore, the next step to expand the boundary of genomics in medicine is the creation, evaluation, and/or refinement of clinical guidelines based on genetics for these conditions.

DATA AVAILABILITY

UKB data are available for download (<https://www.ukbiobank.ac.uk/>). Analysis results for significant associations are available in Table S1. The HNP data are available to qualified researchers upon reasonable request and with permission of the Institute for Health Innovation (IHI) and Helix. Researchers who would like to obtain the raw genotype data related to this study will be presented with a data user agreement which requires that no participants will be re-identified and no data will be shared between individuals or uploaded onto public domains. The IHI encourages and collaborates with scientific researchers on an individual basis. Examples of restrictions

that will be considered in requests to data access include but are not limited to (1) whether the request comes from an academic institution in good standing and will collaborate with our team to protect the privacy of the participants and the security of the data requested, (2) type and amount of data requested, (3) feasibility of the research suggested, (4) amount of resource allocation for the IHI and Renown Hospital required to support the collaboration. Any correspondence and data availability requests should be addressed to JG at (Joe.Grzymski@dri.edu) or Craig Kugler (Craig.Kugler@dri.edu).

Received: 29 April 2021; Revised: 19 July 2021; Accepted: 19 July 2021;
Published online: 13 August 2021

REFERENCES

- Centers for Disease Control and Prevention. Tier 1 genomics applications and their importance to public health. 2019. <https://www.cdc.gov/genomics/implementation/toolkit/tier1.htm>.
- Murray MF et al. DNA-based screening and population health: a points to consider statement for programs and sponsoring organizations from the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2021;23:989–995 <https://doi.org/10.1038/s41436-020-01082-w>.
- Miller DT, et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. 2021. <https://doi.org/10.1038/s41436-021-01172-3>.
- Buchanan AH, et al. Clinical outcomes of a genomic screening program for actionable genetic conditions. *Genet Med*. 2020;22:1874–1882.
- Patel AP, et al. Association of rare pathogenic DNA variants for familial hypercholesterolemia, hereditary breast and ovarian cancer syndrome, and Lynch syndrome with disease risk in adults according to family history. *JAMA Netw Open*. 2020;3:e203959.
- Abul-Husn NS, et al. Exome sequencing reveals a high prevalence of BRCA1 and BRCA2 founder variants in a diverse population-based biobank. *Genome Med*. 2019;12:2.
- Grzymski JJ, et al. Population genetic screening efficiently identifies carriers of autosomal dominant diseases. *Nat. Med*. 2020;26:1235–1239.
- Abul-Husn NS, et al. Implementing genomic screening in diverse populations. *Genome Med*. 2021;13:17.
- Cirulli ET, et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat Commun*. 2020;11:542.
- Wang Q, Dhindsa RS, Carss K, Harper A, Nag A, Tachmazidou I, et al. Surveying the contribution of rare variants to the genetic architecture of human disease through exome sequencing of 177,882 UK Biobank participants. *Cold Spring Harbor Laboratory*. 2020;2020:12.13.422582.
- Denny JC, Bastarache L, Roden DM. Phenome-wide association studies as a tool to advance precision medicine. *Annu Rev Genomics Hum Genet*. 2016;17:353–373.
- Bastarache L, Hughey JJ, Hebringer S, Marlo J, Zhao W, Ho WT, et al. Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science*. 2018;359:1233–1239.
- Denny JC, Bastarache L, Ritchie MD, Carroll RJ, Zink R, Mosley JD, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol*. 2013;31:1102–1110.
- Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM codes to Phecodes: workflow development and initial evaluation. *JMIR Med Inform*. 2019;7:e14325.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17:122.
- Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47:D766–D773.
- Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. *Nucleic Acids Res*. 2018;46:D754–D761.
- Helix's Variants Pipeline Performance White Paper. 2019. https://cdn.shopify.com/s/files/1/2718/3202/files/Helix_Performance_White_Paper_v4.pdf.
- Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole genome regression for quantitative and binary traits. *Cold Spring Harbor Laboratory*. 2020;2020:06.19.162354.
- Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26:2190–2191.
- Chakera AJ, Steele AM, Gloyd AL, Shepherd MH, Shields B, Ellard S, et al. Recognition and management of individuals with hyperglycemia because of a heterozygous glucokinase mutation. *Diabetes Care*. 2015;38:1383–1392.
- Bonnefond A, Boissel M, Bolze A, Durand E, Toussaint B, Vaillant E, et al. Pathogenic variants in actionable MODY genes are associated with type 2 diabetes. *Nat Metab*. 2020;2:1126–1134.
- Stride A, Shields B, Gill-Carey O, Chakera AJ, Colclough K, Ellard S, et al. Cross-sectional and longitudinal studies suggest pharmacological treatment used in patients with glucokinase mutations does not alter glycaemia. *Diabetologia*. 2014;57:54–56.
- Rudland VL. Diagnosis and management of glucokinase monogenic diabetes in pregnancy: current perspectives. *Diabetes Metab Syndr Obes*. 2019;12:1081–1089.
- Grantham JJ. Clinical practice. Autosomal dominant polycystic kidney disease. *N Engl J Med*. 2008;359:1477–1485.
- Helal I, Reed B, Schrier RW. Emergent early markers of renal progression in autosomal-dominant polycystic kidney disease patients: implications for prevention and treatment. *Am. J. Nephrol*. 2012;36:162–167.
- Cao A, Galanello R. β -thalassemia. *Genet Med*. 2010;12:61–76.
- Thein SL. The molecular basis of β -thalassemia. *Cold Spring Harb Perspect Med*. 2013;3:a011700–a011700.
- Weatherall DJ. Phenotype-genotype relationships in monogenic disease: lessons from the thalassaemias. *Nat Rev Genet*. 2001;2:245–255.
- Kohne E. Hemoglobinopathies: clinical manifestations, diagnosis, and treatment. *Dtsch Arztebl Int*. 2011;108:532–540.
- Taher AT, Musallam KM, Cappellini MD, Weatherall DJ. Optimal management of β thalassaemia intermedia. *Br J Haematol*. 2011;152:512–523.
- Long X, Huang Y, Tan H, Li Z, Zhang R, Linpeng S, et al. Identification of a novel MIP frameshift mutation associated with congenital cataract in a Chinese family by whole-exome sequencing and functional analysis. *Eye*. 2018;32:1359–1364.
- Francis P, Chung JJ, Yasui M, Berry V, Moore A, Wyatt MK, et al. Functional impairment of lens aquaporin in two families with dominantly inherited cataracts. *Hum Mol Genet*. 2000;9:2329–2334.
- Shiels A, Hejtmancik JF. Mutations and mechanisms in congenital and age-related cataracts. *Exp Eye Res*. 2017;156:95–102.
- Lee DJ, Lam BL, Arora S, Arheart KL, McCollister KE, Zheng DD, et al. Reported eye care utilization and health insurance status among US adults. *Arch Ophthalmol*. 2009;127:303–310.
- Meuleners LB, Brameld K, Fraser ML, Chow K. The impact of first- and second-eye cataract surgery on motor vehicle crashes and associated costs. *Age Ageing*. 2019;48:128–133.
- Heruye, S. H. et al. Current Trends in the Pharmacotherapy of Cataracts. *Pharmaceuticals*. 2020;13:15. <https://doi.org/10.3390/ph13010015>.
- Moreau KL, King JA. Protein misfolding and aggregation in cataract disease and prospects for prevention. *Trends Mol. Med*. 2012;18:273–282.
- Guzauskas GF, Garbett S, Zhou Z, Spencer SJ, Smith HS, Hao J, et al. Cost-effectiveness of population-wide genomic screening for hereditary breast and ovarian cancer in the United States. *JAMA Netw Open*. 2020;3:e2022874.
- Hao J, et al. Healthcare utilization and costs after receiving a positive BRCA1/2 result from a genomic screening program. *J Pers Med*. 2020;10:7.

