

Genetic links between ovarian ageing, cancer risk and de novo mutation rates


<https://doi.org/10.1038/s41586-024-07931-x>

Received: 20 June 2022

Accepted: 8 August 2024

Published online: 11 September 2024

Open access

 Check for updates

Stasa Stankovic^{1,10}, Saleh Shekari^{2,3,10}, Qin Qin Huang^{4,10}, Eugene J. Gardner^{1,10}, Erna V. Ivarsdottir^{5,10}, Nick D. L. Owens^{2,10}, Nasim Mavaddat⁶, Ajuna Azad⁷, Gareth Hawkes², Katherine A. Kentistou¹, Robin N. Beaumont², Felix R. Day¹, Yajie Zhao¹, Hakon Jonsson⁵, Thorunn Rafnar⁵, Vinicius Tragante⁵, Gardar Sveinbjornsson⁵, Asmundur Oddsson⁵, Unnur Styrkarsdottir⁵, Julius Gudmundsson⁵, Simon N. Stacey⁵, Daniel F. Gudbjartsson⁵, Breast Cancer Association Consortium*, Kitale Kennedy², Andrew R. Wood², Michael N. Weedon², Ken K. Ong^{1,8}, Caroline F. Wright², Eva R. Hoffmann⁷, Patrick Sulem⁵, Matthew E. Hurles⁴, Katherine S. Ruth², Hilary C. Martin^{4,10}, Kari Stefansson^{5,10}, John R. B. Perry^{1,9,10}✉ & Anna Murray^{2,10}✉

Human genetic studies of common variants have provided substantial insight into the biological mechanisms that govern ovarian ageing¹. Here we report analyses of rare protein-coding variants in 106,973 women from the UK Biobank study, implicating genes with effects around five times larger than previously found for common variants (*ETAA1*, *ZNF518A*, *PNPLA8*, *PALB2* and *SAMHD1*). The *SAMHD1* association reinforces the link between ovarian ageing and cancer susceptibility¹, with damaging germline variants being associated with extended reproductive lifespan and increased all-cause cancer risk in both men and women. Protein-truncating variants in *ZNF518A* are associated with shorter reproductive lifespan—that is, earlier age at menopause (by 5.61 years) and later age at menarche (by 0.56 years). Finally, using 8,089 sequenced trios from the 100,000 Genomes Project (100kGP), we observe that common genetic variants associated with earlier ovarian ageing associate with an increased rate of maternally derived de novo mutations. Although we were unable to replicate the finding in independent samples from the deCODE study, it is consistent with the expected role of DNA damage response genes in maintaining the genetic integrity of germ cells. This study provides evidence of genetic links between age of menopause and cancer risk.

Reproductive longevity in women varies substantially in the general population and has profound effects on fertility and health outcomes in later life^{1,2}. Women are born with a non-renewable ovarian reserve, which is established during fetal development. This reserve is continuously depleted throughout reproductive life, ultimately leading to menopause³. Variation in menopause timing is largely dependent on the differences in the size of the initial oocyte pool and the rate of follicle loss. Natural fertility is believed to be closely associated with menopause timing, and it declines on average ten years before the onset of menopause⁴. The effect of early menopause on infertility is becoming increasingly relevant owing to the secular trend of delaying parenthood to later maternal age at childbirth, especially in Western countries. In addition, normal variation in reproductive lifespan is causally associated with the risk of a wide range of disease outcomes, such as type 2 diabetes mellitus, cancer and impaired bone health, further

highlighting the need for better understanding of the regulators and physiological mechanisms involved in reproductive ageing¹.

The variation in timing of menopause reflects a complex mix of genetic and environmental factors that population-based studies have begun to unravel. Previous genome-wide association studies (GWAS) have successfully identified around 300 distinct common genomic loci associated with the timing of menopause¹. These reported variants cumulatively explain 10–12% of the variance in age at natural menopause (ANM) and 31–38% of the overall estimated single nucleotide polymorphism (SNP) heritability^{1,5,6}. Two-thirds of the GWAS signals implicate genes that regulate DNA damage response (DDR), highlighting the particular sensitivity of oocytes to DNA damage due to the prolonged state of cell cycle arrest across the lifetime^{1,7–13}. Genetic studies for ANM to date have focussed largely on assessing common genetic variation, with little insight into the role of rarer, protein-coding variants.

¹MRC Epidemiology Unit, Wellcome–MRC Institute of Metabolic Science, University of Cambridge, Cambridge, UK. ²University of Exeter Medical School, University of Exeter, Exeter, UK. ³School of Public Health, Faculty of Medicine, University of Queensland, Brisbane, Queensland, Australia. ⁴Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK. ⁵deCODE Genetics/Amgen, Reykjavik, Iceland. ⁶Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. ⁷DNRFCenter for Chromosome Stability, Department of Cellular and Molecular Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ⁸Department of Paediatrics, University of Cambridge, Cambridge, UK. ⁹Metabolic Research Laboratory, Wellcome–MRC Institute of Metabolic Science, University of Cambridge, Cambridge, UK. ¹⁰These authors contributed equally: Stasa Stankovic, Saleh Shekari, Qin Qin Huang, Eugene J. Gardner, Erna V. Ivarsdottir, Nick D. L. Owens, Hilary C. Martin, Kari Stefansson, John R. B. Perry, Anna Murray. *A list of members and their affiliations appears in the Supplementary Information. ✉e-mail: john.perry@mrc-epid.cam.ac.uk; A.Murray@exeter.ac.uk

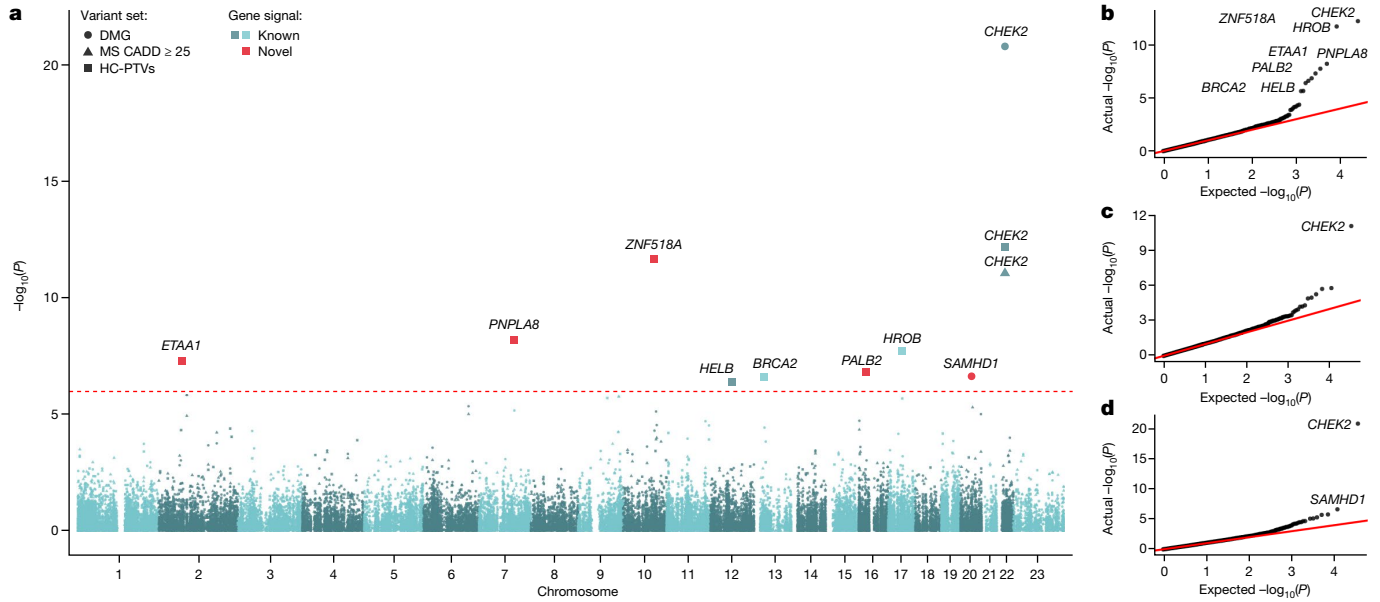


Fig. 1 | Exome-wide associations with ANM. **a**, Manhattan plot showing gene burden test results for ANM from BOLT-LMM in 106,973 female participants. Genes passing exome-wide significance ($P < 1.08 \times 10^{-6}$) are indicated, with the shape signifying the variant class tested and colour indicating the novelty.

MS, missense. **b–d**, QQ plots of P values from BOLT-LMM against expected P values for high-confidence PTVs: $\lambda = 1.047$ (**b**), CADD ≥ 25 missense variants: $\lambda = 1.050$ (**c**) and damaging variants: $\lambda = 1.050$ (**d**).

Initial whole-exome sequencing (WES) analyses in the UK Biobank identified gene-based associations with ANM for *CHEK2*, *DCLRE1A*, *HELB*, *TOP3A*, *BRCA2* and *CLPB*^{1,5}. Here we aimed to explore the role of rare damaging variants in ovarian ageing in greater detail through a combination of enhanced phenotype curation, better-powered statistical tests and assessment of different types of variant class at lower allele frequency thresholds (Supplementary Information). Using these approaches, we identify five genes harbouring variants with large effects, highlighting *ZNF518A* as a major transcriptional regulator of ovarian ageing. Furthermore, we extend these observations to provide initial evidence that women at increased genetic risk of earlier menopause have increased rates of de novo mutations in their offspring.

Exome-wide gene burden associations

Previous studies have focused largely on assessing the role of common genetic variation on ovarian ageing. We sought to better understand the role of rare coding variation in ovarian ageing using WES data available in 106,973 post-menopausal female UK Biobank participants of European genetic ancestry¹⁴. We conducted individual gene burden association tests by collapsing genetic variants according to their predicted functional categories. We defined three categories of rare exome variants with minor allele frequency (MAF) $< 0.1\%$: high-confidence protein-truncating variants (HC-PTVs), missense variants with combined annotation-dependent depletion (CADD) score ≥ 25 , and ‘damaging’ variants (DMG, defined as combination of HC-PTVs and missense variants with CADD ≥ 25). We analysed 17,475 protein-coding genes with a minimum of 10 rare allele carriers in at least one of the masks tested. The primary burden association analysis was conducted using BOLT-LMM¹⁵ (Fig. 1 and Supplementary Table 1). The low exome-wide inflation scores (Fig. 1b–d) and the absence of significant association with synonymous variant burden for any gene indicate that our statistical tests are well calibrated (Extended Data Fig. 1).

We identified rare variation in nine genes associated with ANM at exome-wide significance ($P < 1.08 \times 10^{-6}$; Figs. 1 and 2, Extended Data Fig. 2 and Supplementary Tables 1 and 2). These were confirmed by an independent group of analysts using different quality control and analysis pipelines (Supplementary Tables 1 and 2). Three of these genes

have been previously reported in UK Biobank WES analyses⁵, and we confirm the associations of *CHEK2* (beta = 1.57 years (95% confidence interval (CI): 1.23–1.92), $P = 1.6 \times 10^{-21}$, $n = 578$ damaging allele carriers) and *HELB* (beta = 1.84 years (95% CI: 1.08–2.60), $P = 4.2 \times 10^{-7}$, $n = 120$ HC-PTV carriers) with later ANM and a previously borderline association of *HROB* with earlier ANM (beta = -2.89 years (95% CI: 1.86–3.92),

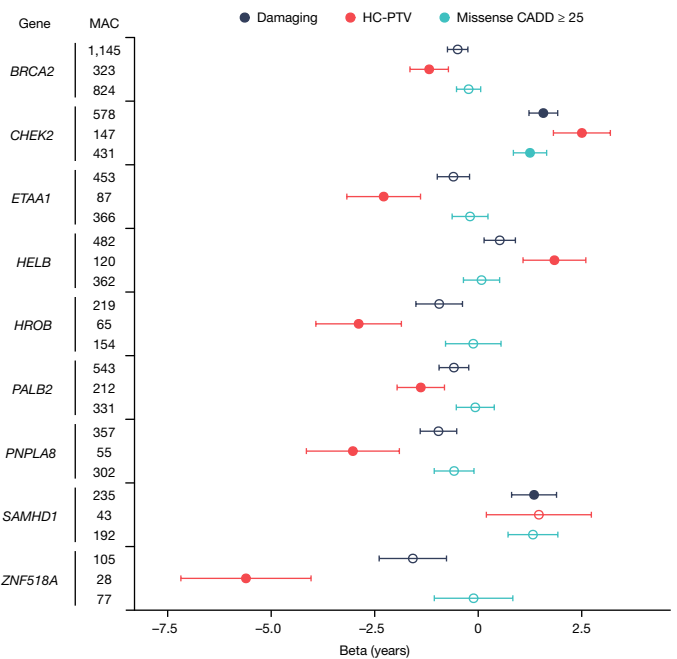


Fig. 2 | Forest plot for gene burden associations with ANM. Exome-wide significant ($P < 1.08 \times 10^{-6}$) genes (filled circles) are displayed; an unfilled circle indicates a nonsignificant association. Points and error bars indicate beta and 95% CI, respectively, for the indicated variant category. Beta values, CIs, minor allele counts (MACs) and P values are derived from BOLT-LMM (values are given in Supplementary Table 2). $n = 115,051$ individuals with ANM are included in the analysis.

$P = 1.9 \times 10^{-8}$, $n = 65$ HC-PTV carriers). In addition, our previous ANM GWAS analyses¹ identified an individual low-frequency PTV variant in *BRCA2*, which we now extend to demonstrate that, in aggregate, *BRCA2* HC-PTV carriers exhibit 1.18 years earlier ANM (beta = -1.18 years (95% CI: 0.72–1.65), $P = 2.6 \times 10^{-7}$, $n = 323$). Rare variants in the remaining five genes (*ETAA1*, *ZNF518A*, *PNPLA8*, *PALB2* and *SAMHDI*) have not previously been implicated in ovarian ageing. Effect sizes of these associations range from 5.61 years earlier ANM for HC-PTV carriers in *ZNF518A* (95% CI: 4.04–7.18, $P = 2.1 \times 10^{-12}$, $n = 28$), to 1.35 years later ANM for women carrying damaging alleles in *SAMHDI* (95% CI: 0.81–1.89, $P = 2.4 \times 10^{-7}$, $n = 235$). This contrasts with a maximum effect size of 1.06 years (median 0.12 years) for common variants (MAF > 1%) identified by previous ANM GWAS¹.

