

# Leveraging Northern European population history: novel low-frequency variants for polycystic ovary syndrome

Jaakko S. Tyrmä  <sup>1,2,3,\*†</sup>, Riikka K. Arffman  <sup>4,†</sup>, Natàlia Pujol-Gualdo  <sup>4,5,†</sup>, Venla Kurra  <sup>6</sup>, Laure Morin-Papunen  <sup>4</sup>, Eeva Sliz  <sup>1,2,3</sup>, FinnGen Consortium, Estonian Biobank Research Team, Terhi T. Piltonen  <sup>4</sup>, Triin Laisk  <sup>5</sup>, Johannes Kettunen  <sup>1,2,3,7,‡</sup>, and Hannele Laivuori  <sup>8,9,10,‡</sup>

<sup>1</sup>Computational Medicine, Faculty of Medicine, University of Oulu, Oulu, Finland; <sup>2</sup>Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland; <sup>3</sup>Biocenter Oulu, University of Oulu, Oulu, Finland; <sup>4</sup>Department of Obstetrics and Gynecology, PEDEGO Research Unit, Medical Research Centre, Oulu University Hospital, University of Oulu, Oulu, Finland; <sup>5</sup>Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu, Estonia; <sup>6</sup>Department of Clinical Genetics, Faculty of Medicine and Health Technology, Tampere University Hospital and Tampere University, Tampere, Finland; <sup>7</sup>Finnish Institute for Health and Welfare, Helsinki, Finland; <sup>8</sup>Department of Obstetrics and Gynecology, Faculty of Medicine and Health Technology, Tampere University Hospital and Tampere University, Tampere, Finland; <sup>9</sup>Medical and Clinical Genetics, University of Helsinki and Helsinki University Hospital, Helsinki, Finland; and <sup>10</sup>Institute for Molecular Medicine Finland, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland

\*Correspondence address. Center for Life Course Health Research, Faculty of Medicine, Aapistie 5A, PO Box 5000, 90014 University of Oulu, Oulu, Finland. E-mail: jaakko.tyrmä@oulu.fi  <https://orcid.org/0000-0002-4757-6563>

Submitted on July 01, 2021; resubmitted on October 07, 2021; editorial decision on October 17, 2021

**STUDY QUESTION:** Can we identify novel variants associated with polycystic ovary syndrome (PCOS) by leveraging the unique population history of Northern Europe?

**SUMMARY ANSWER:** We identified three novel genome-wide significant associations with PCOS, with two putative independent causal variants in the checkpoint kinase 2 (*CHEK2*) gene and a third in myosin X (*MYO10*).

**WHAT IS KNOWN ALREADY:** PCOS is a common, complex disorder with unknown aetiology. While previous genome-wide association studies (GWAS) have mapped several loci associated with PCOS, the analysis of populations with unique population history and genetic makeup has the potential to uncover new low-frequency variants with larger effects.

**STUDY DESIGN, SIZE, DURATION:** A population-based case–control GWAS was carried out.

**PARTICIPANTS/MATERIALS, SETTING, METHODS:** We identified PCOS cases from national registers by ICD codes (ICD-10 E28.2, ICD-9 256.4, or ICD-8 256.90), and all remaining women were considered controls. We then conducted a three-stage case–control GWAS: in the discovery phase, we had a total of 797 cases and 140 558 controls from the FinnGen study. For validation, we used an independent dataset from the Estonian Biobank, including 2812 cases and 89 230 controls. Finally, we performed a joint meta-analysis of 3609 cases and 229 788 controls from both cohorts. Additionally, we reran the association analyses including BMI as a covariate, with 2169 cases and 160 321 controls from both cohorts.

**MAIN RESULTS AND THE ROLE OF CHANCE:** Two out of the three novel genome-wide significant variants associating with PCOS, rs145598156 ( $P=3.6\times 10^{-8}$ , odds ratio (OR)=3.01 [2.02–4.50] minor allele frequency (MAF)=0.005) and rs182075939 ( $P=1.9\times 10^{-16}$ , OR=1.69 [1.49–1.91], MAF=0.04), were found to be enriched in the Finnish and Estonian populations and are tightly linked to a deletion c.1100delC ( $r^2=0.95$ ) and a missense I157T ( $r^2=0.83$ ) in *CHEK2*. The third novel association is a common variant near *MYO10* (rs9312937,  $P=1.7\times 10^{-8}$ , OR=1.16 [1.10–1.23], MAF=0.44). We also replicated four previous reported associations near the genes Erb-B2 Receptor Tyrosine Kinase 4 (*ERBB4*), DENN Domain Containing 1A (*DENND1A*), FSH Subunit Beta (*FSHB*) and Zinc Finger And BTB Domain Containing 16 (*ZBTB16*). When adding BMI as a covariate only one of the novel variants remained genome-

<sup>†</sup>The first three authors contributed equally to this work.

<sup>‡</sup>The last two authors contributed equally to this work.

© The Author(s) 2021. Published by Oxford University Press on behalf of European Society of Human Reproduction and Embryology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

wide significant in the meta-analysis (the EstBB lead signal in *CHEK2* rs182075939,  $P = 1.9 \times 10^{-16}$ , OR = 1.74 [1.5–2.01]) possibly owing to reduced sample size.

**LARGE SCALE DATA:** The age- and BMI-adjusted GWAS meta-analysis summary statistics are available for download from the GWAS Catalog with accession numbers GCST90044902 and GCST90044903.

**LIMITATIONS, REASONS FOR CAUTION:** The main limitation was the low prevalence of PCOS in registers; however, the ones with the diagnosis most likely represent the most severe cases. Also, BMI data were not available for all (63% for FinnGen, 76% for EstBB), and the biobank setting limited the accessibility of PCOS phenotypes and laboratory values.

**WIDER IMPLICATIONS OF THE FINDINGS:** This study encourages the use of isolated populations to perform genetic association studies for the identification of rare variants contributing to the genetic landscape of complex diseases such as PCOS.

**STUDY FUNDING/COMPETING INTEREST(S):** This work has received funding from the European Union's Horizon 2020 research and innovation programme under the MATER Marie Skłodowska-Curie grant agreement No. 813707 (N.P.-G., T.L., T.P.), the Estonian Research Council grant (PRG687, T.L.), the Academy of Finland grants 315921 (T.P.), 321763 (T.P.), 297338 (J.K.), 307247 (J.K.), 344695 (H.L.), Novo Nordisk Foundation grant NNF17OC0026062 (J.K.), the Sigrid Juselius Foundation project grants (T.L., J.K., T.P.), Finska Läkaresällskapet (H.L.) and Jane and Aatos Erkko Foundation (H.L.). The funders had no role in study design, data collection and analysis, publishing or preparation of the manuscript. The authors declare no conflicts of interest.

**Key words:** genome-wide association study / rare variants / polycystic ovary syndrome / checkpoint kinase 2 / myosin X

## Introduction

Polycystic ovary syndrome (PCOS) is a common, multifaceted endocrine disorder. The international evidence-based guideline recommends using the Rotterdam criteria for PCOS diagnosis, requiring the presence of at least two of the following symptoms: oligo- or anovulation, clinical or biochemical hyperandrogenism, or polycystic ovaries seen in ultrasound, after exclusion of related disorders (Teede *et al.*, 2018). The criteria result in a prevalence as high as 18% for PCOS among fertile-aged women and produce several phenotypes (March *et al.*, 2010; Skiba *et al.*, 2018).

PCOS is the most common cause for anovulatory infertility, caused by disrupted follicle development owing to dysregulation of the hypothalamus–pituitary axis. This results in follicle arrest and an increase in the number of antral follicles in the ovaries, as well as a 2- to 3-fold increase in levels of anti-Müllerian hormone (AMH) (Silva and Giacobini, 2021). Ovulatory dysfunction often subsides with age; however, women with PCOS still display higher AMH and later onset of menopause (Piltonen *et al.*, 2005; Li *et al.*, 2016; de Ziegler *et al.*, 2018; Minooe *et al.*, 2018; Forslund *et al.*, 2019). In addition to the reproductive features, PCOS is also characterized by metabolic disturbances such as obesity, insulin resistance and dyslipidemia (Ollila *et al.*, 2016; Lim *et al.*, 2019; Barber and Franks, 2021). Women with PCOS also have an increased risk for endometrial cancer; however, the majority of studies do not indicate a higher susceptibility to other types of cancer (Dumesic and Lobo, 2013; Barry *et al.*, 2014; Gottschau *et al.*, 2015; Ding *et al.*, 2018).

Despite the high prevalence of the syndrome, the origins of PCOS remain unknown. Considering the complex nature of the syndrome, it is likely that both genetic and environmental factors contribute to its development (Abbott *et al.*, 2019; Koivuaho *et al.*, 2019; Moghetti and Tosi, 2021).

Notably, the heritability of PCOS is estimated to be around 70% (Vink *et al.*, 2006; Risal *et al.*, 2019). To elucidate the genetic architecture of PCOS, several genome-wide association studies (GWAS) and meta-analysis studies have been conducted, identifying over 20 susceptibility loci for PCOS (Chen *et al.*, 2011; Shi *et al.*, 2012; Day *et al.*,

2015, 2018; Hayes *et al.*, 2015; Lee *et al.*, 2015; Dapas *et al.*, 2020; Hong *et al.*, 2020; Zhang *et al.*, 2020). The identified loci indicate roles in PCOS for gonadotrophin signalling, folliculogenesis, epithelial growth factor signalling, DNA repair and structure, cell cycle and proliferation, and androgen biosynthesis. However, these common genetic variants explain only around 10% of the heritability (Azziz, 2016). Thus, it has been suggested that rare variants with larger effect sizes may contribute to the heritability of PCOS (Dapas and Dunaif, 2020). Nevertheless, the identification of these may be difficult in data sets with large genetic variations.