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource under application number 40436. Funding was provided to Desert Research Institute (DRI) by the Nevada Governor's Office of Economic Development. Funding was provided to the Renown Institute for Health Innovation by Renown Health and the Renown Health Foundation.

AUTHOR CONTRIBUTIONS

Conceptualization: K.M.S.B., E.T.C., N.L.W., J.J.G., J.T.L. Data curation: Y.N., S.W., M.I., W.L. Formal Analysis: K.M.S.B., E.T.C. Investigation: K.M.S.B., E.T.C., N.L.W. Methodology: K.M.S.B., E.T.C., N.L.W. Resources: Y.N., S.W., M.I., J.J.G., J.T.L., W.L. Software: Y.N., S.W., M.I., W.L. Visualization: K.M.S.B., E.T.C., N.L.W., L.S., E.L. Writing—original draft: K.M.S.B., E.T.C., N.L.W. Writing—review & editing: K.M.S.B., E.T.C., N.L.W., A.B., Y.N., S.W., M.I., L.S., E.L., W.L., J.J.G., J.T.L.

COMPETING INTERESTS

K.M.S.B., E.T.C., A.B., Y.N., W.L., S.W., N.L.W., M.I., L.S., E.L., and J.T.L. are employees of Helix. The other authors declare no competing interests.

ETHICS DECLARATION

The HNP study was reviewed and approved by the University of Nevada, Reno Institutional Review Board (IRB, project 956068-12). The UKB study was performed under protocol 40436 (<https://www.ukbiobank.ac.uk/>). The UKB study was approved by the North West Multicenter Research Ethics Committee, UK (<https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/about-us/ethics>). All participants gave their informed, written consent prior to participation. All data used for research were de-identified.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41436-021-01293-9>.

Correspondence and requests for materials should be addressed to E.T.C.

Reprints and permission information is available at <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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021