We next attempted to replicate these findings using two independent datasets. First, from the Icelandic deCODE study^{16,17}. Despite the substantially smaller sample size ($n = 27,678$ women with ANM), rarity of the alleles we were testing, and minor differences in allele frequency and variant classification, we observed consistent effect estimates for all nine genes that we identified (Supplementary Table 3). This included nominally significant associations at *BRCA2*, *CHEK2*, *ETAA1*, *HROB*, *HELB*, *SAMHDI* and *ZNF518A*. Second, we used data in up to 26,258 women with ANM from the BRIDGES study¹⁸. As this study used a targeted sequencing approach of suspected breast cancer genes, it was informative for only *BRCA2*, *PALB2* and *CHEK2*. Despite the small sample size, for each of these genes we found effect estimates consistent with our discovery analyses, which were maintained when adjusting for cancer status and within women not diagnosed with breast cancer (Supplementary Table 4). Notably, we replicated the novel association with *PALB2*, where the 78 women carrying PTVs experienced menopause 1.78 years earlier on average ($P = 4.6 \times 10^{-4}$). Differences in allele frequency cut-offs had minimal effect on variants included in burden tests, because we only tested predicted deleterious variants and these were mostly rare (less than 0.1%).

We next sought to understand why previous analyses of UK Biobank WES data missed the associations that we report here, and conversely why we did not identify associations with other previously reported genes. Of the seven genes identified by Ward et al.⁵, three were also identified by our study (*CHEK2*, *HELB* and *HROB*), three were recovered when we increased our burden test MAF threshold from 0.1% to 1% (*DCLRE1A*, *RAD54L* and *TOP3A*), and an additional gene fell just below our P value threshold when considering variants with <1% MAF (*CLPB*; $P = 1.2 \times 10^{-5}$). By contrast, our identification of novel associations that were not reported by Ward et al. (*BRCA2*, *ETAA1*, *PALB2*, *PNPLA8*, *SAMHDI* and *ZNF518A*) is probably explained by differences in phenotype preparation, sample size, variant annotation and the statistical model used (see Supplementary Information and Supplementary Table 5).

Overlap with common variant associations

To explore the overlap between common and rare variant association signals for ANM, we integrated our exome-wide results with data generated from the largest reported common variant GWAS of ANM¹.

Five of our nine identified WES genes (*CHEK2*, *BRCA2*, *ETAA1*, *HELB* and *ZNF518A*) mapped within 500 kb of a common GWAS signal (Supplementary Table 6). Notably, we previously reported a common, predicted benign, missense variant (rs35777125-G439R, MAF = 11%) in *ETAA1* associated with 0.26 years earlier ANM¹. By contrast, our WES analysis showed that carriers of rare HC-PTVs in *ETAA1* show a nearly 10-fold earlier ANM (beta = -2.28 years (95% CI: 1.39–3.17), $P = 5.30 \times 10^{-8}$, $n = 87$). Furthermore, three independent non-coding common GWAS signals around 150 kb apart (MAF: 2.8–47.5%, beta: -0.28 to 0.28 years per minor allele) were reported proximal to *ZNF518A*, whereas gene burden testing finds that rare HC-PTV carriers show nearly 20-fold earlier ANM than common variant carriers (beta = -5.61 years (95% CI: 4.04–7.18), $P = 2.10 \times 10^{-12}$, $n = 28$).

ZNF518A is a poorly characterized C2H2 zinc-finger transcription factor, which has been shown to associate with PRC2 and G9A–GLP repressive complexes along with its paralogue *ZNF518B*, suggesting a potential role in transcriptional repression¹⁹. By integrating chromatin immunoprecipitation with sequencing (ChIP-seq) data^{20,21}, we demonstrate that common variants associated with ANM are enriched in the binding sites of *ZNF518A* (Supplementary Table 7 and Supplementary Information), providing further support for the role of this gene in ovarian ageing.

In addition, there were two genes within 500 kb of GWAS loci (*BRCA1* and *SLCO4A1*) that were associated with ANM by gene burden testing at $P < 1.7 \times 10^{-5}$. Effect sizes for common variant associations ranged from 0.07–0.24 years per allele at these loci, whereas gene burden tests for rarer variants at these same loci revealed much larger effect sizes: for *BRCA1*, 2.1 years earlier ANM for PTVs (95% CI: 1.2–3.0, $P = 2.4 \times 10^{-6}$) and for *SLCO4A1*, 1.13 years earlier ANM for damaging variants (95% CI: 0.6–1.64, $P = 1.1 \times 10^{-3}$), with non-overlapping 95% CI between common and rare variant associations for *BRCA1*.

Non-reproductive health and disease effects

Our genetic studies have previously shown that the genetic mechanisms that regulate the end of reproductive life are largely distinct from those that determine its beginning^{22,23}. However, it is noteworthy that the largest reported GWAS for age at menarche identified a common variant signal at the *ZNF518A* locus for later puberty timing in girls (rs1172955, beta = 0.04 years (95% CI: 0.03–0.05), $P = 6.6 \times 10^{-12}$), which appears nominally associated with earlier ANM²² (beta = -0.04 years (95% CI: 0.01–0.06), $P = 6.6 \times 10^{-3}$). To extend this observation, we found that our identified *ZNF518A* PTVs were also associated with later age at menarche (0.56 years (95% CI: 0.14–0.98), $P = 9.2 \times 10^{-3}$). Furthermore, using functional genome-wide association analysis²⁴ and signed linkage disequilibrium profile²⁵ (SLDP), we found that similar to ANM, common variants that associate with puberty in girls were enriched in transcriptional targets of *ZNF518A* (Extended Data Fig. 3 and Supplementary Table 7). These data suggest that loss of *ZNF518A* shortens reproductive lifespan by delaying puberty and reducing age at menopause.

We next explored the effect of ANM-associated genes on cancer outcomes, replicating previously reported associations with PTVs in *BRCA2*, *CHEK2* and *PALB2* and cancer outcomes in male and female subjects^{1,10} (Supplementary Tables 8–10). We also identified a novel association of *SAMHDI* damaging variants and HC-PTVs with ‘all cancer’ in both males (odds ratio (OR) = 2.12 (95% CI: 1.72–2.62), $P = 4.7 \times 10^{-13}$) and females (OR = 1.61 (95% CI: 1.31–1.96), $P = 4 \times 10^{-6}$; Fig. 3 and Supplementary Tables 8–10).

SAMHDI associations with cancer appear to be driven by increased risk for multiple site-specific cancers, notably prostate cancer in men, mesothelioma in both men and women, and suggestive evidence for higher breast cancer susceptibility in women (Fig. 4 and Supplementary Table 11). Although the numbers of mutation carriers diagnosed with each site-specific cancer was small, the majority of these findings persisted using logistic regression with penalized likelihood estimation, which is more robust to extreme case–control imbalance²⁶ (Supplementary Table 11). To replicate this association, we interrogated genetic data in up to 49,981 cancer cases and 337,946 controls from the Icelandic deCODE study^{16,17}. We observed highly similar results (Supplementary Table 12) to those from the UK Biobank, demonstrating increased all-site cancer susceptibility in male (OR = 1.67 (95% CI: 1.18–2.37), $P = 0.004$), female (OR = 1.57 (95% CI: 1.15–2.15), $P = 0.005$) and sex-combined models (OR = 1.61 (95% CI: 1.28–2.03), $P = 6 \times 10^{-5}$). Significant associations were also seen for a number of site-specific cancers, including haematological cancers in men (OR = 4.18 (95% CI: 1.90–9.21), $P = 3.9 \times 10^{-4}$) and prostate cancer (OR = 2.36 (95% CI: 1.14–4.87), $P = 0.02$).

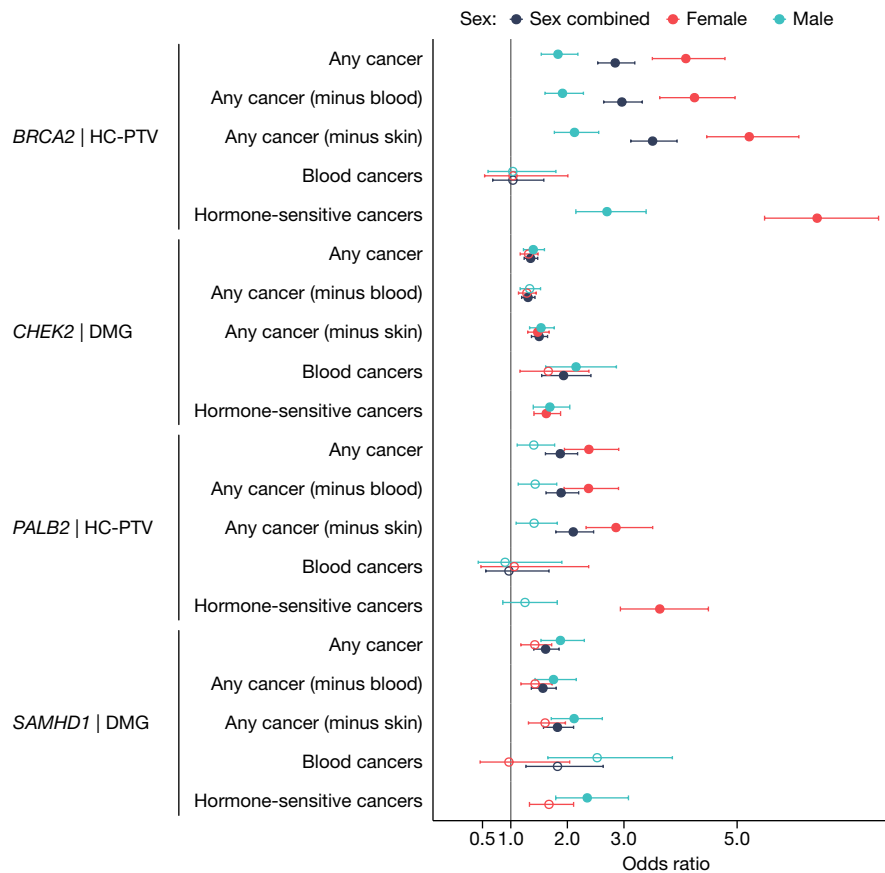


Fig. 3 | Forest plot for ANM WES genes with significant gene burden associations for cancer phenotypes. Exome-wide significant ($P < 1.08 \times 10^{-6}$) genes are displayed, showing sex-stratified and combined results from BOLT-LMM analysis. Hormone-sensitive cancers only were tested in male and female subjects separately (Methods). The presented masks were selected on the basis of the most significant association per gene and cancer type. Points and bars

indicate odds ratio and 95% CI, respectively, for specific genes and their variant categories for each cancer type (values are given in Supplementary Table 10). A filled circle indicates that a result passes a Bonferroni-corrected significance threshold of $P < 1.08 \times 10^{-6}$; an unfilled circle indicates a nonsignificant association. $n = 421,064$ (228,517 female and 192,547 male participants).

Cancer risk-increasing alleles in *SAMHD1* were associated with later ANM, following a similar pattern demonstrated previously for *CHEK2*. This finding is consistent with a mechanism of disrupted DNA damage sensing and apoptosis, resulting in slowed depletion of the ovarian reserve¹. We note, however, that there are other mechanisms of ovarian reserve depletion, and future experimental work should seek to better understand this specific association. In addition, we provide robust evidence for a previously described rare variant association for *SAMHD1* with telomere length²⁷, highlighting that rare damaging variants cause longer telomere length ($P = 1.4 \times 10^{-59}$) (Extended Data Fig. 4 and Supplementary Table 10).

Effects on de novo mutation rate

Of the nine genes that we identified in our exome analysis as associated with ANM, seven are involved in DNA damage repair, further supporting the role of these pathways in ovarian ageing (Supplementary Table 13). For genes that inhibit DNA double-strand break repair, the hypothesis is that they cause premature depletion of the ovarian reserve owing to a failure to repair oocytes with DNA damage¹. This is evidenced by the reported increased numbers of DNA double strand breaks in the oocytes of *Brca1*-deficient mice and of women with *BRCA1* mutations who underwent elective oophorectomy²⁸⁻³⁰. Our current study adds further support for this hypothesis, with heterozygous *BRCA1* and *BRCA2* loss-of-function alleles being associated with 2.1 and 1.18 years earlier ANM, respectively.

We sought to build on these observations by testing the hypothesis that inter-individual variation in these DDR processes would influence the mutation rate in germ cells and thus in the offspring. More specifically, we hypothesized that genetic susceptibility to earlier ovarian ageing would be associated with a higher de novo mutation (DNM) rate in offspring. To test this, we analysed whole-genome-sequenced parent-offspring trios from the 100,000 Genome Project (100kGP)³¹ ($n = 8,809$ with European ancestry) and followed up in trios from the deCODE study^{32,33} ($n = 6,042$) (Extended Data Fig. 5). We calculated a polygenic score (PGS) for ANM in the parents by combining the effect estimates from our previously identified 290 common variants¹ and tested this for association with the phased DNM rate in the offspring, adjusting for parental age and quality control-related covariates. In the 100kGP dataset, we found that maternal genetic susceptibility to earlier ANM was associated with an increased rate of maternally derived DNMs in the offspring (Poisson regression; meta-analysis beta = -0.082 DNMs per s.d. increase in the PGS (95% CI: $-0.126, -0.037$), $P = 0.00033$; Supplementary Table 14). However, this association was not replicated in the deCODE dataset (beta = 0.018 (95% CI: $-0.038, 0.073$), $P = 0.53$), and the estimates from the 100kGP and deCODE data were inconsistent (heterogeneity $P = 0.006$). A meta-analysis of the results across the two cohorts gave a significant effect of maternal ANM PGS on maternally derived DNMs (beta = -0.0426 (95% CI: $-0.0772, -0.0079$) DNMs per s.d. increase in the PGS; standard error = 0.018 , $P = 0.016$). The 100kGP finding was consistent in sensitivity analyses using a two-sample Mendelian randomization framework that can better model the dose-response

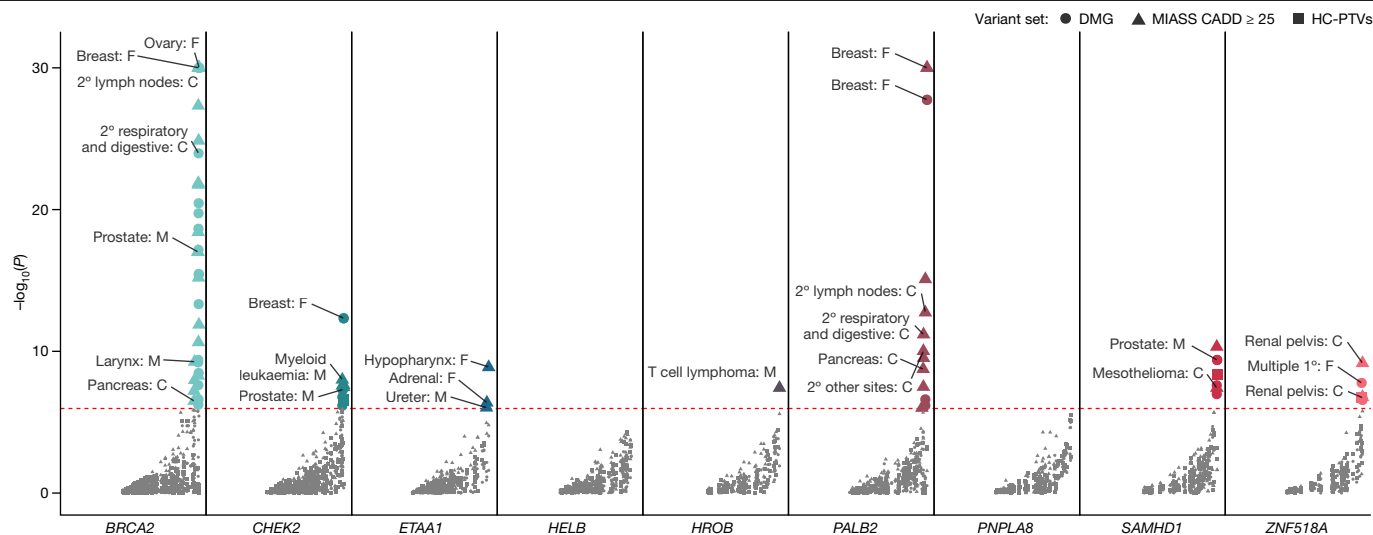


Fig. 4 | Genetic susceptibility to premature ovarian ageing and increased risk for diverse cancer types. Association between loss of ANM genes identified in this study and risk of 90 site-specific cancers among UK Biobank participants. Summary statistics for cancer associations were obtained using a logistic regression with penalized likelihood estimation that controls for case-control

imbalance³³ (Methods). Associations highlighted with text labels passed an exome-wide significance threshold ($P < 1.08 \times 10^{-6}$). The y-axis is capped at $-\log_{10}(P) = 30$ for visualization purposes; uncapped summary statistics are presented in Supplementary Table 11. 1°, primary cancer; 2°, secondary cancer; F, female participants; M, male participants; C, both sexes combined.

