Unraveling the Etiology of Complex Traits by Extending GWAS to Causal Inference and Copy Number Variation Analysis

Daniel Schmitz

UPPSALA UNIVERSITET

Half-Time Thesis

Department of Immunology, Genetics and Pathology,
Science for Life Laboratory, Uppsala University

2021-11-05

Main Supervisor
Åsa Johansson, PhD
Associate Professor
Department of Immunology, Genetics and Pathology. Medical Genetics and Genomics, Science for Life Laboratory, Uppsala University, Sweden

Co-Supervisors
Torgny Karlsson, PhD
Researcher
Department of Immunology, Genetics and Pathology. Medical Genetics and Genomics, Science for Life Laboratory, Uppsala University, Sweden

Adam Ameur, PhD
Bioinformatician
Department of Immunology, Genetics and Pathology, Uppsala Genome Center, Uppsala University, Sweden

Review Committee
Nils Landegren, PhD
Associate senior lecturer/Assistant Professor
Department of Medical Biochemistry and Microbiology, Comparative genetics and functional genomics, Uppsala University, Sweden.

Tomas Bergström, PhD
Researcher
Department of Animal Breeding and Genetics; Molecular genetics, HgenLab, Sveriges Lantbruksuniversitet, Sweden

Susanna Larsson, PhD
Senior lecturer/Associate Professor
Department of Surgical Sciences, Medical epidemiology, Uppsala University, Sweden

© 2021 Daniel Schmitz

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit `http://creativecommons.org/licenses/by/4.0/`.

Abstract

Introduction. Since the release of the first assembly of the human genome in 2001, our knowledge of the genetic architecture of complex traits and diseases has grown steadily. Genome-wide association studies (GWAS), which investigate the association between single-nucleotide polymorphisms (SNPs) and traits or disease risks, have played a major role in this development. Moreover, the emergence of next-generation sequencing (NGS) technology led to an ever-increasing amount of genetic data available to research. These developments have made sequencing and GWAS common approaches for research studies. Recently, Mendelian Randomization (MR) has been popularized as a method to investigate the relationship between traits using readily available GWAS results. Despite these successes, GWAS have not been able to completely explain the heritability that is observed in complex traits. In part, this is due to GWAS being mostly restricted to SNPs. However, large copy-number variations (CNVs) have been shown to be associated with diseases but have largely been absent from GWAS. CNVs are difficult to sequence using the short reads generated by NGS. Novel long-read sequencing technologies have emerged and promise to make CNV detection easier.

Project I. In project I, we aimed to quantify the effect of SNPs on circulating estradiol levels and the effect of estradiol on bone mineral density (BMD). We performed a GWAS using genotyping and estradiol measurements from UK Biobank (UKB) in males ($N = 147,690$) and females ($N = 163,985$). Estradiol was transformed into a binary phenotype (above/below detection limit of $175\,\mathrm{pmol}/\ell$). We then quantified the effect of estradiol on BMD using MR. We found 14 independent loci with genome-wide significant associations ($p < 5 * 10^{-8}$) with estradiol levels in males. One locus was also significant in females and we identified an additional locus specific to females. We found a significant effect of detectable estradiol levels on BMD in males ($p = 1.58 * 10^{-11}$) and females ($p = 7.48 * 10^{-6}$). These results confirm previous research into the effect of estrogens on skeletal health.

Project II. In project II, we aimed to identify the effect of endogenous estradiol on breast, endometrial and ovarian cancer. We performed a MR analysis using the GWAS results from project I, estimating the effect of estradiol on the aforementioned cancers. We found an association between estradiol levels and breast ($p = 0.0074$) and endometrial ($p = 0.0065$) but not ovarian cancer. These findings highlight the effect of the body's own estrogen production on cancer risk. Our results for endometrial cancer confirm previous MR studies.

Project III. In project III, we aimed to identify associations between CNVs and blood plasma protein levels and validate our findings using long-read sequencing. We performed whole-genome sequencing (WGS) in a cohort from Northern Sweden ($N = 1,021$) and called CNVs using CNVnator. We measured 438 plasma protein biomarkers in 892

participants using Olink Protein Extension Assay. Among the 872 participants with genotyping and protein data, we tested for association between copy numbers (CNs) and protein levels. We validated five polymorphic CNVs in 15 individuals by WGS using Pacific Biosciences Single-Molecule Real-Time Sequencing (SMRT) sequencing. A total of 243,987 polymorphic non-overlapping CNV loci were identified. After merging of adjacent loci with consistent CNs, 23,381 remained. Significant associations ($p < 4.79 * 10^{-9}$) were found between the CNs of 30 independent CNVs and the expression level of 17 protein biomarkers. Out of these CNVs, the breakpoints for two CNVs were validated by SMRT sequencing and two CNVs were identified to be clusters of many short repetitive elements. The fifth CNV could not be validated conclusively. Our findings provide insight into the effects of CNVs on protein biomarkers and highlight the application of high-throughput sequencing for CNV detection, including issues with comparability between these technologies.

List of Publications

This thesis is based on the following papers and projects, referred to in the text by their numbers.

I. Schmitz, D., Ek, W. E., Berggren, E., Höglund, J., Karlsson, T. & Johansson, Å. Genome-wide Association Study of Estradiol Levels and the Causal Effect of Estradiol on Bone Mineral Density. *The Journal of Clinical Endocrinology & Metabolism* **XX,** 1–16. ISSN: 0021-972X. `https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117` (July 2021)

II. Johansson, Å., Schmitz, D., Höglund, J., Hadizadeh, F., Karlsson, T. & Ek, W. E. *High oestradiol levels cause an increased risk for breast and endometrial cancer* Uppsala, 2021 *submitted for publication*

III. Characterizing Copy Number Variations using Next- and Third-Generation Sequencing and their Association with Plasma Biomarkers. *In preparation*

IV. Unspecified

Related Publications

The following publications are not part of the main research project.

- Kierczak, M., Rafati, N., Höglund, J., Gourlé, H., Schmitz, D., Ek, W. E., Enroth, S., Ekman, D., Nystedt, B., Karlsson, T. & Johansson, Å. The contribution of rare whole genome sequencing variants to plasma protein levels and to the missing heritability. *Nature Communications* **In prepara.** `https://www.researchsquare.com/article/rs-625433/v1` (June 2021)

Contents

Abstract iii

List of Publications v

1 Introduction 1
- 1.1 Genome-Wide Association Studies . . . 1
- 1.2 Mendelian Randomization . . . 2
- 1.3 Copy Number Variations . . . 4
- 1.4 Sequencing Technology . . . 5
- 1.5 Study Cohorts . . . 6
  - 1.5.1 UK Biobank . . . 6
  - 1.5.2 Northern Swedish Population Health Study . . . 6

2 Project I 7
- 2.1 Background . . . 7
- 2.2 Methods . . . 7
- 2.3 Results and Discussion . . . 8
  - 2.3.1 Genotyping and Estradiol Measurements . . . 8
  - 2.3.2 GWAS Results . . . 9
  - 2.3.3 MR Results . . . 10

3 Project II 11
- 3.1 Background . . . 11
- 3.2 Methods . . . 11
- 3.3 Results and Discussion . . . 12
  - 3.3.1 Cancer-Specific Results . . . 13
  - 3.3.2 Limitations . . . 13

4 Project III 15
- 4.1 Background . . . 15
- 4.2 Methods . . . 15
- 4.3 Results and Discussion . . . 16
- 4.4 Future Work . . . 17

5 Project IV 18
- 5.1 Future Work . . . 18

6 Concluding Remarks 19

Acknowledgements **20**

Acronyms **21**

Bibliography **23**

1 Introduction

Since the first assembly of the human genome was published in 2001, numerous links between genetic variants and complex traits as well as diseases have been established [1]. Traditionally, single-nucleotide polymorphism (SNP) genotyping followed by genome-wide association studies (GWAS) has been the way to go. During recent years however, next-generation sequencing (NGS) and even long-read sequencing have become increasingly mature and affordable. This development allows not only for the genotyping of rare SNPs, but also the detection of large structural variations (SVs). In addition to identifying genetic associations, post-GWAS analyses have been developed. For instance, Mendelian Randomization (MR) studies allow researchers to establish relationships between traits or disease risk through the use of GWAS data.

1.1 Genome-Wide Association Studies

GWAS identify associations between genetic variants and phenotypes. The majority of GWAS investigate SNPs and small insertions and deletions (indels) for associations. However, the method is not restricted to these kinds of variants.

To perform a GWAS, it is first necessary to genotype a large cohort of individuals. Usually, this is done using SNP arrays, which only allow for the detection of previously identified variants but are generally cheap and reliable. Other genotyping approaches which use NGS, such as whole-genome sequencing (WGS) and whole-exome sequencing (WES), have increased in popularity and have recently become feasible for large-scale studies. After genotyping, SNPs that were not present on the genotyping array are often imputed from the ones that were present based on linkage disequilibrium (LD).

Suitable phenotype for GWAS can be quantitative traits, e.g. height or levels of certain proteins in blood, or binary, e.g. the presence of a certain disease. Data can be obtained in a variety of ways depending on the trait to be assessed. The data can stem from measurements of blood samples, physical assessments, questionnaires or registries. In addition to the specific phenotype the study is supposed to investigate, a number of other traits are recorded, which may be included to increase precision in the effect estimates, or which are adjusted for because of an interest in a direct effect of the SNP on the phenotype, independent of the other trait.

After all data was collected, a GWAS identifies associations by applying a linear regression model for quantitative traits or a logistic model for binary traits. This model uses the phenotype as response, each SNP as predictor and all phenotypes that might confound as covariates. To control for multiple testing, an adjusted significance threshold of $p < 5 * 10^{-8}$ is commonly applied. This is equivalent to adjusting for 1 million independent variants [2]. Because of LD, not all variants that pass genome-wide

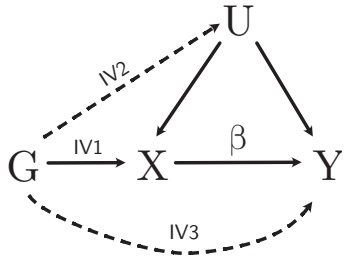


Figure 1.1: Graphical representation of the assumptions of MR. The strength β of the causal effect of X on Y is to be estimated. The genetic variants G affect X (IV1) but neither any confounders U (IV2) nor Y (IV3) directly. Based on illustration from [4].

significance are reported. Instead, variants that are close together, are clumped into one locus and only the variant with the strongest association, the *lead SNP*, is reported.

Over the last two decades, GWAS have enabled us to gain unprecedented insight into the genetic architecture of human diseases and complex traits. Increasingly larger cohorts such as UK Biobank (UKB) contributed to the continuously improving statistical power of GWAS, which accelerated this trend. As of the time of writing, the GWAS Catalog—the largest database of GWAS results—contains 293,040 associations of SNPs and indels to phenotypes [3].

1.2 Mendelian Randomization

In addition to elucidating the genetic architecture of traits and disease risk, GWAS results can also be leveraged to infer associations between traits through an approach called MR (fig. 1.1). MR studies apply a design based on randomized controlled trials to estimate the effect of an exposure X on an outcome Y using genetic variants as instrumental variables (IVs). The underlying reasoning is that if there is a variant that affects the exposure, then it should indirectly and proportionally affect the outcome through the exposure. In order for a variant to be a valid IV, it must meet the following assumptions:

IV1) It is associated with the exposure.

IV2) It is not associated with any confounders to the exposure-outcome relation.

IV3) It affects the outcome exclusively through the exposure.

MR avoids problems of conventional observational studies such as reverse causation because its IVs are assigned at conception. Therefore, an outcome, which manifests later in life, cannot affect the genetic variant.

There are two major designs for MR studies [5]. One-sample MR uses exposure and outcome data from the same cohort. This allows for an easy study design, because you only need to measure an additional phenotype, but it might not always be possible to perform these measurements

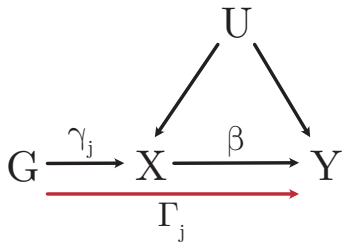


Figure 1.2: Graphical representation of the summary-based MR approach. Γ_j refers to the effect size estimated by a GWAS, which only assesses association, not causality, and not a pleiotropic effect of G on Y. The effect β can be calculated according to $\beta = \frac{\Gamma_j}{\gamma_j}$.

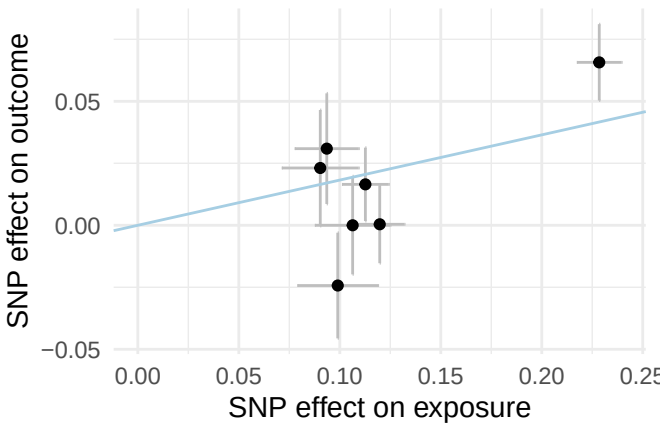


Figure 1.3: Example of a two-sample MR result. The x-axis shows the estimated effect of the IVs on the exposure. The y-axis shows the effect on the outcome. Each point represents one IV. The whiskers correspond to the standard error of each IV's effect estimate The effect of the exposure on the outcome is equal to the slope of the regression line.

Two-sample MR studies on the other hand use GWAS summary statistics from two different cohorts, one for the SNP effects on the exposure and one for the effects on the outcome, respectively. Thanks to platforms like the GWAS Catalog and MR Base, summary statistics to use for two-sample MR studies are easily available. Often, not all genotyped or imputed SNPs are available in all studies. This makes it necessary to find proxy SNPs as replacement for lead SNPs that are not present in both studies' data.

MR studies can either be individual-based, which predict the exposure measurement for each individual based on their alleles and use this prediction as a predictor for the outcome, or summary-based, i.e. use effect estimates from previous GWAS. The summary-based approach has become more popular, thanks to the wide availability of summary statistics online. In short, this approach estim ates β based on the observed effect γ_j of the variant G_j on X and its effect Γ_j on Y (fig. 1.2). If X has a causal effect on Y, then the calculated β should be constant for all G_j (fig. 1.3).

There are multiple algorithms to estimate the causality between X and Y based on summary statistics. They are based on different approaches, which, among other things, affect their their ability to account for outliers or invalid instruments. The most commonly used methods include the *weighted median* and *inverse-variance weighted (IVW)* methods, which apply two different measures of location, i.e., the median and the mean, respectively, to determine the causal effect estimate. Both methods assume no directional pleiotropy, which occurs when the combined pleiotropic effects of the chosen IVs do not add up to zero [6, 10]. However, the MR-Egger method can take directional pleiotropy into account, but its use is discouraged in one-sample studies [8, 12]. Finally, generalized summary-based Mendelian Randomization (GSMR), which allows for the detection and removal of pleiotropic IVs, has become a popular method for MR studies [10].

1.3 Copy Number Variations

Despite the advances in the statistical power of GWAS, only a fraction of observed heritability of many traits can be explained by SNPs. The 1000 Genomes Project estimated in 2015 that 99.9% of variants present in a typical human genome are SNPs and indels [11]. However, they note that another kind of variation, SVs, which are variants that affect more than 50 bp, cover a larger fraction of the genome. Another study from the same year found that SVs account for approximately 13% of genetic variation in humans [12]. Nonetheless, only few studies have investigated associations between SVs and diseases or complex traits.

A big group of SVs that has received more attention are copy-number variations (CNVs). A CNV is a variant that affects the number of copies, i.e. the copy number (CN), of a genomic sequence (fig. 1.4). Short CNVs have well established links to diseases such as Huntington's Disease and Fragile X Syndrome [13, 17]. However, associations of longer CNVs spanning multiple kbp or even Mbp are not well studied.

Despite their clinical significance, research into the effects of CNVs on risk of disease and health-related traits is still lacking. One important reason for the lack of research

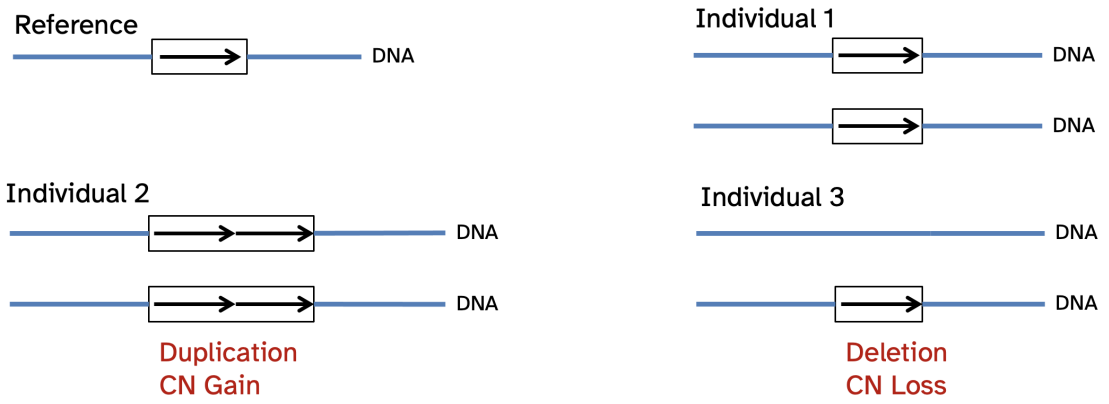


Figure 1.4: Example of a CNV. Each arrow represents a copy of the sequence of interest. There is one copy in the (haploid) reference. In a diploid individual, this corresponds to CN 2 (Individual 1). Individual 2 is homozygous for two copies of the sequence (CN 4). An increase in the number of copies is called a *duplication* or *CN gain*. Individual 3 has lost a copy of the sequence on one chromosome (CN 1). This is called a *deletion* or *CN loss*.

on the effects of CNVs is the fact that they are hard to detect. Most CNV studies use microarray-based technologies such as array comparative genomic hybridization (aCGH). However, these lack the ability to detect novel signals and have a limited resolution. Common array-based workflows can only detect CNVs with a size of at least 8 kbp, on average [15]. Thanks to improving quality and cost of WGS, the interest in CNVs has increased dramatically, as this approach allows for the detection of CNV loci not covered before.

1.4 Sequencing Technology

Sequencing technology has changed a lot over the last two decades. While most of the Human Genome Project's first assembly was sequenced using Sanger sequencing, this technology proved to be too expensive and slow to allow for widespread use for human genomics [1].

Over the following years, technological advances enabled the emergence of a high-throughput sequencing, or NGS. These technologies allow to sequence billions of bases at a time while reducing cost. Over the years, Illumina and their technology based on sequencing by synthesis (SBS) have become the market leader in this segment. Their platform produces short paired-end reads with a usual length of around 150 bp and leverages increasing computational power to align these short reads to the human genome. While Illumina sequencing is highly accurate for the detection of SNPs and indels, they are too short to reliably represent large SVs. Moreover, the need for amplification via polymerase chain reaction (PCR) of the DNA to be sequenced makes Illumina sequencing vulnerable to PCR amplification bias. This leads to low coverage in regions with very

high or low GC content [16].

Several approaches for CNV detection have been developed for short-read alignments. In general, these can be divided into two major groups: depth-based methods (e.g. CNVnator [17]), which identify CNVs through regions of abnormally high or low coverage, and read-based methods (e.g. Manta [18]), which rely on anomalies in fragment size and read-pair orientation.

Recently, third-generation sequencing technologies, in particular Pacific Biosciences (PacBio) Single-Molecule Real-Time Sequencing (SMRT) sequencing and Oxford Nanopore, have matured and become usable for larger research projects. They are able to produce longer reads than previous technologies. SMRT reads usually range between 15 kbp to 20 kbp, while Nanopore can produce reads of up to 4 Mbp [19, 23]. Furthermore, amplification-free protocols have been developed for long-read sequencing, which avoid the issues caused by PCR amplification bias [21]. However, they suffer from low read accuracy, which makes them less suitable for SNP calling [22].

Several SV callers exist for long-read alignments. These include Sniffles, SVIM and PacBio's own PBSV [23, 27]. They differ in how they incorporate SV signatures from single reads and clusters of reads. Furthermore, PBSV is the only caller to directly report CNVs and CNs.

1.5 Study Cohorts

1.5.1 UK Biobank

UKB is a long-time cohort study comprising about 500,000 participants born between 1939 and 1970 that were recruited from 2006 to 2010. Extensive information on lifestyle and anthropometric traits was collected. Participants left blood samples that were used for diverse measurements, including hormone levels, as well as genotyping of approximately 800,000 SNPs.

1.5.2 Northern Swedish Population Health Study

e employed the Northern Swedish Population Health Study (NSPHS), a cohort study consisting of approx. 1000 individuals from two municipalities in northern Sweden, as the base cohort of our study. Blood samples were collected on site and immediately frozen at −70 °C. WGS was performed on an Illumina HiSeq X platform according to manufacturer specification with a target coverage of 30× [25]. After quality control (QC) and mapping to reference genome GRCh37, 1021 individuals remained.

We measured 438 plasma biomarkers measured using Olink protein extension assay (PEA) in 903 individuals. Of these, 982 individuals passed QC for all protein measurements. Overall, there were 872 individuals passing both genotype and protein QC.

2 Project I

Contributions: performed data analysis and interpretation, generated figures, wrote the manuscript

2.1 Background

Estrogen, which is generally known as the primary female sex hormone, is responsible for the female reproductive system's development. Furthermore, it regulates the menstrual cycle and plays a critical role in male sexual function [26, 30]. Among the three major forms of estrogen: estrone, estradiol and estriol, estradiol is the most potent and abundant [28].

Estradiol levels have been associated with several conditions, incl. deep vein thrombosis, cancers and type 2 diabetes (T2D) [29–34]. In particular, declining estradiol levels after menopause have been linked to reduced bone mineral density (BMD) and, in turn, higher risk of osteoporosis [32, 36].

Previous GWAS for estradiol levels have been performed in sex-stratified populations comprising up to 11,000 people, most often of European descent [34–41]. Additionally, a recent study in UKB identified strong sex-specific genetic effects on testosterone but excluded associations with estradiol measurements because of their strong link to age at menopause [39].

Apart from GWAS, the causal effect of hormones on diseases and disease risk has been assessed using MR [38–43]. While previous MR studies identified a beneficial effect of estradiol on BMD in males, there have been no such studies in females due to failure to identify valid instruments for MR analyses [38, 40].

In this project, we aimed to identify variants affecting estradiol. Additionally, we aimed to estimate the causal effect of estrogen on BMD.

2.2 Methods

We used data from UKB to perform this study. We used SNP data imputed using UK10K and 1000 genomes phase 3 reference panels, containing 93,093,070 SNPs overall. After QC, 361,975 individuals remained, of which 167,168 were male and 194,807 were female.

Only measurements that were taken from blood samples given at the first visit at the assessment center were included in the analysis. Estradiol was measured by two-step competitive analysis using a Beckman Coulter Unicel Dxl 800. The assay had a lower detection limit of 175 pmol/ℓ, which is above the normal range for serum estradiol concentrations in postmenopausal females (0 – 73.4 pmol/ℓ) [41]. Because of the resulting

large fraction of measurements below detection limit, estradiol levels were analyzed as a binary phenotype (above/below detection limit).

We performed two sex-stratified GWAS using logistic regression with additive genetic modeling in PLINK 2. We included age, body-mass index (BMI), the first ten genetic principal components (PCs) and the used genotyping array as covariates, as well as a binary indicator for the used genotyping array to control for batch effects. For females, we included hormone-replacement therapy (HRT), oral contraceptive (OC) use (never/ever/current), number of live births, menopausal status and whether they had had a hysterectomy, too. We identified lead SNPs by applying conditional analyses until no significant hits remained. We tested all lead SNPs for sex-specific effects by including an interaction term in the logistic model. We performed four sensitivity analyses. We stratified females into pre- and postmenopausal, excluded all patients with cancer diagnoses and included testosterone and sex hormone-binding globulin (SHBG) levels as covariates. Lastly, we applied a Tobit-I model, which allowed us to incorporate quantitative estradiol measurements where available.

We annotated our lead SNPs to their closest genes and identified possible functional effects using HaploReg version 4.1 [42]. We used data from the Genotype-Tissue Expression project (GTEx) to check for overlap with known expression quantitative trait loci (eQTL) [43]. Lastly, data from the GWAS Catalog was used to search for prior functional annotation of our lead SNPs.

To estimate the effect of estradiol on BMD, we performed a one-sample MR analysis in males and females separately using our GWAS lead SNPs as IVs. Due to the low number of significant associations in the female cohort, we applied a relaxed significance threshold ($p < 10^{-7}$) tp increase the number of available IVs. BMD had been recorded using an ultrasound measurement and converted to T-Scores, i.e. the number of standard deviations the measurement differed from the patient's sex's mean.

The main MR analysis was performed using the gsmr package, version 1.0.8 [10], which implements the GSMR method. We performed sensitivity analyses using the IVW, weighted median and MR-Egger methods implemented in the TwoSampleMR package in R [44]. Additionally, we performed a two-sample MR test using the aforementioned methods with summary statistics for lumbar-spine BMD from the Genetic Factors for Osteoporosis Consortium (GEFOS) [45]. Because GEFOS had used a different imputation panel than UKB, we revised the set of IVs. For each locus, we selected the most significant common SNP in LD with our lead SNP.

2.3 Results and Discussion

2.3.1 Genotyping and Estradiol Measurements

After genotype and estradiol QC, 147,690 males remained, of whom 134,323 had estradiol below and 13,367 above detection limit (175 pmol/ℓ). Slightly more females (163,985) remained, with 126,524 individuals having estradiol levels below detection limit and 37,461 above. Only 9.1% of males and 7.9% of postmenopausal females had detectable estradiol levels, while 71.9% of premenopausal females had measurements.

The low number of individuals with available measurements was due to the estradiol assay having a low sensitivity. A more sensitive method would have been preferable. For future studies aiming to elucidate the genetic factors behind estradiol levels, we would recommend the usage of an assay that can detect lower concentrations of estradiol. In fact, estradiol assays with a detection limit as low as 2 pg/mℓ ($\approx$ 7.34 pmol/ℓ) are available to researchers [46].

2.3.2 GWAS Results

We found 15 loci on 14 chromosomes to be significantly associated ($p < 5 * 10^{-8}$) with estradiol levels, of which 13 were specific to males, one (*MCM8*) specific to females and one (*CYP3A7*) shared between both sexes. We identified one conditional hit each on chromosomes 2 and 15. 12 of our GWAS hits had already established links to estradiol or steroid-hormone metabolism. An additional analysis using a sex-genotype interaction term revealed strong sex-specific effects. Most of the loci we identified have previously established links to steroid-hormone metabolism, including synthesis, conversion, transport and elimination of steroid hormones.

After stratification of the female cohort into pre- and postmenopausal, we identified one locus (*CYP3A7*) with significantly different effects between the two strata. Removal of all participants with previous cancer diagnoses did not lead to a change in our primary GWAS results. Lastly, adjusting our model for testosterone and SHBG levels caused the loci *SHBG* and *FKBP4* to lose genome-wide significance with lower effect estimates. The loci *AR* and *UGT3A1* lost significance, too, but the estimated odds ratios (ORs) did not differ significantly.

Thanks to our secondary analysis using Tobit-I-modeling, we could somewhat ameliorate the problem of missing estradiol measurements. This statistical approach allowed us to make use of estradiol measurements where available while still retaining information about which missing values were below detection limit. This analysis mostly agreed with our primary results. Among males, 11 SNPs and all SNPs that were significant for females replicated. The effect estimates we obtained from our Tobit model were comparable to previous GWAS of estradiol measurements [34, 37, 41]. We argue therefore that Tobit modeling provides a viable approach for GWAS of phenotypes for which high-sensitivity assays might not be available.

