HGG Advances

REPORT

Clinical validation of genomic functional screen data: Analysis of observed *BRCA1* variants in an unselected population cohort

Kelly M. Schiabor Barrett,[1,2,6] Max Masnick,[1,3,6,7] Kathryn E. Hatchell,[1] Juliann M. Savatt,[1] Natalie Banet,[4] Adam Buchanan,[1] and Huntington F. Willard[1,5,*]

Summary

Functional assessment of genomic variants provides a promising approach to systematically examine the potential pathogenicity of variants independent of associated clinical data. However, making such conclusions requires validation with appropriate clinical findings. To this end, here, we use variant calls from exome data and *BRCA1*-related cancer diagnoses from electronic health records to demonstrate an association between published laboratory-based functional designations of *BRCA1* variants and *BRCA1*-related cancer diagnoses in an unselected cohort of patient-participants. These findings validate and support further exploration of functional assay data to better understand the pathogenicity of rare variants. This information may be valuable in the context of healthy population genomic screening, where many rare, potentially pathogenic variants may not have sufficient associated clinical data to inform their interpretation directly.

Introduction

In genomic medicine, DNA-based genetic testing has to date been used largely as a diagnostic tool for individual patients or families to search for explanatory variants in genes known or believed to be related to a phenotype or disease of interest. This practice has led to the characterization of many pathogenic variants in well-studied genes as well as to the discovery of novel gene-disease associations.[1–4]

More recently, large-scale population-based sequencing efforts have revealed that all individuals, regardless of disease status, harbor some degree of rare coding variation genome-wide, including in genes known to be associated with heritable disease.[5–7] These rare variants may increase the risk of disease[8,9] and may provide valuable insights into the assessment of risk if their pathogenicity can be established.

One setting in which such rare variants could be used is in healthy population genomic screening, which uses gene panels, exome sequencing (ES), or genome sequencing to identify potentially pathogenic variants in genes associated with disease across unselected cohorts without prior reference to symptomatic or asymptomatic status.[10–12] The goal of such efforts is to identify predicted pathogenic variants before symptoms arise and as early in life as possible, so that individuals and families can take steps to reduce the likelihood of future disease. Hindering the broad adoption of such an approach, however, are difficulties in assessing the pathogenicity of rare variants in such healthy populations.

Healthy population-based screening and individual indication-based diagnostic testing are inherently different: healthy screening is done at scale and in the absence of a phenotype, with the goal of identifying disease risk rather than determining the specific genetic cause of a known condition.[13–15] In a screening setting, potential variant pathogenicity is typically assessed using any and all available clinical and research data for a given variant, with clinical data from diagnostic testing often providing the strongest evidence of potential pathogenicity. However, many pathogenic variants are truly rare and will only ever be seen in one or a few individual cases or families,[16–18] making it unlikely that informative clinical data from diagnostic testing will be available in such instances. Instead, assessing the pathogenicity of these variants will necessarily rely on predictions from computational, comparative, or functional analyses.[19–21]

Findlay et al.[22] have used such a functional assay to generate (via saturation genome editing [SGE]) and then assess nearly all of the single-nucleotide variants (SNVs; hereafter, variants) endogenously in 13 functionally critical exons of the *BRCA1* gene (MIM: 113705) that harbor all of the known pathogenic missense mutations documented in ClinVar for *BRCA1*-related cancers (MIM: 604370, 614320). This functional screen was performed in a human haploid cell line in which the homology-directed repair pathway, which includes *BRCA1*, had

[1]Geisinger Research, Geisinger Health, Danville, PA, USA; [2]Helix OpCo, San Mateo, CA, USA; [3]The MITRE Corporation, Bedford, MA, USA; [4]Department of Pathology and Laboratory Medicine, Women and Infants Hospital, Warren Alpert Medical School of Brown University, Providence, RI, USA; [5]Genome Medical, South San Francisco, CA, USA
[6]These authors contributed equally
[7]MITRE: Approved for Public Release; Distribution Unlimited. Public Release Case Number 21-3600
*Correspondence: hunt@genomemedical.com
https://doi.org/10.1016/j.xhgg.2022.100086.

been deemed essential,[23] making loss of function (LOF) of *BRCA1*—the relevant mechanism of disease in humans[24]—testable by quantifying cell-growth patterns.[22] Resulting functional scores across all variants tested by Findlay et al. were bimodally distributed, and each variant was previously classified as "functional" (i.e., functionally normal), "non-functional" (here, designated throughout as "functionally abnormal; LOF," as recommended by Brnich et al.[25]), or "intermediate" (could not be classified as functionally normal or abnormal; limited to a small number of variants) based on the score distribution.[22] Variant classification from the SGE functional assay showed high concordance with pathogenicity designations for the subset of variants archived in the ClinVar database.[1]

While, in general, this concordance validates SGE assay classification applied in a phenotype-driven, diagnostic setting, assessing the predicted impact of variants in an unselected cohort of individuals from the general population has not yet been reported. Here, we assess the population-level association between the functionally abnormal; LOF variants identified by the SGE functional assay and *BRCA1*-related cancer phenotypes in the DiscovEHR cohort, an unselected group of patient-participants from the Geisinger MyCode Community Health Initiative (MyCode), for which both ES and extensive, longitudinal clinical data are available for analysis.[26–30]

Subjects and methods

Cohort construction

Participants included in this study are from the Geisinger MyCode Community Health Initiative, a population genomics, discovery research project and biobank that began in 2007 and continues to enroll participants from the Geisinger Health System independent of disease status.[26] For this article, data from 92,453 MyCode participants with ES from the DiscovEHR collaboration with the Regeneron Genetics Center (RGC) were available at the time of initial analysis (fall 2019). Sample preparation and ES methodology are described in detail elsewhere.[29] Genomic variant frequency information for this cohort can be accessed from the DiscovEHR browser (see Web resources). The research outlined, including MyCode participation, was approved by the Geisinger institutional review board (IRB 2006-0258 and 2019-0739). Informed consent was obtained from all of the patient-participants.