relationship of these variants (Supplementary Table 15). These 100kGP Mendelian randomization results were highly concordant, with all models showing a significant result (minimum P value = 6.3×10^{-5}) and no heterogeneity in effect for ANM genetic variants used as instrumental variables (Methods). In both 100kGP and deCODE data, the paternal PGS was not associated with paternally or maternally derived DNMs ($P > 0.05$), or the maternal PGS associated with paternally derived DNMs ($P > 0.05$). Finally, we tested whether rare damaging variants in the nine ANM-associated genes (Fig. 2) were associated with DNM rate in the 100kGP and deCODE study (Supplementary Table 16). After meta-analysis and following multiple testing correction ($P > 0.05 / (2 \times 9)$), none of the nine genes showed significant rare variant associations with DNM rate in either mothers or fathers.

Discussion

Our study extends the number of genes implicated in ovarian ageing through the identification of rare protein-coding variants. Effect sizes ranged from 5.61 years earlier ANM for HC-PTV carriers in *ZNF518A*, to 1.35 years later ANM for women carrying damaging variants in *SAMHD1* compared with a maximum effect size of 1.06 years (median 0.12 years) reported for common variants¹ (MAF > 1%). Several of these effect estimates were comparable to those conferred by *FMRI* premutations, which are currently used as part of the only routinely applied clinical genetics test for premature ovarian insufficiency³⁴. Deleterious variants in three genes (*CHEK2*, *HELB* and *SAMHD1*) were associated with an increase in ANM and therefore represent potential therapeutic targets for enhancing ovarian stimulation in women undergoing in vitro fertilization treatment through short-term apoptotic inhibition. Seven out of the nine ANM genes identified have known roles in DNA damage repair, and—to our knowledge—three of these are linked to ANM for the first time (*PALB2*, *ETAA1* and *HROB*). *PALB2* is involved in *BRCA2* localization and stability, and *PALB2* compound heterozygous mutations result in Fanconi anaemia and predispose to childhood malignancies³⁵. *ETAA1* accumulates at DNA damage sites in response to replication stress^{36,37}, and *HROB* is involved in homologous recombination by recruiting the MCM8–MCM9 helicase to sites of DNA damage to promote DNA synthesis^{38,39}. Homozygous loss of function of *HROB* is associated with premature ovarian insufficiency⁴⁰ and infertility in both sexes in mouse models³⁸.

Novel biological mechanisms of ovarian ageing were revealed by finding associations with two non-DDR genes (*PNPLA8* and *ZNF518A*): *PNPLA8* is a calcium-independent phospholipase^{41,42} and a recessive cause of neurodegenerative mitochondrial disease and mitochondrial myopathy^{43–45}; to our knowledge, an association with reproductive phenotypes has not been described previously. *ZNF518A* belongs to the zinc-finger protein family and is likely to be a transcriptional regulator for a large number of genes¹⁹. We found that female carriers of rare PTVs in *ZNF518A* have shorter reproductive lifespan owing to delayed puberty timing and earlier menopause. Enrichment of GWAS signals at *ZNF518A* binding sites suggests that *ZNF518A* regulates the genes involved in reproductive longevity by repression of regulatory elements distal to their transcription start sites. *ZNF518A* has also recently been demonstrated to have a role in forming heterochromatin at pericentromeric regions, which is essential for proper chromosome segregation during mitosis and meiosis⁴⁶.

Whereas mutation in *SAMHD1* is a common somatic event in a variety of cancers⁴⁷, we demonstrate here that it is also a germline risk factor. Recessive inheritance of *SAMHD1* missense variants and PTVs have been associated with Aicardi–Goutières syndrome, a congenital autoimmune disease⁴⁸. The damaging variants in *SAMHD1* that we identified are associated with increased risk of ‘all cancer’ in men and women, as well as in sex-specific cancers, highlighting *SAMHD1* as a novel risk factor for prostate cancer in men and hormone-sensitive cancers in women. Recent studies have demonstrated association of germline *SAMHD1* coding variants with having two or more primary cancers in the UK Biobank⁴⁹ ($P = 2.4 \times 10^{-7}$), and with breast cancer susceptibility⁵⁰ ($P < 1 \times 10^{-4}$). *SAMHD1* has a role in preventing the accumulation of excess deoxynucleotide triphosphates (dNTPs), particularly in non-dividing cells⁵¹. A regulated dNTP pool is important for the fidelity of DNA repair, thus highlighting additional roles of this gene in facilitation of DNA end resection during DNA replication and repair^{51–56}. *SAMHD1* deficiency leads to resistance to apoptosis^{57,58}, suggesting that delayed ANM might originate from slowed depletion of ovarian reserve due to disrupted apoptosis, analogous to the mechanism for *CHEK2* that has been reported previously.

Previous studies have demonstrated that parental age is strongly associated with the number of de novo mutations in offspring⁵⁹, with the majority of these mutations arising from the high rate of

spermatogonial stem cell divisions that underlie spermatogenesis throughout the adult life of men⁶⁰. We investigated whether women at higher genetic risk of earlier menopause transmit more de novo mutations to offspring. Whereas we found significant evidence for this in 100kGP data, we could not replicate the association between ANM PGS and maternally derived DNMs in deCODE data. We cannot currently explain this finding, other than with the possibility that our result from 100kGP is inflated by ‘winner’s curse’ and that the true effect size is lower than our point estimate. Power calculations suggest that the deCODE dataset is well-powered to replicate the effect if it is equal to our point estimate in 100kGP (>90% power) but has only modest power (around 40%) if the effect is at the lower bound of our 95% CI from 100kGP. Future large studies of whole-genome sequence data in trios will be important to further explore this relationship. If confirmed, this finding could have direct implications for the health of future generations, given the widely reported link between de novo mutations and increased risk of psychiatric disease and developmental disorders^{61–64}. If genetic susceptibility to earlier menopause influences de novo mutation rate, non-genetic risk factors for earlier ANM, such as smoking and alcohol intake, would probably have the same effect, and there is some evidence to support this⁶⁵. Our observations make conceptual sense given that menopause timing appears to be primarily driven by the genetic integrity of oocytes and their ability to sustain, detect, repair and respond to acquired DNA damage¹. These observations also build on earlier work in mice and humans that *BRCA1/2* deficiency increases the rate of double strand breaks in oocytes and reduces ovarian reserve^{28–30}.

A limitation of our work is that a small proportion of maternally phased DNMs could be postzygotic mutations in the child, which did not originate in the maternal germline. We were unable to differentiate between these owing to the modest sequencing coverage of a single tissue per child. A further limitation is that, owing to data availability, analyses have been restricted to women of European ancestry, making it difficult to evaluate how generalizable these findings may be to other populations, as average age of menopause in women from different ancestry groups varies⁶⁶. We anticipate that this will be addressed in future studies as relevant data become available.

Our study of rare coding variation across the genome expands our understanding of the genetic architecture of ovarian ageing. Future genomic studies incorporating rare non-coding variation in addition to experimental work will build on our identified genetic associations to help further our understanding of the underlying biological mechanisms governing ovarian ageing.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-024-07931-x>.

- Ruth, K. S. et al. Genetic insights into biological mechanisms governing human ovarian ageing. *Nature* **596**, 393–397 (2021).
- Perry, J. R. B., Murray, A., Day, F. R. & Ong, K. K. Molecular insights into the aetiology of female reproductive ageing. *Nat. Rev. Endocrinol.* **11**, 725–734 (2015).
- Wallace, W. H. B. & Kelsey, T. W. Human ovarian reserve from conception to the menopause. *PLoS ONE* **5**, e8772 (2010).
- Lambalk, C. B., van Disseldorp, J., de Koning, C. H. & Broekmans, F. J. Testing ovarian reserve to predict age at menopause. *Maturitas* **63**, 280–291 (2009).
- Ward, L. D. et al. Rare coding variants in DNA damage repair genes associated with timing of natural menopause. *HGG Adv.* **3**, 100079 (2021).
- Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
- He, C. et al. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. *Nat. Genet.* **41**, 724–728 (2009).
- Stolk, L. et al. Loci at chromosomes 13, 19 and 20 influence age at natural menopause. *Nat. Genet.* **41**, 645–647 (2009).

- Perry, J. R. B. et al. A genome-wide association study of early menopause and the combined impact of identified variants. *Hum. Mol. Genet.* **22**, 1465–1472 (2013).
- Day, F. R. et al. Large-scale genomic analyses link reproductive aging to hypothalamic signaling, breast cancer susceptibility and BRCA1-mediated DNA repair. *Nat. Genet.* **47**, 1294–1303 (2015).
- Stolk, L. et al. Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nat. Genet.* **44**, 260–268 (2012).
- Perry, J. R. B. et al. DNA mismatch repair gene MSH6 implicated in determining age at natural menopause. *Hum. Mol. Genet.* **23**, 2490–2497 (2014).
- Murray, A. et al. Common genetic variants are significant risk factors for early menopause: results from the Breakthrough Generations Study. *Hum. Mol. Genet.* **20**, 186–192 (2011).
- Szostakowski, J. D. et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat. Genet.* **53**, 942–948 (2021).
- Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
- Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
- Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
- Dorling, L. et al. Breast cancer risk genes—Association analysis in more than 113,000 women. *New Engl. J. Med.* **384**, 428–439 (2021).
- Maier, V. K. et al. Functional proteomic analysis of repressive histone methyltransferase complexes reveals ZNF518B as a G9A regulator. *Mol. Cell. Proteomics* **14**, 1435–1446 (2015).
- Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Davis, C. A. et al. The Encyclopedia of DNA Elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
- Day, F. R. et al. Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet.* **49**, 834–841 (2017).
- Day, F. R. et al. Shared genetic aetiology of puberty timing between sexes and with health-related outcomes. *Nat. Commun.* **6**, 8842 (2015).
- Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
- Reshef, Y. A. et al. Detecting genome-wide directional effects of transcription factor binding on polygenic disease risk. *Nat. Genet.* **50**, 1483–1493 (2018).
- Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38 (1993).
- Codd, V. et al. Polygenic basis and biomedical consequences of telomere length variation. *Nat. Genet.* **53**, 1425–1433 (2021).
- Titus, S. et al. Impairment of BRCA1-related DNA double-strand break repair leads to ovarian aging in mice and humans. *Sci. Transl. Med.* **5**, 172ra21 (2013).
- Miao, Y. et al. BRCA2 deficiency is a potential driver for human primary ovarian insufficiency. *Cell Death Dis.* **10**, 474 (2019).
- Lin, W., Titus, S., Moy, F., Ginsburg, E. S. & Oktay, K. Ovarian aging in women with BRCA germline mutations. *J. Clin. Endocrinol. Metab.* **102**, 3839–3847 (2017).
- Kaplanis, J. et al. Genetic and chemotherapeutic influences on germline hypermutation. *Nature* **605**, 503–508 (2022).
- Jónsson, H. et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
- Halldórsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).
- Sherman, S. L. Premature ovarian failure in the fragile X syndrome. *Am. J. Med. Genet.* **97**, 189–194 (2000).
- Reid, S. et al. Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat. Genet.* **39**, 162–164 (2007).
- Bass, T. E. et al. ETTA1 acts at stalled replication forks to maintain genome integrity. *Nat. Cell Biol.* **18**, 1185–1195 (2016).
- Saldívar, J. C. et al. An intrinsic S/G2 checkpoint enforced by ATR. *Science* **361**, 806–810 (2018).
- Hustedt, N. et al. Control of homologous recombination by the HROB–MCM8–MCM9 pathway. *Genes Dev.* **33**, 1397–1415 (2019).
- Huang, J. W. et al. MCM8IP activates the MCM8–9 helicase to promote DNA synthesis and homologous recombination upon DNA damage. *Nat. Commun.* **11**, 2948 (2020).
- Tucker, E. J. et al. Meiotic genes in premature ovarian insufficiency: variants in HROB and REC8 as likely genetic causes. *Eur. J. Hum. Genet.* **30**, 219–228 (2022).
- Hara, S., Yoda, E., Sasaki, Y., Nakatani, Y. & Kuwata, H. Calcium-independent phospholipase A2γ (iPLA2γ) and its roles in cellular functions and diseases. *Biochim. Biophys. Acta* **1864**, 861–868 (2019).
- Liu, G. Y. et al. The phospholipase iPLA2γ is a major mediator releasing oxidized aliphatic chains from cardiolipin, integrating mitochondrial bioenergetics and signaling. *J. Biol. Chem.* **292**, 10672–10684 (2017).
- Shukla, A., Saneto, R. P., Hebbar, M., Mirzaa, G. & Girisha, K. M. A neurodegenerative mitochondrial disease phenotype due to biallelic loss-of-function variants in PNPLA8 encoding calcium-independent phospholipase A2γ. *Am. J. Med. Genet. A* **176**, 1232–1237 (2018).
- Saunders, C. J. et al. Loss of function variants in human PNPLA8 encoding calcium-independent phospholipase A2γ recapitulate the mitochondriopathy of the homologous null mouse. *Hum. Mutat.* **36**, 301–306 (2015).
- Masih, S., Moirangthem, A. & Phadke, S. R. Homozygous missense variation in PNPLA8 causes prenatal-onset severe neurodegeneration. *Mol. Syndromol.* **12**, 174–178 (2021).
- Ohta, S. et al. Zinc-finger protein 518 plays a crucial role in pericentromeric heterochromatin formation by linking satellite DNA to heterochromatin. Preprint at *bioRxiv*. <https://doi.org/10.1101/2022.09.15.508097> (2022).
- Schott, K. et al. SAMHD1 in cancer: curse or cure? *J. Mol. Med.* **100**, 351–372 (2022).
- Rice, G. I. et al. Mutations involved in Aicardi–Goutières syndrome implicate SAMHD1 as regulator of the innate immune response. *Nat. Genet.* **41**, 829–832 (2009).