CYP3A7 was significant in both males and females, indicating an important role in estrogen metabolism in both sexes. Interestingly, it was also the only gene to have significantly different effects in pre- and postmenopausal females. *CYP3A7* encodes cytochrome P450 3A7 (CYP3A7), which metabolizes a precursor of both androgens and estrogens: dehydroepiandrosterone (DHEA) [47]. Up to 75% of estrogens in premenopausal women are derived from DHEA and after menopause, it is the main precursor of androgens and estrogens [48]. When adjusting for SHBG and testosterone, *CYP3A7*'s effect disappeared in females. These findings point to CYP3A7 fulfilling different functions for steroid-hormone metabolism in males and females.

Interestingly, *ABO*, the gene responsible for the ABO blood groups, was associated with estradiol levels in males [49]. Our effect allele (rs657152-A) is in LD with rs8176719-G,

which is present in individuals that do not have blood type O. This indicates that people with blood type O have higher estradiol levels.

2.3.3 MR Results

We estimated the causal effect of estradiol on BMD using MR in males and for the first time in females. We included up to 16 SNPs in our MR analyses, a large increase from five IVs from previous studies [40]. Our effect estimates were higher in females than in males, indicating that bone metabolism depends more on estradiol in females. This agrees with the rapid decline of BMD after menopause and subsequently the prevalence of osteoporosis in postmenopausal women.

We identified few significant loci in females, which led to only four IVs being included in the MR. Estradiol levels in women vary wildly during the menstrual cycle, making the genetic effect hard to estimate. Furthermore, estrogen levels drop after menopause and are mostly determined by the time that has passed since the last menstruation [50]. Both of these aspects probably limited the power of our GWAS. Moreover, the low number of IVs in our MR analyses could have made the results unstable and increased the risk of them being affected by pleiotropy.

In summary, we identified genetic loci that affect estradiol levels with strong sex-dependent effects. We showed the causal effect of estradiol on BMD, supporting HRT as a preventative treatment of osteoporosis. Our findings confirm established medical research as well as provide insight into the metabolism and function of estrogens.

3 Project II

Contributions: generated figures, performed estradiol GWAS, contributed to and reviewed the manuscript, interpreted data

3.1 Background

Despite its important functions for development and health, estrogen has been associated with a number of diseases. Higher estradiol levels have been associated with an increased risk of breast cancer in pre- as well as postmenopausal women [51–57]. However, a definite causal relationship, i.e. whether high estradiol levels increase the risk of breast cancer or cancer progression causes estradiol levels to rise, has not been established.

Two other major forms of cancer—endometrial and ovarian cancer—have clearly established links to estradiol levels [55, 59]. Progesterone has been shown to have a protective effect against these kinds of cancer, which is why menopausal HRT is often combined with progesterone. The same effect can be observed for OCs because they contain a synthetic form of progesterone—progestin [57, 61].

Despite the clearly established association between the aforementioned cancers and estrogen, it remains unclear whether the body's own production of estrogen has an effect on cancer risk. MR is a commonly used approach to establish such a causal link. In fact, there has been one study that used one SNP as IV to estimate the effect of endogenous estradiol on endometrial cancer [59]. No previous MR studies concerning the effect of estradiol on breast or ovarian cancer have been published.

In this project we aimed to estimate the causal effect of estrogen on women's risk of breast, endometrial and ovarian cancer using the genetic instruments we identified in project I.

3.2 Methods

We used female participants from UKB as the base cohort for this study. Genotype and sample QC were discussed in Project I, as we used the effect estimates for estradiol from that study.

We assessed cancer incidence using diagnoses from hospital stays, death and cancer registries as well as answers from verbal interviews and touchscreen questionnaires. Diagnoses were encoded as codes from the International Classification of Diseases, revision 9 (ICD-9) and revision 10 (ICD-10). Breast cancer was represented by codes 174 (ICD-9) and C50 (ICD-10), ovarian cancer by codes 183 (ICD-9) and C56 (ICD-10) and endometrial cancer by ICD-10 code C541. Most cases were recorded in the cancer register.

However, data from before 1995 was lacking. Therefore, we had to rely on self-reported data for these cases.

We estimated the effects of our estradiol IVs on risk for all three cancers using PLINK 2. We applied a logistic model with additive effects with the following covariates: age, BMI, genetic PCs 1–10, HRT use (never/ever/current), OC use (never/ever/current), number of live births, menopausal status, whether the participant had undergone a hysterectomy and a binary indicator for the used genotyping array.

To infer the effect of endogenous estradiol on cancer risk, we performed a one-sample MR analysis using our computed effect estimates and a two-sample MR analysis, where we used publicly available GWAS data for SNP effects on cancer risk. The main analyses were performed using the `R` package gsmr version 1.0.8 [10]. We used the `R` package MendelianRandomization for sensitivity analyses, using the included methods: robust IVW, weighted median and MR-Egger [60]. Effect estimates for breast cancer were taken from summary statistics published by the Breast Cancer Association Consortium (BCAC), whose study included 122,977 breast-cancer cases and 105,974 healthy individuals of European origin [61]. Endometrial-cancer estimates were taken from a GWAS published by the Endometrial Cancer Association Consortium (ECAC) [62]. After removing individuals from UKB, 12,720 cases and 46,126 controls of European ancestry remained. Summary statistics for ovarian-cancer risk were taken from a study by the Ovarian Cancer Association Consortium (OCAC) including 25,509 cases of epithelial ovarian cancer and 40,491 controls [63].

3.3 Results and Discussion

After QC, 209,877 female UKB participants remained, of which 13,179 had previously received a diagnosis of breast cancer, 1,981 of endometrial cancer and 1,477 of ovarian cancer. Among the four chosen IVs, two were nominally associated ($p_{adj} < 0.05$) with breast cancer and one IV was associated with endometrial cancer. In the summary statistics used for the two-sample MR, one IV was associated with ovarian cancer and one with endometrial cancer. The SNP rs10638101 was missing in BCAC's summary statistics and had therefore be substituted by the proxy rs897797, which was in perfect LD ($R^2 = 1$).

In our primary analysis using gsmr, we identified a causal effect of estradiol on breast-cancer and endometrial-cancer risk, both in the one-sample and the two-sample analyses (table 3.1). We did not detect a causal relationship for ovarian cancer. In the sensitivity analysis, the MR-Egger method reported a significant intercept for the two-sample MR of endometrial cancer, which is an indication of directional pleiotropy. However, MR-Egger controls for this kind of pleiotropy. Given that the effect estimate was significant, and its direction was consistent with all over approaches, we did not see evidence of strong pleiotropy.

	One-Sample		Two-Sample	
	OR (95% CI)	**P Value**	**OR (95% CI)**	**P Value**
Breast Cancer	1.30 (1.07–1.57)	0.0074	1.19 (1.03–1.38)	0.018
Endometrial Cancer	2.01 (1.21–3.31)	0.0065	1.45 (1.14–1.83)	0.0022
Ovarian Cancer	1.55 (0.91–2.65)	0.11	1.20 (0.99–1.46)	0.066

Table 3.1: MR results from our main analysis with GSMR.

3.3.1 Cancer-Specific Results

Our first IV was annotated to *CYP3A7*, which was discussed in section 2.3. It had a nominally significant association to breast and endometrial cancer in UKB, BCAC and ECAC. IV number 2 was *MCM8*, which has a strong association to age at menopause [64]. *MCM8* was nominally associated with breast cancer in UKB and BCAC and with endometrial cancer in ECAC. Although we had not adjusted the estradiol GWAS for age at menopause, an analysis including only postmenopausal women and with age at menopause as covariate did not result in a significantly different effect estimate. This suggests that the effect is not confounded by age at menopause. Our third IV, *ASCL1*, has a previously established association to tumor progression in lung adenocarcinoma and survival in patients with ovarian cancer [65, 69]. It was nominally associated with risk for ovarian cancer in OCAC. Our last IV, *TMEM1150B*, has an established association with age at menopause and menarche [67, 71]. It was significantly associated with breast cancer in UKB and BCAC.

Our analyses showed no significant causal effect on estradiol on ovarian cancer. However, the effect estimates from both the one- and two-sample tests showed a consistent effect direction and the result from the two-sample MR would have passed a one-sided test (table 3.1). Therefore, we cannot conclude that there is no causal effect of estradiol on ovarian cancer. However, we can safely say that the observed effect is weaker than the effects on breast and endometrial cancer [69–74].

3.3.2 Limitations

Our MR study was still limited by the overall small set of four IVs. This made our study vulnerable to directional pleiotropy. In particular, the MR-Egger intercept of our two-sample analysis for endometrial cancer showed indications of pleiotropy. Once larger cohorts or more quantitative estradiol measurements are available, it should be possible to identify more IVs and alleviate this problem. Furthermore, estradiol levels fluctuate wildly in women over the course of the menstrual cycle. This and the strong correlation between estradiol levels in postmenopausal women and time since menopause make the

detection of genetic IVs complicated.

In summary, we identified a causal effect of estradiol on the risk of breast and endometrial cancer using a MR approach in UKB, BCAC and ECAC cohorts. Our findings support prior research regarding the carcinogenic effects of estrogen. Further research on the mechanism by which estrogens influence cancer development is needed, however.

4 Project III

Contributions: analysed the long-read data, generated figures, wrote the manuscript

4.1 Background

Despite the success of NGS, it is difficult to accurately detect CNVs using this approach because the reads generated by NGS are shorter than most CNVs. However, the advent of third-generation sequencing technologies such as PacBio SMRT enables the generation of reads long enough to directly sequence long SVs and improve mapping accuracy.

In this project, the aim was to identify CNVs using WGS and investigate if CNV polymorphisms are associated with levels of plasma protein levels. Moreover, we aimed to explore the application of long-read sequencing for the characterization of CNVs.

4.2 Methods

We called CNVs using CNVnator, a read-depth based method for CNV detection [17]. This tool calls CNVs by dividing the genome into non-overlapping bins and identifying those bins that received abnormally high or low coverage. The bin sizes were determined for each individual separately. To facilitate association testing, we aligned the detected CNVs to non-overlapping genomic windows of 200 bp. If a sample had a CNV overlapping with a given window, we set that window's copy number appropriately. Lastly, adjacent windows that had consistent copy numbers across individuals were merged to make up the final CNVs.

We identified copy number-biomarker associations using a linear regression model (`glm` function in `R` version 4.3.4). We included sex and age as covariates and applied an adjusted significance threshold of $p < 4.79 * 10^{-9}$, which corresponds to the general significance threshold of $p < 0.05$ adjusted for 10,438,413 tests (438 proteins $\times$ 23,831 CNVs).

We selected 15 individuals to be resequenced using PacBio SMRT sequencing based on five CNV that showed strong associations, were highly polymorphic and included both deletions and duplications (table 4.1). The individuals were chosen to capture as many CNs as possible. Long-read sequencing was performed on a PacBio SEQUEL II system in continuous long read (CLR) mode according to manufacturer specification. PacBio's tool chain automatically mapped all reads to GRCh38. Therefore, we had to extract the reads and remap them to GRCh37 using pbmm2 version 1.4.0 to enable comparison between long-read and short-read data. We called SVs using three different tools: SVIM v1.4.2, Sniffles v1.0.12 and PBSV v2.4.0.

Biomarker	Chr	Start	End	Beta	SE	P Value
GPNMB	2	89,613,000	89,613,200	-9.91E-01	1.56E-01	3.33E-10
PD-L2	3	98,411,600	98,411,800	4.09E-01	4.47E-02	3.92E-19
IL-18	5	70,393,100	70,393,300	4.18E-01	5.61E-02	2.38E-13
ST1A1	16	28,613,645	28,613,845	7.86E-01	6.93E-02	1.61E-27
hOSCAR	19	54,558,900	54,559,100	-7.21E-01	6.27E-02	1.61E-28
CD48	1	158,867,600	158,867,800	-2.82E-01	4.71E-02	3.15E-09
FCRLB	1	161,640,580	161,640,780	6.17E-01	9.11E-02	2.42E-11
LY9	1	179,455,600	179,455,800	5.36E-01	8.86E-02	2.17E-09
ICAM-2	3	98,410,600	98,410,800	6.48E-01	3.95E-02	5.67E-53
Siglec-9	3	98,411,800	98,413,400	6.82E-01	4.17E-02	3.48E-52
CD200R1	3	98,411,800	98,413,400	5.08E-01	4.45E-02	3.71E-28
VEGFR-3	3	98,414,600	98,414,800	5.96E-01	4.17E-02	1.10E-41
ICAM-3	3	98,899,900	98,900,100	3.96E-01	5.52E-02	1.45E-12
AMBP	5	745,070	745,270	-2.84E-01	4.51E-02	5.05E-10
MIC-AB	6	32,496,600	32,496,800	6.19E-01	6.69E-02	3.35E-19
CCL19	6	32,522,200	32,522,400	-4.67E-01	5.37E-02	1.93E-17
FR-gamma	11	63,443,100	63,445,300	-1.03E+00	9.83E-02	5.17E-24
FR-gamma	11	67,331,355	67,331,955	-1.23E+00	1.03E-01	1.64E-30
CNTN1	12	45,909,600	45,909,800	-2.85E-01	4.69E-02	1.73E-09
CCL4	17	36,392,670	36,394,670	1.45E-01	2.29E-02	4.65E-10
CCL15	17	39,210,800	39,211,000	-9.21E-01	1.23E-01	1.45E-13
SMPD1	19	35,863,600	35,863,800	3.69E-01	6.08E-02	1.97E-09
MIA	19	41,381,925	41,385,125	-1.23E+00	1.69E-01	8.58E-13
hK11	19	51,508,940	51,510,740	1.41E+00	2.22E-01	3.84E-10
WFDC2	20	44,204,435	44,205,035	4.17E-01	6.81E-02	1.37E-09

Table 4.1: CN-GWAS results. The CNVs in the top section were used to select the 15 individuals for resequencing.

4.3 Results and Discussion

Overall, 872 individuals had both genotyping and protein data which passed QC. CNVnator reported 243,987 CNVs, which after post-processing resulted in 23,831 variants to be included in our analysis. We found 30 CNVs to be associated with 17 biomarkers (table 4.1).

The quality of our long-read sequencing results was mixed. While most samples received high coverage, in three cases less than half of all ZMWs produced high-quality reads. The CNV on chromosome 2 was not called by any of our long-read callers. The CNV on chromosome 3 was only detected by SVIM, which also called all copy numbers in accordance with CNVnator. The CNV on chromosome 5 was not detected in the long-read data. The coverage in this region was spotty at best, in both Illumina and

SMRT sequencing data. This might be caused by the large number of repetitive elements and consequently low mappability in this region. The CNV on chromosome 16, which was exclusively called as a duplication by CNVnator, was not detected by any long-read caller. However, SVIM called many smaller insertions in this region, which mapped well to repetitive elements. This suggests that CNVnator might indeed have picked up on these repetitive elements and merged them into one big CNV down the line. The CNV on chromosome 19 was consistently detected by all callers.

In general, deletions seem easier to detect and verify than duplications.

4.4 Future Work

This project has not concluded fully. To make use of the results we have got so far, it would be interesting to investigate the detected CNVs for effects of known functional elements in the genome. This would help elucidate the pathways through which they affect their associated proteins.

Moreover, the data we alreay have might allow us to identify SNPs that tag or are in LD with CNVs of interest. This might enable the creation of a larger imputation framework for CNVs by use of NGS-derived SNP data.

5 Project IV

5.1 Future Work

The research question that will be investigated in project IV has not been set, yet. There are a few different projects that might be viable. In general, further research into CNVs and/or long-read sequencing, possibly as an extension of project III, sound interesting.

It might be worthwhile to develop a method that allows for easier CNV validation. So far, this is a manual process, which is very time-consuming. Devising such a method would also benefit future research in this field. However, it is difficult to exactly formalize when a CNV can be considered validated. For instance, the CNVs that were called in short-read data but presented as groups of short duplications in long-read data are problematic cases. While the CNVs that were originally called did not validate *per se*, there were in fact smaller variants that just happened to present as one large CNV. One could argue that this can still be considered a positive result that would not be trivial to detect.

We might be able to further elucidate the nature of the detected CNVs by assembling the genomes of the individuals we sequenced using SMRT technology or a combination of both sequencing techniques. No long-read SV caller, except for PBSV, actually reports CNVs as such and consequently, CNVs and CNs have to be inferred from the variants they do report, such as duplications, insertions and breakends. An assembly might allow us to characterize the copy numbers and break points better.

6 Concluding Remarks

We investigated the role of genetic variants in human diseases through methods that supplement, complement and extend conventional GWAS. We identified the effect of genetic variants on estradiol levels and demonstrated the causal effect of estrogen on BMD and certain cancers. Furthermore, we identified CNVs, which are not covered by common GWAS, and characterized their association with plasma protein biomarkers.

While GWAS have enabled groundbreaking research in the field of genetic epidemiology, it is becoming more apparent that there is a large part of heritability of certain traits and disease risk that they cannot explain. To increase their statistical power, increasingly larger cohorts are needed.

Variants such as CNVs are generally hard to detect and quantify. Short-read sequencing, which has driven down cost and allowed for the detection of rare SNPs not covered by SNP arrays, has issues correctly capturing large SVs. This is in part due to their size but also due to many SVs being hard to map because they lie in regions with many repeats or low complexity.

Long-read sequencing promises to ameliorate many problems of short reads but is still too expensive to be used in the same extent. While these sequencing technologies have made considerable advances over the last few years, there are no universally accepted standards like for short-read sequencing and GWAS [72].

The improving accuracy and cost of WGS, along with novel approaches such as haplotype-resolved and telomere-to-telomere assemblies, will make it possible to gain even deeper insights into the genetic architecture of human health [73, 77].

Acknowledgements

I would like to thank my main supervisor *Åsa Johansson*, who made these projects possible and whose advice and and support I can always rely on. I would also like to thank my co-supervisors: *Torgny Karlsson*, who always overdelivers with his solutions for statsitical problems, and *Adam Ameur*, whose sequencing expertise made a lot of this project possible. Many thanks to the remaining members of Åsa Johansson's research group, who have made the last two years a blast. They always have an open ear and are ready help and give advice. I could not have wished for better colleagues.

I would like to acknowledge *Zhiwei Li* and *Nima Rafati*, whose torch I had the honor of carrying forward when taking over the CNV project.

Thank you so much to the people at the *IGP PhD Council*, *Medicinska Doktorandrådet* and *Doktorandnämnden* for allowing me to represent PhD students' interests.

Lastly, I would like to thank all my family and friends both in and outside of Uppsala, who might not have contributed directly to my research but made the last two years as amazing as they were.

Acronyms

aCGH array comparative genomic hybridization. 5

BCAC the Breast Cancer Association Consortium. 12–14

BMD bone mineral density. iii, 7, 8, 10, 19

BMI body-mass index. 8, 12

CLR continuous long read. 15

CN copy number. iv, 4–6, 15, 16, 18

CNV copy-number variation. iii, iv, 4–6, 15–19

CYP3A7 cytochrome P450 3A7. 9

DHEA dehydroepiandrosterone. 9

ECAC the Endometrial Cancer Association Consortium. 12–14

eQTL expression quantitative trait locus. 8

GEFOS the Genetic Factors for Osteoporosis Consortium. 8

GSMR generalized summary-based Mendelian Randomization. 4, 8, 13

GTEx the Genotype-Tissue Expression project. 8

GWAS genome-wide association study. iii, 1–4, 7–13, 16, 19

HRT hormone-replacement therapy. 8, 10–12

ICD International Classification of Diseases. 11, 21

ICD-10 ICD, revision 10. 11

ICD-9 ICD, revision 9. 11

indel small insertion or deletion. 1, 2, 4, 5

IV instrumental variable. 2–4, 8, 10–14

IVW inverse-variance weighted. 4, 8, 12

LD linkage disequilibrium. 1, 8, 9, 12, 17

MR Mendelian Randomization. iii, 1–4, 7, 8, 10–14

NGS next-generation sequencing. iii, 1, 5, 15, 17

NSPHS Northern Swedish Population Health Study. 6

OC oral contraceptive. 8, 11, 12

OCAC the Ovarian Cancer Association Consortium. 12, 13

OR odds ratio. 9

PacBio Pacific Biosciences. 6, 15

PC principal component. 8, 12

PCR polymerase chain reaction. 5, 6

PEA protein extension assay. 6

QC quality control. 6–8, 11, 12, 16

SBS sequencing by synthesis. 5

SHBG sex hormone-binding globulin. 8, 9

SMRT Single-Molecule Real-Time Sequencing. iv, 6, 15, 17, 18

SNP single-nucleotide polymorphism. iii, 1, 2, 4–12, 17, 19

SV structural variation. 1, 4–6, 15, 18, 19

T2D type 2 diabetes. 7

UKB UK Biobank. iii, 2, 6–8, 11–14

WES whole-exome sequencing. 1

WGS whole-genome sequencing. iii, iv, 1, 5, 6, 15, 19

Bibliography

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature 2001 409:6822* **409,** 860–921. ISSN: 1476-4687. https://www.nature.com/articles/35057062 (Feb. 2001).
2. Belmont, J. W. *et al.* A haplotype map of the human genome. *Nature 2005 437:7063* **437,** 1299–1320. ISSN: 1476-4687. https://www.nature.com/articles/nature04226 (Oct. 2005).
3. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research* **47,** D1005–D1012. ISSN: 0305-1048. https://academic.oup.com/nar/article/47/D1/D1005/5184712 (Jan. 2019).
4. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44,** 512–525. ISSN: 14643685. https://pubmed.ncbi.nlm.nih.gov/26050253/ (May 2015).
5. Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations. *Wellcome Open Research* **4.** /pmc/articles/PMC7384151/%20/pmc/articles/PMC7384151/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7384151/ (2019).
6. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology* **37,** 658. /pmc/articles/PMC4377079/%20/pmc/articles/PMC4377079/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4377079/ (Nov. 2013).
7. Bowden, J., Smith, G. D., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* **40,** 304. /pmc/articles/PMC4849733/%20/pmc/articles/PMC4849733/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4849733/ (May 2016).
8. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44,** 512. /pmc/articles/PMC4469799/%20/pmc/articles/PMC4469799/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4469799/ (May 2015).

9. Bowden, J. Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. *International Journal of Epidemiology* **46,** 2097–2099. ISSN: 14643685. https://academic.oup.com/ije/article/46/6/2097/4157383 (Dec. 2017).

10. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nature Communications* **9,** 224. ISSN: 2041-1723. http://www.nature.com/articles/s41467-017-02317-2 (Dec. 2018).

11. Auton, A. *et al.* A global reference for human genetic variation. eng. *Nature* **526,** 68–74. ISSN: 14764687. https://pubmed.ncbi.nlm.nih.gov/26432245%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4750478/ (Oct. 2015).

12. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81. ISSN: 14764687. https://www.nature.com/articles/nature15394 (Sept. 2015).

13. MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72,** 971–983. ISSN: 0092-8674. http://www.cell.com/article/009286749390585E/fulltext%20http://www.cell.com/article/009286749390585E/abstract%20https://www.cell.com/cell/abstract/0092-8674(93)90585-E (Mar. 1993).

14. Verkerk, A. J. *et al.* Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65,** 905–914. ISSN: 0092-8674. http://www.cell.com/article/009286749190397H/fulltext%20http://www.cell.com/article/009286749190397H/abstract%20https://www.cell.com/cell/abstract/0092-8674(91)90397-H (May 1991).

15. Quenez, O. *et al.* Detection of copy-number variations from NGS data using read depth information: a diagnostic performance evaluation. *European Journal of Human Genetics 2020 29:1* **29,** 99–109. ISSN: 1476-5438. https://www.nature.com/articles/s41431-020-0672-2 (June 2020).

16. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology 2011 12:2* **12,** 1–14. ISSN: 1474-760X. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-2-r18 (Feb. 2011).

17. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. eng. *Genome Research* **21,** 974–984. ISSN: 10889051. https://pubmed.ncbi.nlm.nih.gov/21324876%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3106330/%20http://www.genome.org/cgi/doi/10.1101/gr.114876.110. (June 2011).

18. Chen, X. *et al.* Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32,** 1220–1222. ISSN: 14602059. http://dx.doi.org/10.1093/bioinformatics/btv710%20https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv710 (Apr. 2016).

19. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* **37,** 1155. /pmc/articles/PMC6776680/%20/pmc/articles/PMC6776680/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6776680/ (Oct. 2019).

20. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35,** 2193. /pmc/articles/PMC6596899/%20/pmc/articles/PMC6596899/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6596899/ (July 2019).

21. Höijer, I. *et al.* Detailed analysis of HTT repeat elements in human blood using targeted amplification-free long-read sequencing. *Human Mutation* **39,** 1262–1272. ISSN: 1098-1004. https://onlinelibrary.wiley.com/doi/full/10.1002/humu.23580%20https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.23580%20https://onlinelibrary.wiley.com/doi/10.1002/humu.23580 (Sept. 2018).

22. Rang, F. J., Kloosterman, W. P. & de Ridder, J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology* **19.** /pmc/articles/PMC6045860/%20/pmc/articles/PMC6045860/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6045860/ (July 2018).

23. Heller, D. & Vingron, M. SVIM: Structural variant identification using mapped long reads. *Bioinformatics* **35** (ed Birol, I.) 2907–2915. ISSN: 14602059. https://academic.oup.com/bioinformatics/article/35/17/2907/5298305 (Sept. 2019).

24. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* **15,** 461–468. ISSN: 15487105. https://doi.org/10.1038/s41592-018-0001-7 (June 2018).

25. Ameur, A. *et al.* SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. *European Journal of Human Genetics 2017 25:11* **25,** 1253–1260. ISSN: 1476-5438. https://www.nature.com/articles/ejhg2017130 (Aug. 2017).

26. Bates, G. W. & Bowling, M. Physiology of the female reproductive axis. *Periodontology 2000* **61,** 89–102. ISSN: 09066713. http://doi.wiley.com/10.1111/j.1600-0757.2011.00409.x (Feb. 2013).

27. Hess, R. A. *et al.* A role for oestrogens in the male reproductive system. *Nature* **390,** 509–512. ISSN: 0028-0836. http://www.nature.com/articles/37352 (Dec. 1997).

28. Thomas, M. P. & Potter, B. V. *The structural biology of oestrogen metabolism* 2013. /pmc/articles/PMC3866684/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3866684/.

29. Cauley, J. A. *et al.* Elevated serum estradiol and testosterone concentrations are associated with a high risk for breast cancer. *Annals of Internal Medicine* **130,** 270–277. ISSN: 00034819 (Feb. 1999).

30. Rosendaal, F. R., Van Hylckama Vlieg, A., Tanis, B. C. & Helmerhorst, F. M. Estrogens, progestogens and thrombosis. *Journal of Thrombosis and Haemostasis* **1,** 1371–1380. ISSN: 1538-7933. http://doi.wiley.com/10.1046/j.1538-7836.2003.00264.x (July 2003).

31. Vikan, T., Schirmer, H., Njølstad, I. & Svartberg, J. Low testosterone and sex hormone-binding globulin levels and high estradiol levels are independent predictors of type 2 diabetes in men. *European Journal of Endocrinology* **162,** 747–754. ISSN: 08044643 (Apr. 2010).

32. Riggs, B. L., Khosla, S. & Melton, L. J. A Unitary Model for Involutional Osteoporosis: Estrogen Deficiency Causes Both Type I and Type II Osteoporosis in Postmenopausal Women and Contributes to Bone Loss in Aging Men. *Journal of Bone and Mineral Research* **13,** 763–773. ISSN: 08840431. http://doi.wiley.com/10.1359/jbmr.1998.13.5.763 (May 1998).

33. Longo, D. L. *et al. Harrison's principles of internal medicine* (Mcgraw-hill New York, 2012).

34. Pott, J. *et al.* Genetic Association Study of Eight Steroid Hormones and Implications for Sexual Dimorphism of Coronary Artery Disease. *The Journal of clinical endocrinology and metabolism* **104,** 5008–5023. ISSN: 1945-7197. http://www.ncbi.nlm.nih.gov/pubmed/31169883 (Nov. 2019).

35. Chen, Z. *et al.* Genome-wide association study of sex hormones, gonadotropins and sex hormone–binding protein in Chinese men. *Journal of Medical Genetics* **50,** 794–801. ISSN: 0022-2593. http://jmg.bmj.com/lookup/doi/10.1136/jmedgenet-2013-101705 (Dec. 2013).

36. Liu, M. *et al.* TSPYL5 SNPs: association with plasma estradiol concentrations and aromatase expression. *Molecular endocrinology (Baltimore, Md.)* **27,** 657–70. ISSN: 1944-9917. http://www.ncbi.nlm.nih.gov/pubmed/23518928%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3607698%20/pmc/articles/PMC3607698/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3607698/ (Apr. 2013).