Bioinformatics

The *BRCA1* gene region (*BRCA1*: chr17:43044294-43125482 on hg38) was extracted from genomic variant call files (gVCFs), which store exome-wide variant and reference sequencing information for all sites, for each of the 92,453 MyCode participants in the DiscovEHR cohort. Following the method developed by the ExAC Consortium for variant calling and filtering of exomes at scale,[31] the pool of *BRCA1* gVCFs was split into $\sqrt{n}$ groups (304) and each group of gVCF files was joint genotyped using GATK 3.5 haplotype caller.[32,33] The output of this process is a VCF, containing both SNV and insertion and deletion (indel) calls, as well as associated call-quality data, for each sample. Next, GATK 3.5 variant quality score recalibration (VQSR) was used to filter SNVs for quality. Briefly, VQSR is a machine learning algorithm that assigns a well-calibrated probability (VQSLOD) to each variant call using high-quality sets of known variants (e.g., HapMap 3, Omni 2.5M SNP chip array) as training and truth resources. A target sensitivity of 99.6% was used to set the VQSLOD score threshold, and all of the variants with scores at or above this threshold were marked with a PASS flag for downstream filtering.[31]

The VCFs were transferred to a relational database in which genomic coordinates of each variant were converted to build37 using LiftOver,[34] and predicted functional consequence, gnomAD database frequency, and ClinVar information were added as annotations using ANNOVAR.[35] SGE functional scores from data published by Findlay et al.[22] were also added as annotations.

To create the analysis dataset, the annotated VCFs were filtered at the site level using the PASS flag from the VQSR analysis and at the sample level using depth (DP) > 10 and genotype quality (GQ) > 20 and allele frequency (AF) > 0.3 as quality thresholds.[31,36] This combination of site- and sample-level quality filtering removed 208,651 of 1,568,771 SNV calls (186,855 from the DP/GQ filter, and 21,796 from the AF filter).

Phenotyping

All of the MyCode participants in the DiscovEHR cohort were phenotyped from a centralized data warehouse, which holds structured data abstracted from the Geisinger EHR (electronic health record; Epic) and ancillary clinical data systems. We extracted data from this warehouse to phenotype participants for gender, age, and *BRCA1* syndromic cancers (breast, ovarian, pancreatic, and prostate) based on diagnosis codes and the Geisinger tumor registry. *BRCA1*-associated peritoneal cancer was included with ovarian cancer for subsequent analysis. Clinicians (A.B., J.M.S., and N.B.) reviewed the criteria for selecting diagnosis codes, which was based on the "term set" approach described by Williams et al.[37] The same group of clinicians reviewed all of the individual diagnosis codes that were included in the phenotypes. We have included the list of *International Classification of Diseases, Tenth Revision* (ICD-10) diagnosis and tumor registry codes used for the phenotype as a supplement to this article (Data S1). Phenotyping did not include information from outside the Geisinger Health System unless clinicians specifically entered it into the Geisinger EHR. For participants with a relevant diagnosis code, age at diagnosis was determined. All of the relevant phenotypic records were combined with relevant genomic data. The final sample size from this analysis, combining clinical and sequencing data, is 92,453 patient-participants and 1,359,057 *BRCA1* SNVs; when limited to those participants aged 18 years and older, the study cohort consists of 91,659 patient-participants. This adult cohort is referred to as the "DiscovEHR cohort" in the text.

Statistical analysis

To assess potential associations between clinical phenotypes and genomic variant SGE functional screening scores,[22] the combined dataset of phenotypic and genomic data was exported to R for statistical analysis.[38] Variants classified as intermediate in the SGE screen ($n = 4$ unique variants, observed 9 times in this cohort) were not considered in this analysis due to small sample size and lack of a clear pathogenicity prediction. To assess the association between *BRCA1*-related cancer diagnosis and SGE functional score assignment, a χ^2 test was performed comparing the proportion of

Table 1. Characteristics of DiscovEHR cohort, overall and by SGE-classified *BRCA1* variant status

	Participants		Functionally abnormal; LOF		Functionally normal		p[a]
All participants (n, %)	91,659	100.0	29	<0.01	3,513	3.8	
Age at analysis, y (mean, SD)	62	17.8	55.6	17.6	62.2	17.5	0.04
Age at analysis, y (n, %)							
<20	117	0.1	0	0.0	4	0.1	0.70
20–29	3,425	3.7	2	6.9	123	3.5	
30–39	9,260	10.1	5	17.2	331	9.4	
40–49	10,579	11.5	4	13.8	424	12.1	
50–59	14,654	16.0	5	17.2	547	15.6	
60–69	19,274	21.0	5	17.2	758	21.6	
70–79	17,995	19.6	6	20.7	712	20.3	
≥80	16,355	17.8	2	6.9	614	17.5	
Sex (n, %)							
Female	55,637	60.7	19	65.5	2,087	59.4	0.60
Male	36,011	39.3	10	34.5	1,426	40.6	
Unknown	11	0	0	0.0	0	0.0	
Race (n, %)							
White	89,466	97.6	29	100.0	3,351	95.4	0.90
Black or African American	1,519	1.7	0	0.0	130	3.7	
Asian	274	0.3	0	0.0	18	0.5	
Unknown	175	0.2	0	0.0	8	0.2	
Native Hawaiian or other Pacific Islander	122	0.1	0	0.0	3	0.1	
American Indian or Alaskan Native	103	0.1	0	0.0	3	0.1	
Ethnicity (n, %)							
Not Hispanic or Latino	88,427	96.5	26	89.7	3,410	97.1	0.05
Unknown	1,728	1.9	2	6.9	57	1.6	
Hispanic or Latino	1,504	1.6	1	3.4	46	1.3	