The value of studying genetic isolates, such as the Finnish population, has been accepted for decades (Martin *et al.*, 2018). Such populations provide an excellent opportunity to facilitate the discovery of rare variants with larger effects and characterize the genetic basis of complex diseases such as PCOS. The Finnish population originates from a small founder population with several bottleneck events over centuries, followed by genetic drift. These events have led to an enrichment of many low-frequency variants almost absent in most European populations (1000 Genomes Project Consortium *et al.*, 2012; Nelis *et al.*, 2009; Locke *et al.*, 2019). Replication of association results may be difficult when studying isolated populations, but for Finns, the genetically closest Estonian population provides a natural comparison (Nelis *et al.*, 2009; Tambets *et al.*, 2018).

In this study, we first utilized genome-wide association analyses and data from the FinnGen project and the Estonian Biobank (EstBB) to detect novel PCOS-associated variants in these population isolates. Furthermore, as several studies suggest a causal role for obesity in PCOS (Legro, 2012; Brower *et al.*, 2019; Zhao *et al.*, 2020), we examined the influence of BMI on the detected associations with PCOS.

As a result, we unravelled two rare, population-enriched variants located in the checkpoint kinase 2 (*CHEK2*) gene and described one novel variant in the intron of the myosin X (*MYO10*) gene. Additionally, we replicated the previously reported associations for Erb-B2 receptor tyrosine kinase 4 (*ERBB4*), DENN domain containing 1A (*DENND1A*), FSH subunit beta (*FSHB*) and zinc finger and BTB domain containing 16 (*ZBTB16*).

## Materials and methods

This study is reported according to the Strengthening the Reporting of Genetic Association Studies (STREGA) guideline.

### Study cohorts

#### FinnGen

The FinnGen study combines genotype data from the Finnish biobanks with the digital health record data from the Care Register for Health Care (CRCH, from 1968 onwards) and the cancer (1953–), cause of death (1969–), and medication reimbursement (1995–) registries (<https://www.finngen.fi/en>). FinnGen data freeze release 6 (R6) combines the genomic information of 141 355 women (6% of the female Finnish population). In FinnGen, cases of PCOS were defined as women with a record of the following International Classification of Diseases (ICD)-10 code E28.2, ICD-9 code 256.4, or ICD-8 code 256.90. Controls were all women without a PCOS diagnosis, and no other exclusions were made. With this definition, there were 797 cases and 140 558 controls.

Patients and control subjects in FinnGen provided informed consent for biobank research based on the Finnish Biobank Act. Alternatively, older research cohorts, collected prior to the start of FinnGen (in August 2017), were collected based on study-specific consents and later transferred to the Finnish biobanks after approval by the National Supervisory Authority for Welfare and Health, Fimea. Recruitment procedures followed the biobank protocols approved by Fimea. The Coordinating Ethics Committee of the Hospital District of Helsinki and Uusimaa (HUS) approved the FinnGen study protocol (Nr HUS/990/2017).

The FinnGen study was approved by Finnish Institute for Health and Welfare (permit numbers: THL/2031/6.02.00/2017, THL/1101/5.05.00/2017, THL/341/6.02.00/2018, THL/2222/6.02.00/2018, THL/283/6.02.00/2019, THL/1721/5.05.00/2019, THL/1524/5.05.00/2020 and THL/2364/14.02/2020); Digital and population data service agency (permit numbers: VRK43431/2017-3, VRK/6909/2018-3, VRK/4415/2019-3); the Social Insurance Institution (permit numbers: KELA 58/522/2017, KELA 131/522/2018, KELA 70/522/2019, KELA 98/522/2019, KELA 138/522/2019, KELA 2/522/2020, KELA 16/522/2020); and Statistics Finland (permit numbers: TK-53-1041-17 and TK-53-90-20).

The Biobank access decisions for FinnGen samples and data utilized in the FinnGen data freeze R6 include: THL Biobank BB2017\_55, BB2017\_111, BB2018\_19, BB\_2018\_34, BB\_2018\_67, BB2018\_71, BB2019\_7, BB2019\_8, BB2019\_26, BB2020\_1, Finnish Red Cross Blood Service Biobank 7.12.2017, Helsinki Biobank HUS/359/2017, Auria Biobank AB17-5154, Biobank Borealis of Northern Finland\_2017\_1013, Biobank of Eastern Finland 1186/2018, Finnish Clinical Biobank Tampere MH0004, Central Finland Biobank 1-2017 and Terveystalo Biobank STB 2018001.

A full list of FinnGen contributors can be found in [Supplementary Data](#).

#### Estonian Biobank

The EstBB is a volunteer-based biobank with over 200 000 participants, currently including approximately 135 000 women (20% of the female Estonian population). The 150K data freeze was used for the

analyses described in this paper. All biobank participants have signed a broad informed consent form. Individuals with PCOS were identified using the ICD-10 code E28.2, and all of the female biobank participants without this diagnosis served as controls. This included a total of 2812 cases and 89 230 controls. Information on the ICD codes was obtained via regular linking with the National Health Insurance Fund and other relevant databases ([Leitsalu et al., 2015](#)). Analyses in the EstBB were carried out under ethical approval 1.1-12/624 from the Estonian Committee on Bioethics, and Human Research and data release N05 from the EstBB.

### Genotyping and association analyses

#### FinnGen

Sample genotyping in FinnGen was performed using Illumina and Affymetrix arrays (Illumina Inc., San Diego, and Thermo Fisher Scientific, Santa Clara, CA, USA). Genotype calls were made using GenCall or zCall ([Goldstein et al., 2012](#)) for Illumina and the AxiomGT1 algorithm for Affymetrix data. Genotypes with a Hardy-Weinberg Equilibrium (HWE) *P*-value below 1e-6, minor allele count <3, and genotyping success rate <98% were removed. Samples with ambiguous gender, those with high genotype missingness >5% and outliers in the population structure (>4 SD from the mean on the first two dimensions of principal component (PC) analysis) were omitted. Samples were pre-phased with Eagle 2.3.5 ([Loh et al., 2016](#)) using 20 000 conditioning haplotypes. Genotypes were imputed with Beagle 4.1 using the SiSu v3 imputation reference panel, which consisted of 3775 individuals of Finnish ancestry with sequenced whole genomes. The post-imputation protocol is publicly available at <https://dx.doi.org/10.17504/protocols.io.xbfgjw>.

Association analysis was performed using a generalized mixed model as implemented in SAIGE ([Zhou et al., 2018](#)). Included adjustments were age, genotyping batches and the first 10 PCs.

Formatting and preparation of the FinnGen association data for downstream analysis were managed with workflow management software STAPLER ([Tyrmi, 2018](#)).

#### Estonian Biobank

All EstBB participants were genotyped using Illumina GSAv1.0, GSAv2.0 and GSAv2.0\_EST arrays at the Core Genotyping Lab of the Institute of Genomics, University of Tartu. Samples were genotyped and PLINK format files were created using Illumina GenomeStudio v2.0.4. Individuals were excluded from the analysis if their call rate was <95% or if their sex defined by heterozygosity of X chromosomes did not match their sex in the phenotype data. Before imputation, variants were filtered by call rate <95%, HWE *P*-value <1e-4 (autosomal variants only) and minor allele frequency <1%. Variant positions were updated to b37 and all variants were changed to be from the TOP strand using GSAMD-24v1-0\_20011747\_A1-b37.strand.RefAlt.zip files from the <https://www.well.ox.ac.uk/~wrayner/strand/> webpage. Pre-phasing was conducted using Eagle v2.3 software ([Loh et al., 2016](#)) (number of conditioning haplotypes Eagle2 uses when phasing each sample was set to:  $-Kpbwt = 20\,000$ ), and imputation was carried out using Beagle 4.1 with effective population size  $ne = 20\,000$ . The population-specific imputation reference of 2297 whole-genome sequencing samples was used ([Mitt et al., 2017](#)).

Association analysis was carried out using SAIGE (v0.38) software to implement a mixed logistic regression model with a year of birth and 10 PCs as covariates in step I. A total of 2812 cases and 89 230 controls were included in the analyses.

## Meta-analysis

In order to synchronize the build of the datasets, we lifted the FinnGen GWAS summary statistics over to hg37 build using UCSC liftOver (Kent *et al.*, 2002) before running the meta-analyses. METAL software was used to perform inverse-variance-weighted meta-analysis for FinnGen and EstBB GWAS results (Willer *et al.*, 2010). In total, 3609 cases and 229 788 controls were analyzed. High imputation quality markers (INFO score  $> 0.7$ ) were kept from each study prior to the meta-analysis. A total of 24 157 216 markers were included in the analysis. Genome-wide significance was set to  $P < 5 \times 10^{-8}$ . The meta-analyses were conducted independently by two analysts and summary statistics were compared for consistency.

## Functional annotation and gene prioritization

In order to identify plausible candidate genes, we used the FUMA platform (Watanabe *et al.*, 2017). FUMA uses GWAS summary statistics and performs extensive functional annotation and candidate gene mapping using positional, expression quantitative trait loci (eQTL) and chromatin interaction mapping in all genome-wide significant loci. Loci were defined by  $\pm 1000$  kb of the top single nucleotide variant in the region. Gene-based analysis was also performed in this platform using MAGMA (de Leeuw *et al.*, 2015). We prioritized variants that were more likely to have a functional consequence, such as variants in high linkage disequilibrium (LD) ( $r^2 > 0.6$ ) with missense mutations or pathogenic variants. Secondly, we prioritized variants overlapping with regulatory marks, focusing on genes with modified expression or genes that showed chromatin interaction links with the variants. Furthermore, gene functions were examined in GenBank and UniProt portals. In addition, a literature search was performed for the genes of interest to gain further insight into the possible underlying molecular mechanisms. Genes showing relevant functions in relevant tissues or traits with similar PCOS pathophysiology were ultimately considered for gene candidate prioritization.

