HGG Advances

ARTICLE

Rare coding variants in DNA damage repair genes associated with timing of natural menopause

Lucas D. Ward,[1,*] Margaret M. Parker,[1] Aimee M. Deaton,[1] Ho-Chou Tu,[1] Alexander O. Flynn-Carroll,[1] Gregory Hinkle,[1] and Paul Nioi[1]

Summary

The age of menopause is associated with fertility and disease risk, and its genetic control is of great interest. We use whole-exome sequences from 132,370 women in the UK Biobank to test for associations between rare damaging variants and age at natural menopause. Rare damaging variants in five genes are significantly associated with menopause: *CHEK2* ($p = 3.3 \times 10^{-51}$), *DCLRE1A* ($p = 8.4 \times 10^{-13}$), and *HELB* ($p = 5.7 \times 10^{-7}$) with later menopause and *TOP3A* ($p = 7.6 \times 10^{-8}$) and *CLPB* ($p = 8.1 \times 10^{-7}$) with earlier menopause. Two additional genes are suggestive: *RAD54L* ($p = 2.4 \times 10^{-6}$) with later menopause and *HROB* ($p = 2.9 \times 10^{-6}$) with earlier menopause. In a follow-up analysis of repeated questionnaires in women who were initially premenopausal, *CHEK2*, *TOP3A*, and *RAD54L* genotypes are associated with subsequent menopause. Consistent with previous genome-wide association studies (GWASs), six of the seven genes are involved in the DNA damage repair pathway. Phenome-wide scans across 398,569 men and women revealed that in addition to known associations with cancers and blood cell counts, rare variants in *CHEK2* are also associated with increased risk for uterine fibroids, polycystic ovary syndrome, and prostate hypertrophy; these associations are not shared with higher-penetrance breast cancer genes. Causal mediation analysis suggests that approximately 8% of the breast cancer risk conferred by *CHEK2* pathogenic variants after menopause is mediated through delayed menopause.

Introduction

The age at natural menopause (ANM) varies widely between women, and understanding the biology of menopause timing is important because early menopause is associated with risk for cardiovascular disease and osteoporosis, and late menopause is associated with risk for breast cancer.[1] ANM is also strongly related to fertility, because natural fertility ends on average 10 years before menopause.[2]

Genetic and environmental factors both correlate with ANM. Lower socioeconomic status, fewer live births, not using oral contraceptives, and smoking have been consistently associated with earlier ANM.[1] Twin studies and family studies established that ANM is a highly heritable trait.[3,4] Population-based genome-wide association studies (GWAS) have identified many common polymorphisms associated with ANM.[5–11] These ANM-associated common variants are enriched at loci harboring genes involved in the DNA repair and replication checkpoint process, such as *BRCA1* (MIM: 113705), *MCM8* (MIM: 608187), *CHEK2* (MIM: 604373), and *HELB* (MIM: 614539). A recent Mendelian randomization (MR) analysis found an association between common ANM-associated variants and breast cancer risk, but not conversely between common breast-cancer-associated variants and ANM. This observation supports a causal relationship between variation in lifetime endogenous estrogen exposure (resulting from variation in the duration between menarche and menopause) and risk for breast cancer.[11]

GWASs are limited in their ability to identify causal genes because the majority of associated common haplotypes contain no coding variants, and the closest gene to a noncoding variant is not reliably the causal gene in the absence of functional evidence. Indeed, the variants associated with ANM near tumor suppressor genes *BRCA1* and *CHEK2* have not been the same pathogenic variants implicated in breast cancer. To better identify causal genes for ANM by discovering rare coding variants strongly associated with ANM, we used exome sequencing data from 132,370 women in the UK Biobank. After identifying genes where aggregated protein-truncating variants (PTVs) associated with ANM, we then performed phenome-wide analysis of quantitative traits and disease diagnoses in carriers of these variants across 398,569 exome-sequenced individuals in the UK Biobank.

Subjects and methods

UK Biobank study

The UK Biobank consists of approximately 500,000 volunteer participants, who were aged 40–69 years when recruited between 2006 and 2010.[12,13] Both array genotyping and whole-exome sequencing have been performed on most of these participants.[14] Data from genotyping, sequencing, questionnaires, primary care data, hospitalization data, cancer registry data, and death registry data were obtained through application number 26041. Ethical oversight for the UK Biobank is provided by an Ethics and Governance Council which obtained informed consent from all participants to use these data for health-related research.

Variant calling and definition

The source of genetic data for the main analysis was exome sequencing data. DNA from whole blood was extracted and

[1]Alnylam Pharmaceuticals, Cambridge, MA 02142, USA
*Correspondence: lward@alnylam.com
https://doi.org/10.1016/j.xhgg.2021.100079.

sequenced by the Regeneron Genetics Center (RGC) using protocols described elsewhere.[15] Of the variants called by RGC, additional quality-control filters, were applied: Hardy-Weinberg equilibrium (among the white subpopulation) $p > 10^{-10}$ and rate of missing calls across individuals less than 2%. Variants were then annotated using ENSEMBL Variant Effect Predictor (VEP) v.95,[16] using the LOFTEE plug-in to additionally identify high-confidence predicted PTVs (also known as predicted loss of function [pLOF]).[17] Variant effects were scored against all available transcripts in ENSEMBL, and the most severe predicted impact was retained. Because the most severe impact could be in a rarely expressed transcript, to ensure biological relevance of these bioinformatic predictions, we used the "canonical" transcript to create the description of variants in Human Genome Variation Society (HGVS) nomenclature used in the discussion, and the individual strongest-associated missense and frameshift variants referenced in the discussion were checked to ensure that they had these predicted effects relative to the "canonical" transcript according to VEP. Variants were also annotated with the Whole Genome Sequence Annotator (WGSA)[18] to add Combined Annotation Dependent Depletion (CADD) scores[19] to predict deleteriousness of missense variants. For single-variant analyses, a minor allele frequency filter was imposed such that only variants present in at least 10 individuals with phenotype data were retained. For rare variant burden analyses, variants were defined as rare if their minor allele frequency was under 1%. Variants were aggregated in each protein-coding gene as follows: PTV variants were defined as variants with their most severe consequence from VEP as "stop gained," "splice donor," "splice acceptor," or "frameshift," and their confidence from LOFTEE as "HC" (high confidence). Damaging missense variants were defined as variants with their most severe consequence from VEP as "missense" and a CADD score of 25 or greater. Genes were defined as implemented by ENSEMBL v.95 and further filtered to retain only "genes with protein product" as currently defined by the HUGO Gene Nomenclature Committee (HGNC) (accessed January 28, 2021).[20]

A supplementary source of genetic data, used only for single-variant tests, was obtained from array genotyping. Genotypes were called through chip typing and imputation as described previously.[13] Variants were filtered so that imputation quality score (if not directly genotyped) was greater than 0.8, missingness across individuals was less than 2%, and minor allele frequency was at least 10 carriers with data for the phenotype being analyzed.

Participant definition for overall analyses

An initial round of quality control was performed by RGC, which removed subjects with evidence of contamination, discrepancies between chromosomal and reported sex, and high discordance between sequencing and genotyping array data. Genetic relatedness was calculated using the Pedigree Reconstruction and Identification of a Maximum Unrelated Set (PRIMUS) algorithm,[21] and an "unrelated" subpopulation was defined by removing all first- and second-degree relatives, as well as some third-degree relatives; these individuals were removed when performing time-to-event (TTE) analyses and simple linear and logistic regression analyses, and were retained for REGENIE whole-genome ridge regression analyses. Principal-components analysis (PCA) was used to define a European ancestry population as follows: among individuals in the unrelated set who identified as white in the self-reported ethnicity question (Field 21,000), PCA was performed using high-quality common variants using the eigenstrat algorithm[22]; variants for this initial round of PCA were filtered for missingness across individuals <2%, minor allele frequency >1%, excluding regions of long-range linkage disequilibrium (LD),[23] and independence (pairwise LD $r^2 < 0.1$). Principal components (PCs) were then projected onto the related individuals who were held out from the unrelated set, and all individuals greater than three standard deviations away from the mean of PCs 1–6 were removed as ancestry outliers. A final PC estimation was performed in eigenstrat on the remaining individuals, using the unrelated subjects for PC determination and projecting the related individuals onto these final PCs. This resulted in a set of 398,569 European ancestry individuals with exome sequencing data available.

Phenotype sources

The main source of phenotype data (from questionnaires, hospital diagnoses, cancer and death registry, operations, and quantitative traits) was a release of structured data by the UK Biobank Data Showcase on February 1, 2021. Because general practitioner (GP) data were available only in this format as International Classification of Diseases 10th Revision (ICD10) diagnoses with two numerals, to obtain more specific GP data, we downloaded individual GP records as a separate file from the UK Biobank (September 30, 2019, is the most recent update available) and translated diagnoses from Read codes to ICD10 codes at three-numeral resolution.

Participant and phenotype definition for menopause analysis

For menopause analyses, an initial set of women with exome-sequencing data was considered. Self-reported data from the touchscreen questions administered at the initial assessment center visit were used as an initial filter. Women were retained in downstream analyses only if they answered that they had not yet had menopause, or that they had had menopause and knew the age at which it had occurred. Women were excluded if their answer to the menopause question was "not sure – had a hysterectomy" or "not sure – other reason." Women were further asked whether they had ever had a bilateral oophorectomy or a hysterectomy and were excluded if they answered that they did not know. If they reported that they had had either procedure, they were excluded if they were not able to report the age at which they had had the procedure. Women were asked whether they had used hormone-replacement therapy (HRT); women were excluded if they did not know or did not answer the question, and those who reported having used HRT were excluded if they were not able to report the age at which they had first used HRT. Women were excluded if the age they reported having an oophorectomy was earlier or the same age as the age they reported menopause, if the age they reported having a hysterectomy was earlier or the same age as the age they reported menopause, and if the age they reported first using HRT was younger than the age they reported menopause. Premenopausal women were excluded if they had ever used HRT or had a bilateral oophorectomy or hysterectomy.