[a]p values compare predicted functionally abnormal; LOF to predicted functionally normal groups using a t test for continuous variables and a χ^2 test (with a Yates continuity correction) for categorical variables.

BRCA1-related cancer diagnoses for DiscovEHR participants with predicted deleterious *BRCA1* variants to that of DiscovEHR participants with predicted non-deleterious *BRCA1* variants,[22] as well as those without an SGE-screened variant. This analysis was repeated for breast and/or ovarian cancer, limited to female participants.

To evaluate the association between presence of deleterious variants and cancer diagnosis in the context of diagnosis age, we constructed Kaplan-Meier (KM) time-to-event curves (where event = *BRCA1*-related cancer diagnosis) for participants stratified by the presence of predicted deleterious or non-deleterious *BRCA1* variants, or no SGE-screened variant. Participants entered into observation at the earliest of (1) their first clinical encounter in the Geisinger Health System, (2) the date of onset of the earliest problem list item in the Geisinger EHR, or (3) the earliest excision date of a *BRCA1*-related tumor as recorded in the Geisinger EHR. We constructed cumulative risk curves for any *BRCA1*-related cancer for all participants with predicted deleterious *BRCA1* variants and for all participants with predicted non-deleterious *BRCA1* variants. We compared these cumulative risk curves to that of a control group consisting of participants without predicted deleterious or non-deleterious variants using log rank tests. This analysis was repeated for breast and/or ovarian cancer, limited to female participants. To assess the possible misclassification of any potential non-deleterious variants, we created and compared variant-specific time-to-event curves for all non-deleterious variants.

Variant rarity

For many conditions, pathogenic variants are typically rare at the population level. To define rarity in this context, we calculated the highest allowable frequency based on the known prevalence, genetic heterogeneity, allelic heterogeneity, and penetrance of *BRCA1*-related cancer.[12–15] To qualify as "rare," we calculated a maximum allowable frequency (separate of founder mutations) of 0.00023, or present up to 41 times in this cohort (n = 2 × ~90,000 alleles), based on the reported prevalence of inherited

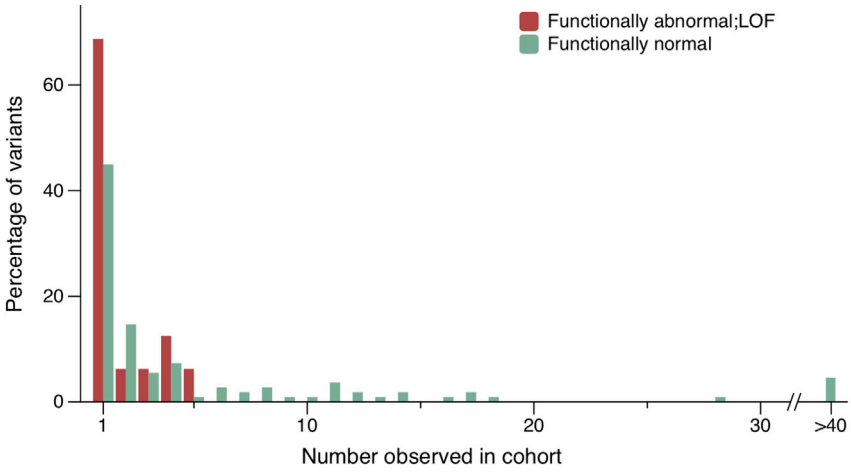

Figure 1. Frequencies of functionally abnormal;LOF and functionally normal *BRCA1* alleles in the DiscovEHR cohort

Along the x axis, the histogram displays variants, grouped by SGE classification and binned by number of occurrences in the cohort. The height of the bar corresponds to the percentage of variants, within each classification, that belong to each frequency bin. In accordance with pathogenicity expectations, variants over 40 occurrences are captured in a single bin.

breast and ovarian cancers (1/330),[16,18] estimated genetic heterogeneity (66%),[16,19] allelic heterogeneity in European populations (as the best match for the MyCode cohort) (0.13),[20] and reported penetrance for women by age 80 (48%).[21]

Results

Table 1 displays the characteristics of the adult DiscovEHR cohort used for this analysis (n = 91,659). Within this cohort, we observed 2,711 unique variants across the entire *BRCA1* locus, 129 (4.8%) of which were screened by Findlay et al. with the SGE functional assay.[22] Of the 129 variants thus informative for the present study, 16 had been determined previously to be functionally abnormal; LOF, and 109 had been determined to be functionally normal (Data S2). The remaining 4 could not be classified ("intermediate" in Findlay et al.[22]) and were not considered further in this study. As shown in Table 1, 29 participants were found to carry 1 of the 16 *BRCA1* variants predicted to be functionally abnormal; LOF, while 3,534 were found to carry 1 of the 109 *BRCA1* variants predicted to be functionally normal.