49. Cavazos, T. B. et al. Assessment of genetic susceptibility to multiple primary cancers through whole-exome sequencing in two large multi-ancestry studies. *BMC Med.* **20**, 332 (2022).
50. Wilcox, N. et al. Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk. *Nat. Genet.* **55**, 1435–1439 (2023).
51. Franzolin, E. et al. The deoxynucleotide triphosphohydrolase SAMHD1 is a major regulator of DNA precursor pools in mammalian cells. *Proc. Natl Acad. Sci. USA* **110**, 14272–14277 (2013).
52. Kumar, D. et al. Mechanisms of mutagenesis in vivo due to imbalanced dNTP pools. *Nucleic Acids Res.* **39**, 1360–1371 (2011).
53. Coquel, F. et al. SAMHD1 acts at stalled replication forks to prevent interferon induction. *Nature* **557**, 57–61 (2018).
54. Daddacha, W. et al. SAMHD1 promotes DNA end resection to facilitate DNA repair by homologous recombination. *Cell Rep.* **20**, 1921–1935 (2017).
55. Mathews, C. K. Deoxyribonucleotide metabolism, mutagenesis and cancer. *Nat. Rev. Cancer* **15**, 528–539 (2015).
56. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle* **7**, 2902–2906 (2008).
57. Bonifati, S. et al. SAMHD1 controls cell cycle status, apoptosis and HIV-1 infection in monocytic THP-1 cells. *Virology* **495**, 92–100 (2016).
58. Kodigepalli, K. M., Li, M., Liu, S. L. & Wu, L. Exogenous expression of SAMHD1 inhibits proliferation and induces apoptosis in cutaneous T-cell lymphoma-derived HuT78 cells. *Cell Cycle* **16**, 179–188 (2017).
59. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
60. Crow, J. F. The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* **1**, 40–47 (2000).
61. Iossifov, I. et al. The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515**, 216–221 (2014).
62. Fromer, M. et al. De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506**, 179–184 (2014).
63. Wang, W., Corominas, R. & Lin, G. N. De novo mutations from whole exome sequencing in neurodevelopmental and psychiatric disorders: from discovery to application. *Front Genet* **10**, 258 (2019).
64. Coe, B. P. et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
65. Garcia-Salinas, O. I. et al. The impact of ancestral, environmental and genetic influences on germline de novo mutation rates and spectra. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.05.17.594464> (2024).
66. Henderson, K. D. L., Bernstein, L., Henderson, B., Kolonel, L. & Pike, M. C. Predictors of the timing of natural menopause in the multiethnic cohort study. *Am. J. Epidemiol.* **167**, 1287–1294 (2008).

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



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

© The Author(s) 2024

Methods

UK Biobank data processing and quality control

To conduct the rare variant burden analyses described in this study, we obtained WES data for 454,787 individuals from the UK Biobank study⁶⁷. Participants were excluded on the basis of excess heterozygosity, autosomal variant missingness on genotyping arrays ($\geq 5\%$), or inclusion in the subset of phased samples as defined in Bycroft et al.⁶⁸. Analysis was restricted to participants with European genetic ancestry, owing to the unknown influence of rare variants on population stratification and limited non-European sample size, leaving a total of 421,065 individuals. Variant quality control and annotation were performed using the UK Biobank Research Analysis Platform (RAP; <https://ukbiobank.dnanexus.com/>), a cloud-based central data repository for UK Biobank WES and phenotypic data. Besides the quality control described by Backman et al.⁶⁷, we performed additional steps using custom applets designed for the RAP. First, we processed population-level variant call format (VCF) files by splitting and left-correcting multi-allelic variants into separate alleles using ‘bcftools norm’⁶⁹. Second, we performed genotype-level filtering applying ‘bcftools filter’ separately for single nucleotide variants (SNVs) and insertions–deletion mutations using a missingness-based approach. Using this approach, we set to missing (./.) all SNV genotypes with depth < 7 and genotype quality < 20 or insertion–deletion genotypes with a depth < 10 and genotype quality < 20 . Next, we applied a binomial test to assess an expected alternate allele contribution of 50% for heterozygous SNVs; we set to missing all SNV genotypes with a binomial test P value $\leq 1 \times 10^{-3}$. Following genotype-level filtering we recalculated the proportion of individuals with a missing genotype for each variant and filtered all variants with a missingness value $> 50\%$. The variant annotation was performed using the ENSEMBL variant effect predictor (VEP) v104⁷⁰ with the ‘--everything’ flag and plugins for CADD⁷¹ and LOFTEE⁷² enabled. For each variant we prioritized the highest impact individual consequence as defined by VEP and one ENSEMBL transcript as determined by whether or not the annotated transcript was protein-coding, MANE select v0.97, or the VEP canonical transcript. Following annotation, variants were categorised on the basis of their predicted impact on the annotated transcript. PTVs were defined as all variants annotated as stop-gained, frameshift, splice acceptor and splice donor. Missense variant consequences are identical to those defined by VEP. Only autosomal or chromosome X variants within ENSEMBL protein-coding transcripts and within transcripts included on the UK Biobank ES assay⁶⁷ were retained for subsequent burden testing.

Exome-wide association analyses in the UK Biobank

To perform rare variant burden tests, we used a custom implementation of BOLT-LMM v2.3.6¹⁵ for the RAP. Two primary inputs are required by BOLT-LMM: (1) a set of genotypes with minor allele count > 100 derived from genotyping arrays to construct a null linear mixed effects model; and (2) a larger set of variants collapsed on ENSEMBL transcript to perform association tests. For the former, we queried genotyping data available on the RAP and restricted to an identical set of individuals included for rare variant association tests. For the latter, and as BOLT-LMM expects imputed genotyping data as input rather than per-gene carrier status, we created dummy genotype files where each variant represents one gene and individuals with a qualifying variant within that gene are coded as heterozygous, regardless of the number of variants that individual has in that gene.

To test a range of variant annotation categories for $MAF < 0.1\%$, we created dummy genotype files for high confidence PTVs as defined by LOFTEE, missense variants with $CADD \geq 25$, and damaging variants that included both high confidence PTVs and missense variants with $CADD \geq 25$. For each phenotype tested, BOLT-LMM was then run with default parameters other than the inclusion of

the ‘ImmInfOnly’ flag. To derive association statistics for individual markers, we also provided all 26,657,229 individual markers regardless of filtering status as input to BOLT-LMM. All tested phenotypes were run as continuous traits corrected by age, age², sex, the first ten genetic principal components as calculated in Bycroft et al.⁶⁸ and study participant ES batch as a categorical covariate (50k, 200k or 450k).

For discovery analysis in the primary trait of interest, ANM, we analysed 17,475 protein-coding genes with the minimum of 10 rare allele carriers in at least one of the masks tested using BOLT-LMM (Supplementary Table 1). The significant gene-level associations for ANM were identified applying Bonferroni correction for the number of masks with $MAC \geq 10$ ($n = 46,251$ masks) in 17,475 protein-coding genes ($P: 0.05/46,251 = 1.08 \times 10^{-6}$) (Supplementary Table 2). Furthermore, to compare and explain potential differences between our WES results and the previously published one⁵, we ran the above approach using $MAF < 1\%$, a cut-off applied by Ward et al. (Supplementary Table 5 and Supplementary Information).

To generate accurate odds ratio and standard error estimates for binary traits, we also implemented a generalized linear model using the statsmodels package⁷³ for Python in a three-step process. First, a null model was run with the phenotype as a continuous trait, corrected for control covariates as described above. Second, we regressed carrier status for individual genes on the residuals of the null model to obtain a preliminary P value. Thirdly, all genes were again tested using a full model to obtain odds ratios and standard errors with the family set to ‘binomial’. Generalized linear models utilized identical input to BOLT-LMM converted to a sparse matrix.

ANM phenotype derivation

ANM was derived for individuals within the UK Biobank, who were deemed to have undergone natural menopause—that is, not affected by surgical or pharmaceutical interventions, as follows.

First, European female participants ($n = 245,820$) who indicated during any of the attended visits having had a hysterectomy were collated (fields 3591 and 2724) and their reported hysterectomy ages were extracted (field 2824) and the median age was kept ($n = 47,218$ and 46,260 with reported ages). The same procedure was followed for participants indicating having undergone a bilateral oophorectomy (surgery field 2834 and age field 3882, $n = 20,495$ and 20,001 with reported ages).

For individuals having indicated the use of hormone-replacement therapy (HRT; field 2814), HRT start and end ages were collated (fields 3536 and 3546, accordingly) across the different attended visits ($n = 98,104$). In cases where the reported chronological HRT age at later attended visits was greater than that at previous visits, the later instances were prioritised, i.e. as they would potentially indicate an updated use of HRT. In cases where different HRT ages were reported, but not in chronologically increasing order, the median age was kept.

Menopausal status was determined using data across instances (field 2724) and prioritizing the latest reported data, to account for changes in menopause status. For participants indicating having undergone menopause, their reported ages at menopause were collated (field 3581) using the same procedure as for HRT ages ($n = 158,264$).

Exclusions were then applied to this age at menopause, as follows:

- Participants reporting undergoing a hysterectomy and/or oophorectomy, but not the age at which this happened ($n = 958$ and 494, accordingly).
- Participants reporting multiple hysterectomy and/or oophorectomy ages, which were more than 10 years apart ($n = 38$ and 23, accordingly).
- Participants reporting multiple HRT start and/or end ages, which were not in chronologically ascending order and were more than 10 years apart ($n = 124$ and 137, accordingly).

Article

- Participants reporting multiple ages at menopause, which were not in chronologically ascending order and were more than 10 years apart ($n = 73$) and participants who reported both having and not having been through menopause and no other interventions ($n = 98$).
- Participants having undergone a hysterectomy or oophorectomy before or during the year they report undergoing menopause.
- Participants starting HRT prior to undergoing menopause and participants reporting HRT use, with no accompanying dates.

The resulting trait was representative of an ANM ($n = 115,051$) and was used in downstream analyses. Two additional ANM traits were also calculated, winsorized one by coding everyone reporting an ANM younger than 34, as 34 used in the discovery analysis as the primary phenotype ($n = 115,051$ total, reduced to 106,973 after covariate-resulting exclusions), and one by only including participants reporting ANM between 40 and 60, inclusive ($n = 104,506$), treated as a sensitivity analysis.

All manipulations were conducted in R (v4.1.2) on the UK Biobank RAP (<https://ukbiobank.dnanexus.com/>).

Replication of rare variant associations

Replication was performed using two study populations: the Icelandic deCODE study^{16,17} and the BRIDGES study¹⁸.

deCODE. The burden test associations are shown for three categories of rare variants with $MAF < 2\%$; (1) loss-of-function (LOF) variants; (2) combination of LOF variants and predicted deleterious missense variants; and (3) combination of LOF variants and missense variants with CADD score ≥ 25 . We furthermore show results for category 3 using a more stringent frequency threshold, 0.1%. We included missense variants predicted to cause LOF by two meta-predictors, MetaSVM and MetaLR⁷⁴, using variants available in dbSNFP v4.1c⁷⁵. We used VEP⁷⁰ to attribute predicted consequences to the variants sequenced. For case-control analyses, we used logistic regression and an additive model to test for association between LOF gene burdens and phenotypes, in which disease status was the dependent variable and genotype counts as the independent variable. Individuals were coded 1 if they carry any predicted LOF in the autosomal gene being tested and 0 otherwise. Age, sex and sequencing status of individuals was used as a covariate in the associations. For the analyses, we used software developed at deCODE genetics and we used linkage disequilibrium (LD) score regression intercepts⁷⁶ to adjust the χ^2 statistics and avoid inflation due to cryptic relatedness and stratification. Quantitative traits were analysed using a linear mixed model implemented in BOLT-LMM¹⁵. To estimate the quality of the sequence variants across the entire set we regressed the alternative allele counts (AD) on the depth (DP) conditioned on the genotypes (GT) reported by GraphTyper⁷⁷. For a well-behaving sequence variant, the mean alternative allele count for a homozygous reference genotype should be 0, for a heterozygous genotype it should be $DP/2$ and for homozygous alternative genotype it should be DP. Under the assumption of no sequencing or genotyping error, the expected value of AD should be DP conditioned on the genotype, in other words an identity line (slope 1 and intercept 0). Deviations from the identity line indicate that the sequence variant is spurious or somatic. We filter variants with slope less than 0.5. Additionally, GraphTyper employs a logistic regression model that assigns each variant a score (AAscore) predicting the probability that it is a true positive. We used only variants that have a AAscore > 0.8 .

BRIDGES. The BRIDGES study included women from studies participating in the Breast Cancer Association Consortium (BCAC; v14) (<http://bcac.ccge.medschl.cam.ac.uk/>). The subset of population or hospital-based studies sampled independently of family history, together with population-matched controls (25 studies) were included in the analyses. ANM (years) was obtained from baseline questionnaire

data. Women were considered as having experienced natural menopause if they indicated that the reason for menopause was reported as 'natural' or 'unknown'. Women were excluded from the analysis if the reason was indicated as either oophorectomy, hysterectomy, chemotherapy, stopping oral contraception or 'any other reason'. Only studies with information on year of birth and age at menopause, and only women with reported age at menopause between ages twenty-five years and sixty years were included. All studies were approved by the relevant ethical review boards and used appropriate consent procedures.

Targeted sequencing of germline DNA from participants for 35 known or suspected breast cancer genes was performed, including the coding sequence and splice sites. Details of library preparation, sequencing, variant calling, and quality control procedures are described in Dorling et al.¹⁸. Carriers of PTVs in more than one of five main breast cancer susceptibility genes (*BRCA1*, *BRCA2*, *ATM*, *CHEK2*, *PALB2*) were excluded. Carriers of pathogenic missense variants (as defined by Dorling et al.¹⁸) in *BRCA1* or *BRCA2* were also excluded.

We carried out burden analyses, assessing the associations between rare variants in aggregate and ANM using linear regression, adjusting for country of origin, breast cancer case-control status and year of birth (categorized as up to 1935, 1936–1945, 1946–1955 or after 1956), and for some analyses body mass index (BMI). For each gene we considered PTVs in aggregate. The primary analyses included covariates to adjust for population, which was defined by country, with the exception of Malaysia and Singapore, in which the three distinct ethnic groups (Chinese, Indian and Malay) were treated as different strata and the UK, which was treated as separate strata (SEARCH, from East Anglia and PROCAS from north-west England). Sensitivity analyses were carried out adjusting for BMI in women with recorded age at BMI, and among women without a diagnosis of breast cancer. Sensitivity analyses were also carried out defining non-carriers as women not harbouring PTVs in the five main genes or pathogenic MSVs in *BRCA1* and *BRCA2*.

