Degree project

1977
UNIVERSITY
OF SKÖVDE

ASSESSMENT OF NON-CODING MUTATIONS IN ENDOMETRIOSIS AND OVARIAN CANCER HISTOTYPES

Master Degree Project (120 credits) in Systems Biology

Second Cycle 45 credits

Spring term 2024

Student: Manuel Andres Rujano Balza, MD

Supervisors: Anna Linder, PhD, University of Gothenburg; Benjamin Ulfenborg, PhD, University of Skövde

Examiner: Dario Melguizo Sanchis, PhD, University of Skövde

Abstract

Endometriosis affects 10% of women of reproductive age and is associated with an increased risk of developing specific ovarian cancer histotypes, such as clear cell ovarian carcinoma (CCOC) and endometrioid ovarian carcinoma (EOC). While most research has focused on exploring direct genomic associations involving cancer-driver genes with inconclusive results, recent genome-wide association studies have delved into non-coding regions to identify additional genetic factors. These studies, however, face challenges in detecting rare and somatic mutations. Advances in next-generation sequencing, accessible datasets, and deep-learning models have enabled comprehensive genomic studies. This study aimed to identify specific non-coding mutations in ovarian endometriosis (OE) and ovarian cancer (OC) that could influence disease development. Using whole exome sequencing (WES) data from 20 OE and OC specimens and a variant calling pipeline called Sarek, 11774 somatic non-coding mutations were identified. Of these, 49.2% (5794 variants) were predicted to be within regulatory elements. Furthermore, 50 common mutations were found in more than three specimens. Unsupervised clustering grouped the results into three categories: the OE group (9 samples) which had specific variants in the *LRRC4B, REPS1*, and *SLC1A1* genes, while variants in *CNN2P1, ANKRD20A4P,* and *ZNF806* were found in both OE and OC groups (20 samples). Additionally, 44 mutations were found in 3 or more OC samples (out of 11), with *VGLL1* and *PRSS56* being the most recurrent across both OC histotypes. In summary, this study has identified specific common mutations prioritized based on their frequency and functional annotation. However, understanding their functional significance will need further experimental validation.

Popular scientific summary

Endometriosis is a disease where tissue similar to the endometrium, the lining inside the uterus, begins to grow outside the uterus. It has important implications for reproductive health and is a common cause of pelvic pain. Additionally, endometriosis has been associated with an increased risk of developing certain types of ovarian cancer.

Traditionally, research has mainly focused on the protein-coding region of the genome, called exons, which make up only 2% of the human genome. Less is known about the non-coding space, which is frequently associated with a lack of functionality. In recent years, regulatory gene regions have been discovered in these non-coding sequences. How a gene regulates its expression affects how its translated protein interacts within a cell and ultimately influences complex systems of an organism.

Understanding how changes in these regulatory elements can lead to diseases requires a combination of advanced biology and computer science methods. This interdisciplinary approach has been made possible by milestones like the Human Genome Project, which began in 1990 and successfully sequenced and mapped all the genes in the human genome. Subsequent advances in bioinformatics have allowed for efficient processing of large amounts of genomic data.

Next-generation sequencing (NGS) has been a fundamental tool in research in recent decades. This technology converts biological data, such as DNA sequences, into digital information. Variants of NGS also capture data at the epigenomic and transcriptomic levels, shedding light on the dynamic nature of the genome. Processing the human genome, which contains about 3 billion base pairs, is a significant challenge in storage and computation. Sophisticated bioinformatics pipelines are accelerating this process, and large-scale international collaborations are making relevant databases more accessible.

Privacy and data security are crucial when handling sensitive personal information like DNA sequences. To protect this information, leading national institutions and health centers must promote best practices in research.

By integrating these advanced tools and collaborative efforts, this study aims to identify specific non-coding mutations associated with endometriosis and ovarian cancer to increase our understanding of these diseases and hopefully offer new perspectives for improved diagnosis and treatment options.

Table of Contents

Abbreviations

CCOC	Clear cell ovarian carcinoma
CI	Confidence interval
EAOC	Endometriosis-associated ovarian cancer
EOC	Endometrioid ovarian carcinoma
eQTL	Expression quantitative loci
GWAS	Genome-wide association study
OC	Ovarian cancer
OE	Ovarian endometriosis
RG	Risk locus target gene
RE	Regulatory element
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TSS	Transcription starting site
VAF	Variant allele frequency
WES	Whole exome sequencing
WGS	Whole genome sequencing
5’ UTR	5’ untranslated region
3’ UTR	3’ untranslated region

Introduction

Association Between Endometriosis and Ovarian Cancer

Endometriosis affects approximately 10% of women of reproductive age and is characterized by the ectopic growth of endometrial tissue (Giudice, 2010). The endometriotic lesions, which are influenced by estrogen and histologically resemble normal endometrial tissue, induce a chronic inflammatory response (Wang *et al.*, 2020). Endometriosis is commonly found in the pelvic peritoneum and ovaries, but it can also extend to the abdominal peritoneum, being a major cause of pelvic pain and infertility (Giudice, 2010).

Epidemiological studies have found an association between endometriosis and a higher risk of developing specific histotypes of epithelial ovarian cancer (OC), such as clear cell ovarian carcinoma (CCOC) and endometrioid ovarian carcinoma (EOC), which are commonly referred to as endometriosis-associated ovarian cancer (EAOC) (Pearce *et al.*, 2012). Observations from a study reported an incidence rate ratio (IRR) of OC in women with confirmed endometriosis of 1.08 (95% CI: 0.87–1.35) after adjusting for age, with a significantly higher IRR for the subtypes CCOC and EOC (2.29 [95%CI: 1.24–4.20] and 2.56 [95% CI: 1.47–4.47], respectively) (Hermens *et al.*, 2020). Whole-exome sequencing (WES) analyses and histopathological studies have identified common somatic mutations in genes such as *ARID1A*, *PIK3CA*, and *KRAS* observed in tumor samples of CCOC and EOC (Anglesio *et al.*, 2015; Yachida *et al.*, 2021). Alterations in these genes were also detected in endometriosis lesions unrelated to cancer and normal endometrial tissue, suggesting a potential clonal origin of the cancer stage from normal endometrium (Anglesio *et al.*, 2017; Suda *et al.*, 2018).

Furthermore, recent studies have started to explore the potential roles of the regulatory elements (REs) affected that could be associated with the development of EAOC (Corona *et al.*, 2020). For instance, a cross-trait meta-analysis of genome-wide association studies (GWAS) on endometriosis and OC susceptibility integrated data from large cohorts of both conditions. This analysis identified single nucleotide variants (SNVs) in regulatory regions common to endometriosis and each OC histotype (Mortlock *et al.*, 2022). These findings suggest the potential involvement of additional mutations in RE during the progression of endometriosis.

Despite the progress made, Linder *et al.* (2024) outlined unanswered questions regarding the role of cancer-driver genes in triggering the malignant progression of OE. Their study was the first to investigate genetic variations in cases of OE followed by a diagnosis of OC and couldn't find a direct relation with cancer-associated genes. This motivates the need to search for additional mutations that may not have been identified in recent GWAS.