## Colocalization analyses

We tested whether the GWAS signals colocalized with variants that affect gene expression using the following pipeline (<https://github.com/eQTL-Catalogue/colocalisation>) (Kerimov *et al.*, 2021). We compared our significant loci to all eQTL Catalogue RNA-Seq datasets containing QTLs for gene expression, exon expression, transcript usage and txrevise event usage; eQTL Catalogue microarray datasets containing QTLs for gene expression; and GTEx v7 datasets containing QTLs for gene expression (Kerimov *et al.*, 2021). We lifted the GWAS summary statistics over to the hg38 build to match the eQTL catalogue and convert the summary statistics to variant call format. For each genome-wide significant ( $P < 5 \times 10^{-8}$ ) GWAS variant, we extracted the 1-Mb radius of its top hit from the QTL datasets. We then ran the colocalization analysis for those eQTL catalogue traits that had at least one cis-QTL within this region with  $P < 1 \times 10^{-6}$ . We

considered two signals to colocalize if the posterior probability for a shared causal variant was 0.8 or higher.

## Conditional analyses

Since considering most significant variants as the causal ones would lead to an underestimation of the total variance explained at each locus, we next performed conditional analyses, which were carried out similarly to the main association testing using SAIGE (Zhou *et al.*, 2018). This approach has been used to identify secondary association signals at a particular locus and involves association analysis conditioning on the primary associated variant at the locus to test for additional significantly associated variants (Yang *et al.*, 2012). We proceeded to test associations using a stepwise analysis, where markers were added to the model until no independent signals were identified.

## Adjusting the GWAS for BMI

In the discovery dataset, an additional association analysis including BMI as a covariate was conducted with a total of 482 PCOS cases (60.5% of the original PCOS sample) and 91 631 controls from FinnGen (65.2% of the original control sample). Similarly, we ran an association analysis including BMI as a covariate for the validation dataset, which contained a total of 2137 PCOS cases (75% of the original PCOS sample size) and 68 690 controls from the EstBB (76.9% of the original control sample size). We then performed a second meta-analysis including the two GWAS adjusted for BMI from both cohorts. This analysis included 2619 cases and 160 321 controls, and a total of 24 461 102 genetic markers were analyzed.

## Interaction analysis

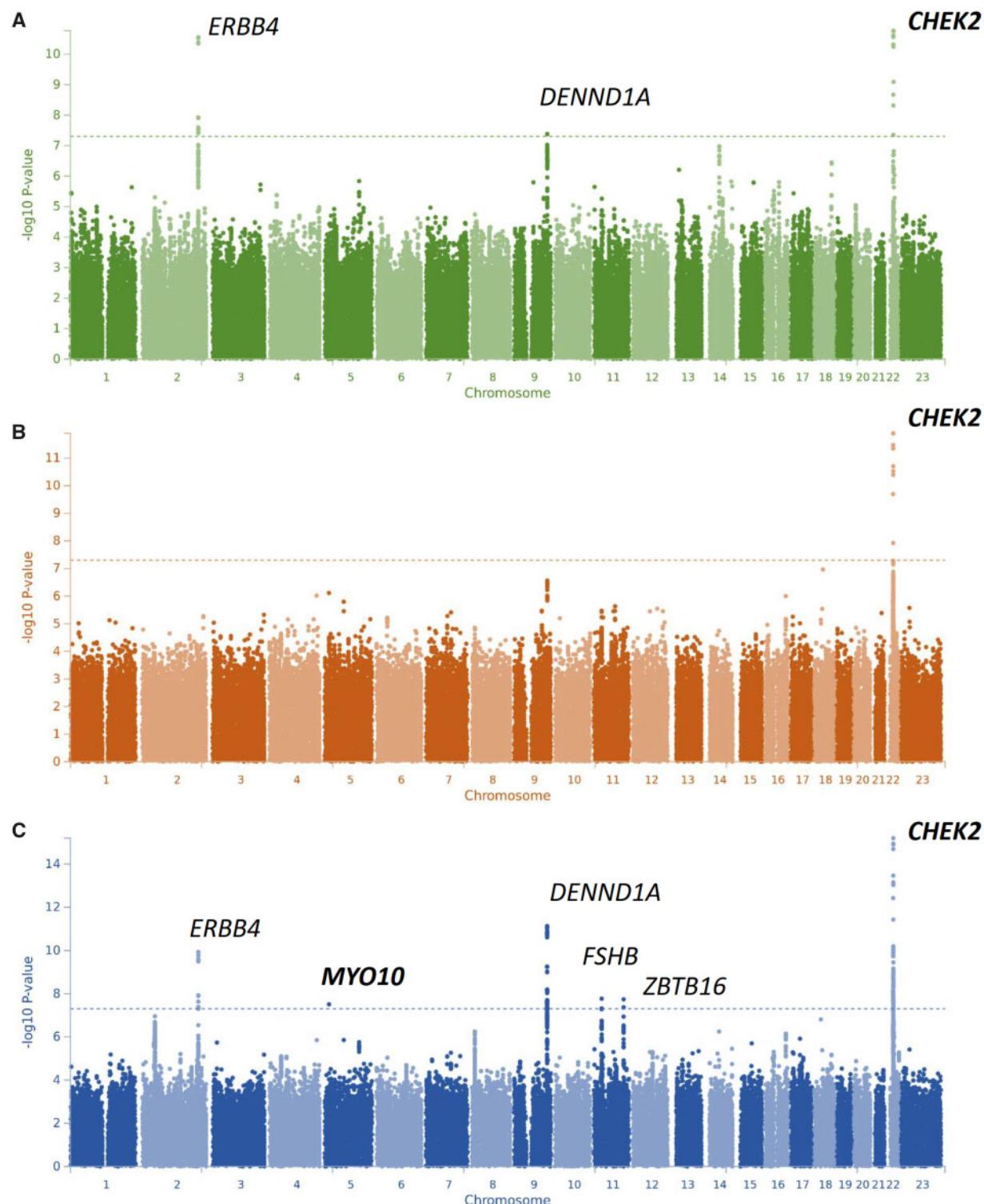
We tested whether an interaction between c.1100delC mutation, obesity and PCOS could be detected, as such a phenomenon has been identified between the mutation carriers in invasive breast cancer (Greville-Heygate *et al.*, 2020). We fitted a logistic model where PCOS was the outcome, the lead variant genotype and BMI formed the interaction term, and the 10 first genetic PCs along with age were added as covariates. This analysis was performed with R version 4.0.5 (R Core Team, 2018).

## Results

### Discovery GWAS identified a rare novel association for PCOS in CHEK2

A discovery GWAS with 797 PCOS cases and 140 558 controls in the FinnGen study uncovered two loci close to *ERBB4* and *DENND1A* that have previously been shown to be associated with PCOS. In addition, a previously unreported large effect association was found in chromosome 22 at 22q11 (Fig. 1A).

The lead variant rs145598156 ( $P = 1.7 \times 10^{-11}$ , odds ratio (OR) = 11.63 [5.69–23.77]) is located in an intronic region 11 kb from the transcription start site (TSS) of *ZNFR3* (Table 1 and Fig. 2A). However, the tight LD spans an area of approximately 2 Mb surrounding the lead variant with many variants in high LD (Fig. 2A). Functional characterization of this locus revealed a frameshift variant, c.1100delC



**Figure 1. Manhattan plot of the results from the age-adjusted genome-wide association studies.** Genome-wide association studies (GWAS) from the Finnish dataset (A), GWAS from Estonian dataset (B) and joint GWAS meta-analysis of polycystic ovary syndrome (PCOS) (C). The novel gene candidates in the six genome-wide significant loci are highlighted in bold. The y axis represents  $-\log_{10}$  P-values for the associations of variants with PCOS from meta-analysis, using an inverse-variance weighted fixed-effects model. The horizontal dashed line represents the threshold for genome-wide significance. *ERBB4* (Erb-B2 Receptor Tyrosine Kinase 4); *DENND1A* (DENN Domain Containing 1A); *FSHB* (FSH Subunit Beta); *ZBTB16* (Zinc Finger And BTB Domain); *MYO10* (myosin X); *CHEK2* (Checkpoint kinase 2).