37. Prescott, J. *et al.* Genome-wide association study of circulating estradiol, testosterone, and sex hormone-binding globulin in postmenopausal women. *PLoS ONE* **7,** 37815. ISSN: 19326203. https://pubmed.ncbi.nlm.nih.gov/22675492/%20www.plosone.org (June 2012).

38. Eriksson, A. L. *et al.* Genetic determinants of circulating estrogen levels and evidence of a causal effect of estradiol on bone density in men. *Journal of Clinical Endocrinology and Metabolism* **103,** 991–1004. ISSN: 19457197. https://academic.oup.com/jcem/article/103/3/991/4794882%20https://pubmed.ncbi.nlm.nih.gov/29325096/ (Mar. 2018).

39. Ruth, K. S. *et al.* Using human genetics to understand the disease impacts of testosterone in men and women. *Nature Medicine* **26,** 252–258. ISSN: 1078-8956. http://www.nature.com/articles/s41591-020-0751-5 (Feb. 2020).

40. Nethander, M. *et al.* Evidence of a Causal Effect of Estradiol on Fracture Risk in Men. *Journal of Clinical Endocrinology and Metabolism* **104,** 433–442. ISSN: 19457197. https://academic.oup.com/jcem/article/104/2/433/5094017 (Feb. 2018).

41. Nakamoto, J., Salameh, W. A. & Carlton, E. in *Endocrinology* (eds Jameson, J. L. & De Groot, L. J. B. T. .-. E. (E.) 2802–2834 (Elsevier, Philadelphia, 2010). ISBN: 978-1-4160-5583-9. http://www.sciencedirect.com/science/article/pii/B9781416055839001556%20https://linkinghub.elsevier.com/retrieve/pii/B9781416055839001556.

42. Ward, L. D. & Kellis, M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research* **40,** D930–D934. ISSN: 03051048. http://compbio.mit.edu/HaploReg. (Jan. 2012).

43. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking* **13,** 307–308. ISSN: 1947-5535 (2015).

44. Hemani, G. *et al.* The MR-base platform supports systematic causal inference across the human phenome. *eLife* **7.** ISSN: 2050084X (May 2018).

45. Estrada, K. *et al.* Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature Genetics* **44,** 491–501. ISSN: 10614036 (Apr. 2012).

46. Travison, T. *et al.* The Heritability of Circulating Testosterone, Estradiol, Estrone, and SHBG Concentrations in Men: The Framingham Heart Study. *Clinical endocrinology* **80,** 277–282. /pmc/articles/PMC3825765/%20/pmc/articles/PMC3825765/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3825765/ (Feb. 2014).

47. Ohmori, S. *et al.* Differential catalytic properties in metabolism of endogenous and exogenous substrates among CYP3A enzymes expressed in COS-7 cells. *Biochimica et Biophysica Acta - General Subjects* **1380,** 297–304. ISSN: 03044165 (May 1998).

48. Simpson, E. R. & Davis, S. R. Minireview: Aromatase and the Regulation of Estrogen Biosynthesis—Some New Perspectives. *Endocrinology* **142,** 4589–4594. ISSN: 0013-7227. https://academic.oup.com/endo/article/142/11/4589/2988522 (Nov. 2001).

49. Ogasawara, K. *et al.* Extensive polymorphism of ABO blood group gene: Three major lineages of the alleles for the common ABO phenotypes. *Human Genetics* **97,** 777–783. ISSN: 03406717. `https://link.springer.com/article/10.1007/BF02346189` (1996).

50. Richardson, H. *et al. Baseline estrogen levels in postmenopausal women participating in the MAP.3 breast cancer chemoprevention trial* in *Menopause* **27** (Lippincott Williams and Wilkins, 2020), 693–700. `/pmc/articles/PMC7469568/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7469568/`.

51. Key, T. Sex hormones and risk of breast cancer in premenopausal women: A collaborative reanalysis of individual participant data from seven prospective studies. *The Lancet Oncology* **14,** 1009–1019. ISSN: 14702045 (Sept. 2013).

52. Kaaks, R. *et al.* Postmenopausal serum androgens, oestrogens and breast cancer risk: the European prospective investigation into cancer and nutrition. *Endocrine-Related Cancer* **12,** 1071–1082. ISSN: 1351-0088. `https://erc.bioscientifica.com/view/journals/erc/12/4/0121071.xml` (Dec. 2005).

53. Zhang, X., Tworoger, S. S., Eliassen, A. H. & Hankinson, S. E. Postmenopausal plasma sex hormone levels and breast cancer risk over 20 years of follow-up. *Breast Cancer Research and Treatment 2012 137:3* **137,** 883–892. ISSN: 1573-7217. `https://link.springer.com/article/10.1007/s10549-012-2391-z` (Jan. 2013).

54. Kaaks, R. *et al.* Serum Sex Steroids in Premenopausal Women and Breast Cancer Risk Within the European Prospective Investigation into Cancer and Nutrition (EPIC). *JNCI: Journal of the National Cancer Institute* **97,** 755–765. ISSN: 0027-8874. `https://academic.oup.com/jnci/article/97/10/755/2544018` (May 2005).

55. Brinton, L. A. & Felix, A. S. Menopausal hormone therapy and risk of endometrial cancer. *The Journal of Steroid Biochemistry and Molecular Biology* **142,** 83–89. ISSN: 0960-0760 (July 2014).

56. Mungenast, F. & Thalhammer, T. Estrogen Biosynthesis and Action in Ovarian Cancer. *Frontiers in Endocrinology* **0,** 192. ISSN: 1664-2392 (2014).

57. Karlsson, T., Johansson, T., Hoglund, J., Ek, W. E. & Johansson, Å. Time-dependent effects of oral contraceptive use on breast, ovarian, and endometrial cancers. *Cancer Research* **81,** 1153–1162. ISSN: 15387445 (2021).

58. Iversen, L., Sivasubramaniam, S., Lee, A. J., Fielding, S. & Hannaford, P. C. Lifetime cancer risk and combined oral contraceptives: the Royal College of General Practitioners' Oral Contraception Study. *American Journal of Obstetrics and Gynecology* **216,** 580.e1–580.e9. ISSN: 10976868 (2017).

59. Thompson, D. J. *et al.* CYP19A1 fine-mapping and Mendelian randomization: Estradiol is causal for endometrial cancer. *Endocrine-Related Cancer* **23,** 77–91. ISSN: 14796821. `https://erc.bioscientifica.com/view/journals/erc/23/2/77.xml` (Feb. 2016).

60. Olena, Y. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *International Journal of Epidemiology* **46,** 1734–1739 (2017).

61. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551,** 92–94. ISSN: 14764687 (2017).

62. O'Mara, T. A. *et al.* Identification of nine new susceptibility loci for endometrial cancer. *Nature Communications* **9.** ISSN: 20411723 (2018).

63. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nature Genetics* **49,** 680–691. ISSN: 15461718 (2017).

64. Chen, C. T. L. *et al.* Meta-analysis of loci associated with age at natural menopause in African-American women. *Human Molecular Genetics* **23,** 3327–3342. ISSN: 14602083 (2014).

65. Miyashita, N. *et al.* ASCL1 promotes tumor progression through cell-autonomous signaling and immune modulation in a subset of lung adenocarcinoma. *Cancer Letters* **489,** 121–132. ISSN: 18727980 (2020).

66. Moore, K. N. *et al.* Genome-wide association study evaluating single-nucleotide polymorphisms and outcomes in patients with advanced stage serous ovarian or primary peritoneal cancer: An NRG Oncology/Gynecologic Oncology Group study. *Gynecologic Oncology* **147,** 396–401. ISSN: 10956859 (2017).

67. Stolk, L. *et al.* Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nature Genetics* **44,** 260–268. ISSN: 10614036 (2012).

68. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics* **48,** 709–717. ISSN: 15461718 (2016).

69. Trabert, B. *et al.* Circulating estrogens and postmenopausal ovarian cancer risk in the women's health initiative observational study. *Cancer Epidemiology Biomarkers and Prevention* **4,** 648–656. ISSN: 10559965 (2016).

70. Key, T. J., Appleby, P. N., Hines, L. M. & Al., E. Circulating sex hormones and breast cancer risk factors in postmenopausal women: reanalysis of 13 studies. *Br J Cancer* **105,** 709–722 (2011).

71. Rodriguez, A. C., Blanchard, Z., Maurer, K. A. & Gertz, J. Estrogen Signaling in Endometrial Cancer: a Key Oncogenic Pathway with Several Open Questions. *Hormonal cancer* **10,** 51–63 (2019).

72. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nature Reviews Genetics 2021 22:9* **22,** 572–587. ISSN: 1471-0064. `https://www.nature.com/articles/s41576-021-00367-3` (May 2021).

73. Logsdon, G. A. *et al.* The structure, function and evolution of a complete human chromosome 8. *Nature 2021 593:7857* **593,** 101–107. ISSN: 1476-4687. `https://www.nature.com/articles/s41586-021-03420-7` (Apr. 2021).

74. Porubsky, D. *et al.* Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology* **39,** 302. /pmc/articles/PMC7954704/%20/pmc/articles/PMC7954704/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7954704/ (2021).

The Journal of Clinical Endocrinology & Metabolism, 2021, Vol. XX, No. XX, 1–16
https://doi.org/10.1210/clinem/dgab507
Clinical Research Article

Clinical Research Article

Genome-wide Association Study of Estradiol Levels and the Causal Effect of Estradiol on Bone Mineral Density

Daniel Schmitz,[1] Weronica E. Ek,[1] Elin Berggren,[1] Julia Höglund,[1] Torgny Karlsson,[1] and Åsa Johansson[1]

[1]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden

ORCiD numbers: 0000-0003-4480-891X (D. Schmitz); 0000-0003-2194-496X (W. E. Ek); 0000-0001-8061-3947 (J. Höglund); 0000-0001-8095-6149 (T. Karlsson); 0000-0002-2915-4498 (Å. Johansson).

Received: 5 March 2021; Editorial Decision: 6 July 2021; First Published Online: 13 July 2021; Corrected and Typeset: 10 August 2021.

Abstract

Context: Estradiol is the primary female sex hormone and plays an important role for skeletal health in both sexes. Several enzymes are involved in estradiol metabolism, but few genome-wide association studies (GWAS) have been performed to characterize the genetic contribution to variation in estrogen levels.

Objective: Identify genetic loci affecting estradiol levels and estimate causal effect of estradiol on bone mineral density (BMD).

Design: We performed GWAS for estradiol in males (n = 147 690) and females (n = 163 985) from UK Biobank. Estradiol was analyzed as a binary phenotype above/below detection limit (175 pmol/L). We further estimated the causal effect of estradiol on BMD using Mendelian randomization.

Results: We identified 14 independent loci associated ($P < 5 \times 10^{-8}$) with estradiol levels in males, of which 1 (*CYP3A7*) was genome-wide and 7 nominally ($P < 0.05$) significant in females. In addition, 1 female-specific locus was identified. Most loci contain functionally relevant genes that have not been discussed in relation to estradiol levels in previous GWAS (eg, *SRD5A2*, which encodes a steroid 5-alpha reductase that is involved in processing androgens, and *UGT3A1* and *UGT2B7*, which encode enzymes likely to be involved in estradiol elimination). The allele that tags the O blood group at the *ABO* locus was associated with higher estradiol levels. We identified a causal effect of high estradiol levels on increased BMD in both males ($P = 1.58 \times 10^{-11}$) and females ($P = 7.48 \times 10^{-6}$).

Conclusion: Our findings further support the importance of the body's own estrogen to maintain skeletal health in males and in females.

ISSN Print 0021-972X ISSN Online 1945-7197
Printed in USA
© The Author(s) 2021. Published by Oxford University Press on behalf of the Endocrine Society.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

Key Words: GWAS, estrogen, bone mineral density, mendelian randomization

Estrogen, the primary female sex hormone, is responsible for the development of the female reproductive system and secondary sex characteristics. Estrogen acts primarily as a growth hormone for the reproductive organs and regulates the menstrual cycle in women (1). Additionally, it plays a critical role in male sexual function (2). There are 3 major forms of estrogen: estrone, estradiol, and estriol. Estradiol is the major and most potent form (3). In females, high estrogen levels have been associated with deep vein thrombosis (4) and increased risk for certain cancers, including breast cancer (5) in women. In males, higher levels have been suggested to be associated with reduced risk of type 2 diabetes (T2D) (6). Estrogen is also an important regulator of bone metabolism in both females and males (7). When estrogen levels drop after menopause, the cells that synthesize bone (osteoblasts) are unable to effectively produce bone mass (8). The loss of ovarian estrogens after menopause has therefore, not surprisingly, been associated with reduced bone mineral density (BMD) and higher risk of osteoporosis (9,10).

Previously genome-wide association studies (GWAS) for estrogen have been performed in cohorts of up to 11 000 individuals, mostly of European descent, both male and female and stratified by sex (11-15). In addition, a GWAS was recently performed for sex hormone levels in UK Biobank (UKB), a study that mainly focused on testosterone (16). They identified strong sex-specific genetic effects on serum testosterone levels but did not consider the results from estradiol in women, due to the strong link between estradiol levels and menopausal status/age of menopause.

In addition to GWAS, previous studies have also addressed the causal effect of hormones in relation to disease risk using Mendelian Randomization (MR) (15-17). MR is an approach inspired by randomized controlled studies that uses genetic variants as instrumental variables (IVs) to estimate a causal relationship between an exposure, estradiol in our case, and an outcome (eg, BMD). The underlying reasoning is that if an exposure is directly causing an increased risk of a disease, then the genetic factors increasing the exposure should also indirectly affect disease risk through their effects on the exposure. However, this method requires that some key assumptions are met: (1) the selected IVs are associated with the exposure, (2) the IVs affect the outcome through the exposure, only, and (3) the IVs are not associated with any known or unknown confounders. Because genotypes are assigned at conception, MR can be used to overcome some of the limitations of observational studies, such as reverse causation, where the (alleged) outcome affects the exposure.

For testosterone, many independent genetic variants have been identified and included in previous MR analyses. They concluded that testosterone increases the risk for polycystic ovary syndrome and T2D in women but decreases the risk for T2D in men (16). Even though the number of GWAS significant findings for estradiol has been quite limited in previous studies (11-15), 2 recent MR studies have been performed, using 5 and 2 single-nucleotide polymorphisms (SNPs), respectively (15,17). Both studies identified a beneficial effect of estradiol on BMD in males (15,17). However, no such study has yet been performed in females due to those previous studies have failed to identify genetic variants to be valid instruments for MR analyses.

By using the largest cohort with estradiol levels measured, including almost half a million participants, we have been able to further highlight novel genes, important for regulating estradiol levels. We have stratified the cohort by sex and menopausal status to enable a comparison of genetic effects across strata. Thanks to the larger sample size in UKB compared to previous studies, we were able to identify enough genetic variants associated with estradiol levels to permit causal effect estimation of estradiol on BMD through MR analysis in females, which has not been done previously.

Material and Methods

Sample

UKB is a cross sectional cohort of about 500 000 participants, born between 1939 and 1970 who were recruited between 2006 and 2010. Detailed information was collected on lifestyle and anthropometric traits and on female specific factors that are likely to effect hormone levels, including use of birth control pills, hormone replacement treatments, pregnancies, and menopause. Participants also provided blood samples that were used to genotype approximately 800 000 single nucleotide polymorphisms (SNPs) across the genome and to measure levels of estradiol. The application to use data from UKB has been approved (project number 41143). UKB has an ethics permit from the National Research Ethics Committee (REC reference 11/NW/0382). Phenotype data for this project were extracted from the UKB database in April 2019. All analyses performed in this study are based on the samples and information collected at the participants initial visit to the

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

assessment center. Ethical approval for the analyses performed in this study has also been approved by the Swedish Ethical Review Authority (Dnr: 2020-04415)

Genotype Data

A total of 438 417 participants had been genotyped using the UKB Axiom array, and another 49 994 participants had been genotyped using the similar (95% common marker content) UK BiLEVE array. SNPs had been imputed using UK10K and 1000 genomes phase 3, as reference panels (18,19). For this project we used imputed data from the third release, which contained a total of 93 093 070 SNPs. We only analyzed SNPs with minor allele frequency > 0.01 and removed SNPs deviating from Hardy-Weinberg (P-value < 1×10^{-20}) and with markers with more than 5% missing genotype data and/or imputation quality <0.3. After quality control (QC), a total of 7 651 231 autosomal SNPs and 220 486 SNPs on the X chromosome remained. We included Caucasian participants clustering with regards to their genetic principal components. We also excluded first- and second-degree relatives (genetic relationship >0.044) using kinship data. Samples with sex discordance, high heterozygosity/missingness, and/or more than 5% missing genotypes were also excluded. After QC and exclusion, 361 975 unrelated Caucasian participants remained (167 168 males and 194 807 females).

Estradiol Measurements

We only included measurements from the blood samples that were collected during the participants initial visit to the assessment center. Estradiol had been measured by 2-step competitive analysis on a Beckman Coulter Unicel Dxl 800. This assay had a lower detection limit of 175 pmol/L. Normal ranges of serum estradiol levels are known to be 36.7 to 183.6 pmol/L in men, 73.4 to 2753.5 pmol/L in premenopausal women, and 0 to 73.4 pmol/L in postmenopausal women (20). This results in a large fraction of the participants, especially men and postmenopausal women, having levels below detection limit. Among the UKB participants, 76 668 had estradiol measurements above detection limit. Estradiol levels were therefore analyzed as a binary phenotype (above/below detection limit) in our primary analyses, in contrast to a previous GWAS for estradiol, which excluded all individuals below detection limit. However, we also performed sensitivity analyses using the Tobit model where quantitative measures (for those above detection limit) are modeled together with the below-detection-limit samples.

GWAS and Sensitivity Analyses

GWAS was performed in males and females separately, using logistic regression with an additive genetic model implemented in PLINK version 2. The following covariates were included in both male and female analyses: age, body mass index, the first 10 genetic principal components (to adjust for population structures and ethnic origins), and a binary indicator variable for UKB Axiom *vs* UK BiLEVE genotyping array to adjust for any array differences. For females, hormone replacement therapy (current, ever, or never), oral contraceptive use (current, ever, or never), number of live births, menopausal status, and whether they have undergone hysterectomy were also included. Women who were unsure about menopausal status were excluded. Information for the covariates were collected from the initial visit to the assessment center, which was the same time as the blood sample for the estradiol measurements were taken.

In the GWAS, we applied a P-value threshold of 5×10^{-8}, which is commonly used in GWAS when using a minor allele frequency cutoff of 1% (21) and corresponds to a correction of 1 million independent tests (SNPs). Manhattan plots were generated using the R package *hudson*. We applied the clump function in PLINK 1.90 to define start and end positions for each locus with the main parameters that determines the clumping being set to R^2=0.1, $p_1 = 5 \times 10^{-8}$ and $p_2 = 1 \times 10^{-4}$. We tested all lead SNPs for significant differences in effect sizes between males and females by including an interaction term in the logistic regression model employed by PLINK.

We subsequently performed conditional analyses for each locus, conditioning on the SNP with the lowest P-value from each locus. If there was any significant conditional SNP, this was repeated by also conditioning on the conditional SNP from the previous analyses until no additional significant associations remained.

Several sensitivity analyses were performed for the GWAS. First, we stratified females into being pre- and postmenopausal. Second, we excluded all participants with a cancer diagnosis prior to assessment (when blood samples were taken). Third, we analyzed males and females with testosterone and sex hormone-binding globulin (SHBG) levels included as covariates, in addition to the other covariates, and with age at menopause as a covariate in the female postmenopausal stratum.

In addition, we replicated our results by estimating the effects of our lead SNPs on quantitative estradiol levels. To properly model estradiol levels as a quantitative response with a lower detection limit, we performed censored regression (Tobit-I) modeling (22). By adopting a model with an integrated error term up to the detection limit (175 pmol/L),

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

censored individuals with estradiol levels below the limit could also be included in the estimation. We used the vglm function from the R package VGAM 1.1-2, with the family argument set to "tobit." Prior to analysis, we transformed estradiol values using the rank-based inverse normal transformation for females and males separately. The corresponding transformed lower detection limit was then given by the maximal (transformed) value for each subset of individuals (males and females) with estradiol levels below 175 pmol/L. The estimated beta coefficient is that of the corresponding latent variable and is here interpreted as the usual additive SNP effect on estradiol level, as if estradiol levels were measured in all individuals. Because it has been recently highlighted that dichotomization of the exposure variable can lead to the IV assumptions in the MR being violated (23), the effect estimates we obtained from this analysis were used for sensitivity MR analyses for males and females separately, as described for the primary analyses in the following discussion.

Annotation of GWAS Lead SNPs

To identify likely target genes for the associated variants, we first identified genes containing the lead SNPs or the closest gene(s) for each individual lead SNP. We also used HaploReg v4.1(24) to identify potential functional effects by the individual lead SNP or any variant in linkage disequilibrium (LD; $R^2 > 0.8$) with the lead SNP. We then checked for overlap between the lead SNPs and expression quantitative trait loci (eQTL) using data from the Genotype-Tissue Expression (GTEx) project (25). We considered our lead SNP to overlap with an eQTL if our lead SNP was in LD ($R^2 > 0.8$) with a lead SNP for an eQTL. To identify overlap with other phenotypes from previously published GWAS and address possible issues with pleiotropy in our MR analyses, we downloaded summary statistics from the GWAS catalog (version 1.0.3 downloaded on November 13, 2020). We defined signals to be overlapping if a lead SNP from a previous GWAS was in LD ($R^2 > 0.8$) with any of our lead SNPs.

Mendelian Randomization

First, we undertook a 1-sample MR approach to investigate the causal effect of estradiol on BMD in UKB in males and females separately. BMD was measured as heel BMD, based on an ultrasound measurement of the calcaneus (heel bone) using the Sahara Clinical Bone Sonometer and automatically converted to T-scores (data field 78 in UKB). The T-score is the number of standard deviations a person's BMD differs from the mean BMD of their respective sex.

The main MR analyses were performed with the R package gsmr (version 1.0.8) (26). The generalized summary data-based MR (GSMR) approach allows both for heteroskedastic and correlated data via the variance-covariance matrix and identifies and removes pleiotropic outliers prior to causal effect estimation. The default LD threshold was used ($R^2 = 0.05$) and the significance threshold for an SNP being identified as an outlier and therefore not belonging to the set of valid instruments was set to the default value of $a = 0.01$. GSMR assumes a fixed effects model with the test statistic following a χ^2 distribution with 1 degree of freedom.

We further performed sensitivity analyses, applying inverse variance weighted, weighted median, and the MR-Egger method, using the TwoSampleMR package implemented in R (27). The MR-Egger intercept was used to assess the possible presence of directional pleiotropy in the set of genetic instruments, where a deviation from zero intercept indicates pleiotropy (28). Both the inverse variance weighted and the MR-Egger methods assume a random effects model, with underdispersion not being allowed. The standard error of the estimate for the weighted median method is estimated using bootstrapping. If heterogeneity is present in the data, these methods would tend to show higher *P*-values as compared GSMR, as GSMR is expected to remove outliers. To get a reasonable number of instruments to include in the MR analysis for females, we used a less strict *P*-value cutoff ($P < 1 \times 10^{-7}$), resulting in a total of 4 IVs that were included in the analysis [see Supplementary Table S3 in (29)]. In the GSMR analyses, the HEIDI (heterogeneity in dependent instrument) outlier flag was disabled in females, as there were fewer (<10) SNPs than recommended for this approach to work properly (26).

A 1-sample MR could be sensitive to weak-instrument bias, which could potentially result in bias toward the confounded association (ie, an inflated MR estimate). We therefore also performed a 2-sample MR, which is not affected by weak-instrument bias to further replicate the MR results. For the 2-sample MR, we used GWAS summary statistics for BMD from the GEFOS consortium (30). We used GWAS summary statistics for lumbar spine BMD from the 2012 data release on www.gefos.org (downloaded November 11, 2020), which was the only data set with summary statistics for males and females separately and that did not include samples from UKB. The 2 sets (for males and females) of genetic variants that had been selected in UKB were tested independently using GEFOS BMD summary statistics for males and females, respectively. The overlap, with regards to SNPs (IVs), was not perfect between UKB and GEFOS. We therefore revised the set of SNPs for the MR in GEFOS

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

slightly. For each of our GWAS locus, we selected the most significant SNP in LD with our lead SNP ($R^2 > 0.6$) for which summary statistics were also available in the GEFOS data as a proxy. However, to avoid too weak instruments in the MR, only proxies that were associated with estradiol levels ($P < 0.0001$) were included in the MR. We harmonized the SNP effects using the harmonize_data function included in the TwoSampleMR package. Finally, we note that any winner's curse bias in our effect estimates of the genetic variants associated with estradiol levels should, on average, drive the estimates of the causal effect of estrogen on BMD towards the null.

Results

A total of 147 690 male participants had passed our genotype and estradiol QC (Fig. 1; Table 1). Of these, 134 323 had estradiol levels below detection limit (175 pmol/L), while 13 367 individuals had measured estradiol levels above or equal to 175 pmol/L (Table 1). The number of females (n = 163 985) was slightly higher compared to males, with 37 461 and 126 524 having estradiol levels above and below detection limit, respectively. Dramatically smaller fractions of males (9.1%) and postmenopausal women (7.9%) had detectable estrogen levels (Table 1) in comparison to the fraction of premenopausal women (71.9%) with detectable estrogen levels. Number of live births, age, and body mass index were significantly lower among women with detectable estradiol in both post and premenopausal strata ($P < 0.05$).

GWAS Results

We identified 15 loci, on 14 different chromosomes, to be significantly associated ($P < 5 \times 10^{-8}$) with estradiol in males or females (Fig. 2; Table 2). Of these, 13 were genome-wide significant only in males, 1 only in females (*MCM8*) and 1 in both males and females (*CYP3A7*). On chromosome 2, and 15, one conditional SNP was identified per locus, located in the vicinity to the primary lead SNPs (Table 2). Among the lead SNPs identified in males, some were located in or close to genes with a well-known role in estradiol metabolism. Among the most significant loci, we found *FAM9A* and *CYP19A1*, which have previously been identified in GWAS for estradiol in males (12,15). One lead SNP, rs45446698, was genome-wide significant in both males and females, mapped to *CYP3A7*. It is also the most significant eQTL for *CYP3A7* in adrenal gland in GTEx

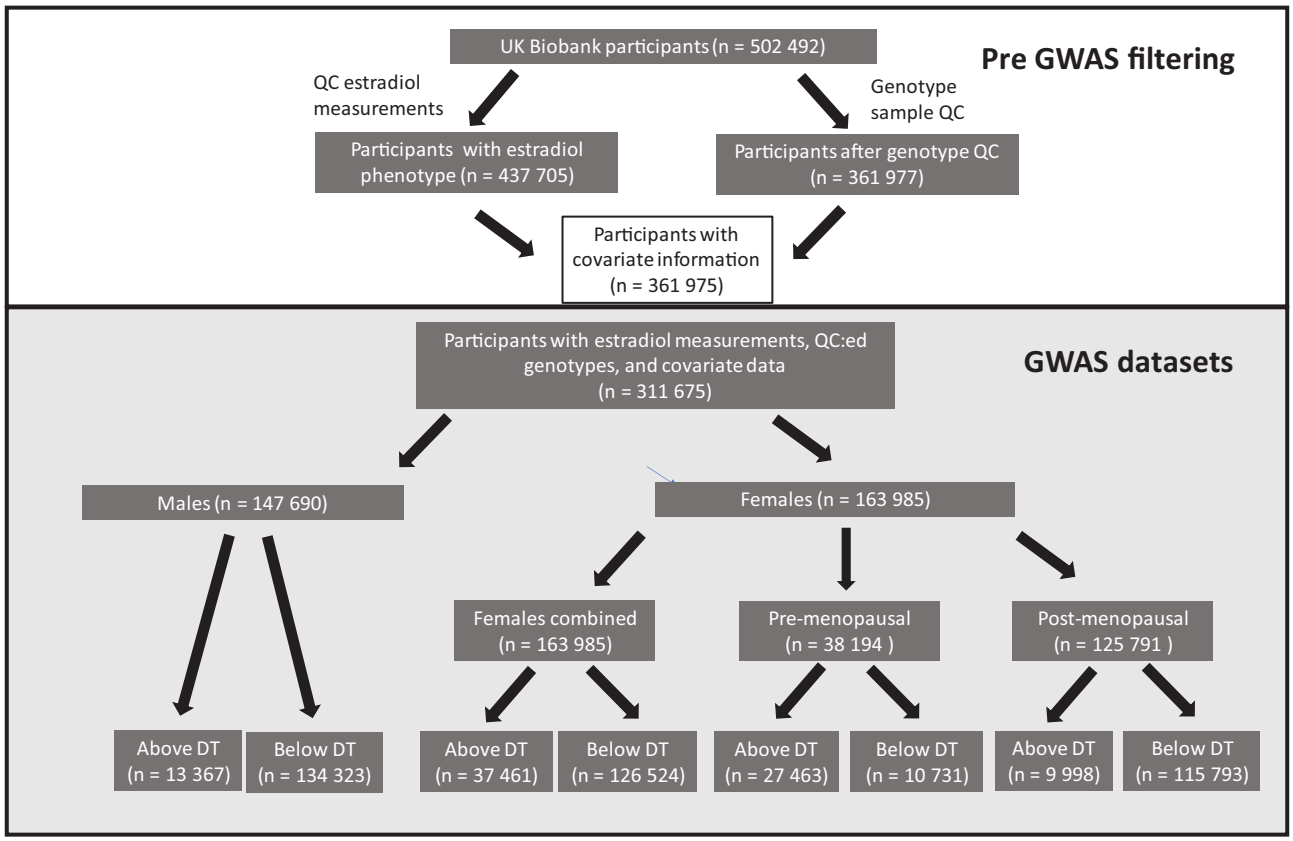


Figure 1. Flowchart of UK Biobank participants included in the four GWAS.
Abbreviations: DT, detection limit (for estradiol measurement); QC, quality control.