Role of Non-coding Mutations in Human Disease

King and Wilson (1975) described how phenotypic differences between species were primarily due to changes in the gene regulation systems rather than changes in the sequence of amino acids. Subsequent research analyzing whole genome sequencing (WGS) data has shown how ~96% of somatic mutations are distributed in the non-coding regions of the human genome (Corona *et al.*, 2020), including introns, intergenic, untranslated regions (UTR), and non-coding functional RNA (Nakagawa & Fujita, 2018). Gene expression relies on a well-organized interaction between transcription factors (TFs) and DNA sequences that serve as their binding sites (Latchman, 1993). These sequences, *cis*-regulatory elements, are typically located in non-coding DNA and include promoters and distal regulatory elements (REs) such as enhancers (Wittkopp & Kalay, 2012;

Khurana *et al.*, 2016). To regulate gene expression, the TFs bind to specific DNA sequences, 'binding motifs', within the regulatory regions (Boeva, 2016) (see Figure 1).

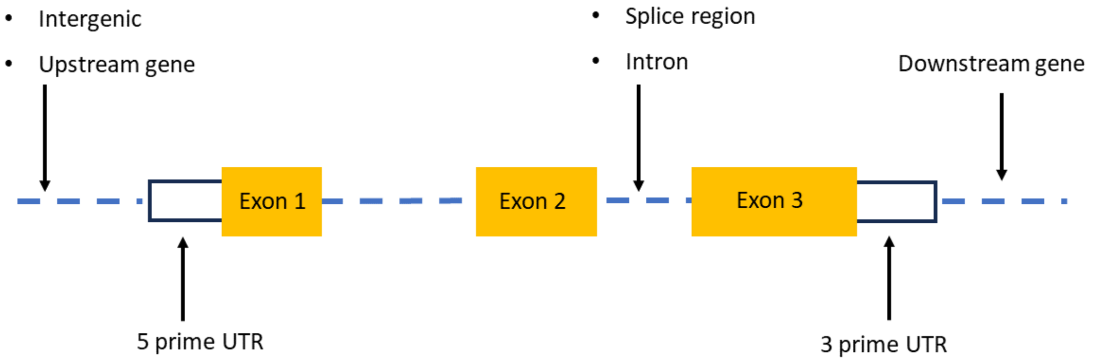


Figure 1. **Illustration of non-coding regions near the gene.** The orange boxes represent the exons or protein-coding sequence, while the arrows point to the areas where the regulatory elements are found.

Mutations in the REs are considered the most common cause of phenotypic divergence (Wittkopp & Kalay, 2012). As technology advances, researchers have expanded the scope beyond studying the protein-coding sequences and Mendelian disorders. Genome-wide association studies have facilitated the identification of SNVs associated with traits/diseases across different populations, with a subset of these SNVs being common (with allele frequency above 1-5% of the population) (Manolio et al., 2009). Moreover, most SNVs reported by large GWAS studies are consistently found within intergenic and intronic regions (Hindorff *et al.*, 2009). Even though GWAS have been important for elucidating variants in non-coding regions, it is limited in detecting rare variants and does not provide mechanistic insights (Ward & Kellis, 2012).

Many studies have demonstrated how seemingly small perturbations of the genome, such as an SNV, can influence changes in the phenotype depending on its location. For instance, a mutation within the *HBB* gene (A>T; rs334) results in an amino acid change, disrupting the functionality and structure of erythrocytes, a defining feature of sickle cell anemia (Steinberg & Sebastiani, 2012). Contrastingly, Claussnitzer *et al.* (2015) described the existence of a mutation within an enhancer located 1.2 Mb away from its targets, able to affect the expression of two genes, *IRX3* and *IRX5*, resulting in a strong pro-obesity effect. The authors suggested that the mutation possibly disrupts a conserved motif as a binding site for the ARID5B repressor. Protein-coding sequences are highly conserved among species; in contrast, the conserved proportion of the total non-coding DNA corresponds to approximately 3% (Drake et al., 2006). This has suggested that conserved non-coding regions (CNCs) are selectively constrained and likely to be the target of functional variants (Claussnitzer *et al.*, 2015).

Computational Methods for Variant Analysis in Cancer

Understanding genetic variants through DNA sequencing is a major focus in cancer research. Somatic mutations are frequently associated with disrupting pathways in tumor cells (Greenman *et al.*, 2007; Martincorena, 2015). Yet, not all somatic mutations contribute to the development of cancer. They can be classified as 'passengers' or 'drivers' based on their role in positive selection within the tumor microenvironment (Greenman *et al.*, 2007). Driver mutations provide a growth advantage to the cancer cell, leading to positive selection, while passenger mutations do not confer growth advantage and hence are not directly related to oncogenesis (Stratton, 2009).

Next-generation sequencing (NGS) data stands out as a powerful tool for detecting somatic mutations because it can generate a high volume of reads. Nevertheless, accurately identifying these mutations can be challenging due to noise and artifacts in the reads (Xu, 2018). Sequencing cancer cells mainly aims to identify somatic mutations (Koboldt, 2020). Hence, this study will focus on somatic variant calling pipelines.

To detect mutations within the sequencing reads, the workflow is commonly divided into three main steps: preprocessing, alignment, and variant calling (Xu, 2018). Subsequent methods involve interpreting the variants by annotating and prioritizing them based on their possible impact (Ward & Kellis, 2012).

The preprocessing step includes quality control checks on the raw FASTQ files to remove low-quality bases and the use of FastQC to assess the overall quality of the reads (Wingett & Andrews, 2018). Alignment involves mapping the reads to their original position using a reference genome. A popular algorithm used for this task is the Burrows-Wheeler Alignment tool (BWA), generating a binary alignment map (BAM) file format (Li & Durbin, 2009). Furthermore, the Genome Analysis Toolkit (GATK) protocols recommend marking duplicated reads, base quality-score recalibration, and indel realignment (Van der Auwera *et al.*, 2013).

For variant detection, there are many tools available. The choice of variant caller typically depends on the specific types of mutations of interest for the analysis, such as SNVs, indels, or structural variants (SV) (Xu, 2018). This selection must also consider the variants' allele frequency (VAF) as a factor. Given the heterogeneity among tumor cells and the nature of somatic mutations, low-frequency variant callers are the preferred algorithms for analyzing cancer samples (Raphael *et al.*, 2014). Tools like MuTect2 (Cibulskis *et al.*, 2013) and Strelka (Saunders *et al.*, 2012) can detect low-frequency variants in both whole-genome and whole-exome sequencing data obtained from matched or non-matched tumor samples. After this, a list of variants is generated in a variant calling file (VCF), and the subsequent methods focus on its interpretation. This phase often starts with prioritizing the variants by their location in the genome using annotators like SnpEff (Cingolani *et al.*, 2012) and Variant Effect Predictor (VEP) (McLaren *et al.*, 2016). Analysis pipelines, such as Sarek, enable running most of these algorithms in a single container, facilitating the overall workflow and reproducibility of the experiments (Garcia *et al.*, 2020).

So far, the methods described here have summarized a standard variant calling pipeline; nevertheless, interpreting non-coding variants presents particular challenges since this includes characterizing the regulatory elements and associating them with genes. To address these challenges, experimental methods such as expression quantitative trait loci (eQTL) analyses have been employed to assess the relationship between genotype and gene expression regulation (GTEx Consortium, 2017). Moreover, understanding epigenetic modifications is becoming increasingly important to provide context to the variants' effects. Using reference profiles from chromatin immunoprecipitation sequencing (ChIP-Seq) of relevant cell lines can benefit these experiments by providing dynamic insights from the regulatory activity of the genome (Ernst *et al.*, 2011; Ward & Kellis, 2012).