While only a subset of rare variants in the genome are expected to be pathogenic, pathogenic variants should, in the absence of positive selection, meet rarity expectations based on the underlying genetic architecture of the associated disease and the known disease prevalence in the reference population.[39] Based on this, we expect pathogenic *BRCA1* variants to have allele frequencies below ~0.00023 (thus, an incidence of <41 copies in this cohort) based on upper-bound estimates of *BRCA1*-related cancer prevalence, genetic heterogeneity, and allelic heterogeneity reported in the literature for European ancestry (see "Variant rarity" above for details).

As shown in Figure 1, high proportions of both predicted functionally abnormal; LOF (11/16, 68.8%) and functionally normal (49/109, 45.0%) variants are singletons in our cohort. No predicted functionally abnormal; LOF variants are seen at counts above the calculated pathogenicity threshold of 41 copies, although a small proportion of predicted functionally normal variants are seen above this threshold. Allele frequencies for the predicted functionally abnormal; LOF variants identified in the cohort are also found below this threshold in gnomAD, a similarly sized, ancestrally matched cohort.[40]

Table 2 shows the proportion of participants with *BRCA1*-related cancer diagnoses for the cohort, stratified by SGE classification[22] as carriers of predicted functionally abnormal; LOF or functionally normal variants and by sex. (Participant characteristics of the DiscovEHR cohort further stratified by cancer diagnosis can be found in Table S1.) For patients harboring a variant predicted to be either functionally abnormal; LOF or functionally normal by the SGE assay,[22] the frequency of cancer diagnosis was compared for (1) all *BRCA1*-related cancer diagnoses (pancreatic, prostate, breast, and ovarian),[24,41] (2) breast and/or ovarian cancer diagnoses, and (3) no cancer diagnosis.

If variants predicted to be functionally abnormal; LOF are in fact pathogenic, then we expect a higher proportion of cancer diagnoses among participants with predicted functionally abnormal; LOF variants compared to participants without such variants. As shown in Table 2, 20.7% of participants carrying predicted functionally abnormal; LOF variants had a *BRCA1*-related cancer diagnosis available in the Geisinger EHR, compared to 7.6% of participants carrying variants predicted to be functionally normal ($p = 0.022$). When limiting analysis to breast and/or ovarian cancer diagnoses in females, 26.3% of patients with predicted functionally abnormal; LOF variants had a diagnosis during follow-up compared to 6.4% of those with predicted functionally normal variants ($p = 0.00243$) (Table 2). (Two of the 1,426 males with a predicted functionally normal variant also had a breast cancer diagnosis; statistical analysis was not performed due to the small number of such cases.)

At the population level, *BRCA1*-related cancer risk and resulting diagnoses increase with age.[42,43] When comparing cumulative risk calculated using time-to-event analysis, participants with a predicted functionally abnormal; LOF variant were found to have an increased cumulative risk of diagnosis of any *BRCA1*-related cancer when compared to participants without any SGE-classified

Table 2. Associations between SGE-classified variant status and *BRCA1*-related cancer diagnosis in the DiscovEHR cohort

		BRCA1-related cancer[a]							
		None		Any			Breast and/or ovarian only		
	All carriers	N	%	N	%	p[b]	N	%	p[b]
All									
Functionally abnormal; LOF variants	29	23	79.3	6	20.7	0.022	5	17.2	0.001
Functionally normal variants	3,513	3,247	92.4	266	7.6		135	3.8	
Females									
Functionally abnormal; LOF variants	19	14	73.7	5	26.3	0.008	5	26.3	0.002
Functionally normal variants	2,087	1,993	92.6	154	7.4		133	6.4	
Males									
Functionally abnormal; LOF variants	10	9	90.0	1	10.0	NA	0	0.0	NA
Functionally normal variants	1,426	1,314	92.1	112	7.9		2	0.1	

[a]*BRCA1*-related cancer diagnoses include breast, prostate, ovarian, and pancreatic malignancies.
[b]p values compare the proportion of predicted functionally abnormal; LOF and predicted functionally normal variant carriers with any cancer diagnosis or a breast/ovarian cancer diagnosis, respectively, to those with no *BRCA1*-related cancer diagnosis using a χ^2 test with a Yates continuity correction. Statistical analysis was not performed for males due to the small number of participants.

variants (p = 0.01) (Figure 2A). This association was even stronger when limited to considering only breast and ovarian cancer in female participants (p = 0.001) (Figure 2C). In contrast, there was no difference in the cumulative risk of diagnosis for patients carrying variants predicted to be functionally normal when compared to those without any SGE-classified variants (p = 0.4 and p = 0.6 for all *BRCA1* cancers and breast and/or ovarian cancers in females, respectively) (Figures 2B and 2D). Looking at cumulative risk by decade for all relevant cancers or specifically for breast and/or ovarian cancers in females, the cumulative risk of cancer for those with predicted functionally abnormal; LOF variants diverges from those without any SGE-classified variants as early as the 30- to 39-year-old age group in our cohort (Figures 2A and 2C).

These findings suggest that, in the aggregate, *BRCA1* variants that are predicted to be functionally abnormal; LOF are associated with increased risk of *BRCA1*-related cancer diagnosis, irrespective of participant age. However, it is not possible to assess the association between specific variants individually and *BRCA1*-related cancer diagnosis due to the range of different variants detected (Figure 1).