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

Table 1. Baseline characteristics in the UKB participants included in GWAS

	Males		Postmenopausal females		Premenopausal females	
	Above DT	Below DT	Above DT	Below DT	Above DT	Below DT
Participants, n (%)[a]	13 499 (9.1)	135 200 (90.1)	10 090 (7.9)	117 158 (92.1)	27 672 (71.9)	10 807 (28.1)
Participants with all covariates, n[a]	13 367	134 323	9998	115 793	27 463	10 731
Estradiol levels (pmol/L), mean (Q1-Q3)	203.8 (188.8-230.8)	NA	302.2 (221.2-475.2)	NA	433.7 (291.7-686.5)	NA
BMI, mean (Q1-Q3)[b]	**28.4 (25.2-30.8)**	**27.8 (25.0-30.0)**	**27.5 (23.6-30.2)**	**27.2 (23.7-29.8)**	**26.2 (22.7-28.7)**	**26.6 (22.9-29.2)**
Age, mean (Q1-Q3)[b]	57.1 (51-64)	57.1 (51-64)	**55.0 (50-60)**	**60.6 (57-65)**	**45.9 (43-48)**	**47.5 (44-50)**
Smoking, n (%)						
Current	1179 (8.73)	11 424 (8.45)	**784 (7.77)**	**7445 (6.35)**	**1862 (6.73)**	**809 (7.49)**
Never	6662 (49.3)	65 993 (48.81)	5635 (55.85)	67537 (57.65)	17 920 (64.76)	6909 (63.93)
Occasional	423 (3.1)	4424 (3.27)	**243 (2.41)**	**1884 (1.61)**	755 (2.73)	281 (2.60)
Previous	5175 (38.3)	52 905 (39.13)	3399 (33.69)	39 842 (34.01)	7083 (25.60)	2788 (25.80)
Had hysterectomy, n (%)[c]	NA	NA	**4868 (48.25)**	**27 841 (23.76)**	**79 (0.29)**	**76 (0.70)**
Hormone replacement therapy (never/current/previous), n[b]	NA	NA	**3474/4586/2024**	**58 112/4606/54 169**	**26 797/403/424**	**10 166/157/463**
Oral contraceptives (never/current/previous), n[b]	NA	NA	**1252/28/8798**	**25 005/87/91 874**	**2442/1090/24 098**	**883/1476/8430**
Number of live births, median (Q1-Q3)[c]	NA	NA	**1.79 (1-2)**	**1.9 (1-3)**	1.58 (0-2)	1.56 (0-2)
Age at menarche, median (Q1-Q3)[b]	NA	NA	**12.62 (12-14)**	**12.54 (12-14)**	**12.68 (12-14)**	**12.59 (12-14)**
Age at menopause, median (Q1-Q3)[b]	NA	NA	**44.62 (44-52)**	**46.53 (46-53)**	NA	NA

Bold cells are significantly ($P < 0.05$) associated with binary estradiol phenotype.
Abbreviations: BMI, body mass index; DT, detection limit; NA, not applicable; Q1, 1st quartile; Q3, 3rd quartile.
[a]These are the actual number of participants analyzed in the GWAS.
[b]Tested with Student's *t* test.
[c]Tested with Fisher's exact test.

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

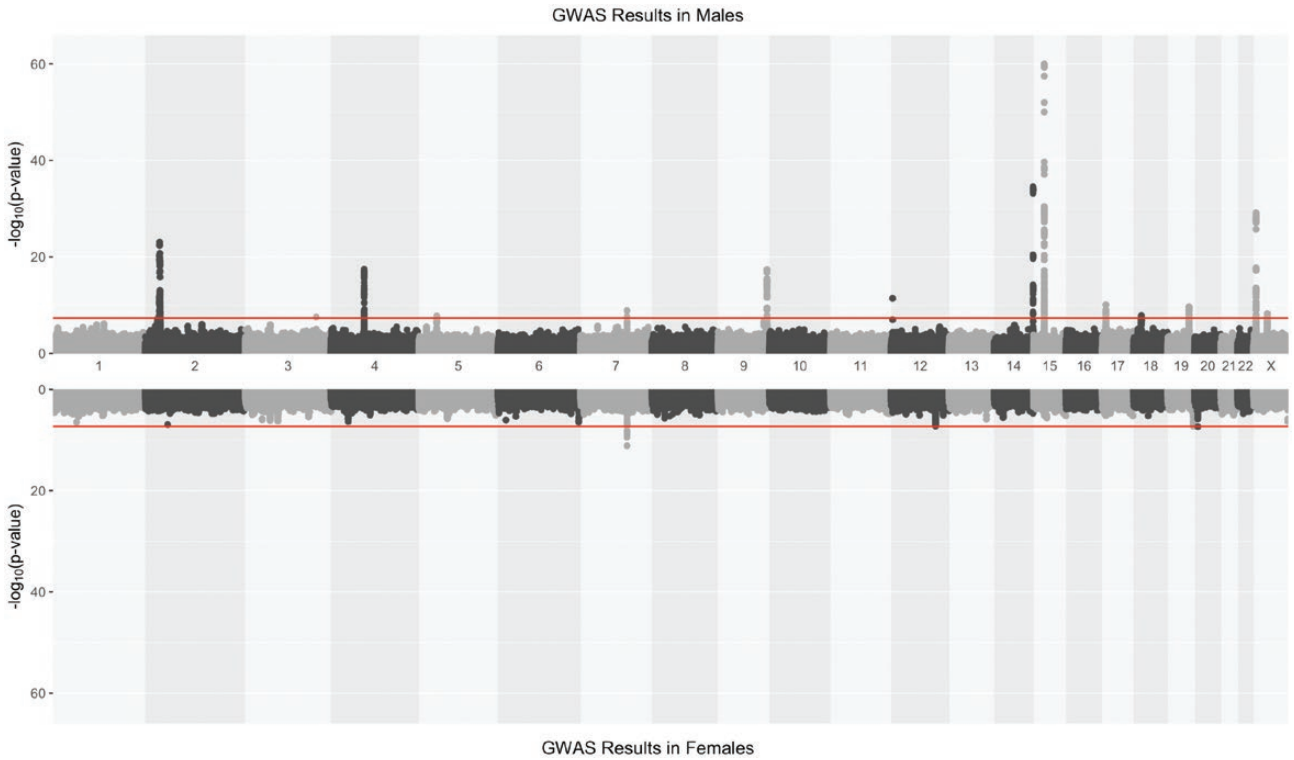


Figure 2. Hudson plot for the estradiol GWAS in males (upper) and females (lower). The red horizontal line indicated the genome-wide significance threshold (5×10^{-8}).

($P = 1.2 \times 10^{-12}$), suggesting that the effect is mediated by the expression level of *CYP3A7*. The minor (G) allele increases the level of *CYP3A7* expression but decreases the estradiol levels. There was also 1 lead SNP in *SHBG*, which codes for SHBG that had been linked to measurements of several other sex hormones (16). Another lead SNP was located on the X chromosome, within the *AR* gene that encodes an androgen receptor. We also found a very strong association on chromosome 2, that overlaps with *SRD5A2*, which codes for 3-oxo-5-alpha-steroid 4-dehydrogenase 2, known to catalyze the conversion of testosterone into androgen or dihydrotestosterone (31). Our most significant SNP is in LD ($R^2 = 0.75$) with rs9282858, which is a missense variant that is predicted to be deleterious (SIFT = 0.03). Here, the minor allele was associated with higher estradiol levels (Table 2). At the same locus on chromosome 2, there was also 1 strong conditional signal, further upstream of the primary signal, closer to *MEMO1*.

There was also an association to rs657152 in *ABO*, a SNP that is in LD ($R^2 = 0.98$) with rs8176719, which tags the O blood group (32). The A allele of rs657152 tags the non-O blood groups, and we found rs657152-A to be associated with lower levels of estradiol, indicating that individuals with a non-O blood group have lower levels of estradiol compared to individuals with the O blood group (Table 2). Two associations on chromosome 4 and 5 respectively, mapped to *UGT2B7* (rs7662029) and *UGT3A1* (rs1073548). UGT3A1 and UGT2B7 are glucuronosyltransferases, which are enzymes that catalyze the glucosidation of lipophilic chemicals, facilitating their elimination (33). Another lead SNP mapped to *SULT2A1*, which encodes a sulfotransferase that is believed to be important for sex-steroid biosynthesis by sulfating hydroxylated steroid hormones and bile acids (34). Another lead SNP on chromosome 14 is a missense variant in *FKBP4,* which encodes the protein FKBP52 that has been suggested to be involved in the formation of steroid hormone nuclear receptors that regulate hormone levels (35).

The remaining lead SNPs from the GWAS in males were mapped to loci with less clear links to estradiol metabolism. These include 1 lead SNP (and 1 conditional SNP) in a locus with a cluster of genes, including *IGHV3-7* and *IGHV6-1*, that encodes parts of immunoglobulin heavy chains that participates in the antigen recognition. Another lead SNP is located in *RBBP8*, which encodes Retinoblastoma-binding protein 8 (RBBP8, CtIP), an endonuclease involved in the repair of double-stranded DNA breaks through homologous recombination as well as regulation of G2/M cell-cycle checkpoints as part of the BRCA1-RBBP8 complex (36). Variation in *RBBP8* genotype and regulation has been reported in multiple types of cancer (37,38). A third SNP was a missense variant in *LINC01324*, a long intervening/

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

Table 2. Results of the GWAS including conditional SNPs

Lead SNP	Chr:position	Primary/ conditional	GWAS strata	Gene	Eff allele	MAF	Males		Females	
							OR (95% CI)	*P*-value	OR (95% CI)	*P*-value
rs112881196[a]	2:31982811	Primary	Males	*SRD5A2*	G	0.039	1.353 (1.275-1.435)	8.52E-24	1.106 (1.041-1.175)[f]	1.13E-03
rs62142080	2:32182528	Conditional	Males	*MEMO1/ SRD5A2*	C	0.396	0.927 (0.903-0.952)	2.87E-08	0.989 (0.965-1.013)[f]	5.66E-01
rs57102816	3:164506246	Primary	Males	*LINC01324*	TA	0.257	1.084 (1.054-1.115)	3.61E-18	1.008 (0.981-1.036[f]	9.45E-03
rs7662029	4:69961912	Primary	Males	*UGT2B7*	G	0.454	0.893 (0.871-0.916)	1.75E-08	0.969 (0.946-0.992[f]	5.28E-03
rs1073548	5:36012617	Primary	Males	*UGT3A1*	G	0.126	1.113 (1.072-1.155)	1.38E-09	1.052 (1.015-1.090)	7.62E-12
rs45446698[b]	7:99332948	Primary	Males/females	*CYP3A7*	G	0.043	0.811 (0.758-0.868)	4.14E-18	0.813 (0.767-0.863)	3.32E-04
rs657152[c]	9:136139265	Primary	Males	*ABO*	A	0.340	0.887 (0.863-0.911)	4.11E-12	0.955 (0.932-0.980)[f]	6.56E-02
rs56196860[d]	12:2908330	Primary	Males	*FKBP4*	A	0.031	1.263 (1.183-1.350)	3.08E-35	0.938 (0.877-1.004)[f]	1.06E-04
rs11160915	14:106514653	Primary	Males	*IGHV3-7*	A	0.444	0.849 (0.827-0.871)	1.01E-60	0.954 (0.931-0.977)[f]	2.27E-03
rs28892005	15:51519945	Primary	Males	*CYP19A1*	A	0.350	0.795 (0.774-0.817)	8.53E-11	0.962 (0.939-0.986)[f]	1.84E-03
rs3751591	15:51606710	Conditional	Males	*CYP19A1*	G	0.167	1.125 (1.088-1.163)	1.47E-08	1.004 (0.972-1.036)[f]	3.37E-01
rs62059839	17:7533015	Primary	Males	*SHBG*	T	0.262	1.098 (1.067-1.130)	2.18E-10	1.044 (1.016-1.072)	4.56E-01
rs113047993	18:20585399	Primary	Males	*RBBP8*	T	0.068	0.858 (0.813-0.904)	1.01E-01	0.977 (0.932-1.024)[f]	4.67E-08
rs62129966	19:48374950	Primary	Males	*SULT2A1*	A	0.164	1.113 (1.077-1.150)	7.49E-30	1.012 (0.981-1.045)[f]	3.48E-01
rs16991615	20:5948227	Primary	Females	*MCM8*	A	0.065	1.043 (0.992-1.097)	5.60E-09	1.139 (1.087-1.193)	9.19E-01
rs5933688[e]	X:8880680	Primary	Males	*FAM9A*	G	0.275	1.118 (1.097-1.140)	8.52E-24	1.013 (0.986-1.040)[f]	1.13E-03
rs114255570	X:67005508	Primary	Males	*AR*	A	0.077	0.899 (0.867-0.931)	2.87E-08	1.002 (0.959-1.047)[f]	5.66E-01

[a]LD with rs9282858, which is missense variant in *SRD5A2* (SIFT = 0.03 deleterious, PolyPhen2 = 0.337 benign).
Abbreviation: MAP, minor allele frequency.
[b]eQTL for CYP3A7.
[c]rs657152 tags the O blood group.
[d]rs56196860 is a missense variant in FKBP4.
[e]rs16991615 is a missense variant in MCM8.
[f]Significant ($P < 0.05$) difference in OR between males and females.

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

intergenic noncoding RNA, which has been established as a predictor for melanoma progression (39).

The lead SNP that was only significant in females (rs16991615) (Table 2) is a missense variant located in *MCM8*. *MCM8* codes for a DNA helicase, which is involved in the repair of double-stranded DNA breaks and recombination (40). Even though very few genome-wide significant SNPs were identified in women, 8 out of 14 of the loci that were genome-wide significant in males were at least nominally associated in females as well ($P < 0.05$) with the same direction of effect (Fig. 3), especially the genes with functions related to hormone metabolism: *UGT2B7*, *UGT3A1*, *CYP3A7*, *CYP19A1*, *SRD5A2*, and *SHBG* replicated in females as well as *IGHV3-7* and *ABO*. However, the effects of all SNPs, except for the SNP in *CYP3A7*, which achieved genome-wide significance in both sexes, were significantly different between males and females, as identified by significant sex-genotype interactions [see Supplementary Table 1 in (29)].

Sensitivity Analysis

The primary aim of the sensitivity analyses was to further verify the associations for the lead SNPs from the primary GWAS. The effect size of the lead SNPs from the primary analyses changed somewhat. When stratifying females into pre- and postmenopause, there was 1 locus, *CYP3A7*, with a significant difference in the odds ratio (OR) between the 2 strata [Fig. 3; also see Supplementary Table 1 in (29)]. For *CYP3A7*, the largest effect was seen in postmenopausal women [OR = 0.703 (0.642-0.770)], followed by males [OR = 0.811 (0.758-0.868)] and premenopausal [OR = 0.907 (0.837-0.984)].

After removing all participants that had a prior cancer diagnosis at the time of assessment, there was no significant difference in OR compared to the primary GWAS, and all lead SNPs were still genome-wide significant, indicating that a prior cancer diagnosis did not influence our primary results [see Supplementary Table 1 in (29)]. However, when adjusting the estradiol GWAS for

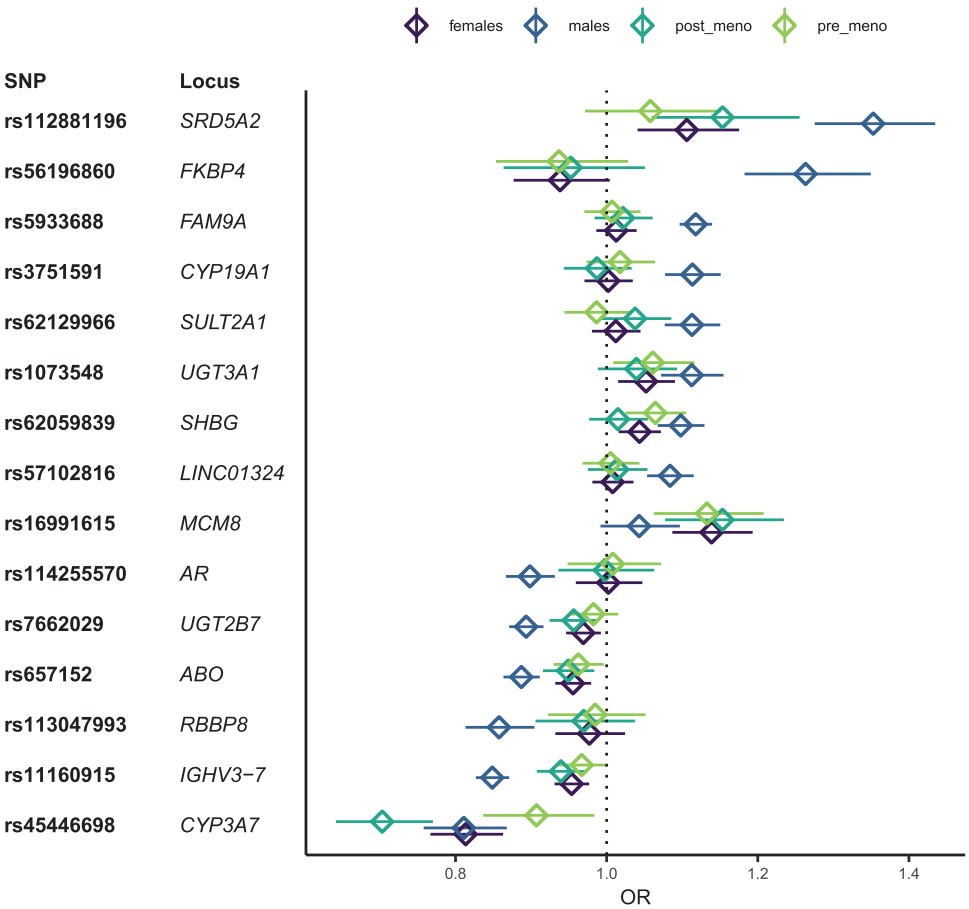


Figure 3. ORs and 95% CIs for the lead SNPs in the male and female strata, as well as the 2 female subgroups: pre- and postmenopause. The lead SNP for each locus that reached genome-wide significance in males or females in the primary analyses are included. Non-overlapping confidence intervals typically indicate a significant difference in ORs well below $P = 0.05$.

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

testosterone and SHBG levels, the lead SNPs at *SHBG* and *FKBP4* were no longer genome-wide significant in males, and the corresponding effects were significantly lower than the previous signals. *UGT3A1* and *AR* lost genome-wide significance for estradiol when adjusting for testosterone and SHBG levels, but the ORs were not significantly different from the primary analyses. This indicates that the signals at *SHBG* and *FKBP4* might have been completely driven by SHBG and testosterone levels, whereas the other loci are more likely to be estradiol specific loci. We observed the same effect for the female-specific locus *MCM8*. There was no significant difference in OR between the primary GWAS in postmenopausal women with and without adjustment for age at menopause, suggesting that the results in females are not strongly confounded by age at menopause as has been discussed previously (16). In our quantitative analysis using the Tobit model [see Supplementary Table 1 in (29)], 13 of the lead SNPs were still genome-wide significant ($P < 5 \times 10^{-8}$), while 3 lead SNPs were just below the threshold (5.2×10^{-8}, 5.7×10^{-8}, and 5.9×10^{-8}, respectively).

GWAS Overlap With Estradiol and Related Phenotypes

We did a systematic check for overlap with previous GWAS results that were available in the GWAS catalog [see Supplementary Table 2 in (29)]. For lead SNPs at *UGT3A1*, *LINC01324*, and *RBBP8*, we found no overlap with previous GWAS signals. Several SNPs at the *AR* locus have previously been associated with male pattern baldness (41), but none of those are in strong LD with our lead SNP ($R^2 < 0.3$). The lead SNP at the other 10 loci all overlapped with at least 1 previous GWAS. There were several overlapping signals with phenotypes that are related to sex hormones. For example, the lead SNPs at *SRD5A2* and *FAM9A* overlap with male pattern baldness as well. The lead SNPs at *CYP3A7* and *SULT2A1* overlap with GWAS hits for several sex hormones and other metabolites, *UGT2B7* overlaps with serum and urinary metabolites, while *SHBG* overlaps with testosterone levels. Our lead SNPs in *FAM9A*, *CYP19A1*, and *CYP3A7* all overlap with previous GWAS for BMD. Other loci overlap with GWAS for more distal traits. For example, the lead SNP in *ABO* overlaps with results from over 100 GWAS, with associations to metabolites, liver enzymes, and proteins as well as cardiovascular diseases, cancer, and T2D. The lead SNP at *FKBP4* overlaps with waist-to-hip ratio and lung function. *MCM8*, the female specific locus, overlaps with traits related to age at menopause. This agrees partly with our sensitivity analyses where the OR decreased slightly from OR = 1.139 (1.087-1.193) in the primary GWAS to OR = 1.126 (1.035-1.225) when adjusted for age at menopause. Still, the OR was significantly different from unity in the sensitivity analyses, suggesting that the locus is probably associated with both age at menopause and estradiol levels.

Mendelian Randomization

A total of 209 043 participants (96 432 males and 112 611 females) had BMD measurements and covariate information available and were included in the MR analysis. Mean value for T-scores in males was −0.085, and for females −0.58. The below-zero average T-scores agree with the average age in UKB participants being higher than the general population and thereby the BMD being lower compared to the mean BMD of respective sex. The instrument variables for the MR analyses were selected from the lead SNPs in the primary GWAS for males and females, respectively. For males, all significant lead SNPs and conditional SNPs were selected, resulting in 16 instruments (Table 3). The estimated effects on estradiol level [see Supplementary Table 3 in (29)] were taken from the primary GWAS for all instruments and for the SNPs that were identified in the conditional analyses. Hence, the estimates for the conditional SNPs differ slightly from the ones in Table 2. Using GSMR, 3 instruments were identified as pleiotropic and were excluded from the primary analysis in males [see Supplementary Table 3 in (29)]. We identified a causal effect of high estradiol levels on BMD in males ($\beta \approx 0.10$,

Table 3. Results of MR analyses

	Males (1-sample MR: UKB)		Males (2-sample MR:UKB-GEFOS)		Females (1-sample MR: UKB)		Females (2-sample MR:UKB-GEFOS)	
MR method	Estimate	*P*-value	Estimate	*P*-value	Estimate	*P*-value	Estimate	*P*-value
GSMR	9.89E-02	1.57E-11	1.82E-01	2.08E-04	3.27E-02	7.48E-06	3.69E-01	1.67E-03
Weighted median	6.61E-02	2.82E-02	2.65E-01	7.44E-05	7.31E-02	1.26E-01	4.19E-01	3.54E-03
IVW	1.48E-01	5.49E-05	1.82E-01	3.47E-03	1.45E-01	9.98E-02	2.45E-01	2.57E-01
MR-Egger	1.42E-01	2.23E-01	4.50E-01	2.73E-02	4.81E-01	2.56E-02	3.96E-01	7.92E-01
MR-Egger intercept	8.87E-04	6.95E-01	−4.01E-02	1.07E-01	−3.63E-02	1.71E-01	−1.29E-02	9.15E-01

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

$P = 1.57 \times 10^{-11}$) (Table 3; Fig. 4A). Similar MR results were obtained using weighted median and MR-Egger, which do not remove pleiotropic SNPs. The MR-Egger intercept was not significantly different from zero, indicating that these results are not strongly influenced by directional pleiotropy.

To get a reasonable number of instruments to be used for the MR analysis in females, we used a less strict *P*-value cutoff ($P < 1 \times 10^{-7}$), resulting in a total of 4 IVs [see Supplementary Table 3 in (29)] that were included in the analysis. We identified a causal effect of high estradiol levels on BMD also in women ($\beta = 0.03$, $P = 7.48 \times 10^{-6}$) (Fig. 4B). All 3 alternative MR methods showed the same direction of effect, although the null hypothesis of no effect could not be rejected (Table 3). The MR-Egger intercept was relatively large (-3.63×10^{-2}), which is an indication of directional pleiotropy but was not significantly different from zero. We investigated each SNP for horizontal pleiotropy using information from the GWAS Catalog (https://www.ebi.ac.uk/gwas/). We did not identify any variants that were in LD with our instruments and associated with correlated traits.

To further validate the MR results, we also performed a 2-sample MR for males and females, respectively, using the instruments selected for estradiol in UKB and the GWAS summary statistics from the BMD GWAS in GEFOS (30). A slightly different set of IVs had to be selected [see Supplementary Table 3 in (29)], since considerably fewer SNPs were analyzed in GEFOS due to their use of the HapMap reference panel for imputations. We included 8 and 4 instruments for the replication in males and females, respectively, none of which were palindromic. The MR estimates were in the same direction as the 1-sample MR (Fig. 4C and 4D), but with a larger effect in both males ($\beta = 0.18$, $P = 0.00012$) and females ($\beta = 0.41$, $P = 3.49 \times 10^{-5}$). Similarly to the 1-sample MR, neither MR-Egger intercept was significantly different from zero.

As estradiol was analyzed as a binary phenotype, the unit for the estimate from the MR can be interpreted as standard units increase in T-score associated with having estradiol levels above detection limit. However, our MR results were further replicated using SNP effects on the quantitative estradiol measurements from the Tobit model [see Supplementary Table 4 in (29)]. Here, the *P*-values were quite comparable to the MR for the binary estradiol exposure. However, the causal estimates are easier to interpreted and indicates that the BMD increases with 0.38 (95% CI 0.30-0.46) and 0.53 (95% CI 0.31-0.74) standard units per 1 standard unit increase in estradiol levels in males and females, respectively.