Deep learning models such as convolutional neural networks (CNN) have demonstrated high efficacy for extracting features from genomic sequences. For instance, models like Basset (Kelley *et al.*, 2016) and DeepSea (Zhou & Troyanskaya, 2015) have designed CNN models to identify relevant binding motifs and predict variant effects based on position-specific enrichments. Similarly, the Sei framework (Chen *et al.*, 2022) is an improved version of DeepSea utilizing a CNN model trained on chromatin profiles from datasets such as ENCODE, the Roadmap Project, and the Cistrome Project to predict the functional role of a variant. This model enables a wide-scale and comprehensive variant analysis by classifying a large list of specific mutations.

Altogether, the goal of this study was to investigate non-coding DNA in OE and OC using some of the latest bioinformatic tools available. The difficulties in diagnosing endometriosis in women motivate the search for new biomarkers associated with the disease.

Aim

This study aims to characterize non-coding mutations in endometriosis and endometriosis-associated ovarian cancer. Specifically, the study will analyze whole exome sequencing data to identify specific non-coding mutations that could clarify the following:

- What mutations are shared within endometriosis samples?
- What genomic variants are shared between endometriosis and cancer?
- What variants are shared within the different cancer histotypes?
- Could specific genetic variants be identified in endometriosis synchronously diagnosed to cancer compared to those with prior diagnosis?

The discovery of new biomarkers could improve diagnostics, enable early detection, and potentially prevent the malignant progression of endometriosis. New perspectives could also help identify novel therapeutic targets. This research aims to contribute valuable knowledge to an understudied disease, hoping to ultimately improve quality of life and save lives.

Materials and Methods

Overview

The study design followed the initial steps for data processing used by Linder *et al*. (2024) and a general approach described by Khurana *et al*. (2016) for the interpretation of non-coding variants in cancer. Sarek version 3.2.3 was used for variant analysis on the FastQ files of whole exome sequencing data (WES). The pipeline included preprocessing, variant calling, and annotation, generating VCF files as output. These were merged and filtered to identify somatic variants. The steps described previously were performed in a Red Hat Enterprise Linux 8.3 secured server provided by Gothenburg University (GU). Further, the outputs were compared with the GTEx dataset. VCF files were processed using the Sei framework, a deep-learning model to categorize the sequences based on their regulatory activity. The steps are illustrated in Figure 2.

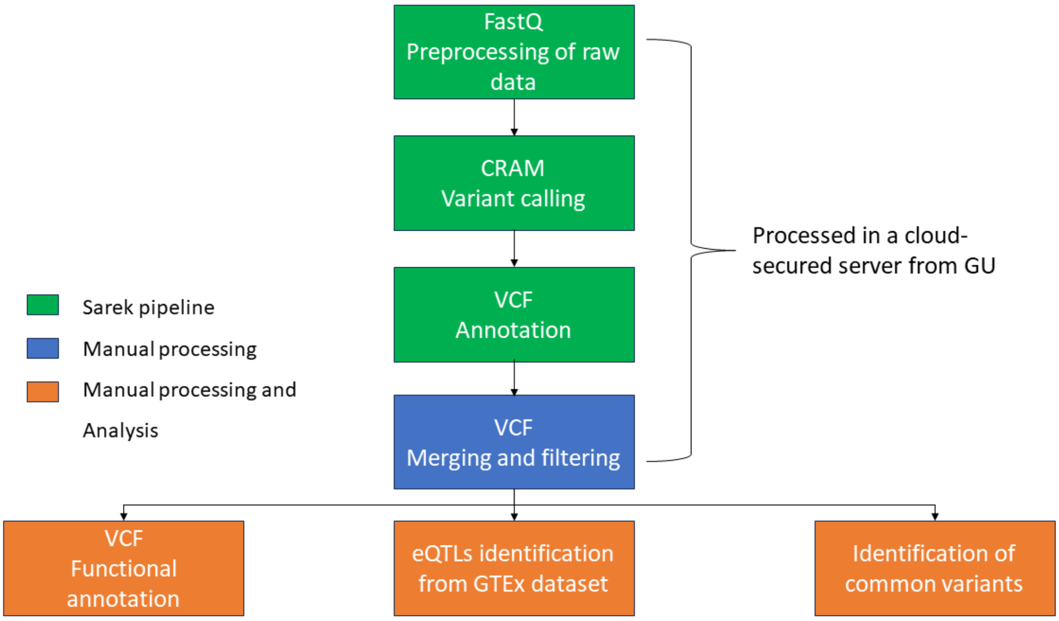

FastQ
Preprocessing of raw data
CRAM
Variant calling
VCF
Annotation
VCF
Merging and filtering
Processed in a cloud-secured server from GU
🟩 Sarek pipeline
🟦 Manual processing
🟧 Manual processing and Analysis
VCF
Functional annotation
eQTLs identification from GTEx dataset
Identification of common variants

Figure 2. **The methods and flow of this study.** Steps colored in green were run from the raw data using Sarek. The steps colored in blue were processed using R to intersect results from different variant callers, merge VCFs from different samples, and then filter variants based on criteria. FastQ, CRAM, and VCF represent the input file format; the steps that do not have these use the resulting tidy data from the merging and filtering step. The steps colored in orange used R or Python scripts to build graphs and run packages used for functional analysis. CRAM: Compressed Reference-oriented Alignment Map. VCF: Variant calling format

Sample selection and whole exome sequencing datasets

This study used whole exome sequencing data from the archival collection of two tissue formalin-fixed and paraffin-embedded (FFPE) specimen groups preserved by Sahlgrenska University Hospital (SUH). For practicality, the groups are named Endo-seq and Synchron-seq throughout the study, as well as a group identifier 'A' or 'B' at the end of each sample name to differentiate the groups, e.g., ovarian endometriosis (OE) 2A for the Endo-seq group and OE-2B for Synchron-seq. The Endo-seq group consisted of 6 specimens of OE surgically removed before the diagnosis of ovarian cancer (OC) and 6 subsequently paired OC specimens removed at the time of diagnosis (Linder *et al.*, 2024). The OC specimens included clear cell ovarian carcinoma (CCOC, n=2) and

Endometrioid ovarian carcinoma (EOC, n=4) histotypes. The median time between OE and OC diagnoses was 11 years (range: 6-27 years). The Synchron-seq group included 3 complete paired sets of OE and OC samples synchronously diagnosed. Additionally, this group included two specimens of EOC without the corresponding OE samples. All samples included a matched germline control either from the blood, fallopian tube, or uterine cervix at the time of cancer diagnosis. Laser Capture Microdissection, DNA extraction, and whole exome sequencing methods were described previously by Linder *et al.* (2024). Details of the cases, containing their corresponding histotypes, ages, and years between diagnoses are provided in Table 1.

Table1. Information on study cohorts

Case	**Age at the time of diagnosis**	**Histotype**	**Years between diagnosis**
Endo-seq group (A)			
OE-1A	52	OE	-
OC-E-1A	65	E-OC	13
OE-2A	48	OE	-
OC-CC-2A	57	CC-OC	9
OE-3A	26	OE	-
OC-E-3A	52	E-OC	26
OE-4A	33	OE	-
OC-CC-4A	42	CC-OC	9
OE-5A	42	OE	-
OC-E-5A	69	E-OC	27
OE-6A	41	OE	-
OC-E-6A	47	E-OC	6
Synchron-seq group (B)			
OE-2B	42	OE	-
OC-CC-2B	42	CC-OC	-
OE-3B	73	OE	-
OC-CC-3B	73	CC-OC	-
OE-9B	50	OE	-
OC-E-9B	50	E-OC	-
OC-E-6B	37	E-OC	-
OC-E-8B	55	E-OC	-

Table 1. OE: ovarian endometriosis, OC: ovarian cancer, E-OC: endometrioid ovarian carcinoma, CC-OC: clear cell ovarian carcinoma, sample names are assigned a letter A or B to categorize the groups: A (Endo-seq) and B (Synchron-seq).