Additional data about potentially menopause-inducing operations were then obtained from inpatient hospital diagnoses and operation codes. Hospital diagnoses for radiotherapy, chemotherapy for cancer or chemotherapy not otherwise specified, prophylaxis for cancer, and follow-up care for those procedures, were obtained corresponding to ICD10 codes Z40.0, Z51.0, Z55.1, Z55.2, Z08.1, Z08.2, Z09.1, and Z09.2 and ICD9 code V672. Operations for chemotherapy, radiotherapy, hysterectomy, or bilateral oophorectomy were obtained corresponding to OPCS

Classification of Interventions and Procedures version 4 (OPCS4) codes X35.2, X37.3, X38.4, X70, X71, X72, X73, X74, X65, X67, Y90.2, Y91, Y92, Q07, Q08, and Q22 and OPCS Classification of Interventions and Procedures version 3 (OPCS3) codes 691, 692, 693, 694, 969, 681, and 961.1. The earliest possible calendar year of menopause was derived by adding the age at menopause to the year of birth. The earliest calendar year for any of these procedures for a woman was defined as the year of the procedure. Postmenopausal women were excluded if they had any of these procedures in the same year or earlier year than the earliest possible calendar year of menopause, and premenopausal women were excluded if they had had any of the procedures.

An additional 2,074 women were excluded from the main analysis because they had an age of menopause less than 40 years, and an additional 43 women were excluded because they reported menopause after 60 years of age or were premenopausal and over 60 when interviewed. These women were used for subsequent replication analysis of extreme phenotypes. These exclusions resulted in a set of 132,370 women for the main analysis of menopause timing (119,986 of whom were unrelated).

A more restrictive subset of menopause data was then constructed as a subset of the main analysis set ("neoplasm- and surgery-free"), which did not rely on the relative timing of potentially menopause-inducing operations, and which excluded anyone with any neoplasm preceding the time of interview. Women were excluded if they reported having had bilateral oophorectomy, hysterectomy, or physician-diagnosed cancer at the time of their initial assessment, regardless of the age at which these had occurred, and if they had any cancer or neoplasm diagnosis in the national cancer registries in the year of the interview or earlier. Women were also excluded based on procedures found in ICD10 or ICD9 hospital diagnoses and OPCS4 or OPCS3 operation codes in a calendar year the same as or earlier than the interview, using the codes listed previously and also adding the following: ICD10 code Z40.0 (prophylactic surgery for risk factors related to malignant neoplasms); OPCS4 codes corresponding to all of chapter Q (operations on the upper female reproductive tract), except for codes purely relating to diagnostic examination: Q18, Q39, Q50, or Q55; and OPCS3 codes corresponding to any operations on the female genital organs (670–739). This "neoplasm- and surgery-free" subset consisted of 100,711 women (91,318 of whom were unrelated).

TTE analysis

The main discovery analysis consisted of a TTE analysis of the association between carrier status of rare variants aggregated per gene and menopause age. For each variant set, a Cox proportional hazards model was constructed using the `survival` R package.[24] A right-censored survival object was created where the status indicator was zero for women who were premenopausal and one for women who were postmenopausal, and follow-up time was the age at interview for premenopausal (censored) women and age at menopause for postmenopausal women. Ties were handled using the default method (Efron approximation). With this survival object as the dependent variable, the independent variable in each model was the presence (coded as one) or absence (coded as zero) of at least one alternate allele from the variant set in each individual; multiple variants and zygosity were not considered for the burden test because of the inability in our unphased sequencing data to distinguish whether carriers of multiple heterozygous alleles should be treated as simple or compound heterozygotes. Covariates were year of birth (to account for secular trends in menopause age[25]) and the first 12 genetic PCs (to account for associations arising purely from population stratification). TTE was performed in the same manner for analysis of subsequent menopause assessed in the subset of women who were premenopausal at enrollment and for whom follow-up interviews were available.

TTE was also performed for individual exome and array-typed variants; these tests were performed as above, except the individual variable was the genotype of the individual variant, coded as 0 (homozygous reference), 1 (heterozygous), or 2 (homozygous alternate.) When both exome and array data were available for a variant, the exome genotype data were used; when only array data were available for a variant, the association was calculated only for the subset of exome-sequenced individuals who also had array data available.

To determine the extent to which linkage disequilibrium was driving multiple associations seen in regional patterns of single-variant association results at the *CHEK2*, *RAD54L*, and *HELB* loci, associations in the region were tested again, including the genotypes of top association signals as covariates in the regression (rs5762534 and rs555607708 at *CHEK2*, rs12142240 and rs12073998 at *RAD54L*, and rs75770066 and rs4430553 at *HELB*).

Burden analysis

Secondary burden analyses were performed for menopause timing: menopause status at interview, as a case-control trait, was tested using logistic regression with age at interview, year of birth, and the first 12 genetic PCs as covariates; and age of menopause among postmenopausal women, as a quantitative trait, was tested using linear regression with year of birth and the first 12 genetic PCs as covariates. These regressions were performed using the `glm` function in R. Genotype was coded as 0 or 1, equivalent to a dominant model. Burden analysis was also performed for two extreme phenotypes held out from the original analysis, primary ovarian insufficiency (POI) and menopause after 60 years of age; all women from the original analysis were controls for these analyses.

Whole-genome ridge regression analysis

Because of known issues with high type 1 error rate when performing association tests between rare variants and rare diagnoses, and because of potential confounding by population stratification even when excluding closely related individuals, REGENIE v.2.2.4[26] was used in addition to linear and logistic regression. REGENIE performs a whole-genome ridge regression taking subject relatedness into account and uses a Firth approximation to estimate p values. REGENIE was performed across the entire European-ancestry population, including related individuals. REGENIE was run with default settings for performing gene-based tests, except that the model type was set to "dominant," for consistency with the way in which variants were collapsed in the main regression analyses.

Pathway analysis

To test for sets of genes enriched for association with menopause, we performed a gene set enrichment analysis procedure on the complete set of p values from the TTE analysis. For each gene with variant sets in the analysis, we retained the smallest p value of the one or two variant sets tested. We obtained Reactome pathways[27] from the Molecular Signature Database[28] (access date: May 5, 2021) and matched to our genes using Entrez IDs. For 1,561 gene sets with at least five genes overlapping with genes in our TTE analysis, we performed a one-sided Mann-Whitney-Wilcoxon test of whether TTE p values

of the gene set were enriched for lower p values relative to all other genes in the TTE analysis. Because power of the TTE analysis for a gene is related to the amount of variation in the gene (both number of variants and their allele frequency spectrum), and this variation is in turn related to biological function due to varying strength of selection and varying gene length, this could lead to confounding in our analysis of p values. We therefore performed a permutation of the per-individual phenotypic data, where participant IDs were assigned to different participants' phenotype information, including year of birth, age, menopause status, and age at menopause, while retaining all original genetic data and PCs for the individual. We then performed the TTE analysis on these permuted data, followed by the gene set enrichment analysis on the resulting p values. Because genes retained their same number of variants and allele frequencies in this analysis, any gene set enrichment owing only to association of biological function with variation patterns ought to still be enriched when analyzing the permuted phenotype data.

Phenome-wide association studies (PheWASs)

For the variant sets significantly associated with menopause timing and also for variant sets canonically associated with breast cancer, phenome-wide association studies were performed between carrier status of variants in each set and 3,489 diagnosis codes and 100 quantitative traits in 398,569 individuals (363,973 of whom were unrelated). Diagnosis codes were obtained from a combination of inpatient hospital diagnoses (Field 41,270), causes of death (Field 40,001 and 40,002), the national cancer registries (Field 40,006), and GP clinical event records, which were available for a subset of individuals (Field 42,040). Diagnosis codes were tested at the ICD10 three-numeral level and also aggregated at the two-numeral level; codes were tested if the first character ranged from A to Q (Chapters 1–17; to exclude nondiagnoses such as laboratory findings and external causes) and if there were greater than 100 cases in the entire biobank. Quantitative traits were selected from a variety of phenotypes available from UK Biobank fields; specifically, a set of 100 non-redundant fields from four categories in the UK Biobank Data Showcase, physical measures, blood counts, blood biochemistry, and urine biochemistry, with at least 100,000 participants, were inverse rank normalized using the RNOmni R package.[29] Regressions were performed using burden tests using the glm function in R and followed up with REGENIE as described above, with age, sex, country of recruitment (England, Scotland, or Wales), availability of GP data, and the first 12 genetic PCs as covariates.

Two sex-stratified analyses were then performed to follow up on PheWAS results. Both analyses used a very restrictive exclusion of anyone with a record of neoplasm from any phenotype source at any time. Among women, because of the focus on following up hematological phenotypes, women who reported that they were menstruating on the day of the assessment were excluded. This resulted in 113,707 men and 76,307 women for further analysis. Regressions were performed using burden tests and REGENIE as described above, with the exception that for women, menopause status was included as a covariate.

A final association analysis was performed testing all ANM-associated variant sets with estrogen levels among the set of women described for the above sex-stratified analysis. This regression was performed using burden tests and REGENIE as described above, including menopause status as a covariate and excluding women menstruating on the day of assessment.

Mediation analysis

Mediation analysis was performed using the R package `mediation`.[30] Diagnosis dates from the cancer registry and inpatient hospitalization diagnosis records were used to limit the analysis to test only breast cancer occurring after 60 years of age as cases, and women who were breast cancer free and over 60 years of age as controls. Women currently younger than 60 years or who died before age 60 were excluded. The analysis was also limited to women who reported an ANM between ages 40 and 60 years at the initial interview, to ensure for simplicity of the model that breast cancer occurred after menopause.