Discussion

We have presented results from a large GWAS for estradiol levels in males and females. Many of the significant SNPs are annotated to genes with functions that are biologically relevant to estrogen metabolism. For example, by converting and transporting androgens and endogens, synthesis or elimination of steroid hormones. In males, we identified 14 loci to be associated with serum estradiol levels, as well as 2 additional independent SNPs on chromosomes 2 and 15. In females, we identified 2 significant loci. One locus within the *CYP3A7* gene overlaps with our findings in males, but the other locus, mapping to *MCM8*, was specific to females. However, a majority (n = 8) of the male loci, was at least nominally significant in females, although the estimated effect sizes were lower. These results suggest that genetic effects on estradiol levels might differ between males and females. However, in premenopausal women, estradiol levels vary during the menstrual cycle, limiting our ability to estimate the genetic effects on estradiol in females. In postmenopausal women, on the other hand, estradiol levels drop rapidly, and it has previously been suggested that the estradiol levels in females are more likely to be determined by the number of years since menopause (42). For those reasons, we performed sensitivity analyses, stratifying females by menopausal status and adjusting for age at menopause in the postmenopausal strata. The only locus that was significantly different between pre- and postmenopausal women was *CYP3A7*. The largest effect was seen in postmenopausal women, followed by males and premenopausal women. *CYP3A7* encodes cytochrome P450 CYP3A7, which metabolizes dehydroepiandrosterone (DHEA), a precursor of both androgen and estrogen synthesis, and shows catalytic activity for conversion of estrone to hydroxyestrogens (43,44). As much as 75% of the circulating estrogens in premenopausal women and 50% of circulating androgens in men are known to be derived from DHEA. After menopause, DHEA is thought to be the major precursor of androgens and estrogens in females (45). Interestingly, the effect of the *CYP3A7* SNP disappeared when adjusting for SHBG and testosterone levels in females, but the effect was unaffected in males. This suggests that *CYP3A7* might have different functions in males and in females, as well as in pre- and postmenopausal women. In postmenopausal women *CYP3A7* appears to be regulating the amount of available testosterone as a precursor for estrogen.

Several other loci, with previous links to estrogen metabolism were also identified. *SULT2A1* encodes a sulfotransferase, which is active on hydroxylated steroid hormones and bile acids (34). Sulfotransferase 2A1, encoded by *SULT2A1*, mediates the sulfation of a wide range of steroids and sterols, including DHEA, a precursor of estrogen and testosterone. *SULT2A1* was only associated with estradiol in males, and this association was similar (slightly stronger) when adjusting for testosterone and SHBG levels. Two of the most well-known loci, which has previously also been associated with estradiol in males in

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

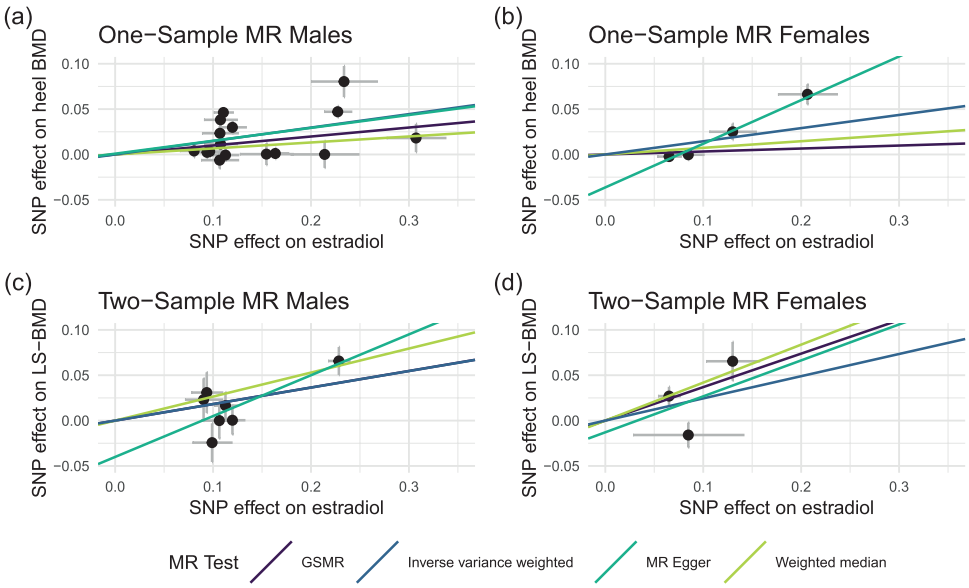


Figure 4. Mendelian Randomization analyses to estimate the causal effect of estradiol on BMD. (A) One-sample MR in UKB males, (B) 1-sample MR in UKB females, (C) 2-sample MR with GEFOS males, and (D) 2-sample MR in GEFOS females. Abbreviation: LS, Lumbar Spine.

GWAS, are *FAM9A* and *CYP19A1*(12,15). Aromatase, encoded by *CYP19A1*, is a key enzyme in the synthesis of estradiol and estrone from testosterone and androstenedione, respectively (46). The association to *CYP19A1* was at least nominally significant, in both the pre- and postmenopausal women, as well as in all sensitivity analyses. The *P*-values were even numerically slightly lower when adjusting for SHBG and testosterone, which agrees with the function of the protein. *FAM9A* is exclusively expressed in testis (47), and it is therefore not surprising that there was no association to *FAM9A* in females. Previously, *FAM9A* has also been associated with testosterone levels (15), and our analyses indicate that the association was partly driven by associations to testosterone or SHBG levels. We could therefore conclude that the SNP at *FAM9A* is likely to affect both testosterone and estrogen metabolism. This signal also overlaps with a GWAS signal for male pattern baldness (41), which further supports the importance of this locus. However, *FAM9A* encodes the protein FAM9A, which has a rather unknown function. Also, the SNP is located between *FAM9A* and *FAM9B*, and the annotation of this locus should therefore be interpreted with care.

We identified a strong association between estradiol levels and the *SHBG* locus. The *SHBG* gene encodes SHBG that has been linked to measurements of several other sex hormones previously (16). SHBG is a glycoprotein that binds to sex hormones, including testosterone and estrogen, in the bloodstream and thereby regulates the amount of bioavailable hormones (48). Only 1% to 2% of sex hormones are unbound and therefore bioavailable (49-51). The association to SHBG in males completely disappeared after adjusting for SHBG levels. In females, there was a nominally significant association to SHBG in the primary GWAS with the same direction of effect as in males. Interestingly, after adjusting for SHBG and testosterone levels, there was still an association ($P = 0.00033$), but the effect was in the opposite direction. Similarly, an opposite effect between males and females has previously been seen for total testosterone, where there is a negative genetic correlation between total testosterone and SHBG in females but a positive correlation in males (16). This clearly illustrates the heterogenous effects between males and females for sex hormones. We also identified 2 associations to the X chromosome in males, 1 mapping to the *AR* gene, which encodes the androgen receptor. In mammals, the action of androgens acts through this single androgen receptor. In different tissues, the androgen receptor is activated by testosterone, or testosterone is first converted into dihydrotestosterone, which is then the activating molecule (52). Therefore, the androgen receptor plays an important role in sex differentiation, mainly through regulating gene expression (53). Interestingly no association ($P > 0.05$) was seen in females, which might be due to the much lower levels of androgens in females. We also found an association to *SRD5A2* which encodes 3-oxo-5-α-steroid 4-dehydrogenase 2 and is located on chromosome 2. This enzyme catalyzes the conversion of testosterone to 5-α-dihydrogentestosterone (31). Thus, the variant in *SRD5A2* might influence estradiol levels in men by affecting the amount of testosterone available for conversion to estradiol. The SNP in *SRD5A2* is nominally significant in females, an association that is more pronounced in

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

postmenopausal women. There is also a conditional signal in males, 200 kb upstream, close to *MEMO1*, that encodes the mediator of ErbB2-driven cell motility 1 (MEMO1). In previous studies, it has been shown that MEMO1 regulates the extranuclear function of the estradiol receptor (54), which is a possible link to estradiol levels. It is therefore possible that the GWAS signal on chromosome 2 indeed represents associations with 2 different genes. Another interesting association is with *FKBP4*, which encodes FKBP52 (FK506-binding protein 4). FKBP52 has been shown to regulate the androgen, glucocorticoid, and progesterone receptor signaling pathways (35) by playing an important role in the formation of steroid hormone nuclear receptors (55). *FKBP4* has been shown to regulate androgen receptor transactivation activity and has also been linked to testosterone levels (56). When adjusting for testosterone and SHBG levels, the *FKBP4* association was weaker and no longer significant.

Two loci, mapping to *UGT2B7* and *UGT3A1*, were also identified. These 2 loci encode uridine diphosphate (UDP)-glycosyltransferases, which are responsible for the addition of sugars to a broad range of lipophilic molecules to improve the water solubility and thereby play a major role in elimination of potentially toxic xenobiotics and endogenous compounds. Both UGT3A1 and UGT2B7 have been shown to metabolize estrogens by glucuronidation (ie, the transferring of glucuronic acid to the metabolites) (33,57,58). This suggests that UDP-glycosyltransferases might play a major role in estradiol metabolisms. Also, *ABO* encodes a glycosyltransferase that is known to modify carbohydrates on the red blood cell antigens. *ABO* is the gene that is responsible for the ABO blood groups (32). The effect allele of our study (the A allele) is in LD with the reference allele of rs8176719 (G), whose alternative allele (a deletion) encodes the O blood group. This indicates that men with a non-O blood type have lower estrogen levels or the other way around—men with blood group O have higher levels. Since *ABO* and its histoblood group properties have been linked to a vast number of diseases and functions, it is not unlikely that our finding is indicative of an association between ABO antigens and estradiol levels. The most well-known association for ABO is that individuals with the O blood group have lower risk of cardiovascular diseases (59), probably due to lower coagulation activity. There are very limited findings of a clear link between ABO and estrogen levels. However, 1 study found an interaction between ABO and hormone replacement therapy in relation to coagulation activity (47). ABO antigens are present on most epithelial and endothelial cells as well as on T cells, B cells, and platelets and are also detectable in most body fluids such as saliva (60). Recently, ABO was linked to coronavirus disease 2019, where the susceptibility to severe acute respiratory syndrome coronavirus 2 has been suggested to be explained by modulation of sialic acid-containing receptors distributed on host cell surface (61). Coronavirus disease 2019 has a higher mortality in males, which has been suggested to be linked to their lower estrogen levels (62,63). We also identified several loci with less clear links to estrogen metabolism including *LINC01324*, *IGHV3-7*, *RBBP8*, and *MCM8*. *MCM8* is the female-specific locus, where the lead SNP is a missense variant. Genetic variants in *MCM8* have previously been associated with premature ovarian failure (64,65), a condition that is characterized by low estrogen levels (55). According to GTEx, *MCM8* is highly expressed in both female and male reproductive organs, despite only being significant in the female GWAS cohort, which further support this locus as being involved in sex hormone metabolism.

We performed MR analysis and estimated a causal effect of estradiol on BMD in males and, to our knowledge, for the first time also in females. Compared to previous studies in males, with not more than 5 SNPs as instruments for estradiol (17), we included as many as 16 SNPs in males and 4 in females. Interestingly, the effect estimate was higher for females than males in the 2-sample MR, which could agree with the fact that BMD decreases more rapidly in females by age. Estrogen can be used as a preventative treatment in postmenopausal women with low BMD (66,67) to prevent osteoporosis, and our results were therefore not surprising.

One of the major strengths of this study is the large cohort from UKB with estradiol measurements available. However, only a subset of the participants had estradiol levels above detection limit. To increase the number of participants to analyze, we created a binary phenotype. Since normal levels of estradiol in males and postmenopausal women have a range below or just above the detection limit, we were not able to capture the full spectrum of individual variation in the cohort. However, we also used a second approach, Tobit-I modeling, which takes into account the quantitative measures for samples above detection limit without disregarding the values below detection limit. While this is a well-established method and frequently used (eg, in econometrics), it has not been widely used in genetic association studies. Here we showed that this approach results in similar *P*-values compared to the binary estradiol variable but provides quantitative effect estimates that are much easier to interpret. Since very few GWAS have been performed for estradiol previously, we only found studies with effects estimates on the quantitative scale for 1 of our GWAS signals: rs28892005 in the *CYP19A6* locus [see Supplementary Table 2B in (29)]. In previous studies, only SNPs in LD with our lead variants had been reported, but the absolute value of the effect sizes ($\beta = 0.107$, $\beta = 0.065$,

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

and β = 0.062) were comparable to our Tobit model estimates for males (β = 0.106) (11,14,15). While many clinical assays have a lower detection limit, we believe that the censored-regression approach using Tobit modeling should be useful for other studies as well. Also, in MR studies, the second MR assumption can be violated by dichotomizing the quantitative phenotype (23). The second MR assumption states that the IVs should be associated with the outcome only through the exposure. However, a change in the underlying quantitative exposure may effect a change in the outcome, without detectable change in the dichotomized, thereby violating this assumption. By re-estimating the causal effects of estradiol on BMD, using the quantitative effect estimates from the Tobit model, we could ensure that a potential violation of the second MR assumption did not dramatically influence our results.

Very few loci were identified to be genome-wide significant in females, which resulted in a low number of IVs available for the MR. This could have been a cause of instability in our MR results, leading to nonsignificant results for some MR approaches. The selection of proxies, for which effect estimates were provided by GEFOS, did not ameliorate this problem. Future studies, where estradiol is measured in an even larger set of women or in women that are more synchronized with regards to timing in relation to menopause and/or ovulation, should be beneficial for identifying additional IVs for females and verifying our MR estimates. In addition, few instruments could also lead to lack of power to detect a significant deviation from zero for the MR-Egger intercept. In fact, we did observe the MR-Egger intercept to deviate from zero, even if it was not a significant deviation, which could possibly indicate that our analyses were affected by directional pleiotropy. However, in a 1-sample MR design, the MR-Egger regression, including the test for directional pleiotropy, has been found to be more strongly affected by weak instrument bias than, for example, the inverse variance weighted method, and it is therefore deemed inappropriate in this setting (68,69). However, through a comparison with the GWAS Catalog (https://www.ebi.ac.uk/gwas/), we did indeed find that many of our GWAS SNPs had previously been associated with different phenotypes. For example, age at menopause, a phenotype that also influences the risk of osteoporosis, is associated with our female-specific locus *MCM8*. This could potentially have led to pleiotropy (68).

In summary, we have identified biologically relevant genetic loci associated with variation in estradiol levels in males and females and found differences in genetic effects between sexes. In addition, we used an MR approach to show that estradiol levels have a causal effect on good bone health in both males and females, supporting the use of estrogen replacement therapy as a good treatment to prevent osteoporosis and fractures in elderly of both sexes. Our findings support previous research on the synthesis of estradiol and its effects on BMD and provide new insights into the genetic components of estrogen regulation and metabolism.

Acknowledgments

This research was conducted using the UKB Resource under application number 41143, following the restrictions on data availability set up by the UKB. We acknowledge all participants and staff involved in UKB for their valuable contribution. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project sens2017538.

Financial Support: The research was funded by the Swedish Research Council (Å.J. 2019-01497), the Swedish Brain Foundation (Å.J. FO2020-0205 and FO2019-0129), the Swedish Heart Lung Foundation (Å.J. 20200687), as well as the Wiberg (W.E.E.), M Borgström (W.E.E.), Hedströms, K och O F (W.E.E.), A and M Rudbergs (W.E.E.) foundations. The funding sources had no influence and took no part in the design or conduct of this research.

Author Contributions: W.E.E. and Å.J. designed the study; E.B., W.E.E., and D.S. performed the data analysis; D.S. and W.E.E. generated the figures; W.E.E., D.S., and Å.J. wrote the manuscript; W.E.E., T.K., D.S., E.B., Å.J., and J.H. interpreted the data, contributed to, and reviewed the manuscript.

Additional Information

Correspondence: Daniel Schmitz, MS, Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. E-mail: daniel.schmitz@igp.uu.se; or Åsa Johansson, PhD, Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. E-mail: asa.johansson@ipg.uu.se.

Disclosure Summary: The authors declare no conflict of interests.

Data Availability: Some data generated or analyzed during this study are included in this published article or in the data repositories listed in references. Supplementary data are freely available on Zenodo (https://doi.org/10.5281/zenodo.4926701). GEFOS summary statistics are available for download from the GEFOS Consortium (http://www.gefos.org/). Summary statistics from other GWAS can be downloaded from the GWAS Catalog (https://www.ebi.ac.uk/gwas/). Restrictions apply to the availability of some data generated or analyzed during this study to preserve patient confidentiality or because they were used under license. Individual-level genetic and clinical data are available for bona-fide researchers from UK Biobank (https://www.ukbiobank.ac.uk/) upon filing an application.

References

1. Bates GW, Bowling M. Physiology of the female reproductive axis. *Periodontol 2000*. 2013;**61**(1):89-102.
2. Hess RA, Bunick D, Lee KH, et al. A role for oestrogens in the male reproductive system. *Nature*. 1997;**390**(6659):509-512.
3. Thomas MP, Potter BV. The structural biology of oestrogen metabolism. *J Steroid Biochem Mol Biol*. 2013;**137**:27-49.

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

4. Rosendaal FR, Van Hylckama Vlieg A, Tanis BC, Helmerhorst FM. Estrogens, progestogens and thrombosis. *J Thromb Haemost*. 2003;**1**(7):1371-1380.
5. Cauley JA, Lucas FL, Kuller LH, Stone K, Browner W, Cummings SR; Study of Osteoporotic Fractures Research Group. Elevated serum estradiol and testosterone concentrations are associated with a high risk for breast cancer. *Ann Intern Med*. 1999;**130**(4 Pt 1):270-277.
6. Vikan T, Schirmer H, Njølstad I, Svartberg J. Low testosterone and sex hormone-binding globulin levels and high estradiol levels are independent predictors of type 2 diabetes in men. *Eur J Endocrinol*. 2010;**162**(4):747-754.
7. Manolagas SC, O'Brien CA, Almeida M. The role of estrogen and androgen receptors in bone health and disease. *Nat Rev Endocrinol*. 2013;**9**(12):699-712.
8. Väänänen HK, Härkönen PL. Estrogen and bone metabolism. *Maturitas*. 1996;**23**(Suppl):S65-S69.
9. Manson JE, Bassuk SS. Menopause and Postmenopausal Hormone Therapy. In: Jameson JL, Fauci AS, Kasper DL, Hauser SL, Longo DL, Loscalzo J, eds. *Harrison's Principles of Internal Medicine*. 20th ed. New York: McGraw-Hill Education; 2018.
10. Riggs BL, Khosla S, Melton LJ 3rd. A unitary model for involutional osteoporosis: estrogen deficiency causes both type I and type II osteoporosis in postmenopausal women and contributes to bone loss in aging men. *J Bone Miner Res*. 1998;**13**(5):763-773.
11. Pott J, Bae YJ, Horn K, et al. Genetic association study of eight steroid hormones and implications for sexual dimorphism of coronary artery disease. *J Clin Endocrinol Metab*. 2019;**104**(11):5008-5023.
12. Chen Z, Tao S, Gao Y, et al. Genome-wide association study of sex hormones, gonadotropins and sex hormone-binding protein in Chinese men. *J Med Genet*. 2013;**50**(12):794-801.
13. Liu M, Ingle JN, Fridley BL, et al. TSPYL5 SNPs: association with plasma estradiol concentrations and aromatase expression. *Mol Endocrinol*. 2013;**27**(4):657-670.
14. Prescott J, Thompson DJ, Kraft P, et al. Genome-wide association study of circulating estradiol, testosterone, and sex hormone-binding globulin in postmenopausal women. *PloS One*. 2012;**7**(6):e37815.
15. Eriksson AL, Perry JRB, Coviello AD, et al. Genetic determinants of circulating estrogen levels and evidence of a causal effect of estradiol on bone density in men. *J Clin Endocrinol Metab*. 2018;**103**(3):991-1004.
16. Ruth KS, Day FR, Tyrrell J, et al; Endometrial Cancer Association Consortium. Using human genetics to understand the disease impacts of testosterone in men and women. *Nat Med*. 2020;**26**(2):252-258.
17. Nethander M, Vandenput L, Eriksson AL, Windahl S, Funck-Brentano T, Ohlsson C. Evidence of a causal effect of estradiol on fracture risk in men. *J Clin Endocrinol Metab*. 2019;**104**(2):433-442.
18. Huang J, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the wellcome trust case control consortium phase 1 data. *Eur J Hum Genet*. 2012;**20**(7):801-805.
19. Walter K, Min JL, Huang J, et al. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;**526**(7571):82-89.
20. Nakamoto J, Salameh WA, Carlton E. Endocrine Testing. In: Jameson JL, De Groot LJBT, eds. *Endocrinology*. 6th ed. Philadelphia: Elsevier; 2010:2802-2834.
21. Xu C, Tachmazidou I, Walter K, Ciampi A, Zeggini E, Greenwood CM; UK10K Consortium. Estimating genome-wide significance for whole-genome sequencing studies. *Genet Epidemiol*. 2014;**38**(4):281-290.
22. Tobin J. Estimation of relationships for limited dependent variables. *Econometrica*. 1958;**26**(1):24-36.
23. Glymour MM, Tchetgen Tchetgen EJ, Robins JM. Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions. *Am J Epidemiol*. 2012;**175**(4):332-339.
24. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;**40**(Database issue):D930-D934.
25. Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) project. *Biopreserv Biobank*. 2015;**13**(5):307-308.
26. Zhu Z, Zheng Z, Zhang F, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun*. 2018;**9**(1):224.
27. Hemani G, Zheng J, Elsworth B, et al. The MR-base platform supports systematic causal inference across the human phenome. *Elife*. 2018;**7**. doi:10.7554/eLife.34408
28. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol*. 2017;**32**(5):377-389.
29. Schmitz D, Ek WE, Berggren E, Höglund J, Karlsson T, Johansson Å. Supplementary data for: Genome-wide association study of estradiol levels, and the causal effect of estradiol on bone mineral density. *Zenodo*. https://doi.org/10.5281/zenodo.4706671, March 3, 2021, preprint: not peer reviewed.
30. Estrada K, Styrkarsdottir U, Evangelou E, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nat Genet*. 2012;**44**(5):491-501.
31. Makridakis NM, di Salle E, Reichardt JK. Biochemical and pharmacogenetic dissection of human steroid 5 alpha-reductase type II. *Pharmacogenetics*. 2000;**10**(5):407-413.
32. Ogasawara K, Bannai M, Saitou N, et al. Extensive polymorphism of ABO blood group gene: three major lineages of the alleles for the common ABO phenotypes. *Hum Genet*. 1996;**97**(6):777-783.
33. Mackenzie PI, Rogers A, Treloar J, Jorgensen BR, Miners JO, Meech R. Identification of UDP glycosyltransferase 3A1 as a UDP N-acetylglucosaminyltransferase. *J Biol Chem*. 2008;**283**(52):36205-36210.
34. Riches Z, Stanley EL, Bloomer JC, Coughtrie MW. Quantitative evaluation of the expression and activity of five major sulfotransferases (SULTs) in human tissues: the SULT "pie." *Drug Metab Dispos*. 2009;**37**(11):2255-2261.
35. Sivils JC, Storer CL, Galigniana MD, Cox MB. Regulation of steroid hormone receptor function by the 52-kDa FK506-binding protein (FKBP52). *Curr Opin Pharmacol*. 2011;**11**(4):314-319.
36. Yu X, Baer R. Nuclear localization and cell cycle-specific expression of CtIP, a protein that associates with the BRCA1 tumor suppressor. *J Biol Chem*. 2000;**275**(24):18541-18549.
37. Wong AK, Ormonde PA, Pero R, et al. Characterization of a carboxy-terminal BRCA1 interacting protein. *Oncogene*. 1998;**17**(18):2279-2285.

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

38. Wu M, Soler DR, Abba MC, et al. CtIP silencing as a novel mechanism of tamoxifen resistance in breast cancer. *Mol Cancer Res*. 2007;**5**(12):1285-1295.
39. Yang S, Xu J, Zeng X. A six-long non-coding RNA signature predicts prognosis in melanoma patients. *Int J Oncol*. 2018;**52**(4):1178-1188.
40. Lee KY, Im JS, Shibata E, et al. MCM8-9 complex promotes resection of double-strand break ends by MRE11-RAD50-NBS1 complex. *Nat Commun*. 2015;**6**:7744.
41. Hagenaars SP, Hill WD, Harris SE, et al. Genetic prediction of male pattern baldness. *PloS Genet*. 2017;**13**(2):e1006594.
42. Richardson H, Ho V, Pasquet R, et al; MAP.3 Investigators. Baseline estrogen levels in postmenopausal women participating in the MAP.3 breast cancer chemoprevention trial. *Menopause*. 2020;**27**(6):693-700.
43. Lee AJ, Conney AH, Zhu BT. Human cytochrome P450 3A7 has a distinct high catalytic activity for the 16alpha-hydroxylation of estrone but not 17beta-estradiol. *Cancer Res*. 2003;**63**(19):6532-6536.
44. Ohmori S, Nakasa H, Asanome K, et al. Differential catalytic properties in metabolism of endogenous and exogenous substrates among CYP3A enzymes expressed in COS-7 cells. *Biochim Biophys Acta*. 1998;**1380**(3):297-304.
45. Simpson ER, Davis SR. Minireview: aromatase and the regulation of estrogen biosynthesis—some new perspectives. *Endocrinology*. 2001;**142**(11):4589-4594.
46. Sohl CD, Guengerich FP. Kinetic analysis of the three-step steroid aromatase reaction of human cytochrome P450 19A1. *J Biol Chem*. 2010;**285**(23):17734-17743.
47. Guimarães DA, dos Santos MS, Gomes KB, et al. Interaction between oral estrogen plus progestogen therapy and ABO blood groups on coagulation activation in postmenopausal women. *Menopause*. 2012;**19**(3):339-345.
48. Hammond GL. Access of reproductive steroids to target tissues. *Obstet Gynecol Clin North Am*. 2002;**29**(3):411-423.
49. Coviello AD, Haring R, Wellons M, et al. A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple Loci implicated in sex steroid hormone regulation. *Plos Genet*. 2012;**8**(7):e1002805.
50. Somboonporn W, Davis SR; National Health and Medical Research Council. Testosterone effects on the breast: implications for testosterone therapy for women. *Endocr Rev*. 2004;**25**(3):374-388.
51. Dunn JF, Nisula BC, Rodbard D. Transport of steroid hormones: binding of 21 endogenous steroids to both testosterone-binding globulin and corticosteroid-binding globulin in human plasma. *J Clin Endocrinol Metab*. 1981;**53**(1):58-68.
52. Davison SL, Bell R. Androgen physiology. *Semin Reprod Med*. 2006;**24**(2):71-77.
53. Mooradian AD, Morley JE, Korenman SG. Biological actions of androgens. *Endocr Rev*. 1987;**8**(1):1-28.
54. Jiang K, Yang Z, Cheng L, et al. Mediator of ERBB2-driven cell motility (MEMO) promotes extranuclear estrogen receptor signaling involving the growth factor receptors IGF1R and ERBB2. *J Biol Chem*. 2013;**288**(34):24590-24599.
55. Hou Q, Gorski J. Estrogen receptor and progesterone receptor genes are expressed differentially in mouse embryos during preimplantation development. *Proc Natl Acad Sci U S A*. 1993;**90**(20):9460-9464.
56. Chen H, Yong W, Hinds TD Jr, et al. Fkbp52 regulates androgen receptor transactivation activity and male urethra morphogenesis. *J Biol Chem*. 2010;**285**(36):27776-27784.
57. Dong M, Owens IS, Sheen YY. Cloning and expression of human liver UDP-glucuronosyltransferase cDNA, UDPGTh2. *Arch Pharm Res*. 1997;**20**(5):459-464.
58. Kallionpää RA, Järvinen E, Finel M. Glucuronidation of estrone and 16α-hydroxyestrone by human UGT enzymes: the key roles of UGT1A10 and UGT2B7. *J Steroid Biochem Mol Biol*. 2015;**154**:104-111.
59. Zhang H, Mooney CJ, Reilly MP. ABO blood groups and cardiovascular diseases. *Int J Vasc Med*. 2012;**2012**:641917.
60. Schachter H, Michaels MA. A quantitative difference in the activity of blood group A-specific N-acetylgalactosaminyltran sferase in serum from A 1 and A 2 human subjects. *Biochem Biophys Res Commun*. 1971;**45**(4):1011-1018.
61. Silva-Filho JC, Melo CGF, Oliveira JL. The influence of ABO blood groups on COVID-19 susceptibility and severity: a molecular hypothesis based on carbohydrate-carbohydrate interactions. *Med Hypotheses*. 2020;**144**:110155.
62. Suba Z. Prevention and therapy of COVID-19 via exogenous estrogen treatment for both male and female patients. *J Pharm Pharm Sci*. 2020;**23**(1):75-85.
63. Li Y, Jerkic M, Slutsky AS, Zhang H. Molecular mechanisms of sex bias differences in COVID-19 mortality. *Crit Care*. 2020;**24**(1):405.
64. AlAsiri S, Basit S, Wood-Trageser MA, et al. Exome sequencing reveals MCM8 mutation underlies ovarian failure and chromosomal instability. *J Clin Invest*. 2015;**125**(1):258-262.
65. Dou X, Guo T, Li G, Zhou L, Qin Y, Chen ZJ. Minichromosome maintenance complex component 8 mutations cause primary ovarian insufficiency. *Fertil Steril*. 2016;**106**(6):1485-1489.e2.
66. Lufkin EG, Wahner HW, O'Fallon WM, et al. Treatment of postmenopausal osteoporosis with transdermal estrogen. *Ann Intern Med*. 1992;**117**(1):1-9.
67. Abdi F, Mobedi H, Bayat F, Mosaffa N, Dolatian M, Ramezani Tehrani F. The effects of transdermal estrogen delivery on bone mineral density in postmenopausal women: a meta-analysis. *Iran J Pharm Res*. 2017;**16**(1):380-389.
68. Bowden J. Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. *Int J Epidemiol*. 2017;**46**(6):2097-2099.
69. Bowden J, Burgess S, Davey Smith G. Response to Hartwig and Davies. *Int J Epidemiol*. 2016;**45**(5):1679-1680.