Preprocessing and variant calling

Paired somatic variant calling for both WES datasets was performed using Sarek version 3.2.3. The Sarek pipeline executed preprocessing steps on FastQ input files automatically in the following order: quality control and trimming using FastQC (Wingett & Andrews, 2018) and FastP (Chen *et al.*, 2018), alignment to the GRCh38 reference genome with BWA-MEM2 (Li & Durbin; 2009), and BAM files processing through GATK's MarkDuplicates, BaseRecalibrator, GATKApplyBQSR (Van der Auwera *et al.*, 2013). Three variant calling tools, FreeBayes, Mutect2, and Strelka, were set in this pipeline, and Singularity was selected as the container platform (Garrison & Marth, 2012; Cibulskis *et al.*, 2013; Saunders *et al.*, 2012).

The genomic loci of variants were annotated with SnpEff in the Sarek pipeline. Each dataset was run independently, and the Synchron-seq cohort was run using two lanes of the same samples. Afterward, the Strelka VCF output files had to be extended by manually calculating genotype (GT), alternative and reference allele depths (ALT AD and REF AD), and alternative allele frequency (AF) since these are not reported in this tool. The latter step was performed in Rstudio using the *VariantAnnotanion* package and applying the formula recommended in the Strelka user guide: AF = ALT AD / (REF AD + ALT AD). Next, the VCF files from the three variant callers were processed with *vcfeval* from RTG Tools version 3.12.1 to select true positive mutations in the VCF files (Cleary *et al.*, 2015).

Three criteria for identifying somatic mutations in test samples were applied: (1) variants marked as PASS in FreeBayes with a quality score over 20 and at least one PASS in either Strelka or Mutect2, (2) an alternative allele frequency above 0.5% and a read depth of at least 5, and (3) the AF and ALT AD in the normal reference sample equal to 0.

The methodology described above follows the same criteria used by Linder *et al.* (2024). The variants found within exonic regions by SnpEff were excluded from the current study (refer to Appendix.1 for a list of the omitted exonic areas). After filtering, the VCF outputs were merged into an Excel-generated dataset.

Identification of eQTLs and functional annotation of the datasets

Gene annotations predicted by SnpEff were compared with eQTLs from all tissues in the GTEx v.8 dataset (GTEx Consortium, 2017). An adapted Python script, based on Chen *et al.* (2022), extracted eQTLs within <10kb of the gene's transcription start site (TSS). These were then included in a separate file to intersect with variants of the datasets generated in the previous section (code links are in Appendix II).

Additionally, the VCF files were used as input for functional annotation with the Sei framework (Chen *et al.*, 2022). A Google Colab T4 runtime was used to run the Sei model scripts following the instructions in the GitHub repository (https://github.com/FunctionLab/sei-framework). The framework assigned a predicted ‘sequence class’ to each SNV based on the absolute effect score of the possible predictions. The extreme values were extracted from columns 3-8 of the outputs, and the resulting files were subsequently merged with the Excel datasets generated in the previous section to add the ‘sequence class’ labels to the labels. These labels were later used to count and identify possible affected regulatory elements.

Identification of common variants

To perform this analysis, the annotated datasets from both cohorts were combined. An exploratory assessment to detect common mutations was performed by manually intersecting subgroups of OE and OC samples from the same cohorts. For instance, OE samples related to EOC were grouped and intersected with OE samples linked to CCOC cases (code details are found in Appendix II).

Recurrent mutations were identified by constructing a contingency table with variants and samples as variables. Subsequently, variants occurring in more than three samples were extracted to form a binary matrix. For visualization of the variants, hierarchical clustering was performed on the matrix using the *pheatmap* R package, version 1.0.12. Three parameters were important to set the pheatmap() function: clustering method: 'single,' clustering distance rows: 'binary,' and clustering distance columns: 'binary' (refer to Appendix II for details regarding the code).

Results

Whole-exome sequencing data processing and annotation

The general statistics table in the MultiQC reports for both study cohorts showed that the library size of most samples was above 40 million reads after trimming and filtering. The report also indicated that most samples had 1% or more genome coverage with 30X depth (see Table 1 from Appendix III). Only three samples, OE-03A, OE-05A, and OC-E-05A, were below these values. Both lanes of the Synchron-seq cohort were checked separately and showed similar quality metrics. The median fraction for duplicated reads and GC% content in both cohorts was 6.3% (1.5-25.7) and 52.15% (49.1-54.4), respectively (refer to Table 2 from Appendix III). The sequence quality histograms reported by FastP displayed an overall good quality of the base calls after trimming and filtering low-quality scores, with Phred scores above 29.8 for the Endo-seq cohort and scores over 35 in the Synchron-seq (see Appendix IV, Figures 1A and 1B). The per-sequence GC content report showed uneven distribution curves in samples from both datasets (Appendix IV, Figure 2).

After applying the filtering criteria, 11774 somatic non-coding mutations were identified in both cohorts (Endo-seq = 6793 and Synchron-seq = 4981), with a median sequence depth of 33 (range 5-591). Assessment of variants per sequence region showed that 99.13% were distributed in seven different types of genomic regions. Most were intron variants, with 59.78% (n=7039) of the total. Variants upstream of genes represented 15.43% (n=1817), while those downstream were 10.73% (n=1263). The remaining variants were distributed among other types: 3' UTR variants (4.04%), 5' UTR variants (2.10%), and variants in intergenic regions (3.57%) (see Figure 3).

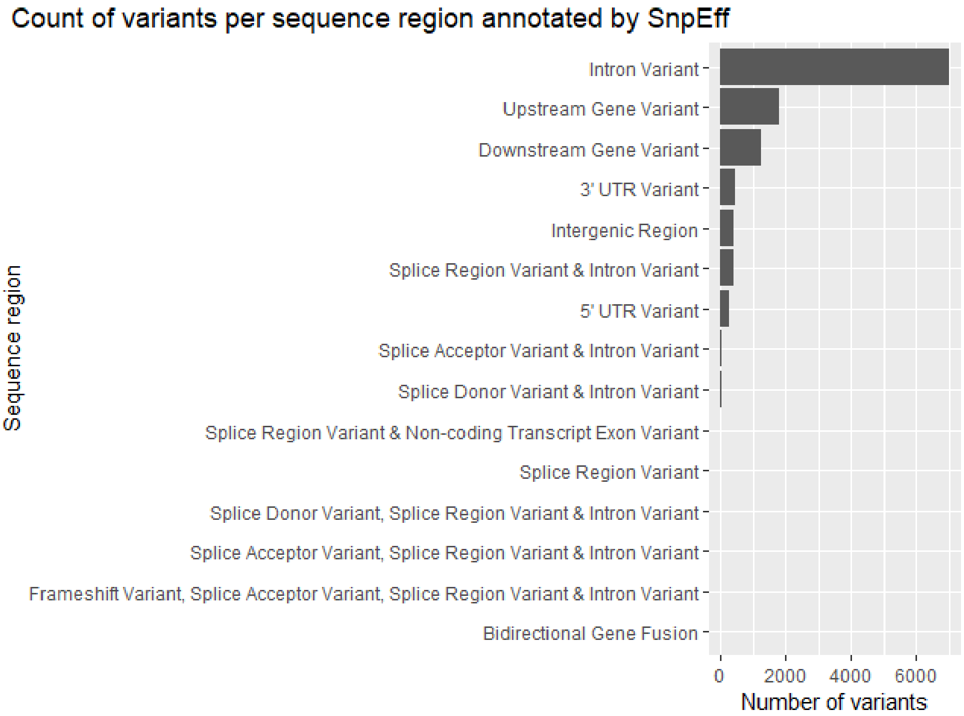


Figure 3. **Distribution of variants across sequence regions annotated by SnpEff.** The bars represent the number of genetic variants found in different genomic regions.