Results

Gene-level associations with ANM

Among UK Biobank participants with exome sequencing data currently available, we identified a subset of European ancestry women for genetic analysis who had not experienced any treatments that would preclude natural menopause and used this subset for a range of analyses (summarized in Figure 1). Of these, 86,056 reported being postmenopausal and reported an age at menopause, while 46,314 reported being premenopausal at their initial interview (Figure S1). We defined two sets of rare exome variants to aggregate for each gene: one set consisting only of PTVs, also known as pLOF variants, and a second set including both PTVs and rare missense variants bioinformatically predicted to be deleterious. We analyzed variant sets in genes that had at least 10 carriers in the subpopulation used for association analysis (12,462 genes with enough PTV carriers and 15,462 genes with enough PTV and missense carriers). We performed three association tests for each gene: a burden test using current menopause status (pre- or postmenopausal) as a case-control trait in a logistic regression model, a burden test using ANM among postmenopausal women as a quantitative trait in a linear regression model, and a TTE analysis using both groups jointly in a Cox proportional hazards model (Manhattan plots, Figure 2; complete results, Table S1; QQ plots and λ_{GC} calculations, Figure S2). In the TTE analyses, aggregated variants in five genes surpassed a conservative threshold Bonferroni corrected for the total number of genes and variant sets tested of $p < 1.8 \times 10^{-6}$: *CHEK2* (both variant definitions), *DCLRE1A* (MIM: 609682) (both variant definitions), and *HELB* (PTVs and missense combined) with later menopause, and *TOP3A* (MIM: 601243) (PTVs and missense combined) and *CLPB* (MIM: 616254) (PTVs and missense combined) with earlier menopause. The largest effect among these associations was for *CHEK2* PTVs, associating with later menopause (Figure 3). At a more lenient threshold correcting only for the number of genes tested for each variant set definition ($p < 4.0 \times 10^{-6}$ for PTVs only and 3.2×10^{-6} for PTVs and missense), two additional genes were associated: *RAD54L* (MIM: 603615) (PTVs and missense combined) with later menopause and *HROB* (MIM: 618611) (PTVs only) with earlier menopause. All seven genes were separately associated

Figure 1. Flowchart of genome-wide analyses of menopause timing

with both ANM among postmenopausal women (burden tested as a quantitative trait with linear regression) and menopause status at the time of the interview among all women (burden tested as a case-control trait with logistic regression) (Table 1). We performed pathway enrichment analysis of the complete p values using 1,561 gene sets from the Reactome database; while no gene set reached a Bonferroni-adjusted $p < 3.2 \times 10^{-5}$, the top enrichment was for genes involved in DNA double-strand break repair (Mann-Whitney-Wilcoxon $p = 7.3 \times 10^{-4}$) (Table S2).

Confirmatory analyses

To account for potential inaccuracy in recalling the timing of hysterectomy and oophorectomy relative to menopause in the interview data, which could have resulted in some residual surgically induced menopause events in our analysis, we created a more conservative set of ANM values that excluded women with any self-reported or registry-recorded cancer or gynecological operations before the date of the questionnaire. All seven genes remained associated with menopause in this subset of women at $p < 0.05$ (Table S3).

Although our discovery analysis excluded closely related individuals, cryptic relatedness cannot be excluded as a source of population stratification. To account for relatedness, we repeated the quantitative trait and case-control analyses using the REGENIE whole-genome ridge regression algorithm,[26] which accounts for genetic relatedness and also provides better control of false discoveries when performing case-control tests on rare variants; because relatedness is explicitly modeled, we did not remove close relatives as we did for the Cox, linear, and logistic regressions. All of these associations remained significant at $p < 0.05$ by REGENIE (Table S4), with the exception of *CLPB* and the case-control menopause status trait ($p = 0.10$).

We sought additional evidence outside of the discovery dataset for these associations in two ways. First, we

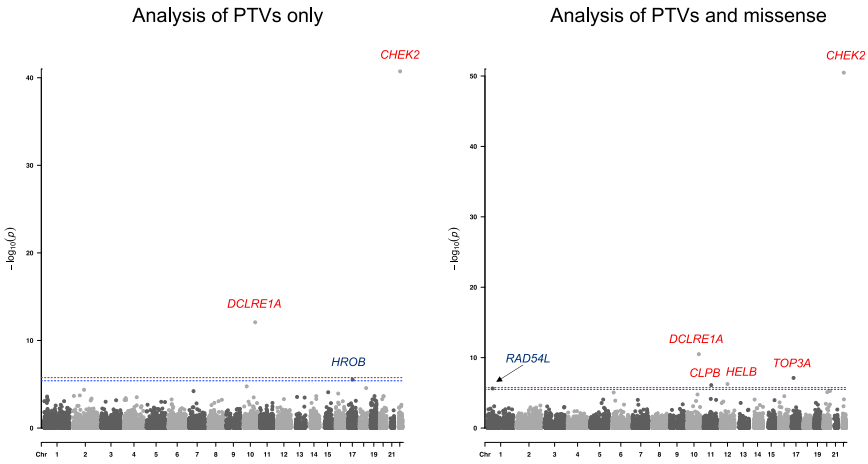

Figure 2. Manhattan plot of genome-wide scan of gene-level tests of menopause timing
p value is from TTE analysis using a Cox proportional hazards model.

performed a TTE analysis in a subset of 5,211 women who were premenopausal at the initial interview and who participated in a follow-up interview between 2 and 14 years later, 3,995 of whom were subsequently postmenopausal and had a known ANM at the time of their most recent follow-up interview. For all nine variant sets associated in the discovery analysis, the 95% confidence interval (CI) in the follow-up analysis was consistent with that of the discovery analysis. For *CHEK2*, *RAD54L*, and *TOP3A*, rare variants were associated with menopause timing in the follow-up interview at $p < 0.05$ (Table S5; Figure S3). Second, we looked at two extreme phenotypes we had excluded from the discovery analysis: 2,074 women with POI, defined as menopause before the age of 40 years, and 286 women with menopause after the age of 60 years, performing a case-control analysis of these two phenotypes with all women in the discovery ANM analysis as the controls. At a threshold of $p < 0.05$ (for both a burden test as a case-control trait with logistic regression and with REGENIE), we detected a depletion of *CHEK2* carriers and an enrichment of *TOP3A* and *CLPB* carriers in women with POI, and an enrichment of *CHEK2* and *HELB* carriers in women with menopause after 60 years of age (Table S6).

Association of individual rare and common variants with ANM

To more thoroughly evaluate genetic mechanisms of association with menopause timing at these five loci, we tested all individual coding variants contributing to these associations with at least 10 carriers (Table S7). The most significant association in *CHEK2* was rs555607708, a frameshift variant (p.Thr367MetfsTer15; commonly known as c.1100del; $p = 1.6 \times 10^{-27}$) well studied as a pathogenic variant in breast and other cancers, and the second most significant association in *CHEK2* was rs587780174, another pathogenic frameshift variant (p.Ser422ValfsTer15; commonly known as c.1263del; $p = 1.7 \times 10^{-9}$); the strongest missense association was with rs28909982 (p.Arg117Gly; $p = 1.3 \times 10^{-4}$). The most significant association in *DCLRE1A* was non-sense (premature stop) variant rs41292634 (p.Arg138Ter; $p = 7.5 \times 10^{-11}$), and the strongest missense association in *DCLRE1A* was missense variant rs11196530-A (p.Ile859Phe; $p = 1.7 \times 10^{-3}$). The most significant association in *TOP3A* was missense variant rs34001746 (p.Leu584Arg; $p = 1.6 \times 10^{-10}$), and the most significant association in *CLPB* was missense variant rs150343959 (relative to the protein produced by canonical transcript ENST00000538039.6, p.Arg598Cys; $p = 8.2 \times 10^{-6}$). The most significant association in *HELB* was missense variant rs140308412 (p.Lys615Glu; $p = 4.3 \times 10^{-5}$). The most significant association in *RAD54L* was missense variant rs28363218 (p.Arg202Cys; $p = 7.1 \times 10^{-5}$), and the most significant association in *HROB* was non-sense rs774881553-T (p.Gln341Ter; $p = 2.7 \times 10^{-3}$). We also re-ran the tests of aggregated variants with a leave-one-out approach to test the extent to which these individual variants contributed to the gene-level associations. For all genes except *TOP3A*, the gene-level association remained significant when holding out the strongest single-variant associations; for *TOP3A*, missense rs34001746 appeared to explain the entire association ($p = 0.50$ for all other missense and PTVs; Table S7).

We then broadened our analysis to 30,571 array-genotyped or exome-sequenced variants within each of the seven gene bodies and 500 kb flanking, focusing on associations that would be considered significant in a GWAS ($p < 5 \times 10^{-8}$; Table S8; Figures S4–S10). No additional variants at *DCLRE1A* or *HROB* were associated with menopause timing. At *TOP3A*, four noncoding variants in addition to missense rs34001746 were associated with similar significance and were in strong linkage disequilibrium (Figure S6). At *CLPB*, one intronic variant rs186195729 had a similar association to missense variant rs150343959 and was in strong linkage disequilibrium (Figure S10). At *RAD54L*, *CHEK2*, and *HELB*, many common and rare variants were significantly associated, consistent with a previous GWAS of ANM by the ReproGen consortium,[11] which identified noncoding associations at all three of these loci and additionally identified coding variants associated at *HELB*. We explored the independence of the associations with the multiple variants at these loci by building TTE models with genotypes of multiple variants as dependent variables.