WES sensitivity analysis using REGENIE

To replicate the primary findings and account for potential bias that could be introduced by exclusively using one discovery approach, a second analyst independently derived the age at menopause phenotype using a previously published method⁷⁸ and conducted additional burden association analysis using the REGENIE regression algorithm (REGENIEv2.2.4; <https://github.com/rgcgithub/regenie>). REGENIE implements a generalized mixed-model region-based association test that can account for population stratification and sample relatedness in large-scale analyses. REGENIE runs in two steps⁷⁹, which we implemented on the UK Biobank RAP. In the first step, genetic variants are aggregated into gene-specific units for each class of variant, called masks. We selected variants in CCDS transcripts deemed to be high confidence by LOFTEE⁷² with $MAF < 0.1\%$ and annotated using VEP⁷⁰. We created three masks, independently of the primary analysis group: (1) LOF variants (stop-gain, frameshift, or abolishing a canonical splice site (-2 or $+2$ bp from exon, excluding the ones in the last exon)) or missense variants with CADD score >30 ; (2) LOF or missense variants with CADD score >25 ; and (3) all missense variants. In the second step, the three masks were tested for association with ANM. We applied an inverse normal rank transformation to ANM and included recruitment centre, sequence batch and 40 principal components as covariates. For each gene, we present results for the transcript with the smallest burden P value. We performed a sensitivity analysis, excluding women who had any cancer diagnosis before ANM (ICD10 C00-C97 excluding C44, ICD9 140-208 excluding 173; $n = 2,585$). The results for the sensitivity analyses performed via REGENIE are available in Supplementary Tables 1 and 2.

Common variant GWAS lookups

Genes within 500 kb upstream and downstream of the 290 lead SNPs from the latest GWAS of ANM¹ were extracted from the exome-wide

analysis. There were a total of 2149 genes within the GWAS regions. Burden tests in these genes with a Bonferroni corrected P value of $<2.3 \times 10^{-5}$ (0.05/2,149) were highlighted. The results are available in Supplementary Table 6.

Phenome-wide association analysis

To test the association of ANM identified genes in other phenotypes, we processed additional reproductive ageing-related phenotypes, including age at menarche, cancer, telomere length and sex hormones. All tested phenotypes were run as either continuous (age at menarche, telomere length and sex hormones) or binary traits (cancer) corrected by age, age², sex, the first ten genetic principal components as calculated in Bycroft et al.⁶⁸, and study participant ES batch as a categorical covariate (either 50k, 200k or 450k). Phenotype definitions and processing used in this study are described in Supplementary Tables 8 and 9. Only the first instance (initial visit) was used for generating all phenotype definitions unless specifically noted in Supplementary Table 8. In case of cancer-specific analysis, data from cancer registries, death records, hospital admissions and self-reported were harmonized to ICD10 coding. If a participant had a code for any of the cancers recorded in ICD10 (C00-C97) then they were counted as a case for this phenotype. Minimal filtering was performed on the data, with only those cases where a diagnosis of sex-specific cancer was given in contrast to the sex data contained in UK Biobank record 31, was a diagnosis not used. For more information on cancer-specific analysis refer to Supplementary Tables 9 and 11.

Cancer PheWAS associations

To test for an association between genes we identified as associated with menopause timing (Fig. 2 and Supplementary Table 2) and 90 individual cancers as included in cancer registries, death records, hospital admissions and self-reported data provided by UK Biobank (for example, breast, prostate, etc.) we utilised a logistic model with identical covariates as used during gene burden testing ($n = 2430$ tests) (Supplementary Tables 9 and 11). As standard logistic regression can lead to inflated test statistic estimates in cases of severe case/control imbalance⁸⁰, we also performed a logistic regression with penalised likelihood estimation as described by Firth²⁶ (Supplementary Table 11). Models were run as discussed in Kosmidis et al.⁸¹ using the `brglm2` package implemented in R. `brglm2` was run via the `glm` function with default parameters other than “family” set to “binomial”, method set to “brglmFit”, and type set to “AS_mean”.

Expression in human female germ cells

We studied the mRNA abundance of WES genes during various stages of human female germ cell development using single-cell RNA sequencing data. We used the processed single cell RNA resequencing datasets from two published studies (Extended Data Figs. 6 and 7 and Supplementary Tables 17 and 18). This included single-cell RNA sequencing data from fetal primordial germ cells of human female embryos (accession code: GSE86146⁸²), and from oocyte and granulosa cell fractions during various stages of follicle development (accession code: GSE107746⁸³). A pseudo score of 1 was added to all values before log transformation of the dataset. The samples from fetal germ cells were categorized into sub-clusters as defined in the original study. The study by Li et al.⁸² identified 17 clusters by performing a t -distributed stochastic neighbour embedding analysis and using expression profiles of known marker genes for various stages of fetal germ cell development. In our analysis we have included four clusters of female fetal germ cells (mitotic, retinoic acid-responsive, meiotic and oogenesis) and four clusters containing somatic cells in the fetal gonads (endothelial, early_granulosa, mural_granulosa and late_granulosa). Software packages for R—tidyverse (<https://www.tidyverse.org/>), pheatmap, (<https://CRAN.R-project.org/package=pheatmap>) and reshape2 (<https://github.com/hadley/reshape>)—were used in processing and visualising the data.

De novo mutation rate analyses in 100kGP

Constructing PGSs. We calculated PGSs in participants from the rare disease programme of the 100kGP v14. There are 77,901 individuals in the aggregated variant calls (aggV2) after excluding participants whose genetically inferred sex is not consistent with their phenotypic sex. We restricted the PGS analysis to individuals of European ancestry, which was predicted by the Genomics England bioinformatics team using a random forest model based on genetic principal components generated by projecting aggV2 data onto the 1000 Genomes phase 3 principal component loadings. We removed one sample in each pair of related probands with kinship coefficient $> 1/(2^{4-5})$ —that is, up to and including third-degree relationships. Probands with the highest number of relatives were removed first. Similarly, we retained unrelated mothers and fathers of these unrelated probands. It left us with 8,089 mother–offspring duos and 8,029 father–offspring duos.

We used the lead variants (or proxies, as described below) for genome-wide significant loci previously reported for ANM¹ to calculate PGS in the parents. In 100kGP, we removed variants with MAF $< 0.5\%$ or missing rate $> 5\%$ from the aggV2 variants prepared by the Genomics England bioinformatics team. For lead variants that did not exist in 100kGP, we used the most significant proxy variants with $LD r^2 > 0.5$ if available in 100kGP. This resulted in a PGS constructed from 287 of the 290 previously reported loci. We regressed out 20 genetic principal components that were calculated within the European subset from the PGS and scaled the residuals to have mean = 0 and s.d. = 1. Higher PGS indicates later ANM.

Calling de novo mutations. De novo mutations (DNMs) were called using the Platypus variant caller in 10,478 parent offspring trios by the Genomics England Bioinformatics team. The detailed analysis pipeline is documented at: <https://research-help.genomicsengland.co.uk/display/GERE/De+novo+variant+research+dataset>. Extensive quality control and filtering were applied as described previously³¹. In brief, multiple filters were applied, including the following:

- the child had a heterozygous genotype and parents were homozygous reference.
- the parents had < 2 reads supporting the alternate allele.
- read depth > 20 in child and parents.
- variant allele fraction (VAF) > 0.3 and < 0.7 in the child.
- no DNMs were clustered (within 20 bp).

Autosomal de novo SNVs (dnSNVs) were phased using reads or read pairs that contained both the dnSNV and heterozygous variants located within 500 bp of it. A DNM was phased to a parent when the DNM appeared exclusively on the same haplotype as its nearby heterozygous variant. About one third of the dnSNVs were phased, of which three quarters were paternally phased (Extended Data Fig. 5 and Supplementary Table 14).

Associating the ANM PGS with DNMs in 100kGP. In association models, we accounted for parental age, the primary determinant of the number of DNMs, and various data quality metrics as described³¹:

- Mean coverage for the child, mother and father (child_mean_RD, mother_mean_RD, father_mean_RD).
- Proportion of aligned reads for the child, mother and father (child_prop_aligned, mother_prop_aligned, father_prop_aligned).
- Number of SNVs called for child, mother and father (child_SNVs, mother_SNVs, father_SNVs).
- Median variant allele fraction of DNMs called in child (median_VAF).
- Median Bayes factor as output by Platypus for DNMs called in the child. This is a metric of DNM quality (median_BF).

We first tested the association between parental PGSs and total de novo autosomal SNV count in the offspring in a Poisson regression with an identity link:

$$\begin{aligned} \text{dnSNVs_total} = & \beta_0 + \beta_1 \text{PGS}(\text{paternal or maternal}) \\ & + \beta_2 \text{paternal_age} + \beta_3 \text{maternal_age} + \beta_4 \text{child_mean_RD} \\ & + \beta_5 \text{mother_mean_RD} + \beta_6 \text{father_mean_RD} \\ & + \beta_7 \text{child_prop_aligned} + \beta_8 \text{mother_prop_aligne} \\ & + \beta_9 \text{father_prop_aligned} + \beta_{10} \text{child_snvs} \\ & + \beta_{11} \text{mother_snvs} + \beta_{12} \text{father_snvs} \\ & + \beta_{13} \text{mediam_VAF} + \beta_{14} \text{median_BF} \end{aligned}$$

We also fitted Poisson regression models to test the association between the PGS of one of the parents and the dnSNVs in the offspring that were phased to the relevant parent. We supplied 0.5 as the starting value for all coefficients when running the `glm()` function in R with an identity link. Using a different starting value (for example, 0.2 and 10) did not change the coefficient estimates.

The paternal model included paternal PGS, age and data quality metrics that are related to the proband and the father:

$$\begin{aligned} \text{dnSNVs_paternal} = & \beta_0 + \beta_1 \text{paternal_PGS} + \beta_2 \text{paternal_age} \\ & + \beta_3 \text{child_mean_RD} + \beta_4 \text{father_mean_RD} \\ & + \beta_5 \text{child_prop_aligned} + \beta_6 \text{father_prop_aligned} \\ & + \beta_7 \text{child_snvs} + \beta_8 \text{father_snvs} \\ & + \beta_9 \text{median_VAF} + \beta_{10} \text{median_BF} \end{aligned}$$

Similarly, the maternal model was as follows:

$$\begin{aligned} \text{dnSNVs_maternal} = & \beta_0 + \beta_1 \text{maternal_PGS} + \beta_2 \text{maternal_age} \\ & + \beta_3 \text{child_mean_RD} + \beta_4 \text{mother_mean_RD} \\ & + \beta_5 \text{child_prop_aligned} + \beta_6 \text{mother_prop_aligned} \\ & + \beta_7 \text{child_snvs} + \beta_8 \text{mother_snvs} \\ & + \beta_9 \text{median_VAF} + \beta_{10} \text{median_BF} \end{aligned}$$

Finally, as a check, we assessed the association between the maternal PGS and paternally phased dnSNVs, and vice versa, using a Poisson regression with an identity link where the same starting value for coefficients were supplied:

$$\begin{aligned} \text{dnSNVs_paternal} = & \beta_0 + \beta_1 \text{maternal_PGS} + \beta_2 \text{paternal_age} \\ & + \beta_3 \text{child_mean_RD} + \beta_4 \text{father_mean_RD} \\ & + \beta_5 \text{child_prop_aligned} + \beta_6 \text{father_prop_aligned} \\ & + \beta_7 \text{child_snvs} + \beta_8 \text{father_snvs} \\ & + \beta_9 \text{median_VAF} + \beta_{10} \text{median_BF} \end{aligned}$$

$$\begin{aligned} \text{dnSNVs_maternal} = & \beta_0 + \beta_1 \text{paternal_PGS} + \beta_2 \text{maternal_age} \\ & + \beta_3 \text{child_mean_RD} + \beta_4 \text{mother_mean_RD} \\ & + \beta_5 \text{child_prop_aligned} + \beta_6 \text{mother_prop_aligned} \\ & + \beta_7 \text{child_snvs} + \beta_8 \text{mother_snvs} \\ & + \beta_9 \text{median_VAF} + \beta_{10} \text{median_BF} \end{aligned}$$

Rare variant burden in ANM genes in 100kGP

We tested for association between the burden of rare coding variants in ANM-associated genes in mothers and fathers with phased de novo SNVs in their offspring. We extracted high-confidence PTVs annotated by LOFTEE in the nine genes associated with ANM at the exome-wide significance (Fig. 2), as well as damaging missense variants with CADD > 25 in *CHEK2* and *SAMHD1*. Annotations were extracted from VEP105. All variants had MAF < 0.1% in the 100kGP aggV2 dataset and in all sub-populations in gnomAD. We set to missing genotypes with depth < 7 and genotype quality < 20 or heterozygous genotypes with a binomial test $P < 0.001$.

We first regressed out the same covariates as described above in the PGS analysis from maternally phased DNMs using a linear regression. We applied a rank-based inverse normal transformation on the residuals and fitted a linear regression model of the adjusted DNMs on carrier status in unrelated mothers for each of the 9 genes that were associated with ANM, adjusting for 20 genetic principal components. We adjusted paternally phased DNMs and regressed it on gene burden similarly in unrelated fathers.

Mendelian randomization

Instrumental variable selection. Mendelian randomization analysis was applied to test whether the common variants associated with ANM¹ have a causal effect on DNM rates in the offspring (Supplementary Table 15). In this approach, genetic variants that are significantly associated with an exposure, in this case ANM, are used as instrumental variables to test the causality of that exposure on the outcome of interest, in this case DNM rate^{84–86}. For a genetic variant to be a reliable instrument, the following assumptions should be met: (1) the genetic instrument is associated with the exposure of interest; (2) the genetic instrument should not be associated with any other competing risk factor that is a confounder; and (3) the genetic instrument should not be associated with the outcome, except via the causal pathway that includes the exposure of interest^{84,87}. Genotypes at all variants were aligned to designate the ANM PGS-increasing alleles as the effect alleles as described above and this was used as a genetic instrument of interest. The effect sizes of genetic instruments (genotypes in the mother) on maternally phased de novo SNVs in the offspring estimated in 8,089 duos were obtained from Genomics England.

Mendelian randomization frameworks. The Mendelian randomization analysis was conducted using the inverse-variance weighted (IVW) model as the primary model due to the highest statistical power⁸⁸. However, as it does not correct for heterogeneity in outcome risk estimates between individual variants⁸⁹, we applied a number of sensitivity Mendelian randomization methods that better account for heterogeneity⁹⁰. These include Mendelian randomization Egger to identify and correct for unbalanced heterogeneity ('horizontal pleiotropy'), indicated by a significant Egger intercept ($P < 0.05$)⁹¹, and weighted median (WM) and penalized weighted median (PWM) models to correct for balanced heterogeneity⁹². In addition, we introduced the Mendelian randomization radial method to exclude variants from each model in cases where they are recognized as outliers⁹³. The results were considered as significant based on the P value significance consistency across different primary and sensitivity models applied. The results are available in Supplementary Table 15.

