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

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

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

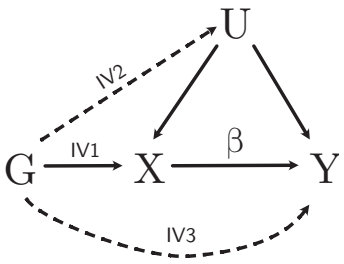

Figure 1.1: Graphical representation of the assumptions of MR. The strength β of the causal effect of X on Y is to be estimated. The genetic variants G affect X (IV1) but neither any confounders U (IV2) nor Y (IV3) directly. Based on illustration from [4].

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

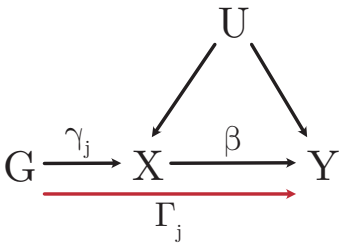

Figure 1.2: Graphical representation of the summary-based MR approach. Γ_j refers to the effect size estimated by a GWAS, which only assesses association, not causality, and not a pleiotropic effect of G on Y. The effect β can be calculated according to $\beta = \frac{\Gamma_j}{\gamma_j}$.

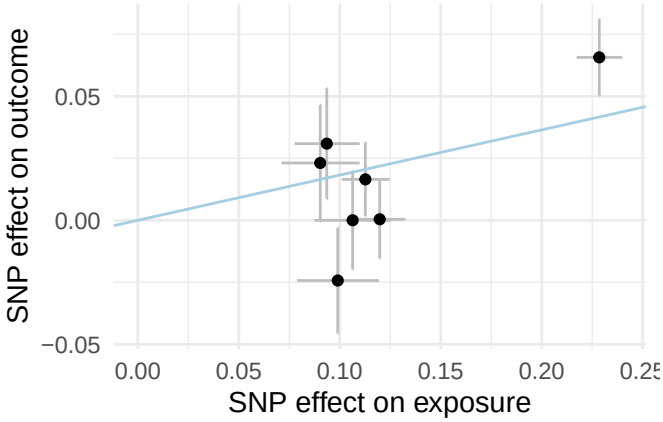

Figure 1.3: Example of a two-sample MR result. The x-axis shows the estimated effect of the IVs on the exposure. The y-axis shows the effect on the outcome. Each point represents one IV. The whiskers correspond to the standard error of each IV's effect estimate The effect of the exposure on the outcome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

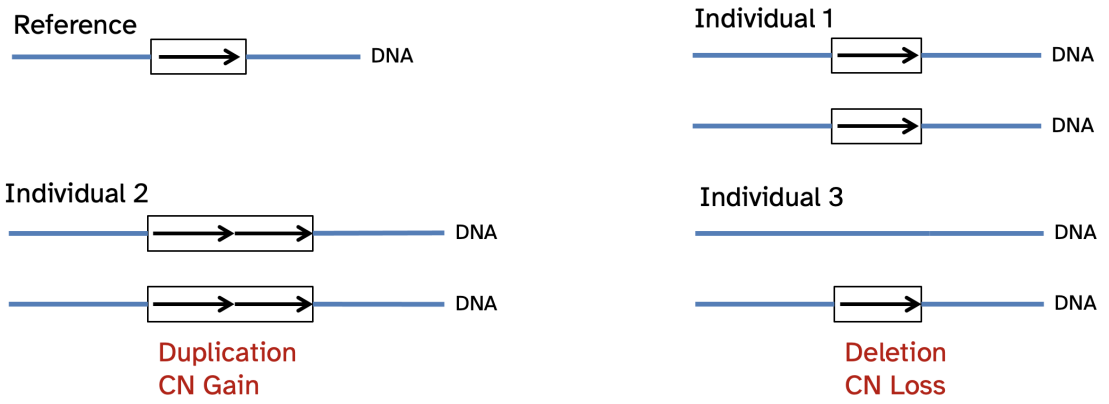

Figure 1.4: Example of a CNV. Each arrow represents a copy of the sequence of interest. There is one copy in the (haploid) reference. In a diploid individual, this corresponds to CN 2 (Individual 1). Individual 2 is homozygous for two copies of the sequence (CN 4). An increase in the number of copies 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 $175\,\mathrm{pmol}/\ell$, which is above the normal range for serum estradiol concentrations in postmenopausal females ($0 - 73.4\,\mathrm{pmol}/\ell$) [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 $2\,\mathrm{pg/m}\ell$ ($\approx 7.34\,\mathrm{pmol}/\ell$) 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).