Downloaded from https://academic.oup.com/jcem/advance-article/doi/10.1210/clinem/dgab507/6320117 by guest on 24 August 2021

High oestradiol levels increase the risk of breast and endometrial cancer

Åsa Johansson[1], Daniel Schmitz[1], Julia Höglund[1], Fatemeh Hadizadeh[1], Torgny Karlsson[1], Weronica E. Ek[1*]

[1] Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University, Box 815, 75108, Uppsala. Sweden

***Corresponding author:** Weronica E. Ek: Weronica.ek@igp.uu.se, BMC, box815, 751 08 Uppsala, +46703519004

Running title: Oestradiol is causal for cancer

Keywords: Mendelian Randomization, Oestradiol, Breast cancer, Endometrial cancer, Ovarian cancer

Word count: 4396

Number of figures and tables: 4

Funding

This work was funded by the Swedish Research Council 2019-01497 (ÅJ), The Swedish Cancer Society 19 0383 Pj (ÅJ), the Marcus Borgström Foundation (WEE), K and O F Hedströms Foundation (WEE), the Åke Wiberg Foundation M19-0349/M20-0057 (WEE), A and M Rudbergs Foundation (WEE), and Vleugels Foundation (WEE).

The funding sources had no influence and took no part in the design or conduct of this research.

The authors declare no potential conflicts of interest

Abstract

Background: High levels of oestrogen are associated with increased risk of breast and endometrial cancer and has been suggested to also play a role in the development of ovarian cancer. Cancerogenic effects of oestradiol, the most prominent form of oestrogen, has been highlighted as a side effect of oestrogen only menopausal hormone therapy. However, whether high levels of endogenous oestrogens, produced within the body, promote cancer development, has not been fully established.

Methods: Here we performed Mendelian Randomization (MR) analyses to estimate the effect of endogenous oestrogen on the risk of developing breast, endometrial and ovarian cancer.

Results: We showed that high levels of oestradiol increase the risk of breast (OR=1.19 [95% CI 1.03-1.38]) and endometrial cancer (OR=1.45 [95% CI 1.14-1.83]) but we could not establish a link to the risk of ovarian cancer (OR=1.20 [95% CI 0.99-1.46]). We thereby suggest that individual variation in oestrogen production is indeed a causal determinant of cancer risk, which agrees with previous research.

Conclusion: This is the first MR study to identify an effect of oestradiol on breast and endometrial cancer using multiple instruments, with effects estimated in women.

Impact: These results can be of importance for future precision medicine interventions to prevent cancer in women.

Introduction

Oestrogens are hormones that are important for the development and control of the primary and secondary sexual characteristics in both genders, and play a key role in the tuning of the reproductive system in women [1,2]. Oestrogen, along with progesterone, broadly contribute to the function and development of a large number of human tissues including endometrium, breast, bone, brain and cardiovascular system [3]. Notably, oestrogen promotes growth in normal endometrial and breast luminal epithelial tissues [3,4]. Several studies have demonstrated that long-lasting exposure or higher levels of oestrogen, as a result of oestrogen medication, early menarche, late menopause, nulliparity, and obesity, is a shared origin of different malignancies including breast, ovarian and endometrial cancer [5–7].

The predominant source of oestrogen production in the premenopausal women are ovaries, while in the postmenopausal women, subcutaneous adipose tissue is the major origin [5]. In premenopausal women, oestrogen is produced by the growing ovarian follicle to stimulate endometrial thickness, so as to prepare it for the implantation of the fertilized egg. At the end of this phase and following a peak in the secretion of oestrogen, ovulation occurs and the corpus luteum will be the source of oestrogen production during the luteal phase. If no fertilization takes place, the corpus luteum becomes inactive and at the end of this phase the oestrogen levels drop and the menstruation starts. Progesterone is exclusively secreted during the luteal phase to oppose the oestrogen-induced growth of the endometrium [4,8,9].

In breast tissue, oestrogen plays an important role in development, proliferation and differentiation [10,11]. Oestradiol, the most potent oestrogen, is the biologically active type of oestrogen in both pre- and post-menopausal women [5]. Previous studies have shown an association between higher risk of breast cancer among women that have high blood levels of oestradiol both before [12,13] and after menopause [14,15]. Still, if high natural oestradiol levels in the body have a causal effect on breast cancer, or whether the high oestradiol levels are just a consequence of cancer progression, has not been clearly described.

Endometrial cancer is the most common gynaecological cancer and is a hormonally-responsive tumour [4]. Endometrial cancer risk increases with use of menopausal hormonal treatment (MHT) that includes oestrogen only. However, this risk can be reduced if the treatment is combined with (opposed by) progesterone [16,17]. Instead, a protective effect on

both endometrial and ovarian cancers have been identified in oral contraceptive users [18,19]. Oral contraceptives contain oestradiol in combination with a synthetic form of progesterone (progestin), or progestin only. The protective effect of oral contraceptives has been suggested to be due to fewer ovulations [20], since the peak in oestrogen secretion that triggers tissue growth in association with the ovulation is depleted during oral contraceptive use. Ovarian cancer is the most fatal of the female gynaecological cancers worldwide with no screening test and late-stage diagnosis [17 21]. Epidemiological studies have suggested a strong role for the oestrogen activity as well as the duration of exposure to it in the initiation, pathogenesis, and progression of ovarian cancer [17].

In addition, driving from experimental data, hormone-associated cancers such as breast, endometrial and ovarian cancer share several pathways for oestrogen regulation [17]. Even though oestrogen has been linked to all three cancer types, there is a lack of knowledge about whether the body's own production of oestrogen promotes the development of breast, endometrial and ovarian cancer. Establishing such causal relations is important in order to determine the pathogenetic mechanisms and to prevent cancers. One difficulty when studying risk factors, such as oestrogen levels on cancer, is to distinguish correlation from causation. Mendelian randomization (MR) is an instrumental variable approach that can be used to disentangle the effect that oestradiol exerts on the risk of developing cancer. In MR, germ-line genetic variants are used as instrumental variables. Thereby, the MR estimate is not affected by reversed causation, since genetic variants are assorted during formation of gametes prior to conception and are not confounded by lifestyle or environmental factors. MR is thereby an ideal approach for estimating causal effects. In MR, the instrumental variables must fulfil three fundamental assumptions in order to be valid: 1) the variant must be associated with the exposure (oestradiol), 2) the variant should not be associated with any potential confounder in the exposure – outcome relation, 3), and the variant should not be directly associated with the outcome (cancer) [22]. Two previous MR studies, both including only one genetic variant, close to the *CYP19A1* gene, that was identified to be associated with oestradiol levels primally in males, identified a causal effect of higher oestradiol levels on endometrial cancer risk [23,24]. MR studies based on a single genetic variant may be greatly biased, as a result of undetectable pleiotropic effects [25]. One previous study used additional genetic variants that had been identified in males as instruments, but conduced, in agreement with the very different production of oestradiol in males and females, that these genetic variants were probably invalid as instruments in females[24]. To our knowledge, no previous

study has been conducted to identify a causal effect of oestradiol on ovarian, breast, or endometrial cancer using multiple instruments for oestradiol levels.

Here, we undertake an MR approach, including four genetic variants that were recently shown to be associated with high oestradiol blood levels in females [26], aiming to establish a causal effect of oestradiol on breast, endometrial and ovarian cancer. We perform a one-sample MR study using data from women in the UK Biobank cohort, as well as a two-sample MR approach using three different independent cancer cohorts, with no overlapping samples with UK Biobank.

Methods

Study samples

UK Biobank

The UK Biobank is a cross-sectional cohort, with both a pro and retrospective study design that includes 502,682 participants, of which 273,404 are women, recruited from all across the United Kingdom (UK). Participants were between 37 and 73 years old at the time of recruitment between 2006 and 2010. Health variables, including cancer diagnosis, have been collected through questionnaires, interviews and hospital records. Participants also answered questions about diet, and lifestyle factors. A total of 820,967 genotyped single nucleotide polymorphisms (SNPs) and up to 90 million imputed variants are available for most participants. For all genetic analyses in UKB the third release of the imputed genetic data was used. The UK Biobank study was approved by the National Research Ethics Committee (REC reference 11/NW/0382). Informed consent to the study was given by all participants. An application for using data from UK Biobank has been approved (application nr: 41143). The UK Biobank analysis performed in this study has also been approved by the Swedish Ethical Review Authority (dnr: 2020-04415).

Oestradiol levels and instrumental variables for oestradiol

Oestradiol levels were measured from blood samples taken at the first assessment in association with the recruitment. Oestradiol was measured by a two-step competitive analysis on a Beckman Coulter Unicel Dxl 800. The instrument variables to be used in the MR analyses were selected from a previous GWAS for oestradiol in UK Biobank [26]. Women with oestradiol levels below 175 pmol/L (detection level) were defined as controls, and

participants with oestrogen above, or equal to175 pmol/L, were defined as cases. Briefly, the GWAS included Caucasian UK Biobank participants clustering with regards to their genetic principal components. First- and second-degree relatives (genetic relationship > 0.044) and samples with sex discordance, high heterozygosity/missingness, and with more than 5% missing genotypes had been excluded. Also, females with unknown menopausal status were removed. After the above filtering, 163,985 females remained in the GWAS of which 37,461 had oestradiol levels above the detection limit (27,463 pre-menopausal and 9,998 post-menopausal). The GWAS was performed using an additive genetic logistic regression model implemented in PLINK 2.0 (2.00-alpha-2-20190429). In total, 4 instrumental variables were identified ($P<1x10^{-7}$) (Table 1). The variance explained by each genetic variant was estimated by calculating the difference between Nagelkerke's pseudo-R^2 for the full model, including both covariates and the SNP, and the reduced model, only including the covariates. The *F*-statistic for each SNP was estimated from the full model by computing the squared ratio of the SNP's beta estimate and its standard error.

Cancer incidences and covariates in UK Biobank

To assess information regarding cancer incidences, data were taken from different categories: main and secondary diagnoses made during a hospital stay, medical conditions assessed both from verbal interviews and touchscreen questionnaires, as well as cause of death and cancer registers. Data taken from the hospital stay, the cancer registry and the death registry were categorised according to the International Classification of Diseases, revision nine (ICD-9) and ten (ICD-10). The ICD10 code C50 (malignant neoplasm of breast) and the ICD9 code 174 (malignant neoplasm of female breast) were used to define breast cancer cases, the ICD10 code C56 (malignant neoplasm of ovary) and the ICD9 code 183 (malignant neoplasm of ovary and other uterine adnexa) were used to define ovarian cancer cases, and the ICD10 code C541 (endometrium) was used to identify endometrial cancer cases. For a detailed description of data-fields and coding used for each cancer and covariate, see Supplementary Table S1-S2. The largest number of cancer cases was identified from the cancer registry data, but since the registry data is sparse before 1995, several additional cancer cases were identified from self-reported data (Supplementary Table S1). General characteristics (Table 1) and information on covariates and potential confounders (Supplementary Table S2) were assessed from data collected during the initial visit to the assessment centre.

To estimate the effect of the oestradiol instrumental variables on the risk for breast, ovarian and endometrial cancer in UK Biobank, we performed a GWAS using a logistic additive genetic model implemented in PLINK 2.0 (2.00-alpha-2-20190429). The following covariates (same as in the oestradiol GWAS) were included: age, BMI, the first ten ancestry derived principal components (to adjust for population structures and ethnic origins), MHT (ever, never and current), oral contraceptive use (ever, never or current), number of live births, menopausal status (pre or post) and hysterectomy (yes or no). In addition, to adjust for the different genotyping chips used in UK Biobank, a binary indicator variable for UK Biobank Axiom versus UK BiLEVE genotyping array was included.

Causal estimation of the effect of oestradiol on breast, endometrial and ovarian cancer

To assess a possible causal effect of high oestradiol levels on the risk of being diagnosed with breast, endometrial or ovarian cancer, we first performed a one-sample MR in UK Biobank. Secondly, we performed a two-sample MR, where the estimates for the oestradiol measurement were extracted from the UK Biobank oestradiol GWAS and the cancer estimates from publicly available GWAS data.

The main MR analyses were performed with the R package gsmr [27] (version 1.0.8), in order to enable causal inference based on multiple SNPs. The HEIDI-outlier flag was disabled as there were too few SNPs (<5) for the procedure to work properly. Note that, all four SNPs are pairwise independent (all located on different chromosomes). We further performed sensitivity analyses using the robust inverse variance weighted (IVW), weighted median, and the MR-Egger methods, included in the "MendelianRandomization" tool implemented in R [28]. The standard IVW method provides a consistent estimate of the causal effect of the exposure on the outcome when all genetic variants are valid instruments [29]. Here, we applied the robust IVW method that is less susceptible to heteroscedasticity and down-weights the effect of possible outlier instruments using robust regression. The two other methods, weighted median and MR-Egger, have been proposed to provide consistent causal estimates for genetic variants under weaker assumptions. The weighted median method only assumes that a majority of the instruments are valid, has a lower bias than the IVW method and owns a reasonable power to detect a causal effect [30]. MR-Egger's intercept test was used to assess the presence of directional pleiotropy [31].

Publicly available GWAS data for two-sample MR analyses

A two-sample MR approach applied for each cancer was done to further support our results from the one-sample MR. In a one-sample MR, any weak-instrument~~s~~ bias is directed towards the confounded association, which could result in inflated MR estimates. In a two-sample MR, this bias is instead directed towards null. Furthermore, the type I error is not inflated in a two-sample setting. In our two-sample MR, the effects of the instruments on cancers were taken from the same oestradiol GWAS as the one-sample MR. However, for breast, endometrial and ovarian cancer respectively (Table 1), the effect sizes and standard errors for the instruments were extracted from publicly available GWAS data, not including UK Biobank participants. For breast cancer, we used summary statistics from the Breast Cancer Association Consortium (BCAC) [32], including 122,977 breast cancer cases and 105,974 healthy controls of European ancestry (downloaded from http://bcac.ccge.medschl.cam.ac.uk/ on 22nd of February 2021). For endometrial cancer, we used data from the Endometrial Cancer Association Consortium (ECAC), which include twelve research cohorts based in Australia, Europe and the USA. From the ECAC cohort, GWAS summary statistics from O'Mara et al (2018) [33], excluding participants from UK Biobank to avoid sample overlap, were used. This restricted ECAC dataset consisted of 12,270 cancer cases and 46,126 controls of European descent. Data from ECAC was available after request from the authors [33]. For ovarian cancer, we used GWAS summary statistics from the Ovarian Cancer Association Consortium (OCAC)[34]. For OCAC, genetic association analysis had been performed for 25,509 epithelial ovarian cancer cases and 40,941 healthy controls. Summary statistics were downloaded from http://ocac.ccge.medschl.cam.ac.uk/data-projects/results-lookup-by-region/, on 26th February 2021.

Data Availability

The data used for this study is available for bona fide researchers from the UK Biobank Resource (http://www.ukbiobank.ac.uk/about-biobank-uk/), and can be accessed by an application to the UK Biobank. The oestradiol data was taken from the supplementary material previously published at https://doi.org/10.5281/zenodo.4926701 [26]. The BCAC data can be downloaded from http://bcac.ccge.medschl.cam.ac.uk/ and the OCAC can be downloaded from http://ocac.ccge.medschl.cam.ac.uk/data-projects/results-lookup-by-region/. The ECAC data was approved after request to the authors. Summary statistics for the published MR study is found in Table 1.

Results

The GWAS for breast, ovarian and endometrial cancer in UK Biobank included 209,877 non-related Caucasian women with complete information for each covariate. In this subset, there were 13,179 breast cancer cases, 1,891 endometrial cancer cases, and 1,477 ovarian cancer cases that were included for each one-sample analysis. We have shown previously that the cancer rate in UK Biobank is similar to the general population in UK for all three cancers studied [18]. The mean age of the UK Biobank subset was 56 years, and most women (74%) had entered menopause. Characteristics for UK Biobank are presented in Table 2. Out of the four instrumental variables selected for oestradiol, two were nominally associated (P<0.05) with breast cancer and one with endometrial cancer in UK Biobank (Table 1). In the BCAC cohort, the SNP rs10638101 was not genotyped, and the proxy rs897797 was therefore included in the two-sample MR for breast cancer. rs897797 was in perfect LD with rs10638101 (R^2=1.0). In addition, one more instrumental variable was associated (P<0.05) with endometrial cancer and one with ovarian cancer (Table 1) in the two-sample cohorts.

Mendelian randomization (MR)

Using our primary MR-method, generalized summary-data-based Mendelian-randomization (GSMR), a significant effect of high oestradiol levels on breast cancer was identified, using both a one-sample approach in UK Biobank (OR=1.30 [95% CI 1.07-1.57]) and a two-sample approach with GWAS summary statistics from BCAC (OR=1.19 [95% CI 1.03-1.38]) (Figure 1, Table 3). Also, for endometrial cancer, we identified a causal effect of oestradiol both in the one-sample MR in UK Biobank (OR=2.01 [95% CI 1.21-3.31]) and in the two-sample approach with cancer estimates from ECAC (OR=1.45 [95% CI 1.14-1.83]) (Figure 1, Table 3). For ovarian cancer, we did not identify a significant effect of oestradiol in the one-sample approach (OR=1.55 [95% CI 0.91-2.65]). Neither did we identify a significant effect in the two-sample approach using GWAS summary statistics from OCAC (OR=1.20 [95% CI 0.99-1.46]) (Figure 1, Table 3).

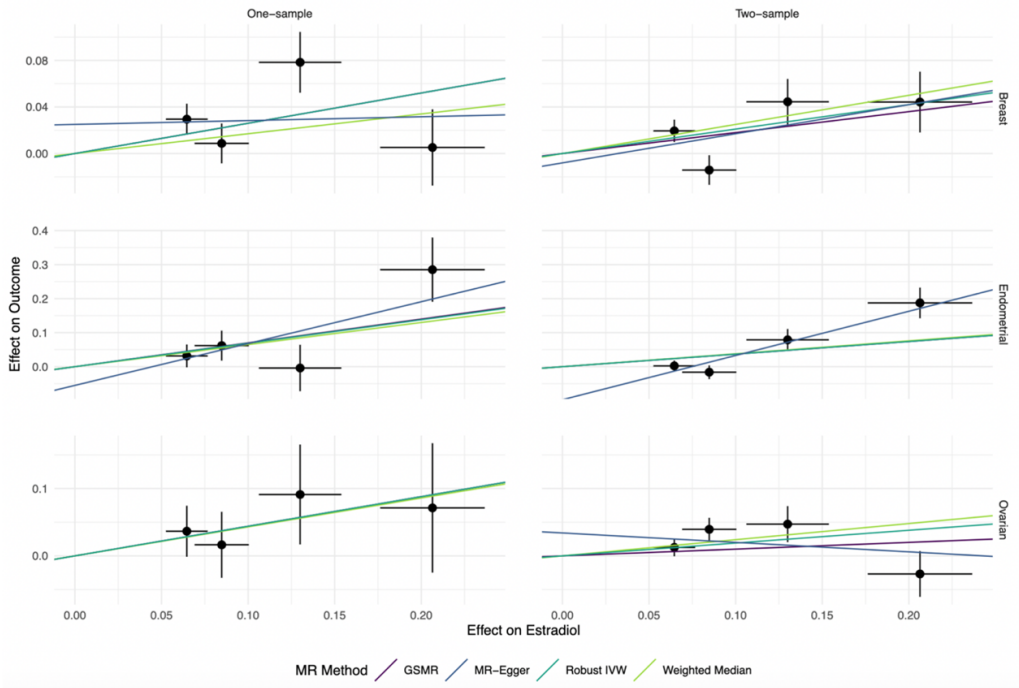


Figure 1. Results from the four different Mendelian randomization methods applied (GSMR, MR-Egger, Robust IVW and Weighted Median) to estimate the causal effect of oestradiol on breast, endometrial and ovarian cancer. The black dots represent the effect size of the SNPs in the GWAS for oestradiol (x-axis) and cancer (y-axis) and the black lines are the standard errors. The lines represent the estimates from the different MR methods.

Sensitivity analyses

As sensitivity analyses, we applied weighted median, robust inverse weighted median and the MR-Egger approach for each cancer, in both the one- and two-sample analyses. Comparing each method, we could see that all three cancers showed the same direction of effect with all methods (Figure 1, Table 3). The MR-Egger intercept for endometrial cancer was significant in the two-sample analysis, which indicates that we might have directional pleiotropy. However, the MR-Egger approach controls for this type of pleiotropy given that the instrument strength independent of direct effect assumption is not violated [31]. Since the MR-Egger beta estimate was still significant with the same direction of effect as in all other methods, we therefore argue that the results from MR-Egger are consistent with the other methods.

Discussion

In this study, we have used a one- and a two-sample MR approach to show that high levels of oestradiol in the body, promotes the development of breast and endometrial cancer.

In a previous MR study, including 6,608 endometrial cancer cases and 37,925 postmenopausal controls, using one single instrument (rs727479) mapped to the *CYP19A1* gene, it was suggested that oestradiol has a causal effect on endometrial cancer risk, but no significant effect was identified for breast cancer [23]. Beside that the previous study was smaller than ours, and might not have had enough power to identify a causal effect for breast cancer, their selected instrument might have been less appropriate to use since it appears to be more strongly associated with oestradiol in males than in females[26], which would bias the causal effect towards null. In addition, any bias due to horizontal pleiotropy may go undetected, by using a single instrument. Another, recent study by Larsson *et al*[24], identified an effect of oestradiol on both ER-positive breast cancer and endometrial cancer, using the same genetic variant, also with the effect on oestradiol being estimated in males. Even though rs727479 was not available in UK Biobank, it is in perfect LD with rs7175531 (R^2=1.0), which we previously showed to be strongly associated with oestradiol in male UK Biobank participants (P=1.32x10^{-60}) [26]. However, in UK Biobank, rs7175531 was only nominally significant, in pre- and post-menopausal women (P=0.020 and P=0.043, respectively). rs7175531 is intronic in *CYP19A1*, which encodes an aromatase that synthesizes endogenous oestrogens from testosterone in adipose tissue [35]. This suggests that this locus might be more important for oestradiol production in males, and not play an as important role in the oestrogen production in females. The *CYP19A1* SNP could thereby be an invalid MR instrument for estimating the causal effect of oestradiol on female-specific phenotypes.

In this MR study, we instead used four oestradiol instruments (SNPs), that have been associated with oestradiol levels in females [26]. These SNPs were annotated to *CYP3A7, ASCL1, TMEM150B* and *MCM8*. The most significant association to high oestradiol levels was identified for the *CYP3A7* SNP (P=7.62x10^{-12}). *CYP3A7* encodes cytochrome P450 CYP3A7, which metabolizes dehydroepiandrosterone (DHEA), which is the main precursor of circulating oestrogens in women [36,37]. The instrument for *CYP3A7* (rs45446698) was nominally significant (P<0.05) for endometrial cancer in both UK Biobank and ECAC, and for breast cancer in BCAC (Table 1). CYP3A7 is mainly expressed in the liver, which is one

of the primary sites of oestrogen metabolism [38]. CYP3A7 is also expressed in the endometrium [39]. The second most significant oestradiol instrument was the *MCM8* SNP, rs16991615 that has previously been shown to be one of the strongest associated loci also for age at menopause [40]. This instrument was nominally significant (P<0.05) for breast cancer both in UK Biobank (P=0.0027) and in BCAC (P=0.024), as well as for endometrial cancer in ECAC (P=0.013) (Table 1). Even though the previous oestradiol GWAS did not adjust for age at menopause, a significant difference in the estimate was not identified when including postmenopausal women only and adjusting for age at menopause [26]. This suggests that this effect is not confounded by age at menopause. An oestradiol GWAS excluding all participants with a cancer diagnosis prior to assessment (when blood samples were drawn), was also performed as a sensitivity analysis. This analysis did not show any significant difference in effect for the instruments, indicating that a prior cancer diagnosis did not influence the results [26]. *ASCL1,* the gene mapping to our third most significant oestradiol instrument, has previously been shown to promote tumour progression in lung adenocarcinoma [41] and overall survival in ovarian cancer patients [42]. The instrument mapped to *ASCL1* was nominally significant for ovarian cancer within the OCAC cohort (P=0.021, Table 1). *TMEM150B* has, except for being associated with age at menopause [43], also been associated with age at menarche [44]. The instrument mapping to *TMEM150B* (rs10638101) was nominally significant for breast cancer in UK Biobank (P=0.026), and its proxy (rs897797) was also nominally significant for breast cancer in BCAC (P=0.043; Table 1).

Even though we used a larger number of instrumental variables in our MR analyses than previous oestradiol MR studies on cancer risk, one of the main limitations of our study is still that the number of instruments is limited. In order for the HEIDI outlier procedure in the GMSR method to work the best, at least ten instrumental variables are needed. This means that we were not able to adjust for pleiotropy by removing instruments that are identified as pleiotropic. However, today, no other studies, with larger sample sizes than UK Biobank exist, and the identification of additional instruments is therefore not possible at this time. For the two-sample MR analyses, the MR-Egger intercept was significantly below zero for endometrial cancer, which could indicate that we have problems with pleiotropy and the results from the current study would benefit from replication, once additional instrumental variables for oestradiol are available. MR-Egger is not appropriate to run for a one-sample approach, and we, therefore, should interpret the MR-Egger results for the one-sample analysis with care [45]. Another limitation of this study is the hormonal fluctuations during the

menstrual cycle and that oestradiol is commonly measured at different time points during the menstrual cycle in different women, but also that the oestradiol measured in blood is the result of the oestrogen produced in the ovaries (in pre-menopausal women), together with the oestradiol produced in other tissues. More detailed measurements of oestradiol during different time points of the menstrual cycle would be beneficial to address the causal effects of oestradiol pre-menopause. Since oestradiol mainly is produced by the ovaries during the reproductive years, and mainly by subcutaneous adipose tissues after menopause [5], another limitation is therefore that the oestradiol GWAS and MR includes both pre- and post-menopausal women.

With regards to very few oestradiol instruments associated with ovarian cancer, it is not surprising that we did not find a causal relationship between high oestradiol levels and risk of ovarian cancer. However, the MR estimates from our one- and two-sample MR are both in the same direction, and at least the two-sample MR would have met the significance thresholds of a one-sided test, testing only if high oestradiol levels increase the risk of ovarian cancer rather than it influencing the risk in any direction (increase or decrease). Therefore, we cannot rule out the possibility of a true causal effect that might be identified in a larger set of ovarian cancer cases. However, the associated effect of oestrogen on the risk of ovarian cancer is not as strong [46] as that previously seen for breast [47], and especially endometrial cancer [4], which agrees with our results. On the other hand, ovarian cancer is a heterogeneous disease with various subtypes which are different concerning their etiology, morphology, biology, pathogenicity and prognosis [48], therefore, future studies should benefit from include larger cohorts and focus on more distinct subtypes.