**Table I** Summary of association results of the genome-wide association meta-analysis of polycystic ovary syndrome.

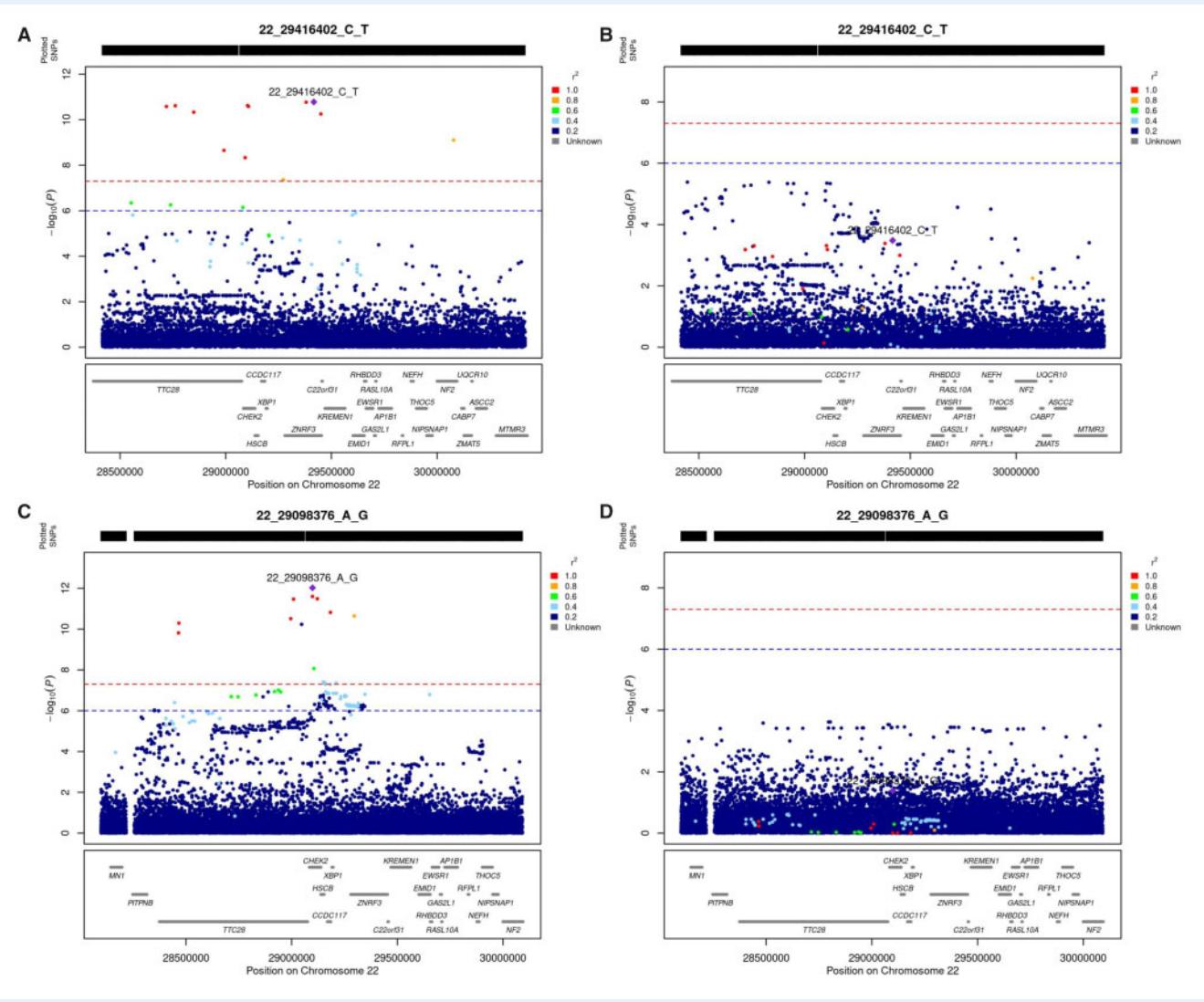
SNP	Chr:BP	Cytoband	EA/NEA	Nearest gene	Candidate gene	Cohort	EAF (%)	OR (95% CI)*	P*	OR (95% CI)**	P**
rs7564590	chr2: 213387900	2q34	T/C	ERBB4	ERBB4	FinnGen	34.50	1.43 (1.29–1.59)	3.0×10 <sup>-11</sup>	1.50 (1.31–1.72)	4.4×10 <sup>-09</sup>
rs9312937	chr5: 16836005	<u>5p15</u>	<u>C/T</u>	<u>MYO10</u>	<u>MYO10</u>	EstBB Meta	35.87 35.56	1.12 (1.06–1.19) 1.19 (1.13–1.25)	1.1×10 <sup>-04</sup> 4.8×10 <sup>-11</sup>	1.13 (1.05–1.20) 1.19 (1.13–1.25)	2.5×10 <sup>-04</sup> 4.6×10 <sup>-09</sup>
rs3945628	chr9: 126535553	9q33	C/T	DENND1A	DENND1A	FinnGen	42.00	1.15 (1.04–1.27)	6.8×10 <sup>-03</sup>	1.06 (0.93–1.20)	3.7×10 <sup>-01</sup>
rs1031002	chr11: 30215261	11p14	A/T	FSHB	FSHB	EstBB Meta	45.47 6.99	1.16 (1.10–1.23) 1.40 (1.27–1.55)	7.7×10 <sup>-07</sup> 2.9×10 <sup>-12</sup>	1.18 (1.10–1.26) 1.15 (1.08–1.22)	1.5×10 <sup>-06</sup> 3.0×10 <sup>-06</sup>
rs672716	chr11: 13952497	11q23	G/A	ZBTB16	ZBTB16	FinnGen	12.00	1.33 (1.14–1.56)	2.3×10 <sup>-04</sup>	1.31 (1.29–2.19)	9.3×10 <sup>-05</sup>
rs145598156	chr22: 29416402	<u>22q12</u>	<u>T/C</u>	<u>ZNFR3</u>	<u>CHEK2</u>	EstBB Meta	12.27 12.21	1.22 (1.12–1.32) 1.24 (1.15–1.34)	5.8×10 <sup>-06</sup> 9.2×10 <sup>-09</sup>	1.16 (1.05–1.27) 1.18 (1.10–1.27)	2.2×10 <sup>-03</sup> 7.7×10 <sup>-05</sup>
rs182075939	chr22: 29098376	<u>22q12</u>	<u>G/A</u>	<u>TTC28</u>	<u>CHEK2</u>	FinnGen	0.79	11.63 (5.69–23.77)	1.7×10 <sup>-11</sup>	13.5 (5.35–34.38)	4.5×10 <sup>-08</sup>
						EstBB Meta	0.37 0.52	1.68 (1.05–2.69) 3.01 (2.02–4.50)	3.2×10 <sup>-02</sup> 3.6×10 <sup>-08</sup>	1.53 (0.90–2.61) 2.61 (2.14–3.07)	1.1×10 <sup>-01</sup> 4.4×10 <sup>-05</sup>
						FinnGen	3.19	1.95 (1.44–2.65)	1.8×10 <sup>-05</sup>	2.18 (1.46–3.24)	1.1×10 <sup>-04</sup>
						EstBB Meta	4.64 4.41	1.64 (1.43–1.88) 1.69 (1.49–1.91)	1.3×10 <sup>-12</sup> 1.9×10 <sup>-16</sup>	1.68 (1.44–1.96) 1.74 (1.5–2.01)	4.6×10 <sup>-11</sup> 4.9×10 <sup>-14</sup>

Meta-analysis results of the genome-wide association studies from Estonian Biobank (EstBB) and FinnGen are shown in italics. Novel associations are underlined. Variant positions (BP) are according to GRCh37/hg19. ERBB4, Erb-B2 Receptor Tyrosine Kinase 4; DENND1A, DENND1A; ZNFR3, ZNFR3; Zinc Finger And BTB Domain; MYO10, myosin X; ZNFR3, Zinc Finger 3; CHEK2, Checkpoint kinase 2; TTC28, Tetra peptide Repeat Domain 28.

EA, effect allele; EAF, effect allele frequency; NEA, non-effect allele; OR, odds ratio; P, P-value; SNP, single-nucleotide polymorphism.

\*OR and P-values of age-adjusted results.

\*\*OR and P-values of age- and BMI-adjusted results.



**Figure 2. Regional plots before and after conditional analyses for lead variants in chromosome 22.** FinnGen lead variant in locus 22q11 (A) along with conditional analysis results with frameshift variant (rs555607708) (B). Regional plot for the Estonian Biobank lead variant in the same locus 22q11 before and after conditional analysis with linked missense variant (rs17879961) are shown in (C) and (D). Regional plots were produced with R-package LocusZooms (<https://github.com/Geeketics/LocusZooms/>).  $r^2$  estimates were generated using LDstore (Benner et al., 2017) with SiSu v3 project WGS data consisting of 3775 individuals with Finnish ancestry.

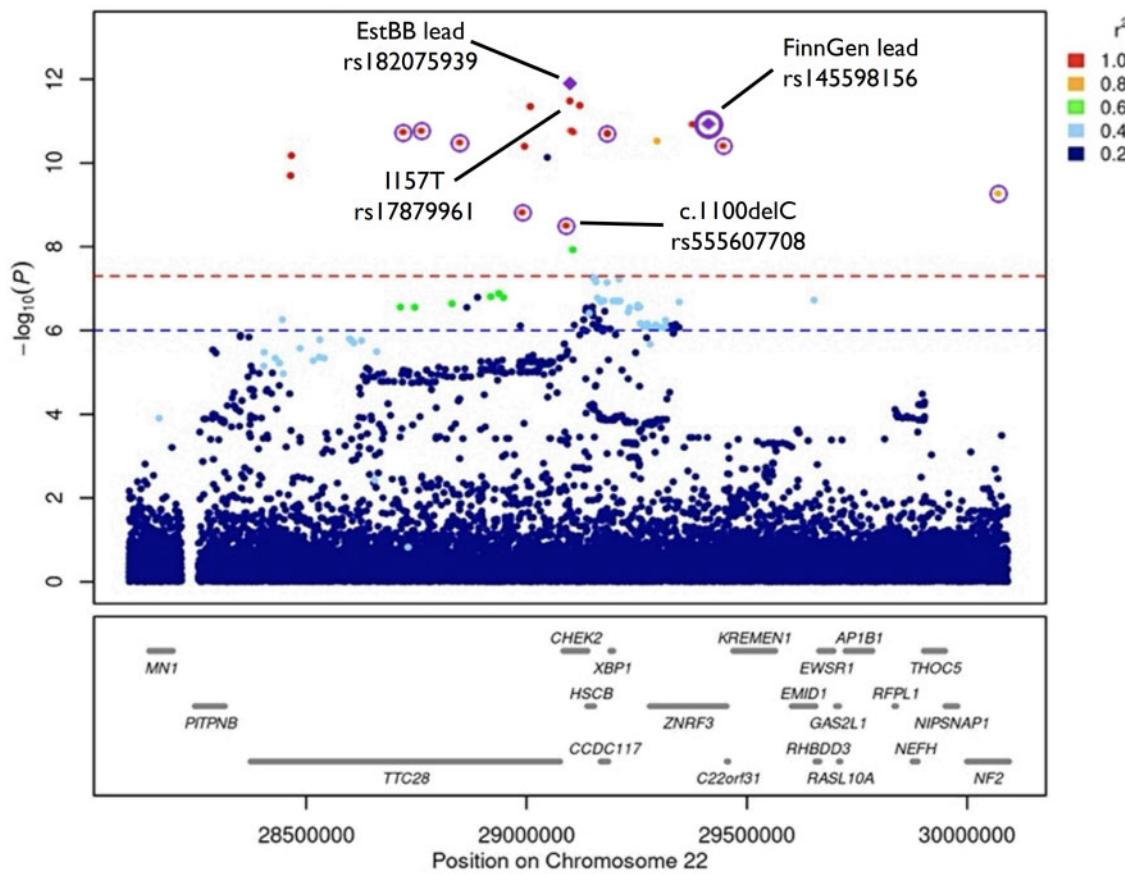
(rs555607708,  $P = 1.68 \times 10^{-9}$ , OR = 13.46 [5.68–31.89]) in *CHEK2*, with a high LD ( $r^2 = 0.95$ ) with the lead variant. Interestingly, the protein-truncating variant c.1100delC is enriched in the Finnish population (AF = 0.008) compared to the Estonian (AF = 0.003) and other European populations (AF = 0.002), according to the gnomAD database (Karczewski et al., 2020). The analysis conditioned on c.1100delC resulted in no genome-wide significant associations in this locus, with a  $P$ -value of  $3.29 \times 10^{-4}$  for the lead variant rs145598156 (Fig. 2B).