Several of the patients with variants designated as functionally abnormal; LOF who developed cancer had clinical features consistent with a pathogenic *BRCA1* variant. One patient had bilateral breast cancer (one of which was of triple negative receptor status), one had an early-onset, triple negative breast cancer, one had an early-onset prostate cancer, and two had high-grade serous ovarian cancer. Reviews of family cancer history among those with variants designated as functionally abnormal; LOF showed that the majority had a family history of breast cancer (16/29), and several had a family history of other *BRCA1*-associated cancers. Given the small numbers, however, we cannot determine whether a family history of *BRCA1*-associated cancers was more common in these individuals.

Because there is a substantial degree of relatedness within the MyCode population,[27] we assessed whether the cancer cases detected in this study represented genetically independent events, as related individuals share genetic variation that is known to contribute to phenotype penetrance and expressivity. Among the 29 participants carrying variants predicted to be functionally abnormal; LOF (Table 1), 5 were found to be genetically related at the first- or second-degree level. However, none of the six cancer cases with predicted functionally abnormal; LOF variants was found to be related, indicating that known relatedness is not responsible for the association we observe between SGE classification and cancer diagnosis. Overall, variants at four different positions were detected in the six unrelated cancer cases, further suggesting that a single variant position is not driving this association.

Discussion

As proactive screening of healthy populations for genetic risk of disease becomes more common, methods to determine variant pathogenicity at scale become essential, as it is unlikely, given the relative rarity of most pathogenic variants for many conditions, that there will be sufficient clinical evidence for the majority of variants discovered in large cohorts.[20] Functional assays, such as the SGE assay developed and used by Findlay et al.,[22] are independent of clinical data and represent a potentially valuable source of information for variant interpretation at scale, provided that the nature of the functional assay can be validated where possible with the available clinical data.[44] Here, we report such a validation, demonstrating an association

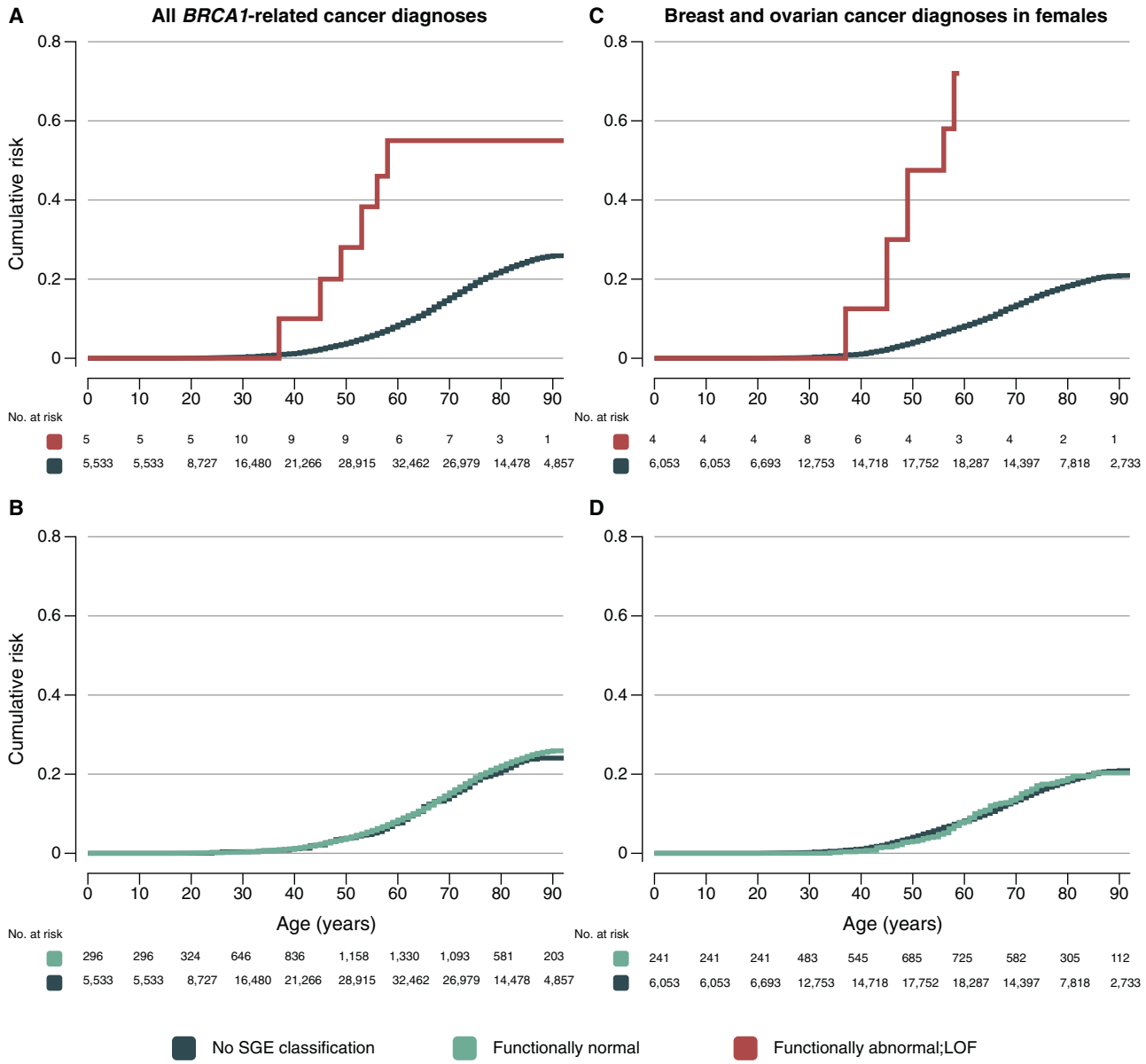

Figure 2. Time-to-event analysis

Age at *BRCA1*-related cancer diagnosis for participants with variants predicted to be functionally abnormal; LOF or functionally normal by results of the SGE assay, compared to participants with non-SGE classified variants.