Out of 416 mutations overlapping with eQTLs in the GTEx dataset, a comparison with genes labeled from SnpEff showed that 71.9% (133 of 185) of intronic variants corresponded to the same genes by both resources. Similarly, most mutations categorized as 3' UTR, splice region, and 5' UTR variants were also linked to an eQTL involving the same transcript predicted. In contrast, downstream and upstream variants showed lower concordance rates, at 33.3% (27 out of 81) and 29.3% (22 out of 75), respectively (see Figure 4.).

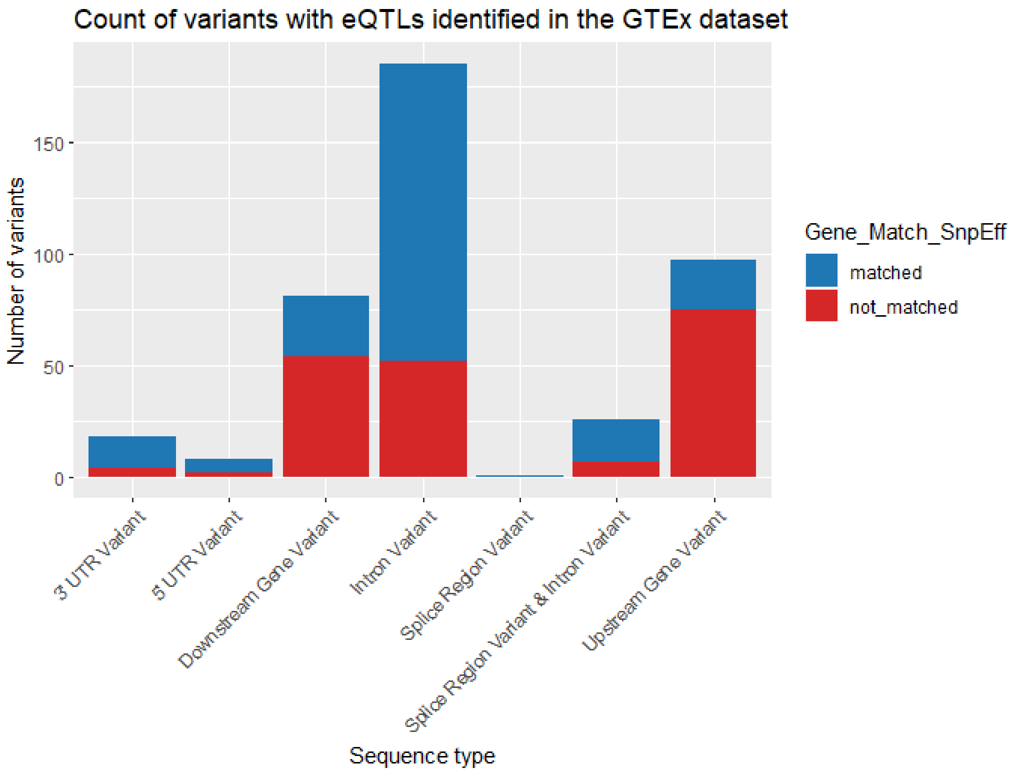


Figure 4. **Distribution of variants with eQTLs identified in the GTEx dataset.** The bar graph displays the number of variants associated with eQTLs across different sequence types. The data differentiates between matched (blue) and not matched (red) genes in relation to the gene labels from SnpEff with the genes affected by eQTLs. The total number of variants identified per cohort was Endo-seq: 399 and Synchron-seq: 17.

Moreover, the Sei model classified 49.2% (n=5794) of the variants as overlapping REs. These variants were distributed as follows: enhancer regions contained the highest number of variants with (2579, 21.9%), followed by polycomb complexes with (1697, 14.4%), and transcription factor binding sites with (1019, 8.65%). CTCF cohesin variants overlapped in 425 regions (3.61%), and promoter sequences had 74 variants (0.629%). The remaining mutations (50.8%) were categorized as variants lacking sufficient signal for classification, transcription sequences, and sequences within heterochromatin-rich regions (see Figure 5).