At *RAD54L*, conditional analysis suggested that the common GWAS variants rs12142240 (reported by ReproGen) and rs12073998 (the strongest common-variant association we found, which is in high linkage disequilibrium

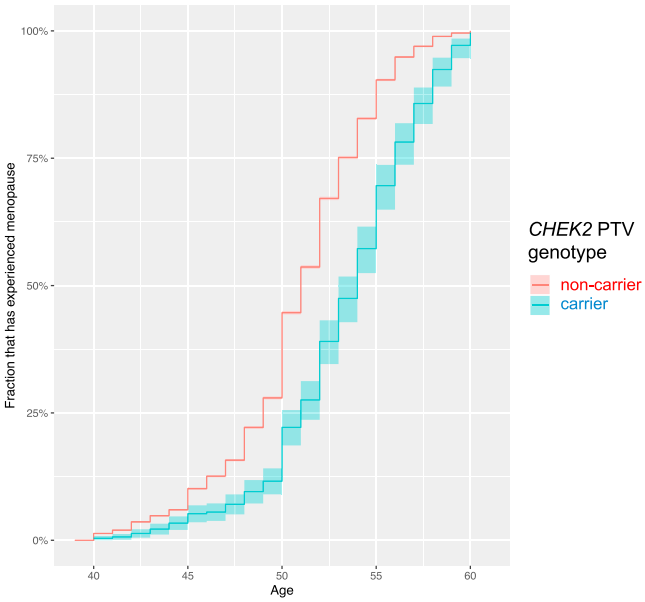

Figure 3. Cumulative incidence of menopause among *CHEK2* PTV carriers and non-carriers between the ages of 40 and 60 years. Shaded area is the 95% confidence interval from a survival model fitted to the data.

with rs12142240) were in high linkage disequilibrium with all of the other genome-wide significant associations at the locus, and rare missense rs28363218 remained nominally significant at $p = 2.6 \times 10^{-3}$, suggesting its independence (Table S9). At *CHEK2*, conditioning on the genotype of the common GWAS variant rs5762534 (reported by ReproGen) reduced the number of additional genome-wide significant SNP associations from 34 to 7, while improving the significance of the association with rare frameshift rs555607708 (Table S10). Conditioning on both rs5762534 and rs555607708, other common and rare variants remain genome-wide significant, suggesting many independent association signals at *CHEK2*. At *HELB*, conditioning on the genotype of the strongest common GWAS association (rs75770066, reported by ReproGen) increased the significance of the association of a common variant haplotype spanning the locus, which includes missense variant rs4430553 (p.Leu191Pro); conditioning on both of these variants, the strongest rare coding association with rs140308412 remained significant ($p = 1.0 \times 10^{-4}$; Table S11), confirming many independent associations at *HELB* as described previously.[11]

Other associations with menopause timing-associated genes

To better understand the full spectrum of consequences of rare coding variants in these seven genes, we performed PheWASs for association with 3,489 diagnosis codes and 100 quantitative traits across 398,569 individuals using burden tests and REGENIE (summarized in Figure 4). Although *DLCRE1A*, *RAD54L*, *TOP3A*, *HELB*, *CLPB*, and *HROB* had no associations reaching phenome-wide significance (for the number of variant sets and phenotypes tested, REGENIE $p < 1.5 \times 10^{-6}$), *CHEK2* was associated with many diagnoses and traits (Tables S12 and 2).

Besides breast cancer and related diagnoses, rare damaging variants in *CHEK2* were also associated with prostate cancer, which has been previously reported.[31] *CHEK2* variants were also associated with myeloid leukemia, consistent with reports that *CHEK2* coding variants were associated with myelodysplastic syndrome.[32] We also found associations with benign neoplasms: carcinoma *in situ* of the breast and leiomyoma of the uterus (uterine fibroids). In addition to these neoplasms, we also detected association between *CHEK2* variants and diagnosis of polycystic ovary syndrome (PCOS), prostate hyperplasia, and seborrheic keratosis. Among quantitative traits, *CHEK2* variants are associated with many hematological parameters, most strongly with increased lymphocyte count and increased platelet crit. Multiple hematological associations with *CHEK2* were noted in a recent study of 49,960 exome-sequenced UK Biobank participants, which the authors noted could be secondary to cancer treatment.[15] To test whether these associations were shared with other breast cancer genes, we performed PheWASs on rare damaging variants in four other moderate- to high-penetrance breast cancer genes (*BRCA2* [MIM: 600185], *BRCA1*, *PALB2* [MIM: 610355], and *ATM* [MIM: 607585])[33] and found that the hematological associations and non-breast cancer disease associations, with the exception of carcinoma *in situ* of the breast and prostate cancer, were specific to *CHEK2* (Tables 3 and S13; Figure 5).

We further tested the non-neoplasm associations in two sex-stratified sub-analyses: an analysis of men who have never been diagnosed with a neoplasm, and an analysis of women from the original menopause analysis who have never been diagnosed with a neoplasm, which also conditioned on menopause status as a covariate and excluded women menstruating on the day of the assessment. We tested 14 hematological traits that were associated at $p < 1.5 \times 10^{-6}$ with either of the two *CHEK2* variant sets and tested them in each subset against both variant sets; in men, 23/28 tests remained significant at $p < 0.05$ by REGENIE, and in women, 24/28 remained significant. We also tested two disease diagnosis associations in each subset. Prostate hyperplasia remained associated with *CHEK2* among neoplasm-free men and PCOS remained associated with *CHEK2* among neoplasm-free women. Seborrheic keratosis remained associated with *CHEK2* in the analysis of women, but not in the subset of men. These associations in the sex-stratified, neoplasm-free analysis suggest that the original associations were not secondary to cancer treatment or diagnosed neoplasms, and that the hematological associations were not related to menopause status, menstruation, or sex (Tables S14 and S15). Because of the association between estrogen exposure and menopause timing, we also tested all the ANM-associated variant sets against estradiol levels, performing this analysis in the neoplasm-free subset of women, excluding women menstruating the day of assessment, and including

Table 1. Results for nine variant sets significantly associated with menopause timing

Gene	Variant set	n carriers total (n_{unrel})	n carriers postmenopause (n_{unrel})	Model	Phenotype	Statistic	95% CI	p value
CHEK2	PTV and damaging missense	1,820 (1,667)	1,059 (976)	Cox	TTE	HR = 0.62	[0.58, 0.66]	3.32×10^{-51}
				linear	ANM	1.49 y	[1.24, 1.73]	1.54×10^{-32}
				logistic	meno. status	OR = 0.37	[0.30, 0.45]	2.64×10^{-20}
CHEK2	PTV only	835 (765)	467 (431)	Cox	TTE	HR = 0.52	[0.47, 0.57]	1.87×10^{-41}
				linear	ANM	2.22 y	[1.86, 2.59]	2.04×10^{-32}
				logistic	meno. status	OR = 0.26	[0.19, 0.36]	7.02×10^{-17}
DCLRE1A	PTV only	1,327 (1,213)	828 (769)	Cox	TTE	HR = 0.77	[0.72, 0.83]	8.42×10^{-13}
				linear	ANM	0.61 y	[0.34, 0.89]	1.34×10^{-5}
				logistic	meno. status	OR = 0.53	[0.41, 0.68]	7.51×10^{-7}
DCLRE1A	PTV and damaging missense	3,164 (2,894)	2,004 (1,852)	Cox	TTE	HR = 0.86	[0.82, 0.90]	3.18×10^{-11}
				linear	ANM	0.35 y	[0.17, 0.53]	1.38×10^{-4}
				logistic	meno. status	OR = 0.69	[0.59, 0.81]	5.32×10^{-6}
TOP3A	PTV and damaging missense	2,473 (2,239)	1,605 (1,452)	Cox	TTE	HR = 1.15	[1.09, 1.21]	7.59×10^{-8}
				linear	ANM	−0.45 y	[-0.65, −0.25]	1.14×10^{-5}
				logistic	meno. status	OR = 1.26	[1.05, 1.51]	1.10×10^{-2}
HELB	PTV and damaging missense	1,099 (1,003)	675 (618)	Cox	TTE	HR = 0.82	[0.75, 0.88]	5.68×10^{-7}
				linear	ANM	0.63 y	[0.32, 0.93]	6.52×10^{-5}
				logistic	meno. status	OR = 0.49	[0.37, 0.65]	4.60×10^{-7}
CLPB	PTV and damaging missense	1,536 (1,394)	1,039 (950)	Cox	TTE	HR = 1.17	[1.10, 1.25]	8.05×10^{-7}
				linear	ANM	−0.50 y	[-0.75, −0.25]	8.84×10^{-5}
				logistic	meno. status	OR = 1.34	[1.06, 1.68]	1.34×10^{-2}
RAD54L	PTV and damaging missense	2,107 (1,908)	1,353 (1,222)	Cox	TTE	HR = 0.87	[0.82, 0.92]	2.37×10^{-6}
				linear	ANM	0.35 y	[0.13, 0.57]	1.92×10^{-3}
				logistic	meno. status	OR = 0.70	[0.58, 0.86]	4.98×10^{-4}
HROB	PTV only	92 (86)	70 (66)	Cox	TTE	HR = 1.78	[1.40, 2.27]	2.91×10^{-6}
				linear	ANM	−2.27 y	[-3.20, −1.33]	2.15×10^{-6}
				logistic	meno. status	OR = 3.97	[1.40, 11.25]	9.34×10^{-3}

Three tests are shown for each variant set: (1) time-to-event (TTE) analysis using Cox regression; resulting hazard ratio (HR) greater than 1 means carriers experienced earlier menopause (meno.) and less than 1 means carriers experienced later menopause; number of events analyzed is N carriers postmenopause; (2) quantitative trait analysis of age of natural menopause (ANM) among postmenopausal women, using linear regression; resulting statistic is the modeled association of carrier status with ANM; and (3) case-control analysis of menopause status at interview (pre- versus postmenopausal), using logistic regression; resulting statistic is the odds ratio (OR), where values greater than 1 mean carriers were more likely to be postmenopausal at time of interview, and values less than 1 mean carriers were more likely to be premenopausal at time of interview.

menopause status as a covariate in the linear model. No association was detected between estradiol levels and any of the ANM-associated variant sets (Table S16).