To investigate the influence of BMI on PCOS, we ran an additional association analysis, including BMI as a covariate. In this analysis, the FinnGen lead variant rs145598156 remained genome-wide significant ( $P = 4.5 \times 10^{-8}$ , OR = 13.5 [5.35–34.38]) (Table 1 and Supplementary Fig. S2).

A recent study has suggested that patients with invasive breast cancer carrying the c.1100delC mutation are more likely to be obese, though this is not the case for the general population (Greville-Heygate et al., 2020). Thus, when we tested for such an interaction between PCOS, c.1100delC, and obesity using a logit regression model, a  $P$ -value of 0.066 for c.1100delC-BMI interaction was obtained (OR 1.04, 95% CI 0.99–1.09).

### Validation GWAS detected an independent association in *CHEK2*

A validation GWAS was performed in the EstBB, including 2812 cases and 89 230 controls. The validation also uncovered a genome-wide significant association ( $P = 1.3 \times 10^{-12}$ , OR = 1.64 [1.34–1.88]) in the



**Figure 3. Checkpoint kinase 2 variants described.** Independent FinnGen and Estonian Biobank (EstBB) GWAS associations overlapping the checkpoint kinase 2 (*CHEK2*) gene are plotted on a single LocusZooms figure. Genome-wide significant variants in FinnGen data are denoted with purple circles; Estonian Biobank-specific variants are not circled.

22q11 region. The lead variant rs182075939 was an intron variant located 22 kb from the TSS of *TTC28* (Figs 1B and 2C). Functional annotation revealed a tightly linked missense variant rs17879961 ( $r^2=0.83$ ,  $P=4.23\times 10^{-12}$ ), known as I157T, in *CHEK2*, which has been shown to alter *CHEK2* ability to bind p53, *BRCA1* (breast cancer gene 1) and *Cdc25A* proteins (Falck *et al.*, 2001a,b). The EstBB lead variant rs182075939 presents a higher allele frequency in Estonians (AF=0.048) compared to Finns (AF=0.029) and other European populations (AF=0.0025) according to gnomAD (Karczewski *et al.*, 2020). The analysis conditioned on I157T resulted in no genome-wide significant associations in this locus, with a  $P$ -value of 0.04 for the lead variant rs182075939 (Fig. 2D).

When also adjusting the GWAS for BMI, the EstBB lead variant rs182075939 remained genome-wide significant ( $P=4.6\times 10^{-11}$ , OR=1.68 [1.44–1.96]) (Table 1 and Supplementary Fig. S2).

Interestingly, even though the association signals found in the EstBB and FinnGen data sets overlap with each other (Fig. 3), they seem to be part of independent haplotypes with an  $r^2$  value below 0.05 between the lead variants. The lead variant of FinnGen data had a  $P$ -value of 0.031 in the EstBB. The EstBB lead variant had a  $P$ -value of  $1.8\times 10^{-5}$  in FinnGen (Table 1).

We also tested if conditioning the discovery GWAS results with I157T and validation GWAS with c.1100delC would affect the significance of the lead variants. Conditioning the discovery analysis on I157T had a minimal effect on the genome-wide significant associations in this locus, with a  $P$ -value of  $9.09\times 10^{-12}$  for the lead variant rs145598156. Similarly, when the validation GWAS in the EstBB was conditioned on c.1100delC, the  $P$ -value of the lead variant rs182075939 was only modestly affected ( $P=9.16\times 10^{-13}$ ).

### Meta-analysis confirmed and expanded novel associations with PCOS in *CHEK2* and *MYO10*

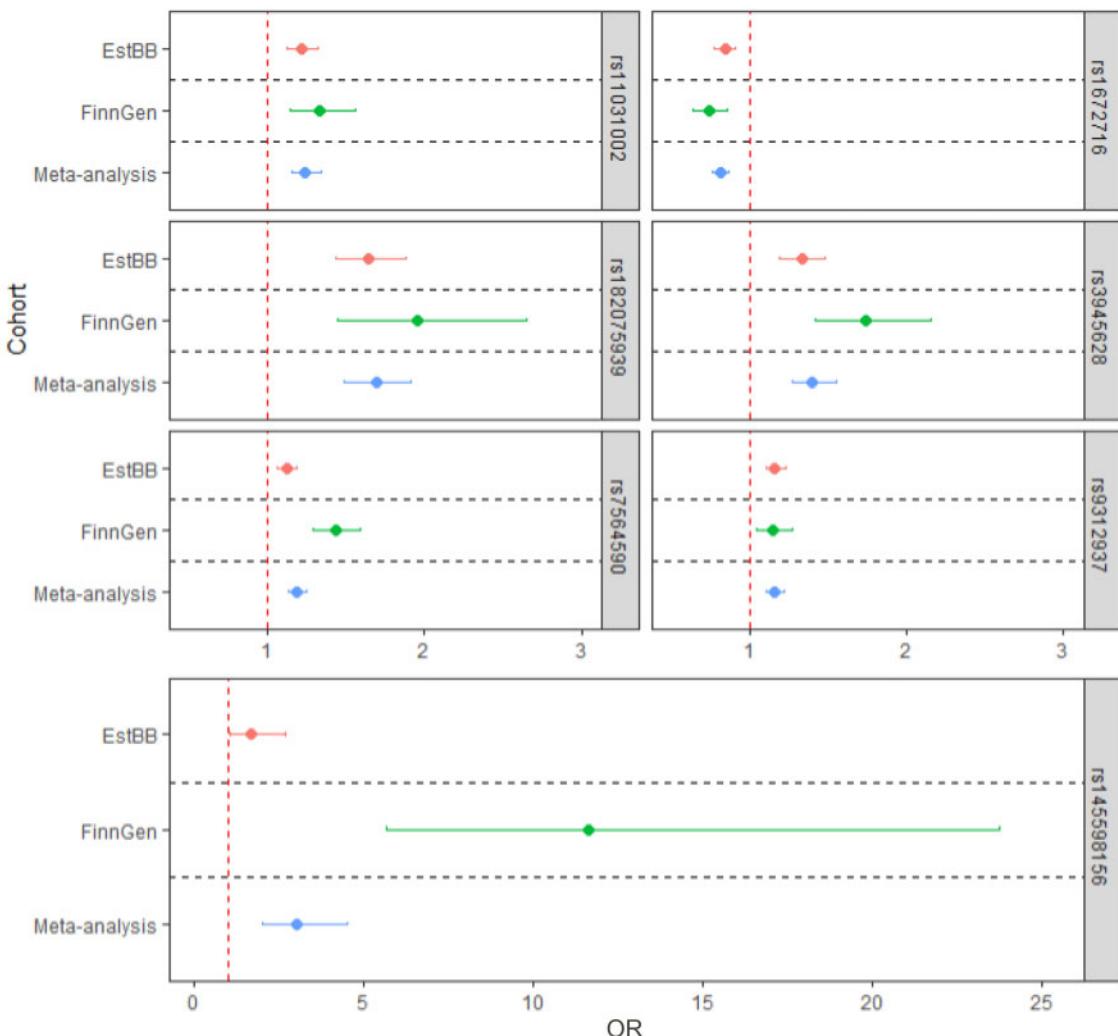
A meta-analysis was performed for the FinnGen and EstBB GWAS incorporating a total of 3609 women with PCOS and 229 788 controls. In the meta-analysis, the FinnGen lead variant on chromosome 22 rs145598156 had a  $P$ -value of  $3.6\times 10^{-8}$  with significant heterogeneity between cohorts ( $p^{het}=9.58\times 10^{-6}$ ), while the EstBB lead variant rs182075939 showed a  $P$ -value of  $1.9\times 10^{-16}$  in the meta-analysis results without significant heterogeneity between cohorts ( $p^{het}=0.3$ ). When the FinnGen and EstBB results were conditioned for the

c.1100delC and 1157T variants and the results were meta-analyzed, there were no additional genome-wide significant signals in the *CHEK2* locus.

The meta-analysis also revealed three more variants associating with PCOS, in addition to the three detected in the FinnGen and EstBB GWAS separately (Table 1 and Supplementary Fig. S1). Two of the additional signals were in chromosome 11 and have been previously shown to be associated with PCOS: rs11031002 is located near *FSHB*, and rs1672716 is an intron variant of *ZBTB16*. The third new association peak in the meta-analysis (rs9312937,  $P = 1.7 \times 10^{-8}$ , OR = 1.16 [1.10–1.22], AF = 0.44) was a common variant in an intronic region of chromosome 5, located 100 kb from the TSS of the *MYO10* gene, which to our knowledge has not previously been associated with PCOS. A total of two potentially causal genes were suggested by chromatin interaction data from 21 different tissues/cell types, with *MYO10* being the closest one, while no significant eQTL associations were detected using FUMA (Watanabe et al., 2017) in this locus.

The average effect sizes of the novel alleles described in chromosome 22 (OR = 1.69–3.01) (Table 1) were higher than the effects observed for alleles associated with PCOS in the rest of the common variants described (OR = 1.06–1.40), which could be explained by the often-observed inverse relationship between allele frequency and effect size (Manolio et al., 2009). Moreover, we observed consistency in the direction of effects between the three datasets analyzed (discovery, validation and joint meta-analysis) (Fig. 4). We further assessed the robustness of our PCOS definition by comparing the effects sizes between the lead variants in the replicated loci presented in the non-NIH Rotterdam criteria (Day et al., 2018) to our association results. We conclude that our results based on ICD codes alone are robust, as the effects are in the same direction and do not present significant heterogeneity ( $p_{het} = 1$ ) compared to those using non-NIH Rotterdam criteria (Supplementary Fig. S3) (Day et al., 2018).