Analyses of DNMs in deCODE

Identifying DNMs. The genome of the Icelandic population was characterised by whole-genome sequencing of 63,460 Icelanders using Illumina standard TruSeq methodology to a mean depth of $35 \times$ (s.d. $8 \times$) with subsequent long-range phasing¹⁶. Analyses of DNMs were restricted to individuals with at least $20 \times$ coverage.

DNM candidates were called in 9,643 trios in a similar manner to that described previously^{32,33}, by comparing the genotypes of the parents and offspring. In brief, we defined a DNM candidate with permissive cut-offs for the genotype of the proband requiring that allele balance is greater than 25% and that there be at least 12 reads at the position (supporting either the reference or alternative allele). For the genotypes of the parents, we required at least 12 reads, maximum of one read supporting the alternative allele and the allelic balance to be less than 5%. Likely (N_{LIK}) and possible carriers (N_{POSS}) of the DNM allele outside the descendants of the parent pair were defined as before³². We restricted to DNM candidates with fewer than 50 likely carriers and either fewer than 10 possible carriers or with a ratio $N_{\text{LIK}}/N_{\text{POSS}}$ greater than 80%.

We tuned the DNM candidate filtering by using segregation of DNM candidates in three-generation families (2,042 probands), following³². We restricted to instances of the DNM candidates where we see both of the proband's haplotypes at a locus transmitted to the offspring of the proband (see fig. 1c in ref. 32). In brief, if the DNM is a true germline variant then the allele of the DNM candidate should be present in the offspring who inherited the haplotype on which it lies. On the other hand, if it is absent from the children, despite both the haplotypes of the proband having been transmitted to at least one offspring, this suggests that the DNM candidate is a false-positive DNM call (see more detailed description in ref. 32). As before³², we fitted the generalized additive model with a logistic link using the *mgcv* R package⁹⁴, using various functions of the following quality metrics as covariates, as indicated in the code:

- AAScore: prediction probability from Graphtyper that the variant is a true positive.
- Carrier_regression_beta: slope from the alternative allele depth regression for the sequence variant (described in the burden method section from deCODE).
- Carrier_regression_alpha: intercept from the alternative allele depth regression for the sequence variant (described in the burden method section from deCODE).
- Proband_het_AB: the allelic balance of the proband.
- MaxAAS: the maximum read support for the sequence variant across all individuals.
- Alignment_Alt_Reads: the number of reads supporting the alternative allele. This covariate and the following covariates were derived by identifying the reads in the BAM files supporting the de novo allele.
- Alignment_Alt_Unique_Positions: the unique number of starting positions for the reads supporting the alternative allele.
- Alignment_Alt_Soft_clipped: the number of soft clipped bases (S in CIGAR string).
- Alignment_Alt_Matched_bases: the number of matched bases (M in CIGAR string).
- Alignment_Alt_Score_diff: the difference in the alignment score between the best and the second best hit as reported by BWA mem.
- Alignment_Alt_Pair_sw_nm: the pairwise mismatches between reads supporting the alternative allele using the Smith Waterman implementation in SeqAn⁹⁵.
- Alignment_Alt_Pair_align: the number of bases in the pairwise alignments.

We fitted the following formula within the *gam()* function in the *mgcv* R package⁹⁴:

```
threegen_consistent_hs-
l(cut(alignment_Alt_Unique_Positions,c(-1,2,4,8,10,Inf)))+
s(l(AAScore))+
s(Carrier_regression_beta)+
s(Carrier_regression_alpha)+
l(ifelse(alignment_Alt_Reads>0,
(alignment_Alt_Score_diff/alignment_Alt_Reads)>10,
FALSE))+
l(ifelse(alignment_Alt_Pair_align>0,
(alignment_Alt_Pair_sw_nm/alignment_Alt_Pair_align)>0.05,
TRUE))+
l(ifelse(alignment_Alt_Matched_bases>0,
alignment_Alt_Soft_clipped/alignment_Alt_Matched_bases>0.5,
TRUE))+
s(Proband_het_AB)+
s(ifelse(MaxAAS>15,
16,
MaxAAS))+
l(NPOSS==0)
```

We then took the model learned using informative DNMs in these three-generation families and applied it to all remaining DNM

candidates. We retained candidate DNMs for which the predicted probability of being a real DNM based on this model was at least 50%.

To validate the false-positive detection rate of the DNMs, we also used the genotype consistency between pairs of monozygotic twins. We found that 3.8% of DNMs are unobserved in the monozygotic twin of the proband. These could either be false-positive DNM calls or high frequency post-zygotic mutations that differ between pairs of monozygotic twins⁹⁶.

To mirror the analysis of 100kGP, we phased the DNMs using read-backed phasing as previously described³².

Associating the ANM PGS with DNMs in deCODE. We calculated the raw ANM PGS per individual by multiplying the effect estimate by the count of the effect allele and summing the product across SNPs. We rank-transformed the raw PGS distribution across individuals to a standardized normal distribution. We fitted the following Poisson regression with identity link to assess the association between the mother's PGS and the number of maternally phased DNMs:

$$\text{dnSNVs_maternal} - \text{maternal_PGS} + \text{maternal_age} \\ + \text{paternal_coverage} + \text{maternal_coverage} \\ + \text{child_coverage} + \text{GAM_Predict_Mean}$$

where GAM_Predict_Mean is the average probability of the DNMs in the probands being real, calculated using the method described above.

We also fitted the following models as negative controls:

$$\text{dnSNVs_paternal} - \text{paternal_PGS} + \text{paternal_age} \\ + \text{paternal_coverage} + \text{maternal_coverage} \\ + \text{child_coverage} + \text{GAM_Predict_Mean}$$

$$\text{dnSNVs_paternal} - \text{maternal_PGS} + \text{paternal_age} \\ + \text{paternal_coverage} + \text{maternal_coverage} \\ + \text{child_coverage} + \text{GAM_Predict_Mean}$$

To fit these models, we randomly chose one offspring per family.

ANM genetic variants and DNMs in deCODE. We analysed variants with MAF < 2%. This was a higher threshold than had been used in 100kGP since our analyses of ANM associations indicated that this threshold appeared to be better powered within deCODE, possibly because some deleterious variants have risen to a higher frequency due to the Icelandic bottleneck. We focused on PTVs in the nine genes associated with ANM at the exome-wide significance (Fig. 2), as well as damaging missense variants with CADD > 25 in *CHEK2* and *SAMHD1*.

For maternally and paternally phased DNMs, we adjusted the number of DNM for parental ages at conception and normalized the trait using rank-based inverse normal transformation. A linear regression model was used to test for association between the transformed DNM rate and the burden genotypes, assuming the variance-covariance matrix to be proportional to the kinship matrix.

We used an inverse-variance weighted approach to meta-analyse the results from 100kGP and deCODE. A Bonferroni correction of 18 tests (9 genes and 2 parents) was applied.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in discovery analyses are available upon application from the UK Biobank study and Genomics England. Research on the

de-identified patient data used in this publication can be carried out in the Genomics England Research Environment subject to a collaborative agreement that adheres to patient led governance. All interested readers will be able to access the data in the same manner that the authors accessed the data. For more information about accessing the data, interested readers may contact research-network@genomicsengland.co.uk or access the relevant information on the Genomics England website: <https://www.genomicsengland.co.uk/research>. The deCODE dataset was used for replication purposes and only summary level results for the specific findings are provided.

67. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
68. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
69. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
70. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
71. Rentsch, P., Schubach, M., Shendure, J. & Kircher, M. CADD–Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. *Genome Med.* **13**, 31 (2021).
72. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
73. Seabold, S. P. J. Statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* (eds. van der Walt, S. & Millman, J.) 92–96 (2010).
74. Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* **37**, 235–241 (2016).
75. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 103 (2020).
76. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
77. Halldórsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
78. Ruth, K. S. et al. Events in early life are associated with female reproductive ageing: a UK Biobank study. *Sci. Rep.* **6**, 24710 (2016).
79. Mbatchou, J. et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
80. Ma, C., Blackwell, T., Boehnke, M. & Scott, L. J. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **37**, 539–550 (2013).
81. Kosmidis, I. K. P. E. C. S. N. Mean and median bias reduction in generalized linear models. *Stat. Comput.* **30**, 43–59 (2019).
82. Li, L. et al. Single-cell RNA-seq analysis maps development of human germline cells and gonadal niche interactions. *Cell Stem Cell* **20**, 858–873.e4 (2017).
83. Zhang, Y. et al. Transcriptome landscape of human folliculogenesis reveals oocyte and granulosa cell interactions. *Mol. Cell* **72**, 1021–1034.e4 (2018).
84. Smith, G. D. & Ebrahim, S. ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
85. Burgess, S., Foley, C. N., Allara, E., Staley, J. R. & Howson, J. M. M. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat. Commun.* **11**, 376 (2020).
86. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Smith, G. D. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
87. Burgess, S. et al. Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res.* **4**, 186 (2020).
88. Slob, E. A. W. & Burgess, S. A comparison of robust Mendelian randomization methods using summary data. *Genet. Epidemiol.* **44**, 313–329 (2020).
89. Bowden, J. et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).
90. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* **28**, 30–42 (2017).
91. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
92. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
93. Bowden, J. et al. Improving the visualization, interpretation and analysis of two-sample summary data Mendelian randomization via the radial plot and radial regression. *Int. J. Epidemiol.* **47**, 1264–1278 (2018).
94. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B* **73**, 3–36 (2011).
95. Döring, A., Weese, D., Rausch, T. & Reinert, K. SeqAn: An efficient, generic C++ library for sequence analysis. *BMC Bioinformatics* **9**, 11 (2008).
96. Jonsson, H. et al. Differences between germline genomes of monozygotic twins. *Nat. Genet.* **53**, 27–34 (2021).

Acknowledgements This work was funded by the Medical Research Council (Unit programmes: MC_UU_12015/2, MC_UU_00006/2, MC_UU_12015/1, and MC_UU_00006/1). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. This research was conducted using the UK Biobank Resource under application 9905 (University of Cambridge) and 9072 and 871 (University of Exeter). A full list of funding sources and acknowledgements can be found in the Supplementary Information. This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

Author contributions All authors were involved in reviewing and editing the manuscript. The study was conceived by A.M., J.R.B.P., H.C.M. and K.S. Lead analysts were S. Stankovic, Q.Q.H., E.V.I. and S. Shekari, with additional analyses by E.J.G. and N.D.L.O. Exome sequencing pipelines were developed by S. Stankovic, E.J.G., G.H., K.A.K., R.N.B., F.R.D., Y.Z., K.K., A.R.W., M.N.W., C.F.W., K.S.R., K.K.O., J.R.B.P. and A.M. Analyses of de novo mutation rate in 100kGP were carried out by Q.Q.H. and overseen by M.E.H. and H.C.M. Replication of exome sequencing findings and de novo mutation rates in deCODE was carried out by E.V.I., H.J., T.R., V.T., G.S., A.O., U.S., J.G., S.N.S., D.F.G., P.S. and K.S. Replication of exome sequencing in Breast Cancer Association consortium was carried out by N.M. Human ovarian expression analyses were led by A.A. and E.R.H.

Competing interests J.R.B.P. and E.J.G. are employees of Insmad Innovation UK and hold stock/stock options in Insmad. J.R.B.P. has also received consultancy fees from Fertility Health and WW International and holds research funding from GSK. The other authors declare no competing interests.

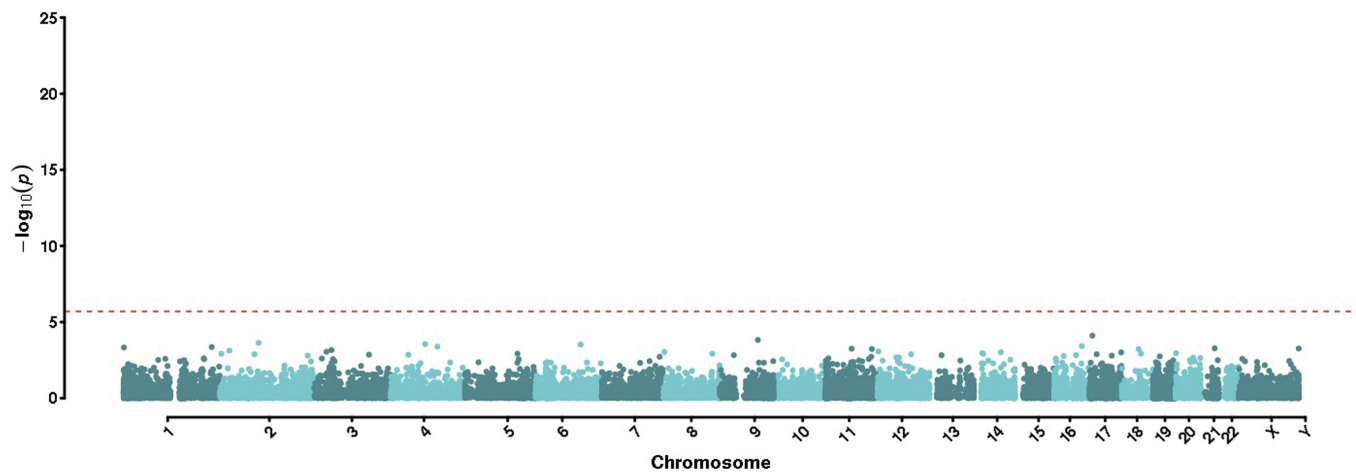
Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-024-07931-x>.