(A and B) All *BRCA1*-related cancers for all participants, for functionally abnormal; LOF ($p = 0.01$) and functionally normal ($p = 0.4$) variants, respectively.

(C and D) Only breast and ovarian cancer diagnoses in female participants, for functionally abnormal; LOF ($p = 0.001$) and functionally normal ($p = 0.6$) variants, respectively. The p values are from log rank tests for differences between the 2 time-to-event curves in each panel.

between patients harboring *BRCA1* variants predicted to be either functionally normal or functionally abnormal; LOF and *BRCA1*-related cancer diagnoses in a large, unselected cohort of 91,659 patient-participants.

Providing support for increased cancer risk among participants carrying predicted functionally abnormal; LOF variants, we found an overabundance of *BRCA1*-related cancer cases in such participants when compared to those with variants predicted by the SGE assay to be functionally normal (Table 2). In addition, the cumulative risk of *BRCA1*-related cancer diagnosis was higher for participants carrying predicted functionally abnormal; LOF variants compared to participants without an SGE-classified variant (Figure 2). We note that the associations reported here can be accounted for entirely by breast or ovarian cancer in females. There are no significant associations involving pancreatic or prostate cancer (data not shown).

Limited to breast or ovarian cancer for females, cumulative risk of diagnosis by age 80 in our cohort for participants carrying predicted functionally abnormal; LOF variants was

72.0%, >4-fold greater than the risk for participants without an SGE-classified variant (18.3%). These cumulative risks are in line with recent estimates from an independent prospective cohort of breast and ovarian cancer risk in >2,000 *BRCA1* pathogenic variant carriers.[42]

Clinically validated functional assessments of genes with association to heritable disease provide a promising opportunity to assess potential variant pathogenicity and to communicate disease risk for individuals harboring variants when little or no clinical evidence of disease is available to predict pathogenicity. The SGE screen[22] identified >400 functionally abnormal; LOF missense variants in the *BRCA1* gene, a class of variants in which computational and comparative predictions often fall short for the inference and prediction of potential or likely pathogenicity.[45,46] Highlighting this potential, 7 of the 16 variants classified as predicted functionally abnormal; LOF in our cohort are difficult to interpret for disease risk without this functional assessment because they are either absent from ClinVar (1 variant) or carry a ClinVar assertion of variant of unknown significance or conflicting pathogenicity (6 variants). Furthermore, these 7 variants are located in 6 different screened exons, emphasizing the importance of the breadth of assay when performing variant interpretation at scale.

The results of the present study provide initial clinical evidence of the potential for leveraging functional data for use in variant interpretation. This conclusion notwithstanding, this study has a number of limitations. While the DiscovEHR cohort analyzed here is relatively large by current standards (n = 91,659), only a small number of relevant variants were detected across both functionally abnormal; LOF and functionally normal classifications (125 variants in 3,568 participants). Given the comprehensive EHR data associated with the cohort, this was sufficient to validate the approach in aggregate across the *BRCA1* gene, but, as anticipated for many rare variants, it was insufficient to assess the validity of any single variant prediction, which would ultimately be of much greater potential value.[44,47] Such a capability will be necessary in time to optimally leverage functional data at the level of individual patients.

This limitation will be even more striking for populations that are far more heterogeneous than the relatively homogeneous Geisinger cohort studied here.[48–50] The average age of functionally abnormal; LOF variant carriers is eight years younger than those counterparts with functionally normal variants and nearly seven years younger than the average age of all participants in the DiscovEHR cohort. This age imbalance suggests that there could be a survivorship bias in the DiscovEHR cohort. Although a limitation in this study, this potential bias highlights the potential for improved clinical and personal utility of sequencing earlier in the life course. Note also that age at analysis did not include deceased patients, who were not available for the DiscovEHR cohort. Full understanding of the impact of rare variants on inherited disease, such as *BRCA1*-related cancers explored here, will require extensive data sharing and collaboration from laboratories and populations around the globe to reach the full potential of population-based screening for precision health.

Data and code availability

Genomic and clinical data used for this study are available only with approval for access from the MyCode Governing Board and the Geisinger IRB. The analysis code is available on request from the corresponding author.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100086.

Acknowledgments

We acknowledge Stacy Joseph for invaluable support in project management and manuscript preparation. In addition, we would like to acknowledge Steven Ney for data analysis contributions.

Author contributions

K.M.S.B., M.M., K.E.H., and H.F.W. designed the study. K.M.S.B., M.M., and K.E.H. developed strategies for data collection and analysis. M.M. performed the data analysis and prepared the tables and figures. J.M.S., N.B., and A.B. provided expert interpretation of the *BRCA1* variants. K.M.S.B. and K.E.H. drafted the initial manuscript. All of the authors contributed to and approved the final manuscript.