Mediation analysis of *CHEK2* effects on menopause and breast cancer

The association of *CHEK2* variants with both breast cancer and delayed menopause raised the question of the extent to which this pleiotropy was mediated by the well-established association between delayed menopause and breast cancer. To distinguish between biological and mediated pleiotropy, we performed causal mediation analysis. For simplicity of the model, we used only the subset of women who were postmenopausal at the time of recruitment and who reported an ANM. To further ensure that breast cancer followed menopause and that genetic effects on ANM did not confound the association between the age covariate and follow-up time, we considered only breast cancer cases versus breast cancer-free controls after the age of 60 years. In this subset, we detected significant associations of *CHEK2* genotype with ANM, *CHEK2* genotype with breast cancer, and ANM with breast cancer (while controlling for *CHEK2* genotype), suggesting both a direct effect and an indirect effect (via menopause delay) of *CHEK2* genotype on breast cancer. Using PTVs only, which have the largest estimated effect on breast cancer, we tested the

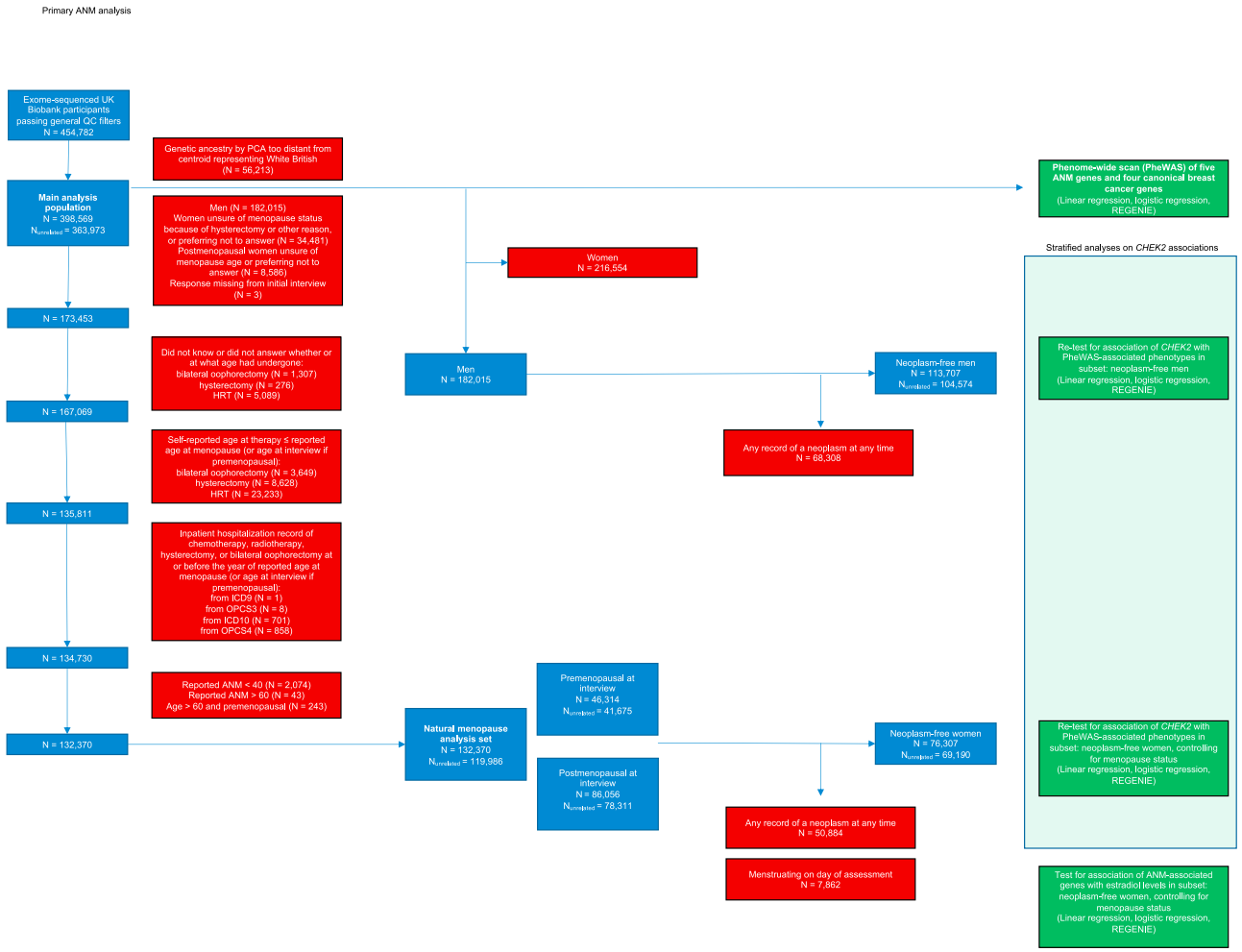

Figure 4. Flowchart of phenome-wide analyses of ANM-associated and breast-cancer-associated genes

significance of the unstandardized indirect effect using 1,000 bootstrapped samples and estimated the bootstrapped indirect effect as 7.8% of the total effect (95% CI: 4.5%–16.5%; $p < 2 \times 10^{-16}$; Figure 6). The proportion was similar when considering both PTVs and missense variants (7.1% mediated; 95% CI: 4.2%–14.1%; $p < 2 \times 10^{-16}$; Figure S11).

Discussion

Genetic architecture of menopause timing

Deciphering the genetic control of menopause is important for understanding the causal relationship between menopause and associated disease risks. Association studies have revealed many loci influencing menopause timing, but the majority of association signals has consisted of common noncoding variants. In addition, these studies have looked at only ANM in postmenopausal women, discarding potentially useful information from the dichotomous trait of whether women are post- or premenopausal at a given age. To increase power and improve causal gene discovery, we performed TTE analysis including both pre- and postmenopausal women and used a large cohort of exome sequences from the UK Biobank to interrogate rare damaging variants. The largest GWAS of menopause timing currently in the GWAS catalog, from the ReproGen consortium, identified 44 associated loci, 29 of which included at least one gene involved in DNA damage repair.[11] However, common coding variants identified causal genes unambiguously at only five loci: *EXO1* (MIM: 606063), *REV3L* (MIM: 602776), *HELB*, *BRCA1*, and *MCM8*, all of which encode genes involved in DNA damage repair. Other associated haplotypes harbored coding variants in *PRIM1* (MIM: 176635), *FANCI* (MIM: 611360), *GSPT1* (MIM: 139259), *NLRP11* (MIM: 609664), and *SLCO4A1* (MIM: 612436), but there were additional potential causal genes at these loci; of these, two are involved in DNA damage repair. The largest GWAS of menopause timing in non-European ancestry women, performed in Japanese women, identified additional associated loci with coding variants in *GNRH1* (MIM: 152760), *ZCCHC2*, and *ZNF518A* (MIM: 617733).[7] Our analysis identified rare coding variant associations with menopause timing in six additional genes: *CHEK2*, *TOP3A*, *DCLRE1A*, *RAD54L*, *HROB*, and *CLPB*. A recent larger GWAS from the ReproGen consortium identified coding associations in *CHEK2*, but not the other five genes.[34]

Table 2. Selected associations of *CHEK2* damaging variants with diagnoses and quantitative traits

Phenotype	p	Effect	N
Platelet crit	2.12×10^{-49}	beta = 0.18 SD [0.16, 0.21]	5,311 carriers measured
Leukocyte count	1.28×10^{-33}	beta = 0.16 SD [0.13, 0.19]	5,311 carriers measured
C50: malignant neoplasm of breast	1.42×10^{-28}	OR = 1.93 [1.74, 2.17]	393 cases/221 expected
Neutrophil count	4.08×10^{-23}	beta = 0.13 SD [0.11, 0.16]	5,304 carriers measured
Mean corpuscular hemoglobin	5.55×10^{-23}	beta = −0.12 SD [−0.15, −0.10]	5,311 carriers measured
Mean corpuscular volume	1.19×10^{-20}	beta = −0.12 SD [−0.14, −0.09]	5,311 carriers measured
C61: malignant neoplasm of prostate	1.26×10^{-13}	OR = 1.72 [1.50, 1.98]	256 cases/162 expected
D05: carcinoma *in situ* of breast	6.20×10^{-13}	OR = 2.17 [1.80, 2.64]	114 cases/54 expected
D25: leiomyoma of uterus	1.49×10^{-9}	OR = 1.48 [1.31, 1.68]	299 cases/206 expected
N40: hyperplasia of prostate	5.15×10^{-9}	OR = 1.41 [1.26, 1.59]	408 cases/314 expected
Erythrocyte count	3.47×10^{-8}	beta = 0.06 SD [0.04, 0.08]	5,311 carriers measured
C92: myeloid leukemia	1.24×10^{-7}	OR = 3.02 [2.13, 4.45]	32 cases/11 expected
L82: seborrheic keratosis	3.42×10^{-7}	OR = 1.31 [1.19, 1.46]	458 cases/358 expected
E282: polycystic ovarian syndrome	9.05×10^{-7}	OR = 2.92 [2.02, 4.30]	30 cases/10 expected

Associations shown are for *CHEK2* PTVs and missense variants aggregated; all associations are also significant when considering *CHEK2* PTVs only. Cases observed are counted among the carriers; the expected value of case carriers is based on the prevalence of the disease and carrier frequency. p value and effect were obtained from REGENIE.

Five of these six genes are involved with various aspects of DNA damage repair.

The strong link between DNA damage repair genes and menopause timing is not surprising given the role of DNA damage repair and surveillance at every stage of the development of oocytes.[11,35] The size of the initial oocyte pool at birth, along with the rate of atresia of oocytes through life, influences the age at which the oocyte pool is depleted to a number low enough to trigger amenorrhea (approximately 1,000 remaining oocytes). The meiosis that occurs in oocytes necessitates programmed double-stranded breaks (DSBs) that must be repaired through the homologous recombination pathway; oocytes that do not properly repair DSBs after this first phase of meiosis undergo apoptosis. Oocytes rest for decades in an arrested meiotic state until ovulation, during which they accumulate additional DNA damage from exogenous insults; the mouse ortholog of *CHEK2* has been shown to cause

Table 3. Testing *CHEK2*-associated phenotypes with four other breast cancer genes

Phenotype	Associated breast cancer genes
Platelet crit	*CHEK2*
Leukocyte count	*CHEK2*
C50: malignant neoplasm of breast	*BRCA2*,[a] *BRCA1*,[a] *PALB2*,[a] *CHEK2*, *ATM*
Mean corpuscular hemoglobin	*CHEK2*
Lymphocyte count	*CHEK2*
Neutrophil count	*CHEK2*
D05: carcinoma *in situ* of breast	*BRCA2*, *PALB2*, *CHEK2*, *ATM*
C61: malignant neoplasm of prostate	*BRCA2*, *CHEK2*, *ATM*
D25: leiomyoma of uterus	*CHEK2*
Erythrocyte count	*CHEK2*
E28.2: polycystic ovarian syndrome	*CHEK2*
N40: hyperplasia of prostate	*CHEK2*
C92: myeloid leukemia	*CHEK2*
L82: seborrheic keratosis	*CHEK2*

For uniformity of comparison, for all genes, PTVs only were aggregated and tested. Genes were listed if the association was $p < 0.05/56$.
[a]The 95% confidence interval of the estimated effect is greater and nonoverlapping with the corresponding interval for *CHEK2*.