Although oestradiol is involved in regulating different features of human physiology in both genders, the underlying mechanisms through which oestrogen exerts its carcinogenic effects are not yet fully established and different hypotheses have been postulated so far. One hypothesis is that oestradiol has toxic effects and promotes mutations in the DNA [49] or that the oestrogen metabolites can cause genomic instability and mutations that can result in malignancy [50,51]. However, it has also been proposed that oestradiol can restrain the DNA damage response, which is a cellular mechanism to repair DNA damage, and thereby, may play a further role in the development and progression of the neoplasms [49]. Another hypothesis is that the cancerogenic effects of oestradiol mainly take place through binding to one of its two receptors (i.e., oestrogen receptor (ER) α and β). While the oestradiol-ERβ complex inhibits

cell proliferation, the binding of the hormone to ERα in the oestradiol-ERα complex induces cell proliferation, which has been suggested as a mechanism of the neoplastic effect of oestradiol [2,4]. In agreement with this, it has been shown that oestrogen signalling via ERα induces hyperproliferation and carcinogenic transformation in breast, ovarian and endometrial cells [5,52].

By identifying and replicating a possible causal link between oestradiol and endometrial, as well as for breast cancer, using a more valid approach compared to previous research, our results further support carcinogenic effects of oestrogen in these tissues. A deeper understanding of such possibly causal relation is also of importance for novel interventions to prevent cancer in women.

Author contributions

WEE and ÅJ designed the study; WEE performed the MR analysis and cancer GWASes, DS generated the figures and performed the oestradiol GWAS; WEE and TK performed the statistical analysis; WEE, TK, JH, DS, ÅJ, FH wrote the manuscript; DS, JH, TK, WEE, FH and ÅJ interpreted the data, contributed to and reviewed the manuscript. All authors declare no conflicts of interest.

Acknowledgements

We acknowledge participants and staff at UK Biobank. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Centre for Advanced Computational Science (UPPMAX) under project sens2017538. The breast cancer genome-wide association analyses were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie, de la Science et de l'Innovation du Québec' through Genome Québec and grant PSR-SIIRI-701, The National Institutes of Health (U19 CA148065, X01HG007492), Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710) and The European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935). All studies and funders are listed in Michailidou et al (Nature, 2017). The endometrial cancer genome-wide association analyses were supported by the National Health and Medical Research Council of Australia (APP552402, APP1031333, APP1109286, APP1111246 and APP1061779, APP1173170),

the U.S. National Institutes of Health (R01-CA134958), European Research Council (EU FP7 Grant), Wellcome Trust Centre for Human Genetics (090532/Z/09Z) and Cancer Research UK. OncoArray genotyping of ECAC cases was performed with the generous assistance of the Ovarian Cancer Association Consortium (OCAC), which was funded through grants from the U.S. National Institutes of Health (CA1X01HG007491-01 (C.I. Amos), U19-CA148112 (T.A. Sellers), R01-CA149429 (C.M. Phelan) and R01-CA058598 (M.T. Goodman); Canadian Institutes of Health Research (MOP-86727 (L.E. Kelemen)) and the Ovarian Cancer Research Fund (A. Berchuck). We particularly thank the efforts of Cathy Phelan. OncoArray genotyping of the BCAC controls was funded by Genome Canada Grant GPH-129344, NIH Grant U19 CA148065, and Cancer UK Grant C1287/A16563. All studies and funders included in ECAC are listed in O'Mara et al (2018). We also like to acknowledge Tracy O'Mara at QIMR Berghofer for sharing the ECAC data, after removing the UK Biobank.

References

1. Wise, P. M., Suzuki, S. & Brown, C. M. Estradiol: A hormone with diverse and contradictory neuroprotective actions. *Dialogues in Clinical Neuroscience* **11**, 297–303 (2009).
2. Ascenzi, P., Bocedi, A. & Marino, M. Structure-function relationship of estrogen receptor α and β: Impact on human health. *Struct. Relatsh. estrogen Recept. α β Impact Hum. Heal.* 299–402 (2006).
3. Hilton, H. N., Clarke, C. L. & Graham, J. D. Estrogen and progesterone signalling in the normal breast and its implications for cancer development. *Molecular and Cellular Endocrinology* **466**, 2–14 (2018).
4. Rodriguez, A. C., Blanchard, Z., Maurer, K. A. & Gertz, J. Estrogen Signaling in Endometrial Cancer: a Key Oncogenic Pathway with Several Open Questions. *Horm. cancer* **10**, 51–63 (2019).
5. Zhao, H., Zhou, L., Shangguan, A. & Bulun, S. Aromatase expression and regulation in breast and endometrial cancer. *J. Mol. Endocrinol.* **57**, 19–33 (2016).
6. Van Weelden, W., Massuger, L. & Pijnenborg, J. Anti-estrogen treatment in endometrial cancer: A systematic review. *Front. Oncol.* **9**, 359 (2018).
7. Salehi, F., Dunfield, L., Phillips, K. P., Krewski, D. & Vanderhyden, B. C. Risk factors for ovarian cancer: An overview with emphasis on hormonal factors. *Journal of Toxicology and Environmental Health - Part B: Critical Reviews* **11**, 301–321 (2008).
8. Groothuis, P., Dassen, H., Romano, A. & Punyadeera, C. Estrogen and the endometrium: Lessons learned from gene expression profiling in rodents and human. *Hum. Reprod. Updat. Oxford Acad.* **13**, 405–417 (2007).
9. Reed, B. & Carr, B. The Normal Menstrual Cycle and the Control of Ovulation. *Endotext. MDText.com, Inc* (2000).
10. Anderson, E., Clarke, R. B. & Howell, A. *Estrogen Responsiveness and Control of Normal Human Breast Proliferation*. *Journal of Mammary Gland Biology and Neoplasia* **3**, 23–35 (Kluwer Academic/Plenum Publishers, 1998).
11. Garvin, S., Nilsson, U. W., Huss, F. R. M., Kratz, G. & Dabrosin, C. Estradiol increases VEGF in human breast studied by whole-tissue culture. *Cell Tissue Res.* **325**, 245–251 (2006).
12. Group, E. H. and B. C. C. *et al.* Sex hormones and risk of breast cancer in premenopausal women: A collaborative reanalysis of individual participant data from

seven prospective studies. *Lancet Oncol.* **14**, 1009–1019 (2013).

13. Kaaks, R. *et al.* Serum sex steroids in premenopausal women and breast cancer risk within the European Prospective Investigation into Cancer and Nutrition (EPIC). *J. Natl. Cancer Inst.* **97**, 755–765 (2005).
14. Zhang, X., Tworoger, S. S., Eliassen, A. H. & Hankinson, S. E. Postmenopausal plasma sex hormone levels and breast cancer risk over 20 years of follow-up. *Breast Cancer Res. Treat.* **137**, 883–892 (2013).
15. Kaaks, R. *et al.* Postmenopausal serum androgens, oestrogens and breast cancer risk: The European prospective investigation into cancer and nutrition. *Endocr. Relat. Cancer* **12**, 1071–1082 (2005).
16. Brinton, L. A. & Felix, A. S. Menopausal hormone therapy and risk of endometrial cancer. *J. Steroid Biochem. Mol. Biol.* **142**, 83–89 (2014).
17. Mungenast, F. & Thalhammer, T. Estrogen biosynthesis and action in ovarian cancer. *Front. Endocrinol. (Lausanne).* **5**, (2014).
18. Karlsson, T., Johansson, T., Hoglund, J., Ek, W. E. & Johansson, Å. Time-dependent effects of oral contraceptive use on breast, ovarian, and endometrial cancers. *Cancer Res.* **81**, 1153–1162 (2021).
19. Iversen, L., Sivasubramaniam, S., Lee, A. J., Fielding, S. & Hannaford, P. C. Lifetime cancer risk and combined oral contraceptives: the Royal College of General Practitioners' Oral Contraception Study. *Am. J. Obstet. Gynecol.* **216**, 580.e1-580.e9 (2017).
20. Berchuck, A. & Scildkraut, J. Oral contraceptive pills. Prevention of ovarian cancer and other benefits. *N C Med J.* **58**, 405–407 (1997).
21. Stewart, C., Raylea, C. & Lockwood, S. Ovarian Cancer: An Integrated Review. Seminars in Oncology Nursing. W.B. *Saunders* **35**, 151–156 (2019).
22. Didelez, V. & Sheehan, N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat. Methods Med. Res.* **16**, 309–330 (2007).
23. Thompson, D. J. *et al.* CYP19A1 fine-mapping and Mendelian randomization: estradiol is causal for endometrial cancer. *Endocr. Relat. Cancer* **23**, 77–91 (2016).
24. Larsson, S. C. *et al.* Serum Estradiol and 20 Site-Specific Cancers in Women: Mendelian Randomization Study. *J. Clin. Endocrinol. Metab.* (2021).
25. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
26. Schmitz, D. *et al.* Genome-Wide Association Study of Estradiol Levels, and the Causal

Effect of Estradiol on Bone Mineral Density. *JCEM* (2021). doi:10.1101/2021.07.01.21259826

27. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat. Commun.* **9**, (2018).
28. Olena, Y. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).
29. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* **37**, 658–665 (2013).
30. Bowden, J., Smith, D. G. & Burgess, S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Int. J. Epidemiol.* **44**, 512–525 (2015).
31. 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).
32. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
33. O'Mara, T. A. *et al.* Identification of nine new susceptibility loci for endometrial cancer. *Nat. Commun.* **9**, (2018).
34. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).
35. Flote, V. G. *et al.* Gene variations in oestrogen pathways, CYP19A1, daily 17β-estradiol and mammographic density phenotypes in premenopausal women. *Breast Cancer Res.* **16**, (2014).
36. Lee, A. J., Conney, A. H. & Zhu, B. T. Human Cytochrome P450 3A7 Has a Distinct High Catalytic Activity for the 16α-Hydroxylation of Estrone but not 17β-Estradiol. *Cancer Res.* **63**, 6532 LP – 6536 (2003).
37. Ohmori, S. *et al.* Differential catalytic properties in metabolism of endogenous and exogenous substrates among CYP3A enzymes expressed in COS-7 cells. *Biochim. Biophys. Acta - Gen. Subj.* **1380**, 297–304 (1998).
38. Tsuchiya, Y., Nakajima, M. & Yokoi, T. Cytochrome P450-mediated metabolism of estrogens and its regulation in human. *Cancer Lett.* **227**, 115–124 (2004).
39. Schuetz, J. D., Kauma, S. & Guzelian, P. S. Identification of the fetal liver cytochrome

CYP3A7 in human endometrium and placenta. *J. Clin. Invest.* **92**, 1018–1024 (1993).

40. Chen, C. T. L. *et al.* Meta-analysis of loci associated with age at natural menopause in African-American women. *Hum. Mol. Genet.* **23**, 3327–3342 (2014).
41. Miyashita, N. *et al.* ASCL1 promotes tumor progression through cell-autonomous signaling and immune modulation in a subset of lung adenocarcinoma. *Cancer Lett.* **489**, 121–132 (2020).
42. Moore, K. N. *et al.* Genome-wide association study evaluating single-nucleotide polymorphisms and outcomes in patients with advanced stage serous ovarian or primary peritoneal cancer: An NRG Oncology/Gynecologic Oncology Group study. *Gynecol. Oncol.* **147**, 396–401 (2017).
43. 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).
44. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).
45. Bowden, J. Misconceptions on the use of MR-Egger regression and the evaluation of the InSIDE assumption. *Int. J. Epidemiol.* **46**, 2097–2099 (2017).
46. Trabert, B. *et al.* Circulating estrogens and postmenopausal ovarian cancer risk in the women's health initiative observational study. *Cancer Epidemiol. Biomarkers Prev.* **4**, 648–656 (2016).
47. Key, T., Appleby, P., Hines, L. & Al., E. Circulating sex hormones and breast cancer risk factors in postmenopausal women: reanalysis of 13 studies. *Br J Cancer* **105**, 709–722 (2011).
48. Kossai, M., Leary, A., Scoazec, J.-Y. & Genestine, C. Ovarian Cancer: A Heterogeneous Disease. *Karger Publ.* **85**, 41–49 (2018).
49. Pescatori, S. *et al.* A Tale of Ice and Fire: The Dual Role for 17β-Estradiol in Balancing DNA Damage and Genome Integrity. *Cancers (Basel).* **13**, (2021).
50. Cavalieri, E., Rogan, E. & Zahid, M. Critical depurinating DNA adducts: Estrogen adducts in the etiology and prevention of cancer and dopamine adducts in the etiology and prevention of Parkinson's disease. *Int. J. Cancer* **141**, (2017).
51. Wen, C., Wu, L., Fu, L., Wang, B. & Zhou, H. Unifying mechanism in the initiation of breast cancer by metabolism of estrogen (Review). *Molecular Medicine Reports* **16**, 1001–1006 (2017).
52. Li, H. H. *et al.* Estradiol 17β and its metabolites stimulate cell proliferation and antagonize ascorbic acid-suppressed cell proliferation in human ovarian cancer cells.

Reprod. Sci. **21**, 102–111 (2014).

Tables

Table 1. Summary GWAS results for each instrument variable included in the MR analysis. BCAC; Breast Cancer Association Consortium, OCAC; Ovarian Cancer Association Consortium, ECAC; Endometrial Cancer Association Consortium. Significant values (P<0.05) are highlighted in bold.

						Breast Cancer				**Ovarian Cancer**				**Endometrial Cancer**			
SNP	**Effect allele/ freq***	**Chromosome position: base pair**	**Delta R2*** / F-statistics**	**Oestradiol Levels OR (95% CI)**	**P-value**	**UK Biobank**		**BCAC**		**UK Biobank**		**OCAC**		**UK Biobank**		**ECAC**	
						OR (95% CI)	**P-value**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**
rs45446698	T/0.96	Chr7: 99332948	0.00026 / 45.16	1.23 (1.16-1.30)	$7.62x10^{-12}$	1.01 (0.94-1.07)	0.87	1.05 (0.99-1.10)	0.09	1.07 (0.89-1.30)	0.51	0.97 (0.91-1.04)	0.43	**1.33 (1.11-3.02)**	**0.0037**	**1.21 (1.10-1.32)**	**$3.09x10^{-5}$**
rs4764934	C/0.82	Chr12: 103469376	0.00015 / 26.42	1.09 (1.06-1.12)	$6.07x10^{-8}$	1.01 (0.96-1.04)	0.61	0.99 (0.96-1.01)	0.26	1.02 (0.92-1.12)	0.74	**1.04 (1.01-1.08)**	**0.021**	1.06 (0.98-2.65)	0.16	0.98 (0.95-1.02)	0.42
rs10638101	A/0.51	Chr19: 55827728	0.00016 / 28.20	1.07 (1.04-1.09)	$6.28x10^{-8}$	**1.03 (1.00-1.06)**	**0.026**	Missing		1.04 (0.96-1.12)	0.33	1.01 (0.99-1.04)	0.34	1.03 (0.97-1.10)	0.34	1.00 (0.97-1.03)	0.88
rs897797**	T/0.50	Chr19: 55831723	0.00016 / 28.52	1.07 (1.04-1.09)	$7.89x10^{-8}$			**1.02 (1.00-1.04)**	**0.043**								
rs16991615	A/0.07	Chr20: 5948227	0.00016 / 28.94	1.14 (1.09-1.19)	$4.67x10^{-8}$	**1.08 (1.08-1.09)**	**0.0027**	**1.05 (1.01-1.09**	**0.024**	1.10 (0.95-1.27)	0.22	1.05 (0.96-1.15)	0.078	1.00 (0.87-1.14)	0.95	**1.08 (1.02-1.15)**	**0.013**

* Allele frequency for effect allele

** Proxy SNP, included since rs10638101 was not genotyped in BCAC

*** Delta R^2 denotes the difference in Nagelkerke's pseudo-R^2 between the full model, including both covariates and SNP, and the reduced model, only including covariates.

Table 2. Characteristics in the UK Biobank participants

	Breast Cancer		**Ovarian Cancer**		**Endometrial Cancer**	
	Cases	**Controls**	**Cases**	**Controls**	**Cases**	**Controls**
Participants, N (%)	13179 (6.77)	181628	1477 (0.71)	193330	1891 (0.97)	192916
Entered menopause (%)	11200 (84.98)	131990 (72.67)	1313 (88.90)	141877 (73.39)	1719 (90.90)	141471 (73.33)
BMI*, mean (Q1 – Q3)	27.19 (23.74-29.80)	27.02 (23.42-29.80)	27.60 (23.94-30.25)	27.03 (23.44-29.63)	29.75 (24.92-33.44)	27.00 (23.43-29.60)
Age, mean (Q1 – Q3)	59.19 (55-61)	56.50 (50-63)	59.28 (54-65)	56.66 (51-63)	60.75 (58-65)	56.64 (50-71)
Smoking N (ever/never)	7472/5657	107989/73032	637/835	78052/114626	1145/737	114316/21150
Had hysterectomy, N (%)	2882 (21.87)	34214 (18.83)	816 (55.25)	36280 (18.77)	1010 (53.41)	36086 (18.71)
Had menopausal hormone therapy, N (%)	5462 (41.44)	17415 (9.59)	783 (53.01)	76094 (39.36)	868 (45.90)	76009 (39.40)
Oral contraceptives, N ever users (%)	10412 (79.00)	150074 (82.63)	1083 (73.32)	159403 (82.45)	1275 (67.42)	159211 (82.53)
Number of live births, median (Q1 – Q3)	2 (1-2)	2 (1-2)	2 (1-2)	2 (1-2)	2 (1-2)	2 (1-2)
Age at menopause, median (Q1 – Q3)	46.12 (46-52)	46.49 (46-53)	45.59 (45-52)	46.47 (46-53)	46.64 (48-52)	46.46 (46-53)

*Body mass index

Table 3. Mendelian Randomisation results for each cancer and method. BCAC; Breast Cancer Association Consortium, ECAC; Endometrial Cancer Association Consortium, OCAC; Ovarian Cancer Association Consortium.

	Breast Cancer				**Ovarian Cancer**				**Endometrial Cancer**			
	UK Biobank		**BCAC**		**UK Biobank**		**OCAC**		**UK Biobank**		**ECAC**	
Method	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
GSMR	1.30 (1.07-1.57)	0.0074	1.19 (1.03-1.38)	0.018	1.55 (0.91-2.65)	0.11	1.20 (0.99-1.46)	0.066	2.01 (1.21-3.31)	0.0065	1.45 (1.14-1.83)	0.0022
Weighted Median	1.18 (0.92-1.51)	0.18	1.28 (1.07-1.54)	0.006	1.54 (0.83-2.86)	0.17	1.26 (0.97-1.64)	0.080	1.91 (0.99-3.68)	0.051	1.46 (1.06-2.02)	0.021
Robust IVW	1.30 (1.00-1.68)	0.051	1.24 (0.92-1.67)	0.15	1.55 (1.11-2.16)	0.009	1.21 (0.87-1.69)	0.25	2.00 (1.13-3.53)	0.017	1.44 (0.89-2.33)	0.133
MR-Egger	1.03 (0.49-2.19)	0.93	1.28 (0.66-2.58)	0.46	1.49 (0.40-5.58)	0.55	0.87 (0.45-1.70)	0.69	3.41 (0.65-17.96)	0.15	3.68 (1.73-7.85)	0.001
MR-Egger intercept	1.03 (0.95-1.10)	0.52	0.99 (0.93-1.06)	0.82	1.00 (0.88-1.14)	0.95	1.03 (0.97-1.10)	0.31	0.95 (0.81-1.11)	0.50	0.91 (0.84-0.98)	0.009

Characterizing Copy Number Variations using Next- and Third-Generation Sequencing and their Association with Plasma Biomarkers

Daniel Schmitz[1], Zhiwei Li[2], Nima Rafati[3] and Åsa Johansson[1]

[1]Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala University

[2]Twelve Bio ApS, Copenhagen, Denmark

[3]National Bioinformatics Infrastructure Sweden, Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University

Abstract

Structural variations (SVs), including copy number variations (CNVs), affect approximately 20 million bases in a typical human genome. Recently, there has been an increasing interest in the role of CNVs in human development and diseases. However, most commonly used CNV detection methods, such as array comparative genomic hybridization, lack the ability to detect novel signals. This study aims to explore the role of CNVs in regulating plasma protein

biomarkers. We identified CNVs, using CNVnator, from high-coverage (30x) Illumina next-generation sequencing data, in more than 1,000 individuals. The identified CNVs were summarized and filtered as a population copy number matrix to 23,381 non-overlapping CNV regions (CNVRs) that were polymorphic in at least three individuals. Using a total of 438 plasma protein biomarkers, that were available in 872 individuals with WGS data, we conducted linear regression analyses . We identified CNVs at 19 CNVRs to be significantly associated with 22 plasma proteins ($\boldsymbol{p} < \mathbf{4.79} \times \mathbf{10^{-9}}$). Lastly, we selected a set of five polymorphic CNVRs for validation using Pacific Bioscience SMRT sequencing in 15 samples. Two CNVRs replicated and we identified two more CNVRs to be clusters of many short repetitive elements. Our findings provide insight into the involvement of CNVs on human disease as well as the application of novel sequencing approaches for SV detection.

Introduction

Genome-wide association studies (GWAS) have enabled the discovery of thousands of associations between single-nucleotide polymorphisms (SNPs) and common traits as well as diseases. Despite their success, GWAS can only explain a small part of observed heritability. For instance, a recent study estimated the variance of 32 complex traits explained by common SNPs to lie between 9.8% and 48.9%[1]. In 2015, the 1000 Genomes project consortium found 99.9% of all identified variants to be SNPs or short indels[2]. However, they also emphasize the importance of structural variants (SVs), which are rarer but cover more bases, with a typical genome containing 2,100 to 2,500 SVs covering 20 million bases. A meta-analysis of 55 population studies using the Database of Genomic Variants (DGV) estimated CNVs to cover 4.8 – 9.5% of the genome and observed complete deletions of about 100 genes without apparent phenotypic effects from the copy number variations (CNVs)[3] . Another study identified SVs in 2,504 human genomes to DGV and found that 43% of their CNVs were novel discoveries. Furthermore, they also observed multiple breakpoints in the CNV regions (CNVRs) across the population likely to be caused by individual mutational events.

Traditionally, array-based methods, such as SNP arrays and array comparative genomic hybridization (aCGH), have been used to detect genetic variations, including CNVs. Their high accuracy and reasonable price made them a popular choice for association studies. However, their limitation to previously identified polymorphisms makes it less powerful to identify novel signals. Furthermore, their resolution is limited, with common arrays being

limited to CNVs of at least 8 kbp in size[5]. Recent developments in sequencing technologies have enabled unprecedented insights into the genetic architecture of the human genome. With improvements in data quality and cost, whole-genome sequencing (WGS) has become more popular in large-scale genomic studies, including the identification of SVs. Thanks to continuously improving cost and quality of high-throughput sequencing, interest in these variants has been invigorated. There is, however, no widely accepted pipeline to detect SVs with high recall for NGS data[6].

Even though SVs identification is state-of-the-art when searching for genetic causes of monogenetic diseases, few association studies have considered the effect of SVs on common diseases and complex traits. Previously, CNVs have been associated with evolutionary fitness and embryonic lethality[7], psychiatric disorders[8,9], Crohn's diseases, type 1 diabetes, and multiple developmental diseases[9]. A previous study using data from UK Biobank found CNVs in 28 genes to be associated with 13 blood biomarkers[10]. Several studies linked both germline and somatic CNVs to multiple types of cancer, including an integrated analysis of CNVs, SNPs and expression data[11–14].

Biomarkers are well-studied traits in GWAS as they are often measured in large cohorts and are quantitative measures which increases the power to find associations. More specifically, protein biomarkers, expressed by one single gene, is usually more or less monogenic, which is beneficial for statistical power in studies in smaller cohorts[15,16]. In this project, we focused on characterizing CNVs from high coverage WGS data of over 1,000 individuals from the

72 Northern Sweden Population Health Study (NSPHS)[17]. We called CNVs using CNVnator, and
73 tested for association between copy number polymorphisms (CNPs) and the variation in
74 protein levels for a large set of proteins (N=438) that has been selected to be established or
75 exploratory biomarkers of disease[18]. Subsequently, we re-sequenced 15 individuals from our
76 cohort using Pacific Bioscience (PacBio) Single-Molecule Real Time (SMRT) technology, to
77 verify the CNPs.

78

79 Materials and methods

80 Study cohort

81 The *Northern Swedish Population Health Study* (NSPHS) was a cohort study conducted in two
82 municipalities in the region of Norrbotten, Sweden. Blood samples were taken and
83 immediately frozen at -70°C. WGS was performed at SciLifeLab in Stockholm using Illumina
84 technology to 30x coverage, and mapped to GRCh37, as described previously[19,20]. After
85 variant calling and QC 1,021 samples remained for analysis.

86

87 Protein biomarkers had been measured in 903 individuals using the Olink Protein Extension
88 Assay (PEA) and five Proseek panels (CVD2, CVD3, NEU1, ONC2, INF1), as described
89 previously[21]. In short, PEA is an affinity-based assay that uses oligonucleotide-labelled pairs
90 of antibodies that bind to the target proteins in close proximity to each other. If both
91 antibodies bind, they produce a PCR target sequence, which can be quantified using

92 standard real-time PCR. The analysis was performed on plates with 96 wells, allowing for 92
93 individuals as well as three positive and one negative controls per batch, which serve to
94 determine the lower detection limit and normalize the protein measurements. Signals below
95 detection limit were removed and the remaining measurements were normalizes using the
96 rank-based inverse normal transformation ($\mu = 0, \sigma = 1$). A total of 438 biomarkers and 892
97 samples and passed protein QC and 872 samples passed both genotyping and biomarker QC.

98

99 ## CNP calling using CNVnator

100 CNV calling was done using CNVnator. We first used CNVnator to estimate the optimal bin
101 size for each sample, as the lowest value of (70, 85, 100, 150, 200, 250 bp) at which the ratio
102 of the bins' read depth (RD) mean values to the standard deviation was between 4 and 5.
103 The reason for having the mean value of the converted RD signal 4~5 times greater than its
104 standard deviation is to preserve enough statistical power for detecting deletions by t-test
105 between the regional and global read depth (RD) signal while detecting the variations with
106 smallest bin size possible to enable higher breakpoints resolution. CNVs were then identified
107 in each sample separately and filtered in accordance with the recommendations set by
108 CNVnator's authors[18] . CNVnator provides P-values for each detected CNV calculated from a
109 one-sample t-test between the local and global RD signal. In the initial CNV calling, CNVs are
110 considered high-quality detections if they pass the significance threshold $p < 0.05$ after
111 Bonferroni correction. Additionally, we excluded CNVs where the fraction of reads with
112 ambiguous alignments (q_0) was 0.5 or more.

113 After detecting CNVs in all samples, a population CNP matrix was created using bedtools. We
114 split the genome into non-overlapping windows of 200 bp each and identified all windows
115 where any sample had a high quality CNV detected. Then, we recorded all CNVs that were
116 detected in each of the windows for all samples. We applied less stringent QC for inclusion
117 of copy numbers in the CNP matrix and only applied the q_0 threshold. Samples for which no
118 CNV had been identified within a 200bp window were assigned CN=2 (wildtype), and
119 samples that failed QC were set as missing. Finally, adjacent 200bp windows with identical
120 genotypes across all samples were merged. The merged windows represented the final CNV
121 regions (CNVRs) used in our downstream analyses.

122

123 ## Association analysis

124 We have previously shown that we need a minimum of four individuals with one copy of the
125 minor allele to reach the genome wide significance threshold in a GWAS in the cohort[19].
126 We therefore excluded all CNVRs, for which less than three individuals had a copy number
127 different from 2. To test for association between the CNVRs and the biomarker levels, we
128 used a linear regression model in the glm function in R version 4.3.4 with the CNs in the
129 CNVRs as predictors and biomarker measurements as responses. We included age and sex as
130 covariates. We applied a Bonferroni adjustment for multiple testing and adjusted for the
131 number of biomarkers $\times$ the number of CNVRs analysed.

132

133 ### Resequencing with PacBio SMRT

134 We selected 15 individuals for resequencing with Pacific Bioscience SMRT technology. We
135 performed whole-genome sequencing on a PacBio SEQUEL II system in continuous long read
136 (CLR) mode. Libraries were prepared according to manufacturer specification. The resulting
137 reads underwent standard QC procedures and were automatically mapped to human
138 genome assembly GRCh38. These mappings were delivered as binary alignment map (BAM)
139 files. We extracted the reads from the mappings using BEDTools bam2fastq version 2.29.2
140 and aligned them to GRCh37 (the same build as the Illumina data) using pbmm2 version
141 1.4.0. We then called structural variation using SVIM v1.4.2, Sniffles v1.0.12 and PBSV v2.4.0
142 and visualized the coverage and reads supporting the CNVs in the regions of interest using
143 samplot v1.1.0[22]. Variants called by SVIM, Sniffles and PBSV were considered to correspond
144 to the previously called CNVs if they overlapped reciprocally by at least 50%.