In colocalization analyses, all posterior probabilities for a shared causal variant were lower than 0.8. Thus, we did not find enough



**Figure 4. Forest plot of effect estimates for the seven lead variants associated with PCOS.** The odds ratios (dots) and 95% CI (whiskers) are shown for the two included cohorts and the meta-analysis.

evidence that two association signals in the genome-wide association analysis and gene expression are consistent with a shared causal variant.

## Discussion

In this study, we found two independent novel associations for PCOS on 22q11.2. In both cases, the lead single nucleotide polymorphisms had tightly linked variants, a frameshift (c.1100delC) and a missense (I157T), in the *CHEK2* gene. A novel association was also detected in an intron of *MYO10*. We were also able to replicate signals commonly reported in PCOS GWAS—*DENND1A*, *ERBB4* (*HER4*), *ZBTB16* and *FSHB*—in our North-European populations.

*CHEK2* rs555607708 (c.1100delC), the likely association-driving variant in FinnGen, is a Finnish-enriched variant with a 3.7-fold enrichment compared to non-Finnish, non-Estonian Europeans and with an enrichment of 1.7 compared to Estonians (Mars et al., 2020). Similarly, I157T, the likely association-driving variant in the EstBB, has a substantially higher allele frequency in the Estonian (0.048) and Finnish (0.029) populations, compared to the non-Finnish, Northwestern European population (0.0025), according to the gnomAD database (Karczewski et al., 2020). The enrichment of the alleles likely allowed us to detect the associations with PCOS in the Finnish and Estonian populations, whereas in populations with lower minor allele frequencies, much larger study populations would need to be used.

*CHEK2* is a mediator of DNA damage signalling in response to double-stranded DNA breaks. *CHEK2* can be considered an important factor in the quality control of cells. If *CHEK2* function is disturbed, DNA repair is imbalanced, which can lead to genomic instability and tumorigenesis (Mustoza et al., 2020).

Whereas the association of *CHEK2* c.1100delC with a moderate-risk breast cancer predisposition is well recognized (Meijers-Heijboer et al., 2002), the pathogenic role of I157T remains controversial (Schutte et al., 2003; Kilpivaara et al., 2004; Muranen et al., 2016). Several studies have shown the pathogenic impact of c.1100delC on breast cancer risk in the Finnish population (Kuusisto et al., 2011; Hallamies et al., 2017; Mars et al., 2020). There are currently no studies evaluating the pathogenic role of c.1100delC or I157T in Estonians, which underlines the need for further research assessing the impact of these variants in this population.

An interaction between BMI and PCOS-associated variants has previously been suggested (Wojciechowski et al., 2012), and interestingly, the c.1100delC variant in *CHEK2* has recently been shown to predispose particularly obese carriers to the development of breast cancer (Greville-Heygate et al., 2020). Although our results did not support such an association between c.1100delC-related PCOS risk and obesity, a replication of this analysis with larger sample size is needed.

Epidemiological studies show an increased risk for endometrial cancer in women with PCOS. However, this does not apply to other gynaecological cancers like ovarian, cervical or breast cancer (Barry et al., 2014; Gottschau et al., 2015; Hart and Doherty, 2015; Harris and Terry, 2016; Ding et al., 2018). Nevertheless, three recent studies utilizing a Mendelian randomization approach have suggested a modest but significant causal effect between PCOS and breast cancer (Wu et al., 2020; Wen et al., 2021; Zhu et al., 2021). The fact that the risks

do not seem to translate into clinical findings is notable and may indicate, for example, more efficient DNA repair systems in women with PCOS, a feature also associated with later onset of menopause (Day et al., 2018; Ruth et al., 2021).

Interestingly, *CHEK2* also plays a crucial role in foetal oocyte attrition, a phenomenon through which 80% of the initial ovarian oocyte reserve is lost during foetal development in mammals (Tharp et al., 2020). Deletion of *Chk2* in mice leads to a maximized ovarian reserve at postnatal day 2 (Tharp et al., 2020) and reduced follicle atresia, a higher number of ovulated metaphase II oocytes, and higher AMH levels at 13.5 months (Ruth et al., 2021). It was also reported that a *CHEK2* loss-of-function allele is associated with later menopausal age in humans (Ruth et al., 2021). This would be in line with women with PCOS, as they also present with an increased ovarian reserve, higher AMH levels, even at later reproductive years, and delayed menopause (Piltonen et al., 2005; de Ziegler et al., 2018; Minooee et al., 2018; Forslund et al., 2019; Ward et al., 2021). A specific association between menopause-delaying alleles and PCOS has also been previously demonstrated (Day et al., 2015). In a recent preprint work, Ward et al. also found that *CHEK2* was associated with the age of menopause. When conducting a genome-wide association study (PheWAS) on their associations, an aggregate of *CHEK2*-damaging variants also associated with PCOS, which is in line with our findings (Ward et al., 2021). Our study also detected the previously reported associations with PCOS for *ERBB4*, *DENND1A*, *FSHB* and *ZBTB16*. Interestingly, *ERBB4* has also recently been linked to proper oocyte maturation and high AMH in mice (Veikkolainen et al., 2020). Thus, the present study reinforces the links between PCOS, abnormal follicle development and high levels of AMH.

This study also presents an interesting novel association in an intronic region of *MYO10*. The *MYO10* gene codes for an atypical myosin, which is involved in filopodia formation, phagocytosis and cargo transport in cells (Sousa and Cheney, 2005). Genetic variation in *MYO10* has previously been linked to type 2 diabetes (Salonen et al., 2007) and traits of metabolic syndrome (Zhang et al., 2013). Interestingly, the identified variant seems to be associated with the age at menarche (Kichae et al., 2019), indicating a reproductive function for *MYO10*. Although a metabolic link between *MYO10* and PCOS seems likely, further research is needed to characterize the role of *MYO10* in PCOS.

As previous studies have suggested a causal role for obesity in PCOS (Brower et al., 2019; Zhao et al., 2020), we reran the association analyses adjusting for BMI. A reduction in the significance of several associations was expected owing to the limited availability of BMI data (60% in FinnGen and 75% in EstBB). Two of the replicated (*FSHB*, *ZBTB16*) and two of the novel associations (*MYO10* and *CHEK2*) did not reach genome-wide significance after adjustment. As the effect sizes remain largely unchanged when adjusted for BMI, the statistical significance of the associations was diluted by the reduction in sample size. Thus, we mainly focused on age-adjusted associations and acknowledge that larger sample sizes are needed to further explore the interplay between BMI and PCOS-related genetic factors.

Overall, it is important to note that complex LD patterns between association signals might eclipse more distant causal genes. To infer plausible shared causal variants between PCOS-related genetic variants and gene expression, we conducted colocalization analyses without

significant findings. This might be explained by the low sample size in gene expression panels that study tissues of interest in PCOS, such as reproductive tissues. Thus, further functional studies are warranted to characterize the regulatory functions of the uncovered loci (Peltonen et al., 2000; Lim et al., 2014; Martin et al., 2018; Prohaska et al., 2019).

The main strength of this study was the use of the two large, comprehensive genetic data sets, FinnGen and the EstBB, which have been extensively linked to national registers, such as the CRCH in Finland and the Estonian Health Insurance Fund registries in Estonia, as well as with other relevant databases (Leitsalu et al., 2015). Both populations are genetically well-characterized (Salmela et al., 2008). Furthermore, our main discovery of the two rare PCOS-associated variants near *CHEK2* underlines the value of using study populations with a distinct genetic makeup. The interplay of past demographic events may result in regionally varying genetic architectures for medical conditions (Peltonen et al., 2000; Martin et al., 2018). When alleles enriched in such populations are causal or linked to causal variation, increased statistical power is present, enabling their detection in an association analysis (Lim et al., 2014; Prohaska et al., 2019).

The register-based approach is also a limiting factor, as the health register-based prevalence of PCOS is very low in our study populations (0.57% for FinnGen and 3.15% for the EstBB), plausibly reflecting underdiagnosis of the syndrome. We were unable to validate the ICD codes, as the FinnGen dataset does not contain identifying information of the subjects; however, the coverage and accuracy of the Finnish CRCH have been validated in several studies, and they have been shown to be excellent (Sund, 2012). The CRHC diagnoses are hospital-based, and thus the PCOS cases were diagnosed by specialized doctors. The validity of the PCOS diagnosis is also supported by the fact that we were able to replicate four previously reported signals, *ERBB4*, *DENND1A*, *FSHB* and *ZBTB16*. In addition, there was consistency in the direction of effects between our association results and the non-NIH Rotterdam-criteria association results presented in the largest European GWAS meta-analysis to date (Day et al., 2018), which adds robustness to our approach. Given the register-based approach, we could not assess in more detail the different PCOS phenotypes; however, a previous study indicated that women with PCOS diagnosed by a physician using different diagnostic criteria are genetically similar (Day et al., 2018).

In conclusion, we identified two rare population-enriched variants located in *CHEK2* that are significantly associated with PCOS. The findings emphasize the benefits of utilizing isolated populations in genetic studies of complex diseases and advance the understanding of genetic factors underlying PCOS.

## Supplementary data

Supplementary data are available at *Human Reproduction* online.

## Data availability

The GWAS meta-analysis summary statistics that support the findings of this study are available for download from the GWAS Catalog at

ebi.ac.uk/gwas/ with accession ID numbers GCST90044902 and GCST90044903.

## Acknowledgements

We thank all FinnGen and EstBB participants for offering us the valuable resources. We also acknowledge the Estonian Biobank Research team members Andres Metspalu, Tõnu Esko, Mari Nelis and Lili Milani.