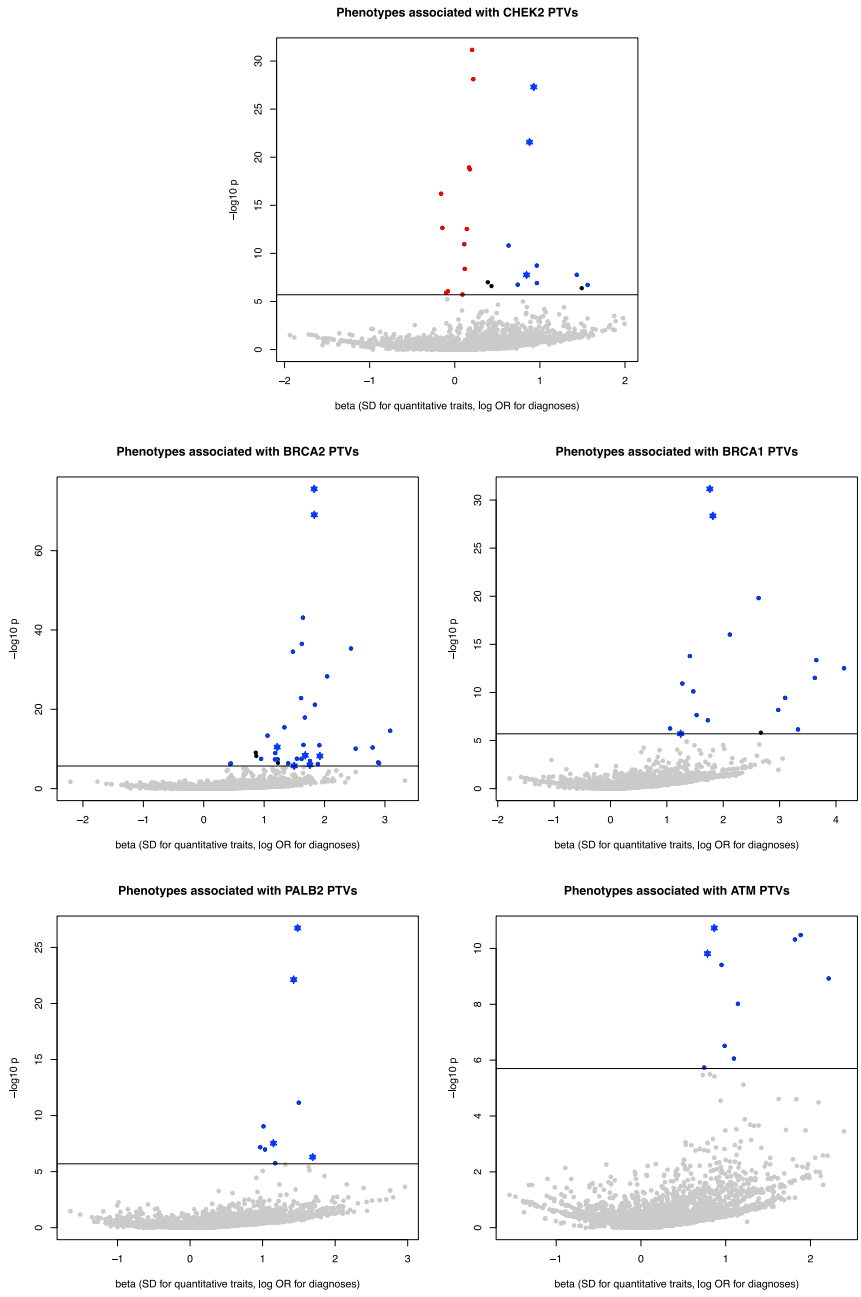

Figure 5. Results of phenome-wide analyses of PTVs in *CHEK2* and four other canonical breast cancer genes

Volcano plots for each gene show beta and p value from REGENIE. Colored in red are hematological quantitative traits, colored in blue are cancer and neoplasms (ICD10 codes beginning with letter C or D), and colored in black are other traits.

oocytes to undergo apoptosis when they have experienced damage from ionizing radiation.[35,36]

MR analysis of menopause-delaying alleles supports a causal role of these variants in breast cancer risk, mediated through prolonged exposure to endogenous estrogen.[11] The plausibility of this causal mechanism is strongly supported by the fact that HRT is a risk factor for breast cancer.[37] A recent report identified common-variant associations at *CHEK2* and *BRCA2* with ANM and followed up these associations by testing for an association between *CHEK2* and *BRCA2* pathogenic variants and menopause timing among 45,351 exome-sequenced women; they found associations with later menopause ($p = 6.8 \times 10^{-5}$) and earlier menopause ($p = 8 \times 10^{-13}$), respectively.[34] Although we find an association between *BRCA2* PTVs and earlier menopause in our primary analysis (hazard ratio [HR] = 1.27; 95% CI: 1.12–1.44; $p = 2.8 \times 10^{-4}$), this association is not genome-wide significant and is no longer nominally significant when excluding any women with a history of neoplasms or gynecological surgery (HR = 1.16; 95% CI: 0.99–1.37; p = 0.07), suggesting that either the primary association could be confounded by prophylactic or therapeutic surgery in *BRCA2* carriers or that in this subgroup there is still a bona fide association with *BRCA2* that we are underpowered to detect. The association we observe between *CHEK2* rare damaging variants and later menopause timing is strong and genome-wide significant even at the level of individual variants and is seen in additional data not used in the discovery analysis. Through causal mediation analysis, we show for the first time that although the predominant effect of *CHEK2* pathogenic variants is directly on breast cancer risk, that risk would be slightly less if it were not for their menopause-delaying effect.

Novel *CHEK2* biology

In addition to identifying associations between rare damaging variants in *CHEK2* with menopause timing and replicating known associations with breast cancer,[33] prostate cancer,[31] and myeloid leukemia,[32] we find associations with many hematological measurements, such as increased leukocyte counts and platelet crit, and associations with several diagnoses: PCOS, uterine fibroids, prostate hyperplasia, and seborrheic keratosis. These hematological associations were first reported in an earlier analysis of the UK Biobank exome data and were hypothesized to be secondary to cancer treatment[15]; however, we find that they exist in individuals with no diagnosed neoplasm. They are unlikely to be secondary to menopause timing, because they persist in both men and women, and they are seen among women even when correcting for menopause status at the time of

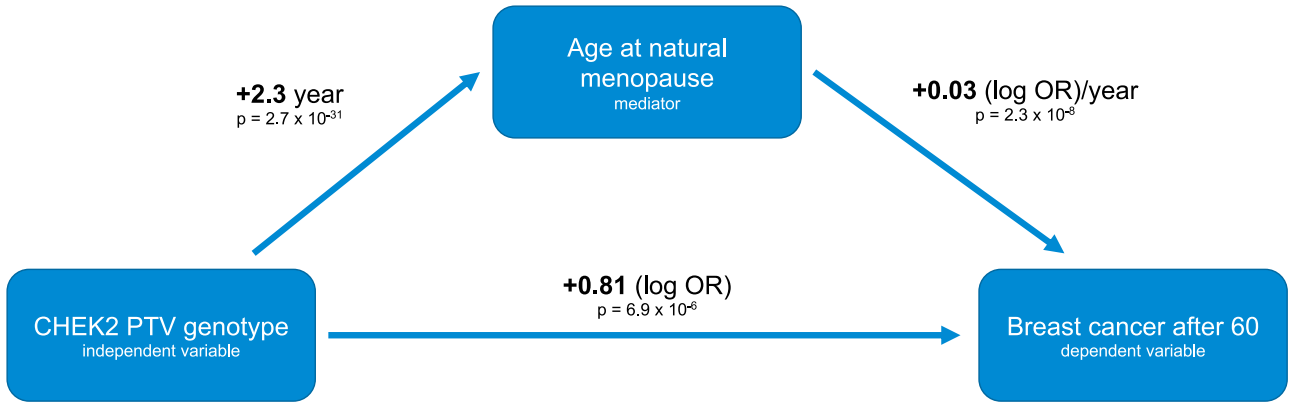

Figure 6. Mediation analysis of the proportion of *CHEK2* PTV's breast cancer effect mediated by delaying ANM
The bottom arrow shows the modeled total effect of *CHEK2* genotype increasing risk for breast cancer after 60 years of age (from a logistic regression model). The upper left arrow shows the modeled amount by which *CHEK2* delays ANM (using linear regression among postmenopausal women). The upper right arrow shows the modeled effect of delaying ANM by 1 year on risk for breast cancer after 60 years of age (using logistic regression).

the blood assay. It is more likely, therefore, that these hematological changes arise through the same mechanisms that predispose to myelodysplastic syndrome in *CHEK2* mutation carriers.[32] After *TP53*, *CHEK2* was identified as a second causal gene for Li-Fraumeni syndrome (MIM: 609265), which involves risk for many cancers.[38,39] A common feature of the non-cancer diseases we find associated with *CHEK2* is that they all involve benign hyperproliferation of tissue. Although *CHEK2* acts as a brake on proliferation in response to DNA damage, it may serve as a more general negative regulator of mitosis in somatic cells even in the absence of DNA damage.

Genetics and molecular biology of other ANM-associated genes

HELB (helicase, DNA, B) encodes a helicase enzyme essential for DNA replication.[40] Both common and rare variants at *HELB* were found to associate with menopause timing in the ReproGen study,[11] and GWASs of hematological traits have associated common variants at the locus with basophil count.[41,42]

DCLRE1A (DNA crosslink repair protein 1a, also known as the homolog of *S. cerevisiae* SNM1) encodes a protein that repairs inter-strand crosslinks[43]; it has not been implicated in any Mendelian syndromes or GWASs of complex traits. Non-sense variants such as rs41292634 and missense variants such as rs11196530 both contribute to the ANM association we identify with *DCLRE1A*, which was not detected in prior GWASs of ANM. Although rs41292634 has been reported as a candidate SNP in *BRCA1*- and *BRCA2*-negative cancer families,[44] *DCLRE1A* has not been detected in larger exome-sequencing studies.