145

146 ## Results

147 ### CNV Detection

148 We selected different optimal bin sizes for each sample. The mean optimal bin size was 92
149 bp. We expected the resolution of CNVnator to be around the bin size for each sample,
150 which is around 92 bp. We observed that the size of structural variations reported by
151 CNVnator ranged from 140 bp to 20 Mbp. Our observation of the smallest variations
152 reported by CNVnator in the population is 140 bp, which agreed with our expectation. An
153 average 774 deletions and 641 insertions were detected per sample. The final CNP matrix

154 contained genotypes (CNs) of 1,021 individuals in 243,987 windows. After merging of

155 adjacent loci with consistent genotypes, 23,381 CNVRs remained.

156

157 ## Association Analysis

158 In the CNV-phenotype association analysis with 438 plasma proteins, we detected 17

159 biomarkers with at least one association passing the significance threshold ($p <$

160 $\frac{0.05}{438 \times 243987} \approx 4.68 \times 10^{-10}$), when adjusting for all 243,987 200-bp windows (Figure 1). Since

161 some of the 17 biomarkers are clustered in the nearby regions in chromosome 3, 6, 17 and

162 19 only nine peaks of significant associations between CNVs and 17 protein biomarkers. The

163 nine peaks consist of 382 significant 200-bp windows. Considering that the 200-bp windows

164 were not independent and could be merged into 23,381 independent CNVRs, a more liberal

165 adjustment for multiple testing might be appropriate, which resulted in a total of 19

166 significant CNVRs (Table 1) that are significantly associated with the 22 protein markers.

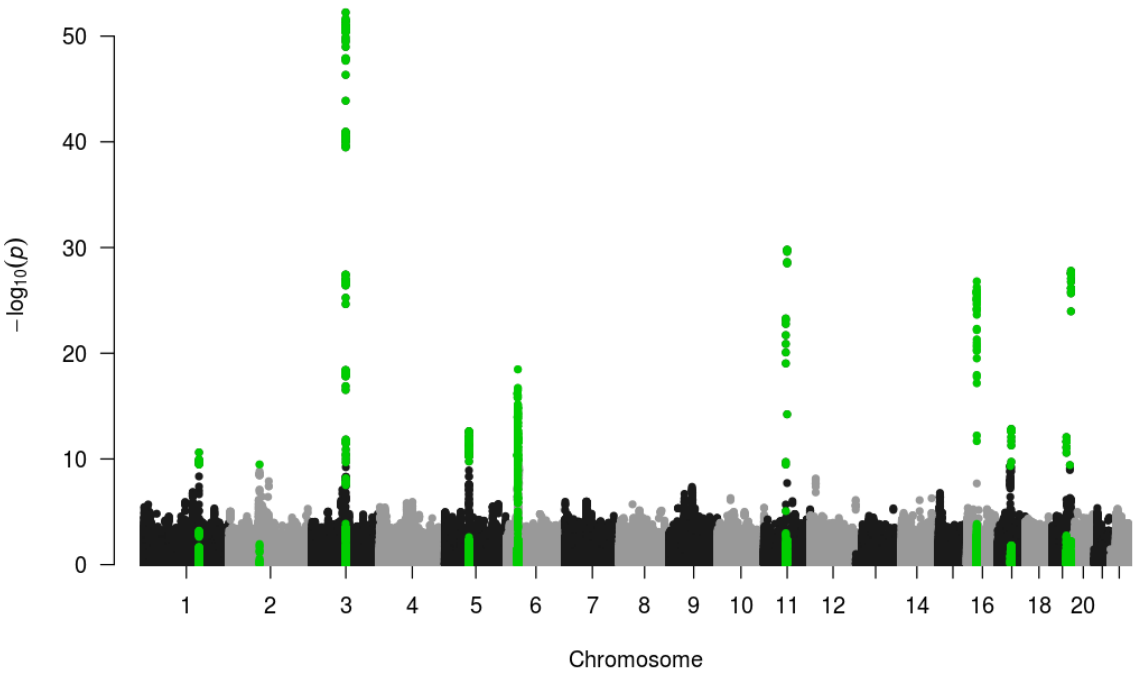


Figure 1. Manhattan plot for the CNPs-biomarker association with 200-bp CNV windows. A total of 17 protein biomarkers had at least one hit passing the adjusted significant threshold $\boldsymbol{p < 4.68 \times 10^{-10}}$. However, a total of 382 significant 200-bp windows, distributed over nine loci, were associated with any of the 17 biomarkers. The CNPs with at least one hit are highlighted in green.

172 ## SMRT Sequencing

173 We selected five of the CNVR regions (Table 1) which were highly polymorphic in the
174 population and strongly associated with any of the biomarkers for validation using long-read
175 sequencing. In addition, theses five CNV-associations did not overlap with, or were more
176 strongly associated, than previous GWAS findings and were of an optimal size to be detected
177 by long read sequencing (close to 5000bp). The 15 individuals that were prioritized for long
178 read sequencing, were selected to display different alleles of the five CNVRs. The number of
179 high-quality reads – and therefore the coverage – varied wildly between samples. The
180 system reported that less than half of all reactions produced high-quality reads for three
181 samples. These sample were therefore excluded from further analyses. In general, deletions
182 were consistently called among the three CNV-callers used, whereas duplications could not
183 be replicated between the callers. SVIM called the most of our target CNVs in agreement
184 with CNVnator , while PBSV detected the fewest.

185

186 The CNV on chromosome 2 was not detected by any of our callers in the long-read
187 sequencing data. The coverage data in this region also lacked evidence for CNVs in the
188 samples where CNVnator called a CNV in the short-read sequencing data. This led us to
189 suspect this CNV to be an artifact. Interestingly, the region downstream of this CNV, as well
190 as a short one within the CNV, received no coverage at all in any sample. This might have
191 affected CNVnator's ability to accurately detect structural variation.

192

193 The CNV on chromosome 3 was detected by SVIM, only, which called it consistently with the
194 Illumina data set. Neither PBSV or Sniffles called variants in this region consistent with
195 previous results. An optical inspection of the region revealed strong evidence for a large
196 deletion where CNVnator called this CNV.

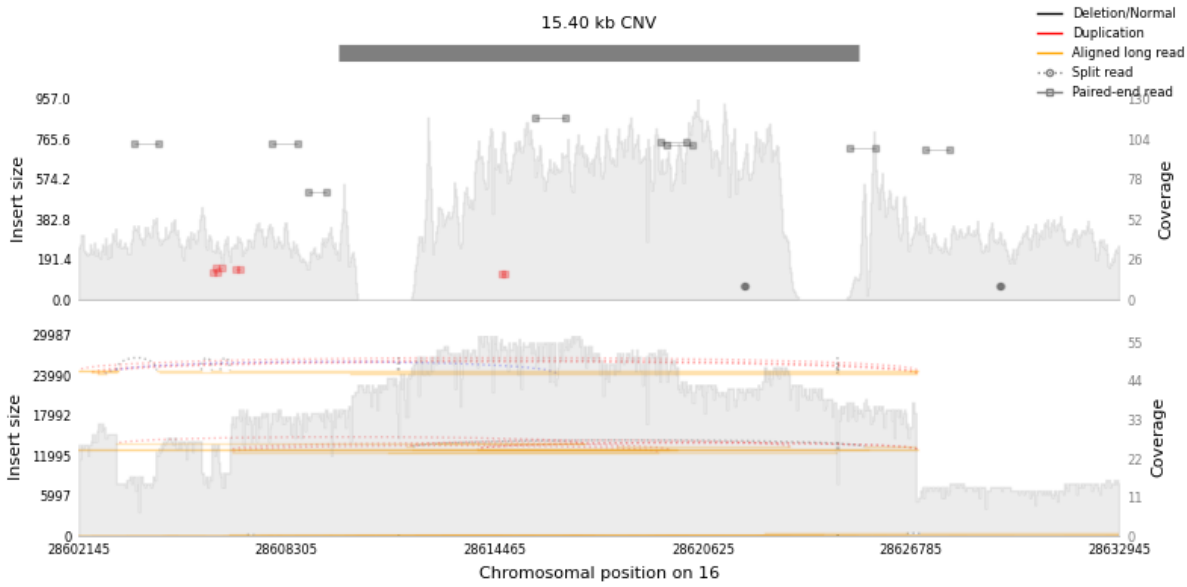
197

198 The CNV on chromosome 5 could not be replicated with any caller in the long-read data. The
199 per-base coverage in this region turned out to be highly variable with many low-quality
200 alignments. This may be due to the high frequency of repetitive elements, in this region.
201 Genome Multitool reported a low mappability of this region, as well.

202

203 The CNV on chromosome 16 could not be replicated in the long-read data (Figure 2).
204 However, SVIM called several small insertions in this region, corresponding to CN gains
205 reported by CNVnator. Additionally, the flanking regions of this CNV exhibited low mapping
206 quality. Both observations might be caused by the repetitive elements in this region, which
207 map well to them (Figure 3). This provides a possible explanation for the reported CNV.
208 CNVnator might have called these small insertions, which were then merged into a single
209 larger window by the analysis pipeline.

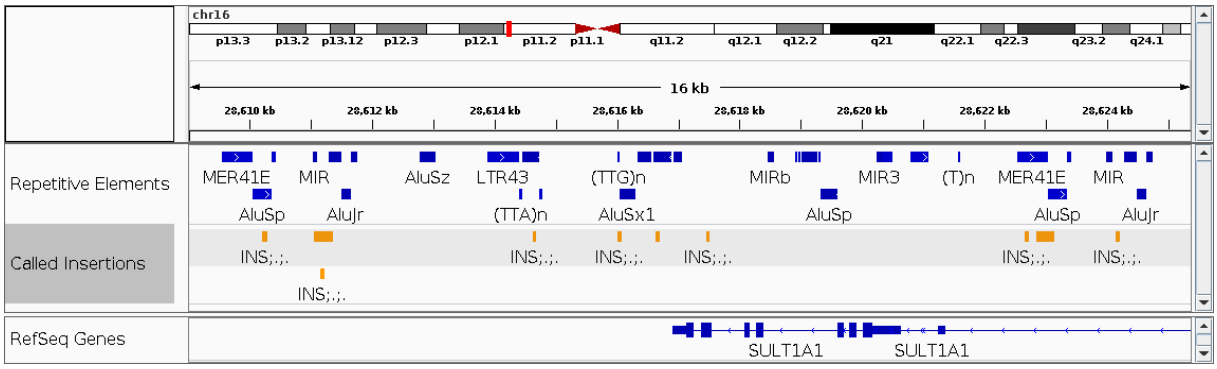
210



211

212 *Figure 2: Coverage plot of the CNV on chromosome 16 in one individual. The top shows Illumina, the bottom SMRT data. The*
213 *gray area represents the per-base coverage in the area. Lines in the plot show read-level evidence for SV. In this individual,*
214 *this CNV was called as a duplication by CNVnator, which is illustrated by the higher coverage. However, there was no clear*
215 *evidence a duplication event with the same breakpoints in the long-read sequencing data*

216



217

218 *Figure 3. Repetitive elements on chromosome 16 and insertions called by SVIM. The insertions mostly map to the repetitive*
219 *elements reported by RepeatMasker. This suggests that CNVnator actually detected these smaller CNVs and merged them*
220 *because of its binning approach.*

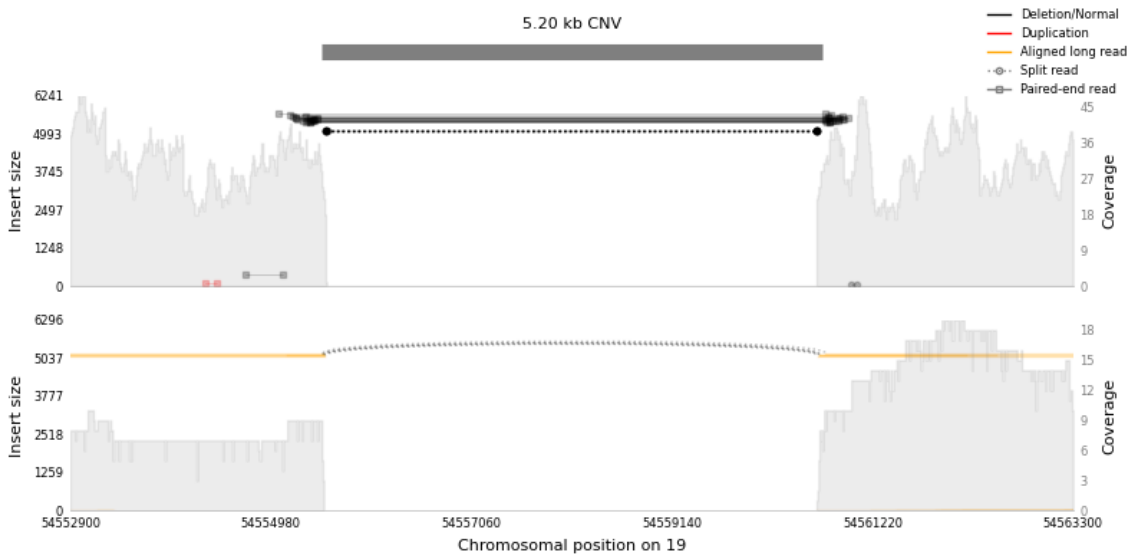


Figure 4. Coverage plot of the CNV on chromosome 19 in one individual. The top shows Illumina data and the bottom SMRT data. This was called as a homozygous deletion (CN 0). There is very clear evidence of this in both short- and long-read data.

The CNV on chromosome 19 was consistently detected by all callers. SVIM confirmed its presence, including zygosity, in all samples. Sniffles and PBSV called the variant in accordance with the Illumina data in all but two samples. A visual inspection of the area also confirmed its presence (Figure 4,Figure 5)

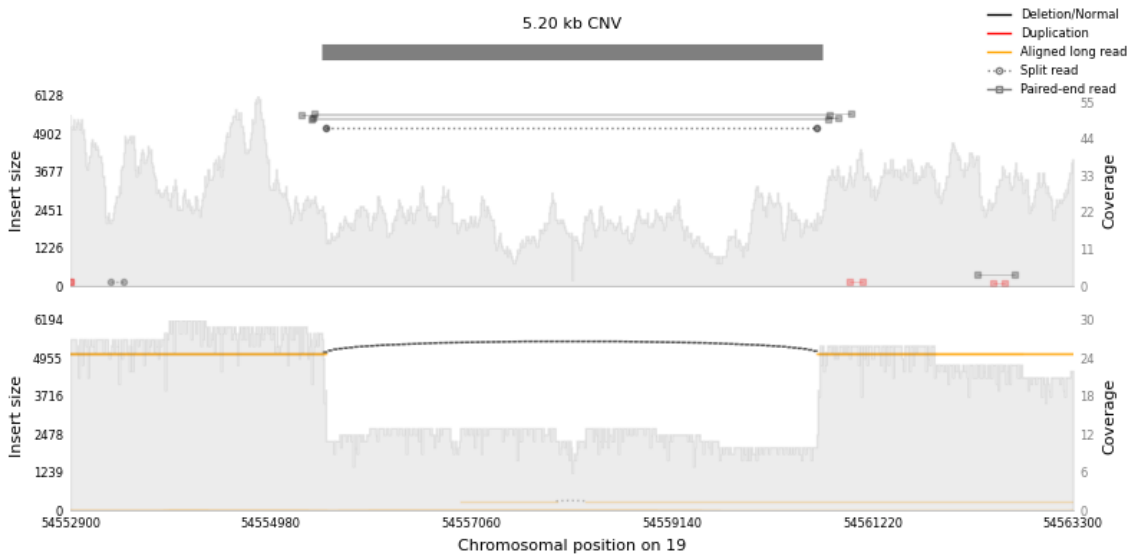


Figure 5. Coverage plot of the CNV on chromosome 19 in one individual. This was called as a heterozygous deletion (CN1). There is clear evidence in both long- and short-read data.

Discussion

We have identified CNVs in a population-based cohort and identified associations between CNVs and protein biomarkers. For the 872 samples with both genotype and phenotype data, we identified a total of 19 CNVR regions to be associated with 22 protein biomarkers. This is clearly a smaller number compared to our previous GWAS in the same cohort[19,21]. However, this agrees with that polymorphic CNVs are numerically fewer compared to SNPs that are commonly used in GWAS.

241 There have not been many studies focusing on identifying CNPs using high-throughput
242 sequencing and downstream CNV-phenotype association analysis previously. A recent study
243 developed a novel pipe-line for CNV discovery at population level with 1,364 individuals and
244 tested for association with 275 protein biomarkers[23]. They report four significant
245 associations ($\boldsymbol{p < 1.79 \times 10^{-6}}$, resolution: 15kb). Only one of our significant CNVRs (chr6:
246 30994100 — 35629600) overlapped with the result of their study. However the proteins
247 measured by the two studies are not identical which might explain the low degree of
248 overlap. By identifying 19 CNVs to be associated with the expression level of protein
249 biomarkers, our study contributes to increasing the number of CNV-biomarker associations
250 compared to previous studies.

251

252 Our long-read sequencing approach showed that CNV calling results from short- and long-
253 read technologies may not agree in many cases. While the deletions among our target CNVs
254 mostly could be validated, the characterization of the duplications was not immediately
255 clear. For instance, the target CNV on chromosome 16 manifested as many smaller insertion
256 events, rather than one large duplication as indicated form the short read sequencing. The
257 binning approach employed by CNVnator might have been responsible for them appearing
258 as a single variant in the short-read data. SMRT sequencing provided a way to accurately
259 resolve this region. On the other hand, we could not detect the target CNV on chromosome
260 2 in the long-read data, despite the presence of a strong association with levels of GPNMB,
261 which would suggest that there is indeed an underlying genetic effect causing this
262 association.

263

264 One limitation in the current study is that we did not include other SVs such as inversions
265 and translocations. A recent study of haplotype-resolved SVs discovery in the human
266 genome integrated long-read, short-read, strand-specific sequencing technologies and
267 numerous variations calling algorithms in three parents-child trios and detected an average
268 of 156 inversions out of 27,622 SVs per sample[24]. The lower frequency of other SVs and lack
269 of long-read strand-specific information cause difficulty in detecting them by only WGS data.

270

271 In this project, we focus on CNV discovery based on the alignments of short-read WGS reads
272 to the human genome reference. Although the current human genome references (GRCh38
273 and GRCh37) claim to resolve 99% of the human euchromatic genome, a study constructed a
274 de novo assembly of two Swedish genomes by long-read sequences and reported around
275 10 Mbp novel sequences missing from the GRCh38 mainly located in the centromeric or
276 telomeric regions[25]. The misalignments of the reads from the unresolved regions on the
277 current human genome reference can limit the discovery of true signals and lead to false
278 positive discoveries.

279

280 **Author contributions.** ÅJ conceived the study. NR and ÅJ planned and designed the study.
281 DS and ZL performed analyses and DS produced the figures. DS, ZL and ÅJ wrote the first
282 draft if the manuscript and all authors have read and critically reviewed the manuscript.

283

Competing interests.

The author declares no conflict of interests.

Acknowledgement

We acknowledge all the participants and staff involved in NSPHS for their valuable contribution. The NSPHS was funded by the Foundation for Strategic Research (UG) and the European Commission FP6 (UG). Sequencing was funded by the Science for Life Laboratory (SciLifeLab), Swedish Genomes Program, which has been made available by support from the Knut and Alice Wallenberg Foundation. Sequencing was performed by NGI (National Genomics Infrastructure). Protein measurements were carried out by Olink Proteomics AB in Uppsala, Sweden. The computations and data handling were enabled by resources in project SNIC 2018/8-372, sense2016007 provided by the Swedish National Infrastructure for Computing (SNIC) at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. This work was also funded by the SciLifeLab´s Technology Development Project (TDP) program, the Swedish Medical Research Council (2019-01497) and the Swedish Heart-Lung foundation (nr. 20200687).

References

1. Nolte, I. M. *et al.* Missing heritability: Is the gap closing? An analysis of 32 complex traits in the Lifelines Cohort Study. *Eur. J. Hum. Genet.* **25**, 877–885 (2017).

306 2. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74
307 (2015).
308 3. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes.
309 *Nature* **526**, 75–81 (2015).
310 4. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of
311 the human genome. *Nat. Rev. Genet.* **16**, 172–183 (2015).
312 5. Quenez, O. *et al.* Detection of copy-number variations from NGS data using read depth
313 information: a diagnostic performance evaluation. *Eur. J. Hum. Genet. 2020 291* **29**, 99–
314 109 (2020).
315 6. Guan, P. & Sung, W. K. Structural variation detection using next-generation sequencing
316 data: A comparative technical review. *Methods* **102**, 36–49 (2016).
317 7. Beckmann, J. S., Estivill, X. & Antonarakis, S. E. Copy number variants and genetic traits:
318 Closer to the resolution of phenotypic to genotypic variability. *Nat. Rev. Genet.* **8**, 639–
319 646 (2007).
320 8. Malhotra, D. & Sebat, J. CNVs: Harbingers of a rare variant revolution in psychiatric
321 genetics. *Cell* **148**, 1223–1241 (2012).
322 9. Huang, J., Ellinghaus, D., Franke, A., Howie, B. & Li, Y. 1000 Genomes-based imputation
323 identifies novel and refined associations for the Wellcome Trust Case Control
324 Consortium phase 1 Data. *Eur. J. Hum. Genet.* **20**, 801–805 (2012).
325 10. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK
326 Biobank. *Nat. Genet.* **53**, 185–194 (2021).

327 11. Momtaz, R., Ghanem, N. M., El-Makky, N. M. & Ismail, M. A. Integrated analysis of SNP,
328 CNV and gene expression data in genetic association studies. *Clin. Genet.* **93**, 557–566
329 (2018).
330 12. Park, R. W., Kim, T.-M., Kasif, S. & Park, P. J. Identification of rare germline copy number
331 variations over-represented in five human cancer types. *Mol. Cancer 2015 141* **14**, 1–12
332 (2015).
333 13. Chattopadhyay, A. *et al.* CNVIntegrate: the first multi-ethnic database for identifying
334 copy number variations associated with cancer. *Database* **2021**, 1–12 (2021).
335 14. Brezina, S. *et al.* Genome-wide association study of germline copy number variations
336 reveals an association with prostate cancer aggressiveness. *Mutagenesis* **35**, 283–290
337 (2020).
338 15. Enroth, S., Johansson, Å., Enroth, S. B. & Gyllensten, U. Strong effects of genetic and
339 lifestyle factors on biomarker variation and use of personalized cutoffs. *Nat. Commun.* **5**,
340 4684 (2014).
341 16. Enroth, S., Bosdotter Enroth, S., Johansson, Å. & Gyllensten, U. Effect of genetic and
342 environmental factors on protein biomarkers for common non-communicable disease
343 and use of personally normalized plasma protein profiles (PNPPP). *Biomarkers* **20**, 355–
344 364 (2015).
345 17. Igl, W., Johansson, A. & Gyllensten, U. The Northern Swedish Population Health Study
346 (NSPHS)--a paradigmatic study in a rural population combining community health and
347 basic research. *Rural Remote Health* **10**, 1363 (2010).

348 18. Abyzov, A., Urban, A. E., Snyder, M. & Gerstein, M. CNVnator: An approach to discover,
349 genotype, and characterize typical and atypical CNVs from family and population
350 genome sequencing. *Genome Res.* **21**, 974–984 (2011).
351 19. Höglund, J. *et al.* Improved power and precision with whole genome sequencing data in
352 genome-wide association studies of inflammatory biomarkers. *Sci. Rep.* **9**, 16844 (2019).
353 20. Ameur, A. *et al.* SweGen: a whole-genome data resource of genetic variability in a cross-
354 section of the Swedish population. *Eur. J. Hum. Genet. 2017 2511* **25**, 1253–1260 (2017).
355 21. Enroth, S. B. S. *et al.* Systemic and specific effects of antihypertensive and lipid-lowering
356 medication on plasma protein biomarkers for cardiovascular diseases. *Sci. Rep.* **8**, 5531
357 (2018).
358 22. Belyeu, J. R. *et al.* Samplot: a platform for structural variant visual validation and
359 automated filtering. *Genome Biol.* **22**, 161 (2021).
360 23. Png, G. *et al.* Population-wide copy number variation calling using variant call format
361 files from 6,898 individuals. *Genet. Epidemiol.* **44**, 79–89 (2020).
362 24. Chaisson, M. J. P. *et al.* Multi-platform discovery of haplotype-resolved structural
363 variation in human genomes. *Nat. Commun.* **10**, 1–16 (2019).
364 25. Ameur, A. *et al.* De novo assembly of two swedish genomes reveals missing segments
365 from the human GRCh38 reference and improves variant calling of population-scale
366 sequencing data. *Genes* **9**, 486 (2018).
367

Table 1. CNVRs detected to be significantly associations with a biomarker when adjusting for the number of CNVRs identified. The table show the summary statistics for the most significant 200-bp window in each CNVR.
* CNVRs selected for validation using long-read sequencing.

CNVR #	Biomarker	Chr	Start	End	Size	Lead CNV	Beta	SE	P Value
1	CD48	1	158867600	158867800	200	1:158867600-158867800	-0.282444	0.04714985	3.1473E-09
2	FCRLB	1	161640580	161642980	2400	1:161640580-161640780	0.61738819	0.09112364	2.4227E-11
3	LY9	1	179455600	179455800	200	1:179455600-179455800	0.53598439	0.0885526	2.1686E-09
4*	GPNMB	2	89610400	89613200	2800	2:89613000-89613200	-0.9906326	0.15575469	3.3349E-10
5*	PD-L2	3	98410600	98414800	4200	3:98411600-98411800	0.40933966	0.04469867	3.9228E-19
5*	ICAM-2	3	98410600	98414800	4200	3:98410600-98410800	0.64827303	0.03946845	5.665E-53
5*	Siglec-9	3	98410600	98414800	4200	3:98411800-98413400	0.68180291	0.04171774	3.4758E-52
5*	CD200R1	3	98410600	98414800	4200	3:98411800-98413400	0.50824674	0.04448441	3.708E-28
5*	VEGFR-3	3	98410600	98414800	4200	3:98414600-98414800	0.59647014	0.04166388	1.1011E-41
5*	ICAM-3	3	98899100	98902300	3200	3:98899900-98900100	0.39646955	0.05517257	1.4524E-12
6*	AMBP	5	737870	746270	8400	5:745070-745270	-0.2838274	0.0451137	5.0537E-10
7	IL-18	5	70303300	70395300	92000	5:70393100-70393300	0.41790501	0.05609865	2.3753E-13
8	MIC-AB	6	30994100	35629600	4635500	6:32496600-32496800	0.61926004	0.06689785	3.3464E-19
9	CCL19	6	32501000	32541400	40400	6:32522200-32522400	-0.4665064	0.05366892	1.9287E-17
10	FR-gamma	11	63442300	67332555	3890255	11:67331355-67331955	-1.2325907	0.10305024	1.6409E-30
11	FR-gamma	11	67330155	67332555	2400	11:63443100-63445300	-1.0251322	0.09829072	5.1728E-24
12	CNTN1	12	45903400	45909800	6400	12:45909600-45909800	-0.2854559	0.04690077	1.7331E-09
13*	ST1A1	16	28609845	28625245	15400	16:28613645-28613845	0.78565085	0.06931645	1.6051E-27
14	CCL4	17	36387670	36399670	12000	17:36392670-36394670	0.14458813	0.02294042	4.649E-10
15	CCL15	17	39203400	39211600	8200	17:39210800-39211000	-0.9207193	0.12255953	1.4473E-13
16	SMPD1	19	35863600	35863800	200	19:35863600-35863800	0.3688639	0.0607791	1.9663E-09
17	MIA	19	41381725	41387525	5800	19:41381925-41385125	-1.2302332	0.1693139	8.578E-13
17	hK11	19	51508740	51510940	2200	19:51508940-51510740	1.40588574	0.22183536	3.8361E-10
18*	hOSCAR	19	54555500	54560700	5200	19:54558900-54559100	-0.7211762	0.06273603	1.6099E-28
19	WFDC2	20	44204035	44206635	2600	20:44204435-44205035	0.41745715	0.06811548	1.3725E-09