Declaration of interests

K.M.S.B. is an employee at Helix. M.M.'s affiliation with The MITRE Corporation is provided for identification purposes only and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions, or viewpoints expressed by M.M. K.E.H. is an employee of and shareholder in Invitae. H.F.W. is an employee at Genome Medical.

Received: August 30, 2021
Accepted: January 6, 2022

Web resources

Online Mendelian Inheritance in Man: http://www.omim.org.

BRCA1 SGE screen details: https://sge.gs.washington.edu/BRCA1/

DiscovEHR browser: http://www.discovehrshare.com/

References

1. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067.
2. Strande, N.T., Riggs, E.R., Buchanan, A.H., Ceyhan-Birsoy, O., DiStefano, M., Dwight, S.S., Goldstein, J., Ghosh, R., Seifert, B.A., Sneddon, T.P., et al. (2017). Evaluating the clinical

validity of gene-disease associations: an evidence-based framework developed by the Clinical Genome Resource. Am. J. Hum. Genet. *100*, 895–906.
3. Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D., and Cooper, D.N. (2017). The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum. Genet. *136*, 665–677.
4. Rehm, H.L., Berg, J.S., and Plon, S.E. (2018). ClinGen and ClinVar - enabling genomics in precision medicine. Hum. Mutat. *39*, 1473–1475.
5. Campbell, C.D., and Eichler, E.E. (2013). Properties and rates of germline mutations in humans. Trends Genet. *29*, 575–584.
6. Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E.V., Mort, M., Phillips, A.D., Shaw, K., Stenson, P.D., Cooper, D.N., et al. (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. Am. J. Hum. Genet. *91*, 1022–1032.
7. Telenti, A., Pierce, L.C.T., Biggs, W.H., di Iulio, J., Wong, E.H.M., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C., et al. (2016). Deep sequencing of 10,000 human genomes. Proc. Natl. Acad. Sci. U S A *113*, 11901–11906.
8. Schork, N.J., Murray, S.S., Frazer, K.A., and Topol, E.J. (2009). Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. *19*, 212–219.
9. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
10. Grzymski, J.J., Elhanan, G., Morales Rosado, J.A., Smith, E., Schlauch, K.A., Read, R., Rowan, C., Slotnick, N., Dabe, S., Metcalf, W.J., et al. (2020). Population genetic screening efficiently identifies carriers of autosomal dominant diseases. Nat. Med *26*, 1235–1239.
11. Goldfeder, R.L., Wall, D.P., Khoury, M.J., Ioannidis, J.P.A., and Ashley, E.A. (2017). Human genome sequencing at the population scale: a primer on high-throughput DNA sequencing and analysis. Am. J. Epidemiol. *186*, 1000–1009.
12. Haer-Wigman, L., van der Schoot, V., Feenstra, I., Vulto-van Silfhout, A.T., Gilissen, C., Brunner, H.G., Vissers, L.E.L.M., and Yntema, H.G. (2019). 1 in 38 individuals at risk of a dominant medically actionable disease. Eur. J. Hum. Genet. *27*, 325–330.
13. Marzuillo, C., De Vito, C., D'Andrea, E., Rosso, A., and Villari, P. (2014). Predictive genetic testing for complex diseases: a public health perspective. QJM *107*, 93–97.
14. Biesecker, L.G. (2019). Genomic screening and genomic diagnostic testing—two very different kettles of fish. Genome Med. *11*, 75.
15. Kotze, M.J., Lückhoff, H.K., Peeters, A.V., Baatjes, K., Schoeman, M., van der Merwe, L., Grant, K.A., Fisher, L.R., van der Merwe, N., Pretorius, J., et al. (2015). Genomic medicine and risk prediction across the disease spectrum. Crit. Rev. Clin. Lab. Sci. *52*, 120–137.
16. Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.
17. Kobayashi, Y., Yang, S., Nykamp, K., Garcia, J., Lincoln, S.E., and Topper, S.E. (2017). Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. Genome Med. *9*, 13.
18. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.
19. Hassan, M.S., Shaalan, A.A., Dessouky, M.I., Abdelnaiem, A.E., and ElHefnawi, M. (2019). Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity. Genomics *111*, 869–882.
20. Cline, M.S., Babbi, G., Bonache, S., Cao, Y., Casadio, R., de la Cruz, X., Díez, O., Gutiérrez-Enríquez, S., Katsonis, P., Lai, C., et al. (2019). Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. Hum. Mutat. *40*, 1546–1556.
21. Reeb, J., Wirth, T., and Rost, B. (2020). Variant effect predictions capture some aspects of deep mutational scanning experiments. BMC Bioinformatics *21*, 107.
22. Findlay, G.M., Daza, R.M., Martin, B., Zhang, M.D., Leith, A.P., Gasperini, M., Janizek, J.D., Huang, X., Starita, L.M., and Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. Nature *562*, 217–222.
23. Blomen, V.A., Majek, P., Jae, L.T., Bigenzahn, J.W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F.R., Olk, N., Stukalov, A., et al. (2015). Gene essentiality and synthetic lethality in haploid human cells. Science *350*, 1092–1096.
24. Petrucelli, N., Daly, M.B., and Pal, T. (1993). BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer. In GeneReviews, M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, G. Mirzaa, and A. Amemiya, eds. (University of Washington, Seattle).
25. The Clinical Genome Resource Sequence Variant Interpretation Working Group, Brnich, S.E., Abou Tayoun, A.N., Couch, F.J., Cutting, G.R., Greenblatt, M.S., Heinen, C.D., Kanavy, D.M., Luo, X., McNulty, S.M., et al. (2019). Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med. *12*, 3.
26. Carey, D.J., Fetterolf, S.N., Davis, F.D., Faucett, W.A., Kirchner, H.L., Mirshahi, U., Murray, M.F., Smelser, D.T., Gerhard, G.S., and Ledbetter, D.H. (2016). The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. Genet. Med. *18*, 906–913.
27. Staples, J., Maxwell, E.K., Gosalia, N., Gonzaga-Jauregui, C., Snyder, C., Hawes, A., Penn, J., Ulloa, R., Bai, X., Lopez, A.E., et al. (2018). Profiling and leveraging relatedness in a precision medicine cohort of 92,455 exomes. Am. J. Hum. Genet. *102*, 874–889.
28. Buchanan, A.H., Manickam, K., Meyer, M.N., Wagner, J.K., Hallquist, M.L.G., Williams, J.L., Rahm, A.K., Williams, M.S., Chen, Z.-M.E., Shah, C.K., et al. (2018). Early cancer diagnoses through BRCA1/2 screening of unselected adult biobank participants. Genet. Med. *20*, 554–558.
29. Dewey, F.E., Murray, M.F., Overton, J.D., Habegger, L., Leader, J.B., Fetterolf, S.N., O'Dushlaine, C., Van Hout, C.V., Staples, J., Gonzaga-Jauregui, C., et al. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. Science *354*, aaf6814.
30. Buchanan, A.H., Kirchner, H., Schwartz, M.L.B., Kelly, M.A., Schmidlen, T., Jones, L.K., Hallquist, M.L.G., Rocha, H., Betts,