TOP3A (DNA topoisomerase III alpha) encodes a protein critical to DNA synthesis by dissolving double Holliday junctions after homologous recombination-mediated repair of double-stranded DNA breaks.[45,46] It has been implicated in a recessive developmental disorder (MGRISCE2, related to Bloom syndrome [MIM: 618097]) characterized by short stature and microcephaly. We find an ANM association at *TOP3A* that is entirely explained by a rare missense variant rs34001746, reported as likely benign in ClinVar; this association is the strongest rare variant association we detect with earlier menopause. Consistent with the direction we detect, *TOP3A* rs34001746 carriers are enriched in cases of POI not included in our discovery analysis, and carriers who were premenopausal at their initial interview were more likely than non-carriers to be postmenopausal at follow-up interviews.

CLPB (caseinolytic mitochondrial matrix peptidase chaperone subunit B) encodes a mitochondrial chaperone whose exact biological role is unknown; biallelic and some monoallelic coding mutations cause the syndrome methylglutaconic aciduria type VII (MGCA7) (MIM: 616271), characterized by neutropenia and disordered neurological development.[47–49] Common noncoding variants at *CLPB* have been associated by GWASs with a variety of phenotypes: height, scoliosis, blood pressure, and electrocardiographic traits, none of which are related to the manifestations of MGCA7 or related to menopause timing.[50–53] We see no association between the ANM-associated set of damaging missense and PTV variants and neutrophil count at $p < 0.05$, and no associations besides menopause timing are phenome-wide significant. *CLBP* is the gene with the least prior knowledge tying the gene to menopause or the established menopause-linked process of DNA repair; however, its enrichment in women with POI increases the likelihood that the ANM association we observe is not due to chance.

The remaining loci, *RAD54L* and *HROB*, are genome-wide significant in the individual genome-wide scans but do not surpass a strict Bonferroni-corrected threshold for having performed two genome-wide scans. Whether such a conservative threshold is justified is debatable given that variants were shared between the two scans, and they were therefore far from independent tests. We chose to follow up and report on these results in light of prior biological evidence, while noting that replication in a

different cohort would be critical before considering these associations high confidence.

RAD54L (Rad54-like, named after its *S. cerevisiae* ortholog) encodes a protein involved in homologous recombination that binds tightly to Holliday junctions.[54] Although *RAD54L* has been proposed as a gene altered in the germline in breast cancer,[55,56] it has not been associated in modern exome-wide studies of breast cancer[33] and consistent with this, we do not observe an association with breast cancer in the UK Biobank. The ReproGen GWAS of ANM identified a noncoding association at the *RAD54L* locus; we identify an independent rare variant association of later ANM with missense rs28363218, which has not been reported as pathogenic to ClinVar. We find that the association of *RAD54L* with later ANM is confirmed by follow-up interviews of initially premenopausal women.

HROB (homologous recombination factor with OB-fold, recently renamed from *C17orf53*) encodes a protein that is involved in homologous recombination by recruiting the MCM8-MCM9 helicase after DNA damage.[57] HROB-deficient mice are infertile, with ovaries lacking follicles in females and defective sperm production in males. Missense variants in *MCM8* have been previously identified in GWASs of ANM, and non-sense variants in *MCM9* have been implicated in POI (MIM: 616185).[58] Although *HROB* PTVs are exceedingly rare in our study (only 86 carriers), precluding power to replicate in our subgroup analyses, evidence from mouse phenotypes and the human genetics of the epistatic *MCM8* and *MCM9* genes make the association with *HROB* highly biologically plausible.

Limitations

The problems inherent in studies on self-reported recollection of age at menopause have been previously reported.[1,59] The accuracy of recall of age at menopause is known to decrease with duration between menopause and the interview. A bias toward ages divisible by five (40, 45, 50) is clear from the distribution of ages and suggests low accuracy of individual reports of ANM. Oral contraceptive use may mask the onset of menopause but is so widespread that excluding oral contraceptive users was infeasible for our study.

Our exclusion of individuals with a history of neoplasm or gynecological surgery relies on the hospital records, cancer registry, and questionnaire data. Because hospitalization data extend only a limited time in the past and because recollection of medical history is imperfect, some people with prior cancer or surgery may have been included in this subset, allowing for possible confounding.

To minimize artifacts from population stratification, we limited our analysis to a predominantly European ancestry sample, and other ancestry groups were too small in the UK Biobank to perform sufficiently powered analyses of rare variants. More genetically diverse studies are needed to harness the power of variation outside of European ancestry populations. In addition, although it was not an explicit inclusion criterion, our study is likely to be heavily enriched for cisgender women, rather than being inclusive of all people who menstruate and experience menopause; genetic quality control filters require concordance between genetic and reported sex and exclude aneuploidies, and participant exclusions encompass HRT and gynecological surgeries.

Except for the most common pathogenic *CHEK2* variants that have been characterized in the context of breast cancer, the other variants we identify as associating with ANM remain to be tested functionally to validate bioinformatic predictions. Experimental work is needed to verify that these variants indeed disrupt protein function (in the case of missense variants) or result in non-sense-mediated decay of the principal message (in the case of PTVs).

Although we have attempted to use follow-up interviews and extreme phenotypes to obtain additional evidence to confirm the discoveries in data that were held out from the initial analysis, a true replication would involve study of a separate cohort that has both menopause information and rare variant information from exome or whole-genome sequencing. Such independent replication will be critical to confirming the weaker evidence seen for associations with *RAD54L* and *HROB*.

Conclusions

Our study of rare coding variants confirms findings from previous GWASs highlighting a key role of DNA damage repair proteins in genetic determination of menopause timing. In addition to identifying coding variant associations at the *CHEK2*, *HELB*, and *RAD54L* loci previously identified by GWAS, we identify novel associations at *TOP3A*, *DCLRE1A*, *CLPB*, and *HROB* and confirm that menopause timing is the sole phenome-wide significant association for rare variants in these genes. *CHEK2* also appears to be highly pleiotropic beyond its known role in breast cancer and other cancer syndromes, affecting hematological traits, as well as conferring risk for non-cancer disorders involving hyperproliferation of tissue, such as PCOS, prostate hyperplasia, and seborrheic keratosis.

Data and code availability

Individual-level data analyzed in this study are not public due to privacy restrictions but are available to qualified investigators by application to the UK Biobank. All data generated by this analysis are contained in the supplemental data.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2021.100079.

Acknowledgments

This research has been conducted using the UK Biobank resource, application number 26041. We thank the UK Biobank participants for their donations to this resource. Data management and analytics were performed using the REVEAL/SciDB translational analytics platform from Paradigm4.

Declaration of interests

All of the authors are employees and shareholders of Alnylam Pharmaceuticals.

Received: April 16, 2021
Accepted: December 8, 2021

References

1. Gold, E.B. (2011). The timing of the age at which natural menopause occurs. Obstet Gynecol Clin North Am *38*, 425–440.
2. te Velde, E.R., and Pearson, P.L. (2002). The variability of female reproductive ageing. Hum Reprod Update *8*, 141–154.
3. Murabito, J.M., Yang, Q., Fox, C., Wilson, P.W., and Cupples, L.A. (2005). Heritability of age at natural menopause in the framingham heart study. J Clin Endocrinol Metab *90*, 3427–3430.
4. Snieder, H., MacGregor, A.J., and Spector, T.D. (1998). Genes control the cessation of a woman's reproductive life: a twin study of hysterectomy and age at menopause. J Clin Endocrinol Metab *83*, 1875–1880.
5. He, C., Kraft, P., Chen, C., Buring, J.E., Pare, G., Hankinson, S.E., Chanock, S.J., Ridker, P.M., Hunter, D.J., and Chasman, D.I. (2009). Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. Nat Genet *41*, 724–728.
6. Lunetta, K.L., D'Agostino, R.B., Sr, , ., Karasik, D., Benjamin, E.J., Guo, C.Y., Govindaraju, R., Kiel, D.P., Kelly-Hayes, M., Massaro, J.M., et al. (2007). Genetic correlates of longevity and selected age-related phenotypes: a genome-wide association study in the Framingham Study. BMC Med Genet *8* (*Suppl 1*), S13.
7. Horikoshi, M., Day, F.R., Akiyama, M., Hirata, M., Kamatani, Y., Matsuda, K., Ishigaki, K., Kanai, M., Wright, H., Toro, C.A., et al. (2018). Elucidating the genetic architecture of reproductive ageing in the Japanese population. Nat Commun *9*, 1977.
8. Stolk, L., Zhai, G., van Meurs, J.B., Verbiest, M.M., Visser, J.A., Estrada, K., Rivadeneira, F., Williams, F.M., Cherkas, L., Deloukas, P., et al. (2009). Loci at chromosomes 13, 19 and 20 influence age at natural menopause. Nat Genet *41*, 645–647.
9. Hartge, P. (2009). Genetics of reproductive lifespan. Nat Genet *41*, 637–638.
10. Perry, J.R., Corre, T., Esko, T., Chasman, D.I., Fischer, K., Franceschini, N., He, C., Kutalik, Z., Mangino, M., Rose, L.M., et al. (2013). A genome-wide association study of early menopause and the combined impact of identified variants. Hum Mol Genet *22*, 1465–1472.
11. Day, F.R., Ruth, K.S., Thompson, D.J., Lunetta, K.L., Pervjakova, N., Chasman, D.I., Stolk, L., Finucane, H.K., Sulem, P., Bulik-Sullivan, B., et al. (2015). Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. Nat Genet *47*, 1294–1303.
12. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med *12*, e1001779.
13. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. bioRxiv, 166298.
14. Szustakowski, J.D., Balasubramanian, S., Kvikstad, E., Khalid, S., Bronson, P.G., Sasson, A., Wong, E., Liu, D., Wade Davis, J., Haefliger, C., et al. (2021). Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat Genet *53*, 942–948.
15. Van Hout, C.V., Tachmazidou, I., Backman, J.D., Hoffman, J.D., Liu, D., Pandey, A.K., Gonzaga-Jauregui, C., Khalid, S., Ye, B., Banerjee, N., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature *586*, 749–756.
16. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol *17*, 122.
17. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alfoldi, 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.
18. Liu, X., White, S., Peng, B., Johnson, A.D., Brody, J.A., Li, A.H., Huang, Z., Carroll, A., Wei, P., Gibbs, R., et al. (2016). WGSA: an annotation pipeline for human genome sequencing studies. J Med Genet *53*, 111–112.
19. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., and Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res *47*, D886–D894.
20. Tweedie, S., Braschi, B., Gray, K., Jones, T.E.M., Seal, R.L., Yates, B., and Bruford, E.A. (2021). Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Res *49*, D939–D946.
21. Staples, J., Qiao, D., Cho, M.H., Silverman, E.K., University of Washington Center for Mendelian Genomics, Nickerson, D.A., and Below, J.E. (2014). PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. Am J Hum Genet *95*, 553–564.
22. Wang, L., Zhang, W., and Li, Q. (2020). AssocTests: an R package for genetic association studies. J Stat Soft *94*, 1–26.
23. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al. (2008). Long-range LD can confound genome scans in admixed populations. Am J Hum Genet *83*, 132–135, author reply 135–9.
24. Therneau, T.M. (2021). A Package for Survival Analysis in R. https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf.
25. Rodstrom, K., Bengtsson, C., Milsom, I., Lissner, L., Sundh, V., and Bjourkelund, C. (2003). Evidence for a secular trend in menopausal age: a population study of women in Gothenburg. Menopause *10*, 538–543.
26. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet *53*, 1097–1103.

27. Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res *48*, D498–D503.
28. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545–15550.
29. McCaw, Z.R., Lane, J.M., Saxena, R., Redline, S., and Lin, X. (2019). Operating Characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics *76*, 1262–1272.
30. Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. Psychol Methods *15*, 309–334.
31. Cybulski, C., Gorski, B., Huzarski, T., Masojc, B., Mierzejewski, M., Debniak, T., Teodorczyk, U., Byrski, T., Gronwald, J., Matyjasik, J., et al. (2004). CHEK2 is a multiorgan cancer susceptibility gene. Am J Hum Genet *75*, 1131–1135.
32. Janiszewska, H., Bak, A., Skonieczka, K., Jaskowiec, A., Kielbinski, M., Jachalska, A., Czyzewska, M., Jazwiec, B., Kuliszkiewicz-Janus, M., Czyz, J., et al. (2018). Constitutional mutations of the CHEK2 gene are a risk factor for MDS, but not for de novo AML. Leuk Res *70*, 74–78.
33. Breast Cancer Association, C., Dorling, L., Carvalho, S., Allen, J., Gonzalez-Neira, A., Luccarini, C., Wahlstrom, C., Pooley, K.A., Parsons, M.T., Fortuno, C., et al. (2021). Breast cancer risk genes - association analysis in more than 113,000 women. N Engl J Med *384*, 428–439.
34. Ruth, K.S., Day, F.R., Hussain, J., Martinez-Marchal, A., Aiken, C.E., Azad, A., Thompson, D.J., Knoblochova, L., Abe, H., Tarry-Adkins, J.L., et al. (2021). Genetic insights into biological mechanisms governing human ovarian ageing. Nature *596*, 393–397.
35. Gebel, J., Tuppi, M., Sanger, N., Schumacher, B., and Dotsch, V. (2020). DNA damaged induced cell death in oocytes. Molecules *25*, 5714.
36. Bolcun-Filas, E., Rinaldi, V.D., White, M.E., and Schimenti, J.C. (2014). Reversal of female infertility by Chk2 ablation reveals the oocyte DNA damage checkpoint pathway. Science *343*, 533–536.
37. Manson, J.E., Chlebowski, R.T., Stefanick, M.L., Aragaki, A.K., Rossouw, J.E., Prentice, R.L., Anderson, G., Howard, B.V., Thomson, C.A., LaCroix, A.Z., et al. (2013). Menopausal hormone therapy and health outcomes during the intervention and extended poststopping phases of the Women's Health Initiative randomized trials. JAMA *310*, 1353–1368.
38. Bell, D.W., Varley, J.M., Szydlo, T.E., Kang, D.H., Wahrer, D.C., Shannon, K.E., Lubratovich, M., Verselis, S.J., Isselbacher, K.J., Fraumeni, J.F., et al. (1999). Heterozygous germ line hCHK2 mutations in Li-Fraumeni syndrome. Science *286*, 2528–2531.
39. De Silva, D.L., and Winship, I. (2020). Is CHEK2 a moderate-risk breast cancer gene or the younger sister of Li-Fraumeni? BMJ Case Rep *13*, e236435.
40. Taneja, P., Gu, J., Peng, R., Carrick, R., Uchiumi, F., Ott, R.D., Gustafson, E., Podust, V.N., and Fanning, E. (2002). A dominant-negative mutant of human DNA helicase B blocks the onset of chromosomal DNA replication. J Biol Chem *277*, 40853–40861.
41. Astle, W.J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A.L., Mead, D., Bouman, H., Riveros-Mckay, F., Kostadima, M.A., et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. Cell *167*, 1415–1429.e9.
42. Vuckovic, D., Bao, E.L., Akbari, P., Lareau, C.A., Mousas, A., Jiang, T., Chen, M.H., Raffield, L.M., Tardaguila, M., Huffman, J.E., et al. (2020). The polygenic and monogenic basis of blood traits and diseases. Cell *182*, 1214–1231.e1.
43. Demuth, I., and Digweed, M. (1998). Genomic organization of a potential human DNA-crosslink repair gene, KIAA0086. Mutat Res *409*, 11–16.
44. Shahi, R.B., De Brakeleer, S., Caljon, B., Pauwels, I., Bonduelle, M., Joris, S., Fontaine, C., Vanhoeij, M., Van Dooren, S., Teugels, E., et al. (2019). Identification of candidate cancer predisposing variants by performing whole-exome sequencing on index patients from BRCA1 and BRCA2-negative breast cancer families. BMC Cancer *19*, 313.
45. Hanai, R., Caron, P.R., and Wang, J.C. (1996). Human TOP3: a single-copy gene encoding DNA topoisomerase III. Proc Natl Acad Sci U S A *93*, 3653–3657.
46. Martin, C.A., Sarlos, K., Logan, C.V., Thakur, R.S., Parry, D.A., Bizard, A.H., Leitch, A., Cleal, L., Ali, N.S., Al-Owain, M.A., et al. (2018). Mutations in TOP3A cause a Bloom syndrome-like disorder. Am J Hum Genet *103*, 456.
47. Wortmann, S.B., Zietkiewicz, S., Kousi, M., Szklarczyk, R., Haack, T.B., Gersting, S.W., Muntau, A.C., Rakovic, A., Renkema, G.H., Rodenburg, R.J., et al. (2015). CLPB mutations cause 3-methylglutaconic aciduria, progressive brain atrophy, intellectual disability, congenital neutropenia, cataracts, movement disorder. Am J Hum Genet *96*, 245–257.
48. Kanabus, M., Shahni, R., Saldanha, J.W., Murphy, E., Plagnol, V., Hoff, W.V., Heales, S., and Rahman, S. (2015). Bi-allelic CLPB mutations cause cataract, renal cysts, nephrocalcinosis and 3-methylglutaconic aciduria, a novel disorder of mitochondrial protein disaggregation. J Inherit Metab Dis *38*, 211–219.
49. Warren, J.T., Cupo, R.R., Wattanasirakul, P., Spencer, D., Locke, A.E., Makaryan, V., Bolyard, A.A., Kelley, M.L., Kingston, N.L., Shorter, J., et al. (2021). Heterozygous variants of CLPB are a cause of severe congenital neutropenia. Blood. https://doi.org/10.1182/blood.2021010762.
50. Liu, J., Zhou, Y., Liu, S., Song, X., Yang, X.Z., Fan, Y., Chen, W., Akdemir, Z.C., Yan, Z., Zuo, Y., et al. (2018). The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. Hum Genet *137*, 553–567.
51. Evangelou, E., Warren, H.R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C.P., Karaman, I., et al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat Genet *50*, 1412–1425.
52. Akiyama, M., Ishigaki, K., Sakaue, S., Momozawa, Y., Horikoshi, M., Hirata, M., Matsuda, K., Ikegawa, S., Takahashi, A., Kanai, M., et al. (2019). Characterizing rare and low-frequency height-associated variants in the Japanese population. Nat Commun *10*, 4393.
53. Verweij, N., Benjamins, J.W., Morley, M.P., van de Vegte, Y.J., Teumer, A., Trenkwalder, T., Reinhard, W., Cappola, T.P., and van der Harst, P. (2020). The genetic makeup of the electrocardiogram. Cell Syst *11*, 229–238.e5.
54. Kanaar, R., Troelstra, C., Swagemakers, S.M., Essers, J., Smit, B., Franssen, J.H., Pastink, A., Bezzubova, O.Y., Buerstedde, J.M., Clever, B., et al. (1996). Human and mouse homologs of the

Saccharomyces cerevisiae RAD54 DNA repair gene: evidence for functional conservation. Curr Biol *6*, 828–838.

55. Matsuda, M., Miyagawa, K., Takahashi, M., Fukuda, T., Kataoka, T., Asahara, T., Inui, H., Watatani, M., Yasutomi, M., Kamada, N., et al. (1999). Mutations in the RAD54 recombination gene in primary cancers. Oncogene *18*, 3427–3430.
56. Pelttari, L.M., Kiiski, J., Nurminen, R., Kallioniemi, A., Schleutker, J., Gylfe, A., Aaltonen, L.A., Leminen, A., Heikkila, P., Blomqvist, C., et al. (2012). A finnish founder mutation in RAD51D: analysis in breast, ovarian, prostate, and colorectal cancer. J Med Genet *49*, 429–432.
57. Hustedt, N., Saito, Y., Zimmermann, M., Alvarez-Quilon, A., Setiaputra, D., Adam, S., McEwan, A., Yuan, J.Y., Olivieri, M., Zhao, Y., et al. (2019). Control of homologous recombination by the HROB-MCM8-MCM9 pathway. Genes Dev *33*, 1397–1415.
58. Fauchereau, F., Shalev, S., Chervinsky, E., Beck-Fruchter, R., Legois, B., Fellous, M., Caburet, S., and Veitia, R.A. (2016). A non-sense MCM9 mutation in a familial case of primary ovarian insufficiency. Clin Genet *89*, 603–607.
59. Hahn, R.A., Eaker, E., and Rolka, H. (1997). Reliability of reported age at menopause. Am J Epidemiol *146*, 771–775.