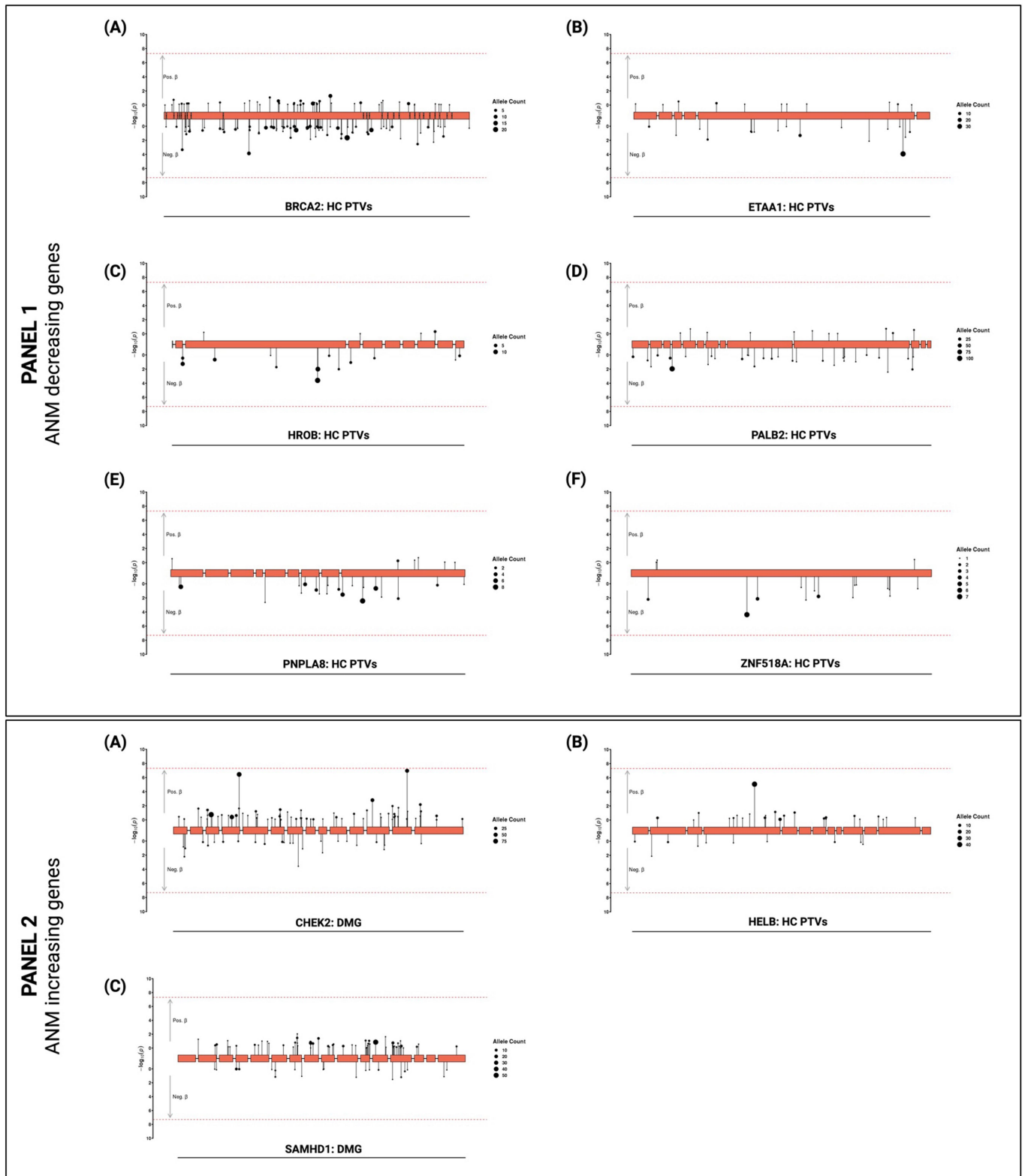
Correspondence and requests for materials should be addressed to John R. B. Perry or Anna Murray.

Peer review information Nature thanks Anne Goriely, Diana Laird and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

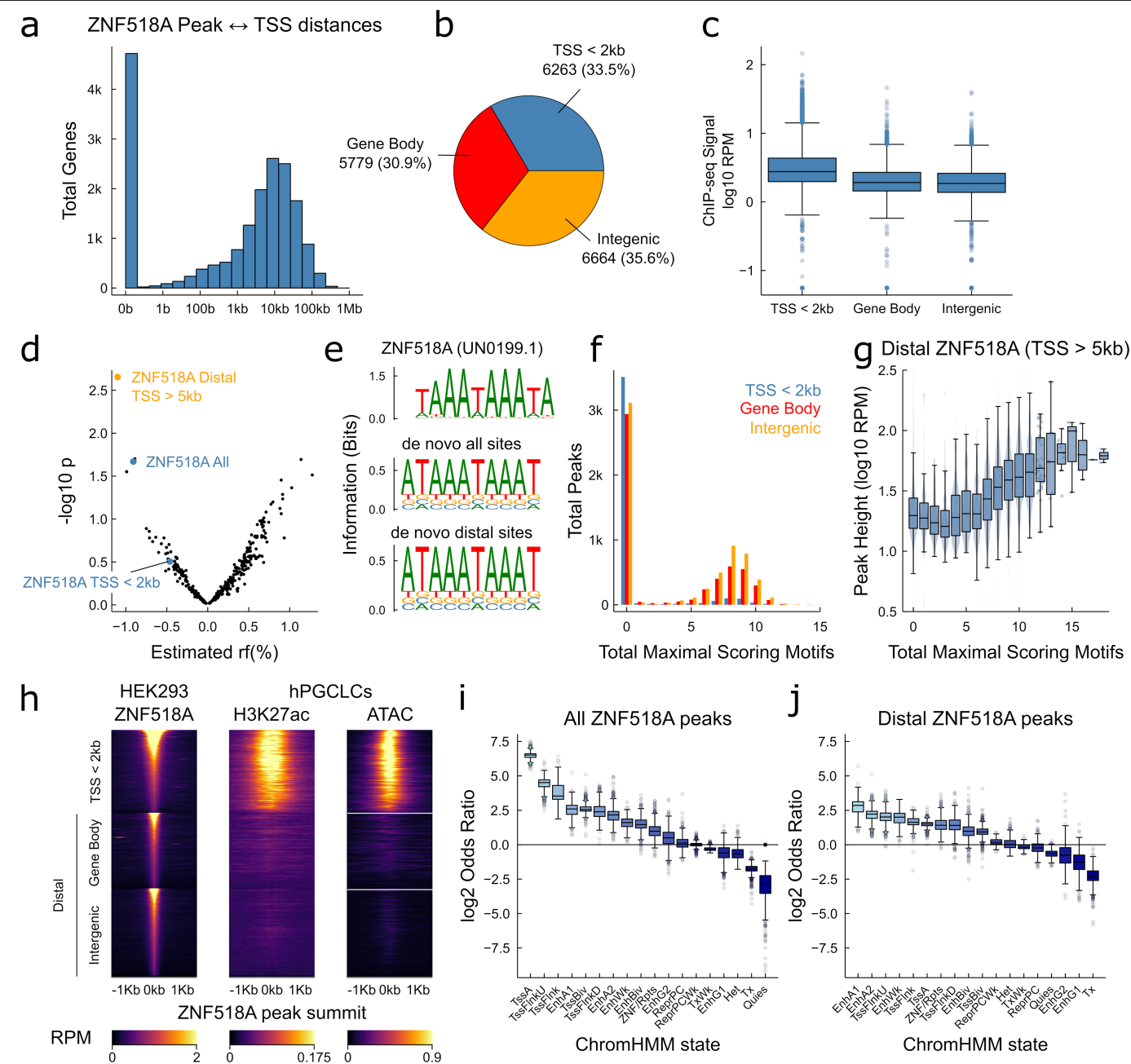


Extended Data Fig. 1 | Exome-wide association results for synonymous variants. Plotted are per-gene burden results for synonymous variants. The red line indicates the exome-wide significant P value after Bonferroni correction of 1.08×10^{-6} .



Extended Data Fig. 2 | Lollipop plots show the variants, clustered for the best performing functional mask in a gene, which went into the gene burden test for ANM using BOLT-LMM. The arrows pointing upwards represent the variants positively associated with ANM, while the ones pointing downwards show the negatively associated variants. The size of the point indicates the allele count in carriers. Panel 1: Variant level associations for ANM decreasing WES genes. (A) *BRCA2* HC PTV mask (genomic size = 84,761 bp, coding sequence = 10,254 bp); (B) *ETAA1*, HC PTV mask (genomic size = 14,757 bp, coding sequence = 2,778 bp); (C) *HROB*, HC PTV mask (genomic size = 20,547 bp,

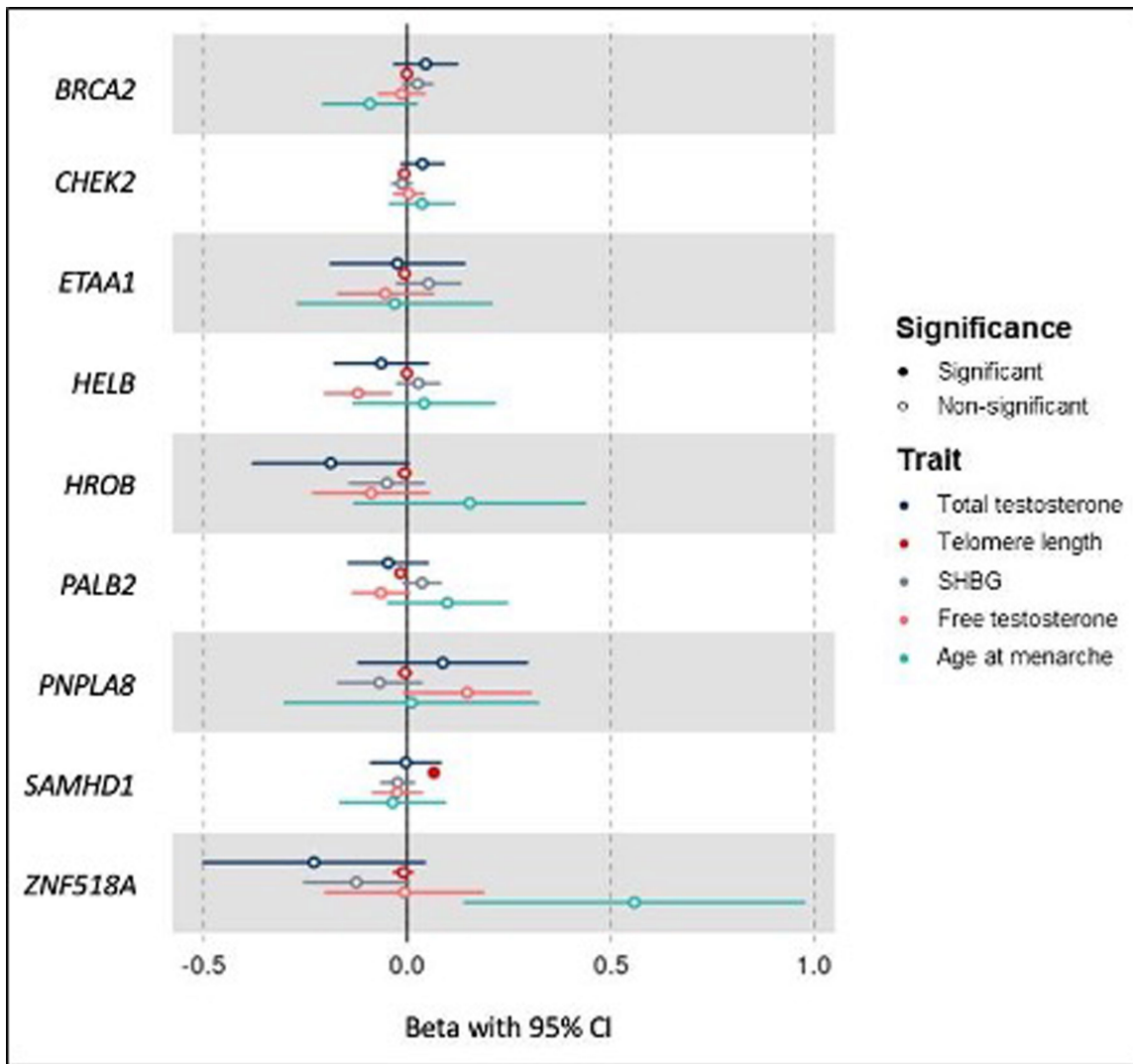
coding sequence = 1,938 bp); (D) *PALB2*, HC PTV mask (genomic size = 38,146 bp, coding sequence = 3,558 bp); (E) *PNPLA8*, HC PTV mask (genomic size = 55,719 bp, coding sequence = 2,346 bp); and (F) *ZNF518A*, HC PTV mask (genomic size = 33,463 bp, coding sequence = 4,449 bp). Panel 2: Variant level associations for ANM increasing WES genes. (A) *CHEK2*, damaging mask (genomic size = 54,078 bp, coding sequence = 1,629 bp); (B) *HELB*, HC PTV mask (genomic size = 35,707 bp, coding sequence = 3,261 bp); and (C) *SAMHD1*, damaging mask (genomic size = 61,480 bp, coding sequence = 1,878 bp).



Extended Data Fig. 3 | Functional analysis of ZNF518A bound loci.

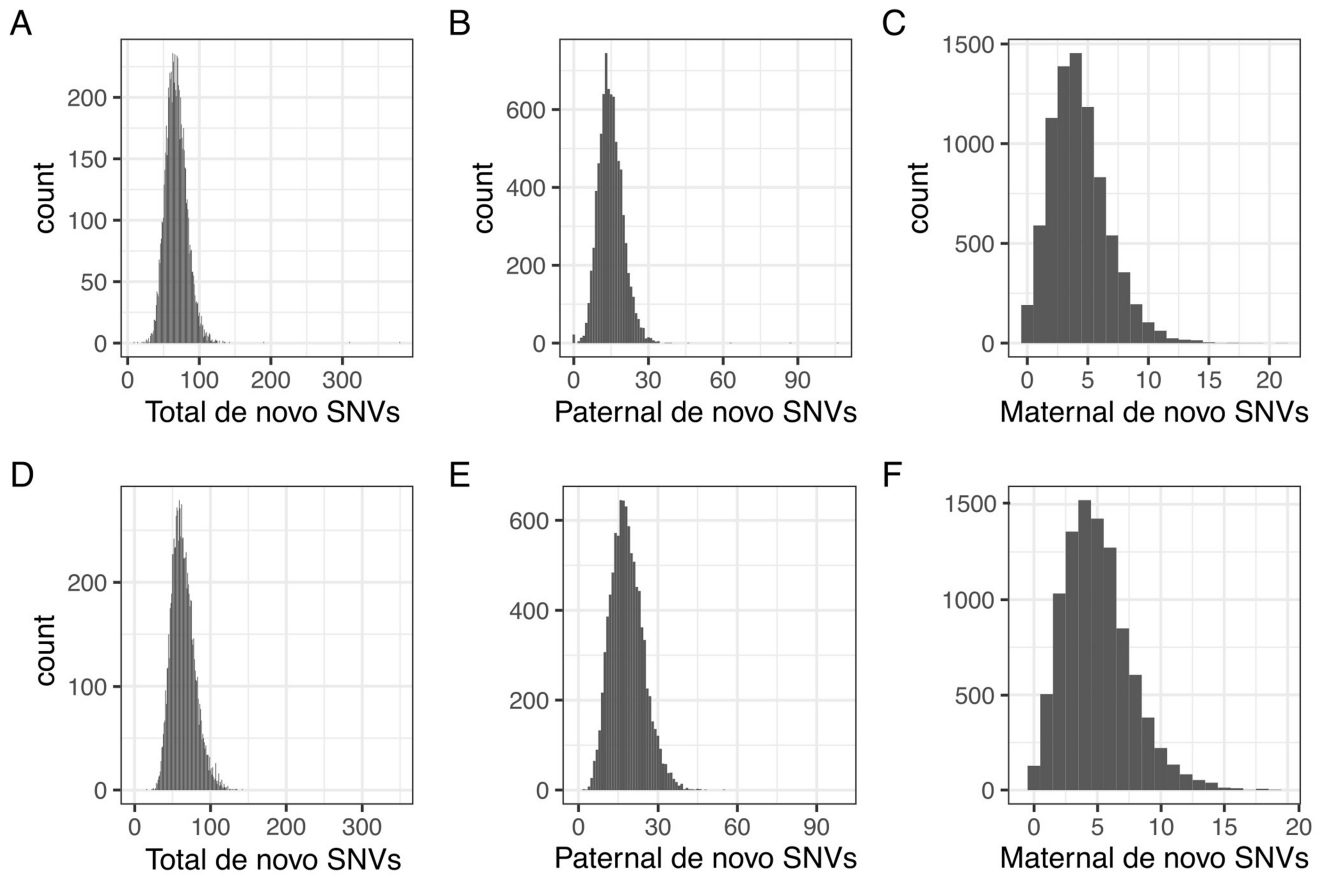
(a) Histogram of log₁₀-scale distances between *ZNF518A* and nearest gene transcription start site (TSS). **(b)** Proportion of *ZNF518A* peaks falling proximal to TSS (TSS < 2 kb), within gene bodies and in intergenic regions. **(c)** Boxplots showing total normalised reads per million (RPM) for every peak for categories TSS < 2 kb, gene body and intergenic - *ZNF518A* peaks have greater signal at proximal to TSS. **(d)** SLDP association between ANM GWAS variants and *ZNF518A* peaks, stratified by all peaks, proximal (<2 kb) from a TSS, and distal (>5 kb) from a TSS. The association between ANM variants and *ZNF518A* peaks appears due to distal *ZNF518A* peaks (either gene body or intergenic, >5 kb TSS) and not proximal TSS binding. Numerical results are reported in Supplementary Table 7. **(e)** De novo motif discovery recovers unvalidated JASPAR motif for *ZNF518A* UN0199.1. Homer enrichment statistics: all sites $P = 10^{-6451}$ motif in 31.2% of targets (1.15% background); distal sites $P = 10^{-4590}$ motif in 47.3% of targets (1.81% background). **(f)** Proportion of maximal scoring instances of UN0199.1 (sequences