M., Schwiter, R., et al. (2020). Clinical outcomes of a genomic screening program for actionable genetic conditions. Genet. Med. *22*, 1874–1882.
31. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
32. DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet *43*, 491–498.
33. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics *11*, 11.10.1–11.10.33.
34. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.
35. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
36. Carson, A.R., Smith, E.N., Matsui, H., Bræ, S.K., Jepsen, K., Hansen, J.-B., and Frazer, K.A. (2014). Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC Bioinformatics *15*, 125.
37. Williams, R., Brown, B., Kontopantelis, E., van Staa, T., and Peek, N. (2019). Term sets: a transparent and reproducible representation of clinical code sets. PLoS One *14*, e0212291.
38. R Core Team (2020). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
39. Whiffin, N., Minikel, E., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K., Ing, A.Y., Barton, P.J.R., Funke, B., Cook, S.A., MacArthur, D., et al. (2017). Using high-resolution variant frequencies to empower clinical genome interpretation. Genet. Med. *19*, 1151–1158.
40. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
41. Manickam, K., Buchanan, A.H., Schwartz, M.L.B., Hallquist, M.L.G., Williams, J.L., Rahm, A.K., Rocha, H., Savatt, J.M., Evans, A.E., Butry, L.M., et al. (2018). Exome sequencing-based screening for BRCA1/2 expected pathogenic variants among adult biobank participants. JAMA Netw. Open *1*, e182140.
42. Kuchenbaecker, K.B., Hopper, J.L., Barnes, D.R., Phillips, K.-A., Mooij, T.M., Roos-Blom, M.-J., Jervis, S., van Leeuwen, F.E., Milne, R.L., Andrieu, N., et al. (2017). Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. JAMA *317*, 2402–2416.
43. Manchanda, R., Lieberman, S., Gaba, F., Lahad, A., and Levy-Lahad, E. (2020). Population screening for inherited predisposition to breast and ovarian cancer. Annu. Rev. Genomics Hum. Genet. *21*, 373–412.
44. Kim, H.-K., Lee, E.J., Lee, Y.-J., Kim, J., Kim, Y., Kim, K., Lee, S.-W., Chang, S., Lee, Y.J., Lee, J.W., et al. (2020). Impact of proactive high-throughput functional assay data on BRCA1 variant interpretation in 3684 patients with breast or ovarian cancer. J. Hum. Genet. *65*, 209–220.
45. Li, J., Zhao, T., Zhang, Y., Zhang, K., Shi, L., Chen, Y., Wang, X., and Sun, Z. (2018). Performance evaluation of pathogenicity-computation methods for missense variants. Nucleic Acids Res. *46*, 7793–7804.
46. Hart, S.N., Hoskin, T., Shimelis, H., Moore, R.M., Feng, B., Thomas, A., Lindor, N.M., Polley, E.C., Goldgar, D.E., Iversen, E., et al. (2019). Comprehensive annotation of BRCA1 and BRCA2 missense variants by functionally validated sequence-based computational prediction models. Genet. Med. *21*, 71–80.
47. Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant interpretation: functional assays to the rescue. Am. J. Hum. Genet. *101*, 315–325.
48. Gurdasani, D., Barroso, I., Zeggini, E., and Sandhu, M.S. (2019). Genomics of disease risk in globally diverse populations. Nat. Rev. Genet. *20*, 520–535.
49. Popejoy, A.B., and Fullerton, S.M. (2016). Genomics is failing on diversity. Nature *538*, 161–164.
50. Abul-Husn, N.S., Soper, E.R., Odgis, J.A., Cullina, S., Bobo, D., Moscati, A., Rodriguez, J.E., CBJL Genomics Team, Regeneron Genetics Center, and Loos, R.J.F., et al. (2019). Exome sequencing reveals a high prevalence of BRCA1 and BRCA2 founder variants in a diverse population-based biobank. Genome Med. *12*, 2.