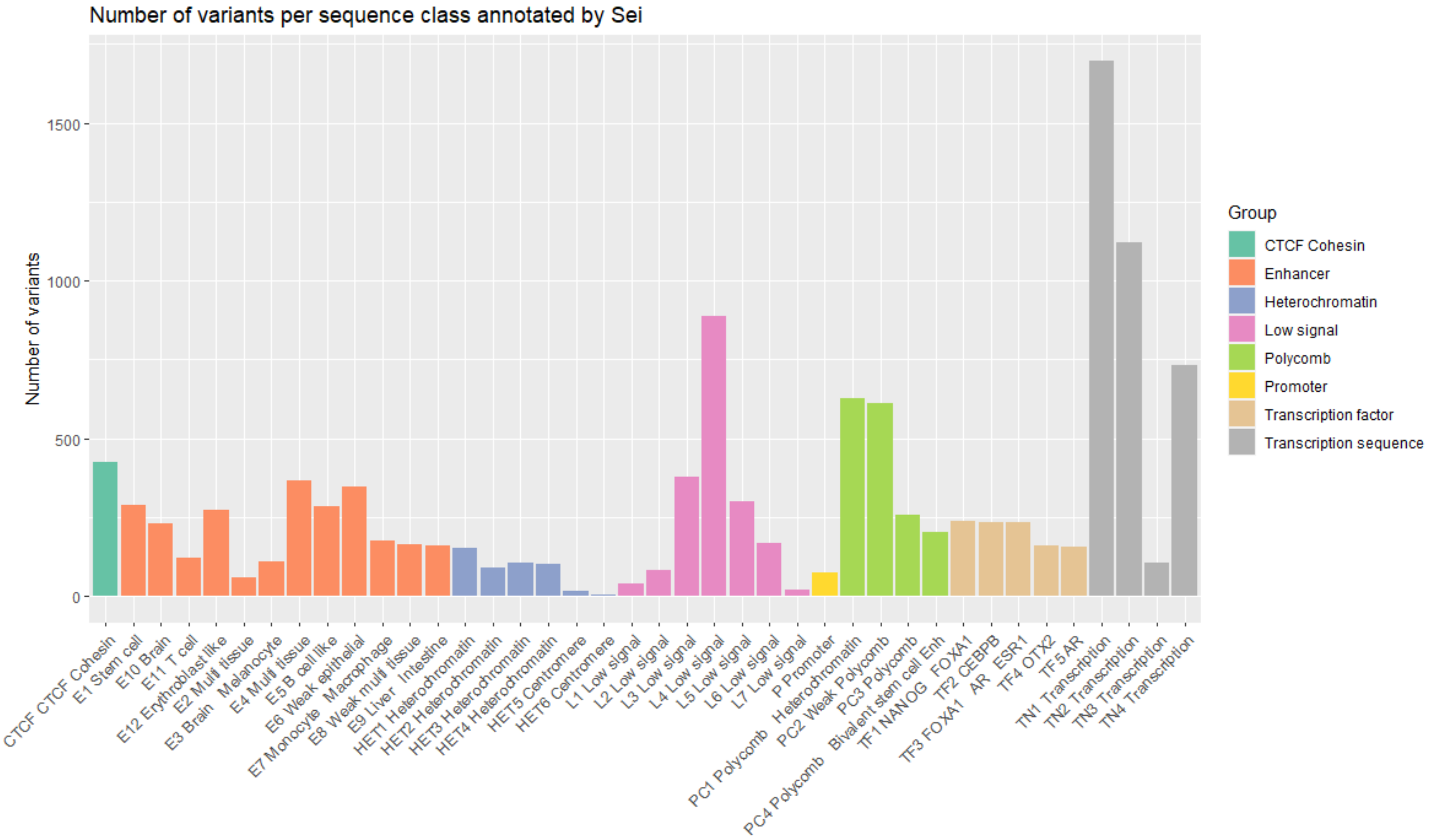


Figure 5. **Number of variants per Sequence Class annotated by Sei.** The bar graph categorizes the number of genetic variants across different sequence classes, with annotations grouped into categories such as CTCF Cohesin, Enhancer (E), Heterochromatin (HET), Low Signal (L), Polycomb (PC), Promoter (P), Transcription Factor (TF) and Transcription Sequence (TN).

Analysis of common variants

The analysis aimed to identify specific mutations and their potential genes affected within the disease cohorts. The existence of common mutations was first identified through the manual intersection of subgroups of samples for both cohorts separately. The subsequent selection of recurrent variants per sample was performed through a contingency table which included both cohorts. A total of 50 different variants, including SNVs and indels, present in more than three samples, which were represented in a heatmap. Using unsupervised hierarchical clustering based on mutation profiles to a large extent reflected the disease group of the samples (see Figure 6).

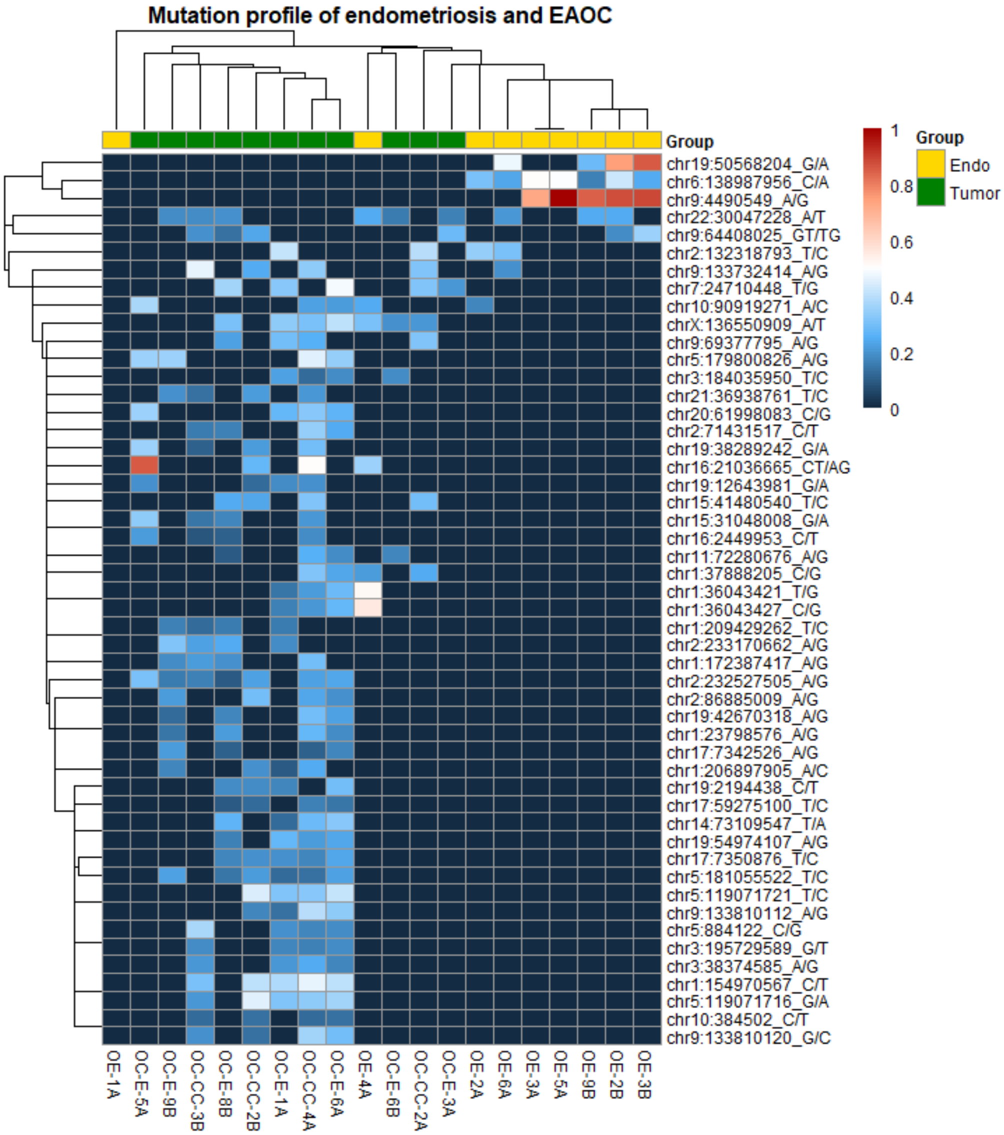


Figure 6. **Heatmap of common mutations in OE and OC with frequency variants in >3 samples per group.** The heatmap shows common mutations in OE and OC, with variants occurring in >3 samples per group. Rows represent variants by chromosomal position, while columns represent samples. The color gradient indicates variant allele frequencies (VAF), from blue (VAF=0) to red (VAF=1). "E" or "CC" identifies OC histotypes, with A or B suffixes distinguishing Endo-seq and Synchron-seq samples.

In the OE group, three variants were consistently identified. An upstream variant, chr19:50568204-G>A–*LRRC4B*, present in 4 OE specimens (Endo-seq=1 and Synchron-seq=3), and two 5'UTR located variants: chr6:138987956-C>A–*REPS1* found in 7 OE samples (Endo-seq=4 and Synchron-seq=3), and chr9:4490549-A>G–*SLC1A1* observed in 5 OE samples (Endo-seq=2 and Synchron-seq=3). The median variant allele frequencies (VAF) for these mutations were: *LRRC4B*: 0.6 (range 0.27-0.86), *REPS1*: 0.31 (range 0.17-0.5), and *SLC1A1*: 0.87 (range 0.72-0.89). These three mutations were labeled as enhancer sequences (refer to Table 1 in Appendix IV for detailed information on the variants).