## Authors' roles

J.S.T., R.K.A., N.P.-G., L.M.-P., T.T.P., T.L., J.K. and H.L. contributed to the conceptualization of the study. J.S.T., N.P.-G., T.L., E.S. and J.K. conducted data curation, formal analysis and validation of the results. J.S.T., R.K.A. and N.P.-G. carried out the research and investigation process, manuscript writing and data visualization. FinnGen and EstBB provided the data resources. V.K. administrated the project and provided feedback. Funding was acquired by T.T.P., T.L., J.K. and H.L. All authors participated in manuscript editing and review and approved the final version.

## Funding

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 813707 (N.P.-G., T.L. and T.P.), the Estonian Research Council grant (PRG687, T.L.), the Academy of Finland grants 315921 (T.P.), 321763 (T.P.), 297338 (J.K.), 307247 (J.K.), 344695 (H.L.), Novo Nordisk Foundation grant NNF17OC0026062 (J.K.), the Sigrid Juselius Foundation project grants (T.L., J.K., T.P.), Finska Läkaresällskapet (H.L.) and Jane and Aatos Erkko Foundation (H.L.). The funders had no role in study design, data collection and analysis, publishing or preparation of the manuscript.

## Conflict of interest

The authors declare no conflicts of interest.

## References

- 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491**:56–65.
- Abbott DH, Dumesic DA, Levine JE. Hyperandrogenic origins of polycystic ovary syndrome—implications for pathophysiology and therapy. *Expert Rev Endocrinol Metab* 2019; **14**:131–143.
- Azziz R. Introduction: determinants of polycystic ovary syndrome. *Fertil Steril* 2016; **106**:4–5.
- Barber TM, Franks S. Obesity and polycystic ovary syndrome. *Clin Endocrinol (Oxf)* 2021; **95**:531–541.
- Barry JA, Azizia MM, Hardiman PJ. Risk of endometrial, ovarian and breast cancer in women with polycystic ovary syndrome: a systematic review and meta-analysis. *Hum Reprod Update* 2014; **20**:748–758.

Benner C, Havulinna AS, Järvelin MR, Salomaa V, Ripatti S, Pirinen M. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am J Hum Genet* 2017; **101**:539–551.

Brower MA, Hai Y, Jones MR, Guo X, Chen YI, Rotter JI, Krauss RM, Legro RS, Azziz R, Goodarzi MO. Bidirectional Mendelian randomization to explore the causal relationships between body mass index and polycystic ovary syndrome. *Hum Reprod* 2019; **34**:127–136.

Chen ZJ, Zhao H, He L, Shi Y, Qin Y, Shi Y, Li Z, You L, Zhao J, Liu J et al. Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nat Genet* 2011; **43**:55–59.

Dapas M, Dunaif A. The contribution of rare genetic variants to the pathogenesis of polycystic ovary syndrome. *Curr Opin Endocr Metab Res* 2020; **12**:26–32.

Dapas M, Lin FTJ, Nadkarni GN, Sisk R, Legro RS, Urbanek M, Hayes MG, Dunaif A. Distinct subtypes of polycystic ovary syndrome with novel genetic associations: an unsupervised, phenotypic clustering analysis. *PLoS Med* 2020; **17**:e1003132.

Day F, Karaderi T, Jones MR, Meun C, He C, Drong A, Kraft P, Lin N, Huang H, Broer L et al. Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet* 2018; **14**:e1007813.

Day FR, Hinds DA, Tung JY, Stolk L, Stykarsdottir U, Saxena R, Bjornes A, Broer L, Dunger DB, Halldorsson BV et al. Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nat Commun* 2015; **6**:8464.

de Leeuw CA, Mooij JM, Heskes T, Posthuma D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015; **11**:e1004219.

de Ziegler D, Pirtea P, Fanchin R, Ayoubi JM. Ovarian reserve in polycystic ovary syndrome: more, but for how long? *Fertil Steril* 2018; **109**:448–449.

Ding DC, Chen W, Wang JH, Lin SZ. Association between polycystic ovarian syndrome and endometrial, ovarian, and breast cancer: a population-based cohort study in Taiwan. *Medicine (Baltimore)* 2018; **97**:e12608.

Dumesic DA, Lobo RA. Cancer risk and PCOS. *Steroids* 2013; **78**:782–785.

Falck J, Lukas C, Protopopova M, Lukas J, Selivanova G, Bartek J. Functional impact of concomitant versus alternative defects in the Chk2-p53 tumour suppressor pathway. *Oncogene* 2001a; **20**:5503–5510.

Falck J, Mailand N, Syljuåsen RG, Bartek J, Lukas J. The ATM-Chk2-Cdc25A checkpoint pathway guards against radioresistant DNA synthesis. *Nature* 2001b; **410**:842–847.

Forslund M, Landin-Wilhelmsen K, Schmidt J, Brännström M, Trimpou P, Dahlgren E. Higher menopausal age but no differences in parity in women with polycystic ovary syndrome compared with controls. *Acta Obstet Gynecol Scand* 2019; **98**:320–326.

Goldstein JI, Crenshaw A, Carey J, Grant GB, Maguire J, Fromer M, O'Dushlaine C, Moran JL, Chambert K, Stevens C et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics* 2012; **28**:2543–2545.

Gottschau M, Kjaer SK, Jensen A, Munk C, Mellemkjaer L. Risk of cancer among women with polycystic ovary syndrome: a Danish cohort study. *Gynecol Oncol* 2015; **136**:99–103.

Greville-Heygate SL, Maishman T, Tapper WJ, Cutress RI, Copson E, Dunning AM, Haywood L, Jones LJ, Eccles DM. Pathogenic variants in CHEK2 are associated with an adverse prognosis in symptomatic early-onset breast cancer. *JCO Precis Oncol* 2020; **4**:PO.19.00178.

Hallamies S, Pelttari LM, Poikonen-Saksela P, Jekunen A, Jukkola-Vuorinen A, Auvinen P, Blomqvist C, Aittomäki K, Mattson J, Nevanlinna H. CHEK2 c.1100delC mutation is associated with an increased risk for male breast cancer in Finnish patient population. *BMC Cancer* 2017; **17**:620.

Harris HR, Terry KL. Polycystic ovary syndrome and risk of endometrial, ovarian, and breast cancer: a systematic review. *Fertil Res Pract* 2016; **2**:14.

Hart R, Doherty DA. The potential implications of a PCOS diagnosis on a woman's long-term health using data linkage. *J Clin Endocrinol Metab* 2015; **100**:911–919.

Hayes MG, Urbanek M, Ehrmann DA, Armstrong LL, Lee JY, Sisk R, Karaderi T, Barber TM, McCarthy MI, Franks S et al. Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. *Nat Commun* 2015; **6**:7502.

Hong SH, Hong YS, Jeong K, Chung H, Lee H, Sung YA. Relationship between the characteristic traits of polycystic ovary syndrome and susceptibility genes. *Sci Rep* 2020; **10**:10479.

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020; **581**:434–443.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res* 2002; **12**:996–1006.

Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, Samovića M, Sakthivel MP, Kuzmin I, Trevanion SJ et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* 2021; **53**:1290–1299.

Kichaev G, Bhatia G, Loh P, Gazal S, Burch K, Freund MK, Schoech A, Pasaniuc B, Price AL. Leveraging polygenic functional enrichment to improve GWAS power. *Am J Hum Genet* 2019; **104**:65–75.

Kilpivaara O, Vahteristo P, Falck J, Syrjäkoski K, Eerola H, Easton D, Bartkova J, Lukas J, Heikkilä P, Aittomäki K et al. CHEK2 variant I157T may be associated with increased breast cancer risk. *Int J Cancer* 2004; **111**:543–547.

Koivuaho E, Laru J, Ojaniemi M, Puukka K, Kettunen J, Tapanainen JS, Franks S, Järvelin M-R, Morin-Papunen L, Sebert S et al. Early childhood BMI rise, the adiposity rebound, associates with PCOS diagnosis and obesity at ages 31 and 46 years—analysis of 46-year growth data from birth to adulthood in PCOS. *Int J Obes (Lond)* 2019; **43**:1370–1379.

Kuusisto KM, Bebel A, Vihinen M, Schleutker J, Sallinen SL. Screening for *BRCA1*, *BRCA2*, *CHEK2*, *PALB2*, *BRIP1*, *RAD50*, and *CDH1* mutations in high-risk Finnish *BRCA1*/2-founder mutation-negative breast and/or ovarian cancer individuals. *Breast Cancer Res* 2011; **13**:R20.

Lee H, Oh JY, Sung YA, Chung H, Kim HL, Kim GS, Cho YS, Kim JT. Genome-wide association study identified new susceptibility loci for polycystic ovary syndrome. *Hum Reprod* 2015; **30**:723–731.

Legro RS. Obesity and PCOS: implications for diagnosis and treatment. *Semin Reprod Med* 2012; **30**:496–506.

Leitsalu L, Haller T, Esko T, Tammesoo ML, Alavere H, Snieder H, Perola M, Ng PC, Mägi R, Milani L et al. Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* 2015; **44**:1137–1147.

Li J, Eriksson M, Czene K, Hall P, Rodriguez-Wallberg KA. Common diseases as determinants of menopausal age. *Hum Reprod* 2016; **31**:2856–2864.

Lim ET, Würtz P, Havulinna AS, Palta P, Tukiainen T, Rehnström K, Esko T, Mägi R, Inouye M, Lappalainen T et al. Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet* 2014; **10**:e1004494.

Lim SS, Kakoly NS, Tan JWJ, Fitzgerald G, Bahri Khomami M, Joham AE, Cooray SD, Misso ML, Norman RJ, Harrison CL et al. Metabolic syndrome in polycystic ovary syndrome: a systematic review, meta-analysis and meta-regression. *Obes Rev* 2019; **20**:339–352.

Locke AE, Steinberg KM, Chiang CWK, Service SK, Havulinna AS, Stell L, Pirinen M, Abel HJ, Chiang CC, Fulton RS et al. Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature* 2019; **572**:323–328.

Loh PR, Danecek P, Palamara PF, Fuchsberger C, A Reshef Y, K Finucane H, Schoenherr S, Forer L, McCarthy S, Abecasis GR et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016; **48**:1443–1448.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A et al. Finding the missing heritability of complex diseases. *Nature* 2009; **461**:747–753.