that exactly match motif consensus) by *ZNF518A* peak category. Many distal peaks contain multiple perfect instances of the motif. **(g)** Boxplots, violin plots and dot plots depicting the relationship between *ZNF518A* ChIP-seq peak height and number of maximal scoring motifs present in peak. A strong relationship between peak height and number of motif instances can be observed. **(h)** Heatmaps depicting *ZNF518A* ChIP-seq, H3K27ac ChIP-seq in hPGCLCs, and chromatin accessibility by ATAC-seq in hPGCLCs. Signal shown over all *ZNF518A* peaks in RPM +/- 1 kb of *ZNF518A* peak summit. *ZNF518A* bound promoters (TSS < 2 kb) are accessible and are marked with H3K27ac, distal regions either in gene bodies or intergenic regions show no H3K27ac or chromatin accessibility, suggestive that *ZNF518A* represses these regulatory regions. **(i, j)** Association shown in odds ratios of ChromHMM states over 833 tissues/cell types from Epimap; boxplots with outliers shown, with each boxplot summarising the distribution of associations over all tissues/cell types for a given chromatin state. **(i)** All *ZNF518A* peaks; **(j)** *ZNF518A* peaks distal from TSS.



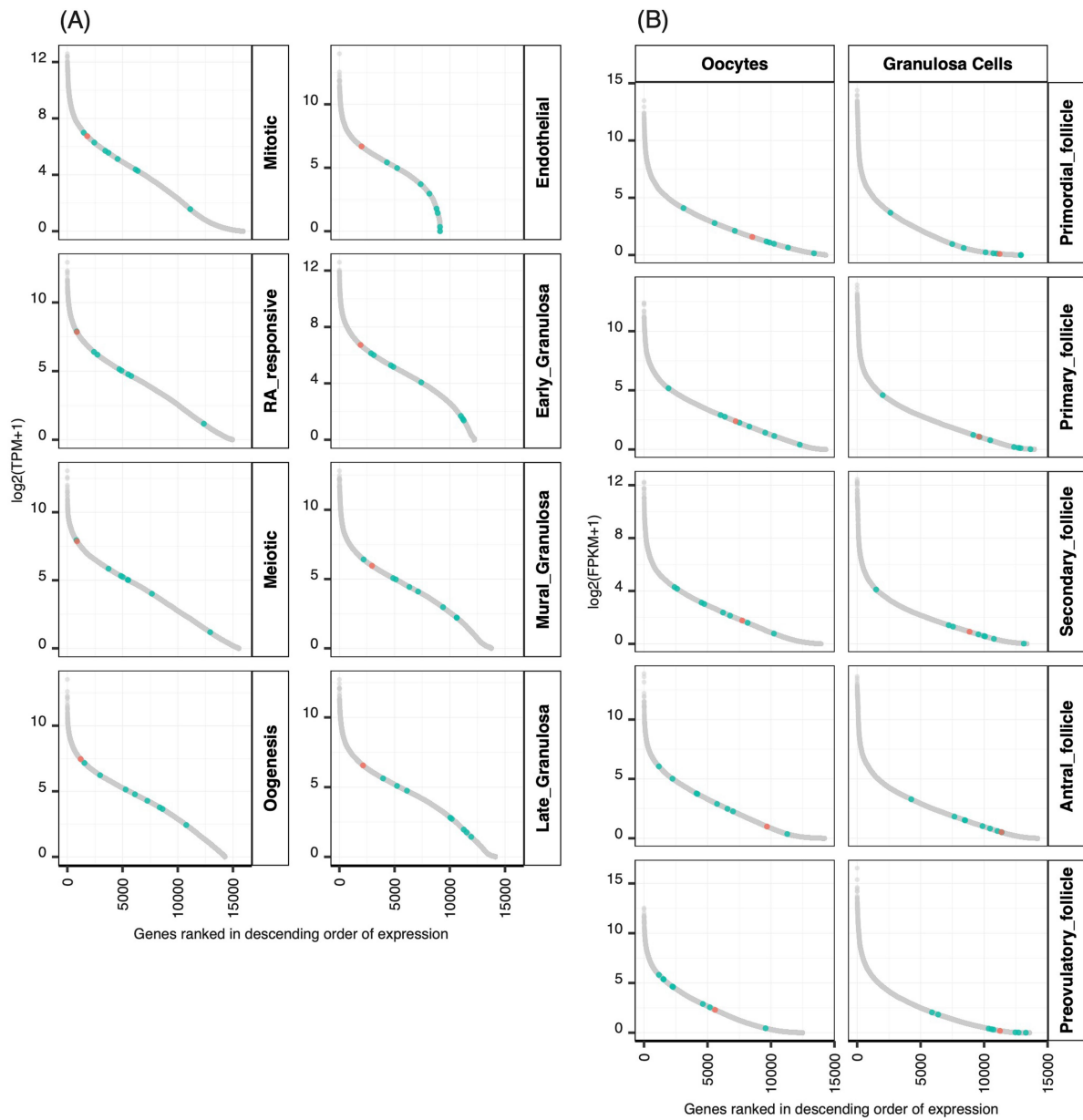
Extended Data Fig. 4 | ANM gene burden associations with reproductive ageing-related traits of interest in females only. The coefficients and 95% CIs were female-specific and plotted for the quantitative traits only. The association

was tested using BOLT-LMM. Male-specific and sex combined associations could be found in Supplementary Table 10.



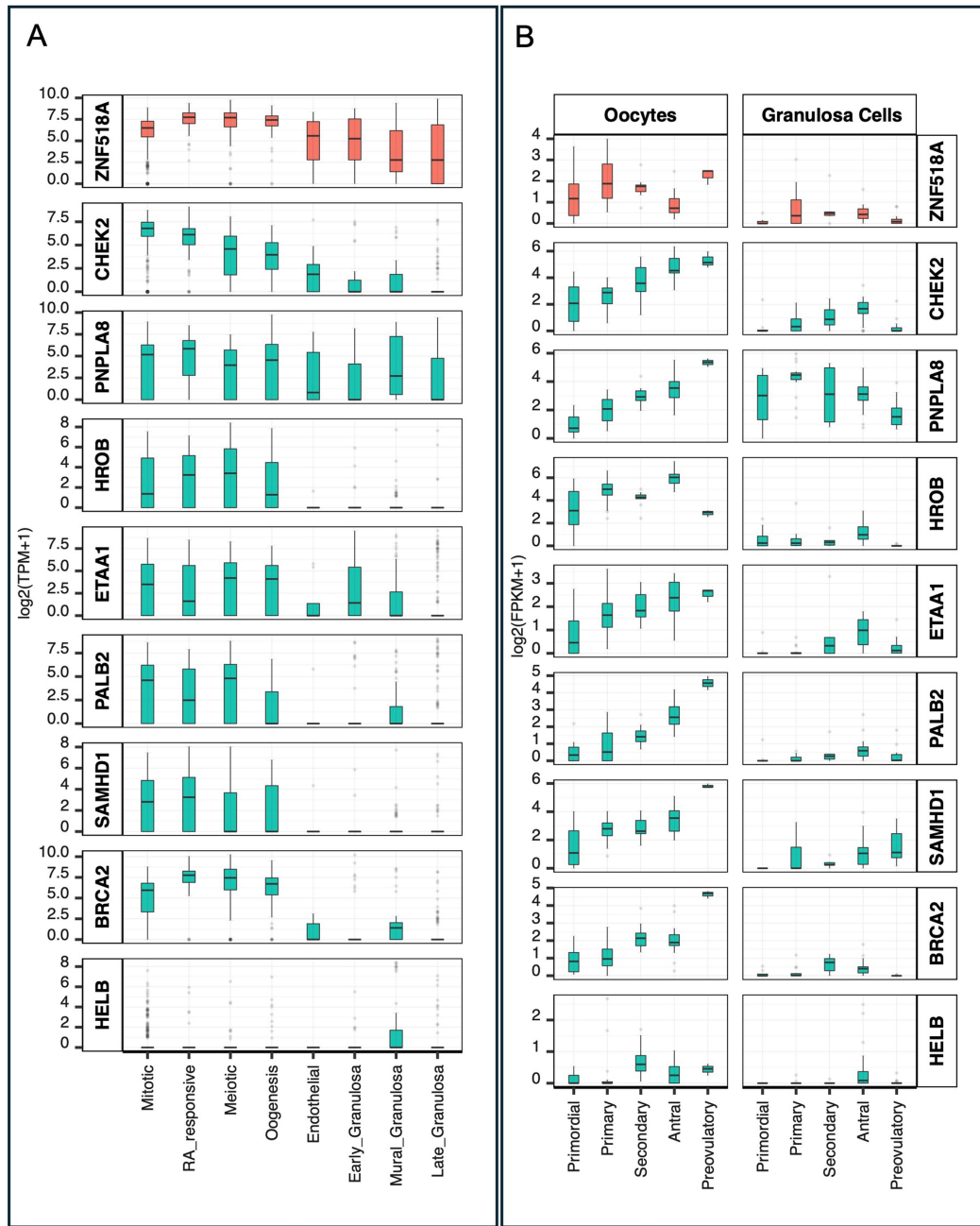
Extended Data Fig. 5 | Distribution of de novo single nucleotide variants (dnSNVs). The histogram shows the number of (A) total dnSNVs, includes both phased (maternal + paternal) and unphased DNMs, (B) paternally derived

dnSNVs and (C) maternally derived dnSNVs in unrelated probands with European ancestry from the 100,000 Genomes Project. (D-F) show similar distributions in individuals from deCODE.



Extended Data Fig. 6 | Expression levels of genes across various stages of female germ cell development. In the X-axis, genes are ranked according to their average expression at each stage (Y-axis) **(A)** in human foetal primordial germ cells and **(B)** in granulosa cells in adult follicles. Genes identified as novel

ANM genes in WES analysis are coloured in green and all other genes in the genome are in grey. *ZNF518A* is depicted in orange for the ease of comparison with other genes (Supplementary Table 17).



Extended Data Fig. 7 | mRNA expression of WES genes during foetal stages and folliculogenesis. Box and whisker plots of mRNA expression of the WES genes at different stages of germ cell development. The plots represent the interquartile range of TPM values, the line at the centre of the box represents the median, error bars indicate the 95% CI and outliers are shown as dots.

(A) Sub-clusters from single foetal cells from week 5 to 26 post-fertilisation are on the X-axis with the average TPM expression values $\log_2(\text{TPM} + 1)$ on the Y-axis. **(B)** Different stages of folliculogenesis in oocytes and granulosa cells are represented on the X-axis with their average expression values $\log_2(\text{FPKM} + 1)$ on the Y-axis (Supplementary Table 18).

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| n/a | Confirmed |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No data was collected
Data analysis	<p>The variant annotation was performed using the ENSEMBL Variant Effect Predictor (VEP) v104 Rare variant association testing was performed in BOLT-LMM v2.3. and REGENIEv2.2.4 De novo mutations (DNMs) were called using the Platypus variant caller. The commit version used: bayesiandenovofilter.py python script: 414ca566f8b2269d0caae726b21f87e03783ca1a (latest at time of analysis).</p>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data used in discovery analyses are available upon application from the UK Biobank study and Genomics England. Research on the de-identified patient data

used in this publication can be carried out in the Genomics England Research Environment subject to a collaborative agreement that adheres to patient led governance. All interested readers will be able to access the data in the same manner that the authors accessed the data. For more information about accessing the data, interested readers may contact research-network@genomicsengland.co.uk or access the relevant information on the Genomics England website: <https://www.genomicsengland.co.uk/research>.

The deCODE dataset was used for replication purposes and only summary level results for the specific findings are provided.

All datasets are available through an application process to individual studies.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	The study was restricted to participants of female sex as defined by genetic markers, as only biological females undergo menopause, the trait being investigated.
Reporting on race, ethnicity, or other socially relevant groupings	We restricted our analyses to individuals of European ancestry as defined by genetic principle component analysis.
Population characteristics	Data from participants aged 40-70 from the UK Biobank were included in our analyses. We did not include disease or treatment characteristics as covariates.
Recruitment	Volunteers to UK Biobank
Ethics oversight	UK Biobank has ethics approval from the North West MREC under application number 21/NW/0157

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All genetic females with natural age at menopause in Uk Biobank
Data exclusions	None
Replication	Replication was performed using two study populations - the Icelandic deCODE study and the BRIDGES study. We were able to replicate the novel genetic associations with ANM , but we were not able to replicate the de novo mutation discovery from 100KGP in deCODE
Randomization	The principle exposure in this study is naturally occurring genetic variants, meaning that we were unable to randomise the individuals in the study. To account for possible confounding we used a linear mixed model and adjusted for technical and demographic covariates.
Blinding	Blinding is not applicable to this study, as it is a genome-wide study of rare coding variants and not a randomised controlled trial. We did not deliver any intervention to participants in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

- | n/a | Involvement |
|-------------------------------------|--------------------------------------------------------|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Antibodies |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Eukaryotic cell lines |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Palaeontology and archaeology |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Animals and other organisms |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Clinical data |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Dual use research of concern |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Plants |

- | n/a | Involvement |
|-------------------------------------|-------------------------------------------------|
| <input checked="" type="checkbox"/> | <input type="checkbox"/> ChIP-seq |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> Flow cytometry |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> MRI-based neuroimaging |

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.