Furthermore, three other variants were identified uniformly distributed across OE and OC groups: chr22:30047228-A>T–*CNN2P1* (VAF: 0.19, range 0.16-0.25), predicted as an enhancer sequence located upstream of the gene and found in 9 samples (OE=4 and OC=5); chr9:64408025_GT>TG–*ANKRD20A4P* (VAF: 0.22, range 0.13-0.35) classified as a transcription factor region and seen in 6 samples (OE=2 and OC=3); and chr2:132318793_T>C–*ZNF806* (VAF: 0.37, range 0.31-0.42) labeled as an enhancer sequence and present in 4 samples (OE=2 and OC=2). All three variants (*CNN2P1*=1, *ANKRD20A4P*=2, *ZNF806*=1) were also found in the paired OE-OC sample (see Table 2 for details).

Table 2. Common mutations within OE and OC specimens.

Chr	Pos	Ref	Alt	Gene	OE (A)	OE (B)	EOC	CCOC	Total
Ovarian endometriosis									
19	50568204	G	A	LRRC4B	1	3	0	0	4
6	138987956	C	A	REPS1	4	3	0	0	7
9	4490549	A	G	SLC1A1	2	3	0	0	5
Ovarian endometriosis and ovarian cancer									
22	30047228	A	T	CNN2P1	2*	2	4*	1	9 (1)
9	64408025	GT	TG	ANKRD20A4P	0	2*	2	2*	6 (2)
2	132318793	T	C	ZNF806	2*	0	1	1*	4 (1)
Ovarian cancer									
9	133732414	A	G	SARDH	1	0	0	4	5
7	24710448	T	G	GSDME	0	0	4	1	5
10	90919271	A	C	ANKRD1	2	0	2	1	5
X	136550909	A	T	VGLL1	1*	0	4	2*	7 (1)
2	232527505	A	G	PRSS56	0	0	4	3	7
5	181055522	T	C	BTNL9	0	0	4	2	6
1	154970567	C	T	SHC1	0	0	2	3	5
5	119071716	G	A	DMXL1-DT	0	0	2	3	5

*Had at least one pair of variants belonging to the same OE-OC case. Numbers in parentheses represent the count of paired cases.

An additional 44 mutations were found in 3 or more OC samples. The following were found in 7 samples and identified among the different OC histotypes: chrX:136550909-A>T–*VGLL1* (VAF: 0.31 [range 0.19-0.41]) distributed in CCOC (n=2), EOC (n=4) and OE (n=1); chr2:232527505-A>G–*PRSS56* (VAF: 0.22 [range 0.09-0.31]) found in CCOC (n=3) and EOC (n=4) (refer to Table 2.). These variants were labeled as transcription sequence and polycomb complex region, respectively. The VGLL1 variant was also found in the paired sample. Gene loci with specific variants found in five or more samples included: *BTNL9* (n=6 [EOC=4, CCOC=2]), *SHC1* (n=5 [EOC=2, CCOC=3]), *GSDME* (n=5 [EOC=2, CCOC=2]), *DMXL1-DT* (n=5 [EOC=2, CCOC=3]), *ANKRD1* [n=5 (OE=2, EOC=2, CCOC=1)], and *SARDH* (n=5 [OE=1, CCOC=4]).

Discussion

Despite advances in studying disease driver genes shared between OE and OC, the functional role of non-coding mutations in the pathogenesis of these two diseases has been less studied. A prior WES study made by Linder *et al.*, (2024) on the Endo-seq cohort reported that 87% of the somatic mutations were found outside the protein-coding region, which aligns with previous WGS data from OC studies reporting a ~98% (Corona et al., 2020). Given the relevance of REs for gene expression, this study aimed to identify non-coding mutations in OE and EAOC using whole-exome sequencing data from two cohorts with different times of OE removal in relation to the diagnosis of OC.

After whole-exome sequencing and applying filtering criteria, a substantial number of somatic mutations were identified, with 49.2% of these predicted to be within regulatory elements. This highlights the potential role of polygenic interactions being affected. Moreover, the use of the GTEx dataset, despite its limitation to non-pathological tissues, suggests that many of these mutations may overlap with eQTLs, emphasizing the role of REs in gene expression.

The higher number of genes from the GTEx dataset that coincided with those predicted by the SnpEff annotator in non-coding regions near the gene locus compared to regions further upstream and downstream suggests that regulatory mutations are more likely to affect genes in closer proximity. These discrepancies could be explained by pseudogenes, RNA genes, and microRNAs labeled by SnpEff or in the GTEx dataset. Studies have also described REs interacting with genes at a long distance, which makes predicting the genes affected even harder (Claussnitzer *et al.*, 2015).

The findings from the Sei annotation included sequences labeled as enhancers, CTCF, promoters, polycomb, and transcription factors, which could suggest their functional relevance as proposed by Chen *et al.* (2022). Similar approaches using Chip-Seq to identify regulatory regions in OE and OC have been previously explored through a GWAS metanalysis (Mortlock *et al.,* 2022) and WGS data (Corona *et al.,* 2020). Even though both studies used Chip-Seq data, the number of complete epigenomic profiles used (11 and 20, respectively) was considerably lower than the number of sequences used for training the Sei model. These differences, plus the possibility to make predictions per individual variant sequence, allowed for predictions from an automatic and wide analysis, enabling an overview of the regulatory DNA in a scalable manner for the current study. However, low signals and other transcription sequence labels were frequent and are still challenging to interpret functionally. The transcription factors available for prediction are limited to five types depending on their enrichment, NANOG FOXA1, CEBPB, FOXA1 AR ESR1, OTX2, and AR, which means the model could benefit from more training data, as well as other improvements proposed by Kundaje and Meuleman (2022). At the same time, there wasn't a notable difference in distribution for specific transcription factor variants (refer to Figure 5.).

Furthermore, the analysis by Mortlock et al. (2022) integrated data from genotyping of blood and saliva samples, focusing on unrelated cohorts of endometriosis and OC histotypes. This contrasts with the current study, which obtained the samples directly from lesions, with both the OE and OC samples coming from the same case.

The analysis of common variants showed a pattern in the distribution of specific SNVs and indels across different sample groups. Notably, variants such as *LRRC4B*, *REPS1*, and *SLC1A1* were frequently observed in the OE group, while *CNN2P1*, *ANKRD20A4P*, and *ZNF806* appeared across both OE and OC groups. *VGLL1* and *PRSS56* had the most recurrent mutations in OC samples. These findings underscore the heterogeneity between the OE and OC groups and further functional studies are required to understand the biological importance of these recurrent variants.

The SLC1A1 variant showed the highest VAF of 0.87 (range 0.73-1), suggesting a positive selection for this SNV. Both this variant and the *REPS1* variants were located at the 5' UTR of their respective genes and classified as enhancers, which in turn could potentially impact gene expression (Dvir *et al.*, 2013). Matsuzaki et al. (2005) previously reported downregulation of *SLC1A1* in stromal cells of eutopic endometrium in patients with endometriosis compared to control endometrium. However, it is noteworthy that the cells selected for the current study come from the epithelium of endometriosis lesions. Furthermore, studies in Alzheimer's disease have reported a negative correlation in *REPS1* expression with immune cell infiltration (Luo *et al.*, 2022), and other lines of studies have described mutated forms of *REPS1* protein as potential neoantigens (Yadav *et al.*, 2014; Shae *et al.*, 2020). Moreover, *LRRC4B* was found in 3 out of 3 Synchron-seq OE samples versus 1 out of 6 of Endo-seq samples, suggesting that these variants could be acquired over time. Leucine-rich repeat (LRR) proteins have been mostly related to the development of the nervous system, and their function in other tissues is lacking evidence (Feng *et al.*, 2020).