March WA, Moore VM, Willson KJ, Phillips DI, Norman RJ, Davies MJ. The prevalence of polycystic ovary syndrome in a community sample assessed under contrasting diagnostic criteria. *Hum Reprod* 2010; **25**:544–551.

Mars N, Widén E, Kerminen S, Meretoja T, Pirinen M, Della Briotta Parolo P, Palta P, Palotie A, Kaprio J, Joensuu H et al. The role of polygenic risk and susceptibility genes in breast cancer over the course of life. *Nat Commun* 2020; **11**:6383.

Martin AR, Karczewski KJ, Kerminen S, Kurki MI, Sarin A, Artomov M, Eriksson JG, Esko T, Genovese G, Havulinna AS et al. Haplotype sharing provides insights into fine-scale population history and disease in Finland. *Am J Hum Genet* 2018; **102**:760–775.

Meijers-Heijboer H, van den Ouwendijk A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M et al. Low-penetrance susceptibility to breast cancer due to *CHEK2*(1100delC) in noncarriers of *BRCA1* or *BRCA2* mutations. *Nat Genet* 2002; **31**:55–59.

Minooe S, Ramezani Tehrani F, Rahmati M, Mansournia MA, Azizi F. Prediction of age at menopause in women with polycystic ovary syndrome. *Climacteric* 2018; **21**:29–34.

Mitt M, Kals M, Pärn K, Gabriel SB, Lander ES, Palotie A, Ripatti S, Morris AP, Metspalu A, Esko T et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel. *Eur J Hum Genet* 2017; **25**:869–876.

Moghetti P, Tosi F. Insulin resistance and PCOS: chicken or egg? *J Endocrinol Invest* 2021; **44**:233–244.

Muranen TA, Blomqvist C, Dörk T, Jakubowska A, Heikkilä P, Fagerholm R, Greco D, Aittomäki K, Bojesen SE, Shah M et al. Patient survival and tumor characteristics associated with *CHEK2*:p.I157T—findings from the Breast Cancer Association Consortium. *Breast Cancer Res* 2016; **18**:98.

Mustafa MK, Tanoue Y, Tateishi C, Vaziri C, Tateishi S. Roles of Chk2/CHEK2 in guarding against environmentally induced DNA damage and replication-stress. *Environ Mol Mutagen* 2020; **61**:730–735.

Nelis M, Esko T, Mägi R, Zimprich F, Zimprich A, Toncheva D, Karachanak S, Piskáčková T, Balaščák I, Peltonen L et al. Genetic structure of Europeans: a view from the North–East. *PLoS One* 2009; **4**:e5472.

Ollila MM, Piltonen T, Puukka K, Ruokonen A, Jarvelin MR, Tapanainen JS, Franks S, Morin-Papunen L. Weight gain and dyslipidemia in early adulthood associate with polycystic ovary syndrome: prospective cohort study. *J Clin Endocrinol Metab* 2016; **101**:739–747.

Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. *Nat Rev Genet* 2000; **1**:182–190.

Piltonen T, Morin-Papunen L, Koivunen R, Perheentupa A, Ruokonen A, Tapanainen JS. Serum anti-Müllerian hormone levels remain high until late reproductive age and decrease during metformin therapy in women with polycystic ovary syndrome. *Hum Reprod* 2005; **20**:1820–1826.

Prohaska A, Racimo F, Schork AJ, Sikora M, Stern AJ, Ilardo M, Allentoft ME, Folkersten L, Buil A, Moreno-Mayar JV et al. Human disease variation in the light of population genomics. *Cell* 2019; **177**:115–131.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.

Risal S, Pei Y, Lu H, Manti M, Fornes R, Pui HP, Zhao Z, Massart J, Ohlsson C, Lindgren E et al. Prenatal androgen exposure and transgenerational susceptibility to polycystic ovary syndrome. *Nat Med* 2019; **25**:1894–1904.

Ruth KS, Day FR, Hussain J, Martínez-Marchal A, Aiken CE, Azad A, Thompson DJ, Knoblochova L, Abe H, Tarry-Adkins J, et al. Genetic insights into the biological mechanisms governing human ovarian ageing. *Nature* 2021; **592**:393–397.

Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A, Sistonen P, Savontaus M, Schreiber S, Kere J et al. Genome-wide analysis of single nucleotide polymorphisms uncovers population structure in northern Europe. *PLoS One* 2008; **3**:e3519.

Salonen JT, Uimari P, Aalto JM, Pirskanen M, Kaikkonen J, Todorova B, Hyppönen J, Korhonen VP, Asikainen J, Devine C et al. Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium. *Am J Hum Genet* 2007; **81**:338–345.

Schutte M, Seal S, Barfoot R, Meijers-Heijboer H, Wasielewski M, Evans DG, Eccles D, Meijers C, Lohman F, Klijn J et al. Variants in *CHEK2* other than 1100delC do not make a major contribution to breast cancer susceptibility. *Am J Hum Genet* 2003; **72**:1023–1028.

Shi Y, Zhao H, Shi Y, Cao Y, Yang D, Li Z, Zhang B, Liang X, Li T, Chen J et al. Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. *Nat Genet* 2012; **44**:1020–1025.

Silva MSB, Giacobini P. New insights into anti-Müllerian hormone role in the hypothalamic-pituitary-gonadal axis and neuroendocrine development. *Cell Mol Life Sci* 2021; **78**:1–16.

Skiba MA, Islam RM, Bell RJ, Davis SR. Understanding variation in prevalence estimates of polycystic ovary syndrome: a systematic review and meta-analysis. *Hum Reprod Update* 2018; **24**:694–709.

Sousa AD, Cheney RE. Myosin-X: a molecular motor at the cell's finger tips. *Trends Cell Biol* 2005; **15**:533–539.

Sund R. Quality of the Finnish Hospital Discharge Register: a systematic review. *Scand J Public Health* 2012; **40**:505–515.

Tambets K, Yunusbayev B, Hudjashov G, Ilumäe A, Roots S, Honkola T, Vesakoski O, Atkinson Q, Skoglund P, Kushniarevich A et al. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol* 2018; **19**:139.

Teede HJ, Misso ML, Costello MF, Dokras A, Laven J, Moran L, Piltonen T, Norman RJ; International PCOS Network. Recommendations from the international evidence-based guideline for the assessment and management of polycystic ovary syndrome. *Hum Reprod* 2018; **33**:1602–1618.

Tharp ME, Malki S, Bortvin A. Maximizing the ovarian reserve in mice by evading LINE-1 genotoxicity. *Nat Commun* 2020; **11**:330.

Tyrim JS. STAPLER: a simple tool for creating, managing and parallelizing common high-throughput sequencing workflows. *bioRxiv* 2018;445056.

Veikkolainen V, Ali N, Doroszko M, Kiviniemi A, Miinalainen I, Ohlsson C, Poutanen M, Rahman N, Elenius K, Vainio SJ et al. Erbb4 regulates the oocyte microenvironment during folliculogenesis. *Hum Mol Genet* 2020; **29**:2813–2830.

Vink JM, Sadrzadeh S, Lambalk CB, Boomsma DI. Heritability of polycystic ovary syndrome in a Dutch twin-family study. *J Clin Endocrinol Metab* 2006; **91**:2100–2104.

Ward LD, Parker MM, Deaton AM, Tu H, Flynn-Carroll A, Hinkle G, Nioi P. Rare coding variants in five DNA damage repair genes associate with timing of natural menopause. *medRxiv* 2021; 2021.04.18.21255506.

Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017; **8**:1826.

Wen Y, Wu X, Peng H, Li C, Jiang Y, Su Z, Liang H, Liu J, He J, Liang W. Breast cancer risk in patients with polycystic ovary syndrome: a Mendelian randomization analysis. *Breast Cancer Res Treat* 2021; **185**:799–806.

Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**:2190–2191.

Wojciechowski P, Lipowska A, Rys P, Ewens KG, Franks S, Tan S, Lerchbaum E, Vcelak J, Attaoua R, Straczkowski M et al. Impact of FTO genotypes on BMI and weight in polycystic ovary syndrome: a systematic review and meta-analysis. *Diabetologia* 2012; **55**:2636–2645.

Wu P, Li R, Zhang W, Hu H, Wang W, Lin Y. Polycystic ovary syndrome is causally associated with estrogen receptor-positive instead of estrogen receptor-negative breast cancer: a Mendelian randomization study. *Am J Obstet Gynecol* 2020; **223**:583–585.

Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; **44**:369–75, S1–3.

Zhang Y, Ho K, Keaton JM, Hartzel DN, Day F, Justice AE, Josyula NS, Pendergrass SA, Actkins K, Davis LK et al. A genome-wide association study of polycystic ovary syndrome identified from electronic health records. *Am J Obstet Gynecol* 2020; **223**:559.e1–21.

Zhang Y, Kent JW, Jr, Olivier M, Ali O, Cerjak D, Broeckel U, Abdou RM, Dyer TD, Comuzzie A, Curran JE et al. A comprehensive analysis of adiponectin QTLs using SNP association, SNP cis-effects on peripheral blood gene expression and gene expression correlation identified novel metabolic syndrome (MetS) genes with potential role in carcinogenesis and systemic inflammation. *BMC Med Genomics* 2013; **6**:14.

Zhao Y, Xu Y, Wang X, Xu L, Chen J, Gao C, Wu C, Pan D, Zhang Q, Zhou J et al. Body mass index and polycystic ovary syndrome: a 2-sample bidirectional Mendelian randomization study. *J Clin Endocrinol Metab* 2020; **105**:dga125.

Zhou W, Nielsen JB, Fritzsche LG, Dey R, Gabrielsen ME, Wolford BN, LeFaive J, VandeHaar P, Gagliano SA, Gifford A et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* 2018; **50**:1335–1341.

Zhu T, Cui J, Goodarzi MO. Polycystic ovary syndrome and breast cancer subtypes: a Mendelian randomization study. *Am J Obstet Gynecol* 2021; **225**:99–101.