CNN2P1 was the variant most commonly identified between the OE and OC groups. Similar to *ANKRD20A4P*, these are pseudogenes, for which it wasn't possible to find literature linking them to OE or cancer at the time of this analysis. However, some lines of research have described possible ways in which pseudogenes could affect the expression of their gene counterpart (Sasidharan *et al.*, 2008).

The *VGLL1* and *PRSS56* mutations were the most common in the OC group. Previous studies have associated *VGLL1* with a possible role in the development of various cancer types, including pancreatic and basal-like breast cancers (Sonnemann, *et al*. 2023). The sequence class prediction and further visualization in the UCSC Genome Browser (Nassar *et al.*, 2023) suggest that this variant is located in a microRNA transcription sequence rather than a specific RE. On the other hand, the biological association of *PRSS56* with OE or cancer remains unknown.

At least three studies analyzing WES data that focused on endometriosis were sequenced using a different exon kit than the one used for both cohorts in this study and described by Linder *et al*. (2024) (Li *et al.*, 2021; Wu *et al.*, 2019; Xiaolei *et al.*, 2014). Moreover, WGS studies on endometriosis have only reported SNVs in exonic regions (Ward *et al.*, 2012). This could partially explain the absence of these non-coding variants in their reports.

Due to their potential relevance in OE diagnosis and the number of mutated samples, the *REPS1* and *SLC1A1* are proposed as candidates for biomarkers. Predicting computationally the impact of genetic variants is still a challenging task. Therefore, future directions should aim to test these findings for experimental validation. Potential methods for functional validation of non-coding mutations could include introducing the mutated sequence into experimental cells, quantifying the effect on transcription using high-throughput RNA sequencing, and demonstrating the biological impact by looking for key features of the disease (Khurana *et al.*, 2016).

This study has many limitations. First, the small sample size may limit the generalizability of the findings. Larger studies, including negative controls (OE cases without EAOC), would be needed to confirm the results. Although numerous variants were found in this study, WES is not optimal for studying non-coding mutations as it targets exonic sequences. WGS would provide a more comprehensive analysis of non-coding regions. Moreover, enhancers and other REs can interact with genes spanning large genomic distances (Claussnitzer *et al.*, 2015). Overall, the exact mechanisms these REs interact with genes are still unknown.

Finally, this study has used some of the newest tools in bioinformatics to characterize mutations identified in OE and OC. This project also complements the previous research conducted by Linder et al. (2024), which was the first to assess genetic variations in cases of OE with a subsequent OC diagnosis. Similar to the latter, the findings here presented did not find a clear genomic association between OE and OC that could suggest the clonality of both groups of cells. The results of this report offer a new perspective on potential biomarkers located in non-coding regions of the genome, an area of research still in its infancy. Despite the limitations discussed, the genetic variants identified here could be relevant for clinical applications, particularly given the ongoing challenges in diagnosing endometriosis. These findings should motivate further analysis and experimental validation.

Ethical aspects, gender perspectives, and impact on the society

The present study will adhere to the Declaration of Helsinki and has obtained approval from the Gothenburg local ethics committee (Dnr 201-15 and T522-17). Data collection from Sahlgrenska University Hospital included women who provided written informed consent. The raw data (FastQ files) was processed on a secured server provided by Gothenburg University.

In addition to the aims already presented, this project also seeks to raise awareness of the gender gap in research on women's health. Endometriosis is a relatively unknown and understudied disease. Confirmation of endometriosis usually involves surgical laparoscopic visualization and histologic analysis, highlighting the need for new non-invasive diagnostic methods.

Moreover, there is no treatment cure for this disease. Early prevention screening approaches to detect endometriosis could have a significant impact on society due to its association with ovarian cancer.

Future perspectives

An interesting finding of this research was the high number of mutations overlapping REs. Added to the lack of specific mutations in common between OE and OC, a polygenic approach to search for other forms of interaction between OE and OC could be implemented. Such methods could include network enrichment analysis to assess potential pathways disrupted. Similar approaches to studying non-coding variants have been applied to study autism cases by Zhou et al. (2019).

Previous works have associated 5' UTR variants that have significant effects on protein expression (Dvir et al., 2013). These regions may be rich in promoter or enhancer sequences that play a significant role in gene expression (Chen et al., 2022). Given their frequency, variants in *REPS1* and *SLC1A1* should be considered for experimental validation. Methods for functional validation could involve introducing the mutated sequence into experimental cells, measuring the impact on transcription using high-throughput RNA sequencing, and demonstrating the biological effects by examining key disease-related features (Khurana et al., 2016).

Acknowledgments

I would like to thank my supervisors, Anna Linder and Benjamin Ulfenborg, for their guidance and patience during this research. Their expertise and feedback were crucial in shaping this work.

I am also thankful to Karin Sundfeldt for allowing me to join the lab for my thesis and for her support in facilitating the needs of this research. I also extend my thanks to Gothenburg University for providing me access to their remote server.

My sincere thanks also go to the friends I have made during this program, especially Patricia, Dhanya, Pablo, and Rajeesh, for their companionship during challenging moments.

I'm especially grateful with my brother, José Gabriel who has supported my journey to Sweden and encouraged me to achieve this goal. I am incredibly fortunate to have you as a guide.

Lastly, I dedicate this thesis to my family—Bethania, José, Vanessa, JG, Tía María, and Sofía. Thank you for your unwavering support and love.

References

Anglesio, M. S., Bashashati, A., Wang, Y. K., Senz, J., Ha, G., Yang, W., Aniba, M.R., Prentice, L.M., Farahani, H., Li Chang, H., Karnezis, A.N., Marra, M.A., Yong, P.J., Hirst, M., Gilks, B., Shah, S.P., & Huntsman, D. G. (2015). Multifocal endometriotic lesions associated with cancer are clonal and carry a high mutation burden. *Journal of Pathology, 236*(2), 201-209.
https://doi.org/10.1002/path.4516

Anglesio, M. S., Papadopoulos, N., Ayhan, A., Nazeran, T. M., Noë, M., Horlings, H. M., Lum, A., Jones, S., Senz, J., Seckin, T., Ho, J., Wu, R.C., Lac, V., Ogawa, H., Tessier-Cloutier, B., Alhassan, R., Wang, A., Wang, Y., Cohen, J.D., Wong, F., Hasanovic, A., Orr, N., Zhang, M., Popoli, M., McMahon, W., Wood, L.D., Mattox, A., Allaire, C., Segars, J., Williams, C., Tomasetti, C., Boyd, N., Kinzler, K.W., Gilks, C.B., Diaz, L., Wang, T.L., Vogelstein, B., Yong, P.J., Huntsman, D.G., & Shih, I. M. (2017). Cancer-Associated Mutations in Endometriosis without Cancer. *New England Journal of Medicine, 376*(19), 1835-1848.
https://doi.org/10.1056/NEJMoa1614814

Boeva V. (2016). Analysis of Genomic Sequence Motifs for Deciphering Transcription Factor Binding and Transcriptional Regulation in Eukaryotic Cells. *Frontiers in genetics, 7*, 24.
https://doi.org/10.3389/fgene.2016.00024

Chen, K.M., Wong, A.K., Troyanskaya, O.G., & Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat Genet 54*, 940–949 (2022).
https://doi.org/10.1038/s41588-022-01102-2

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics (Oxford, England), 34(17), i884–i890.
https://doi.org/10.1093/bioinformatics/bty560

Cibulskis, K., Lawrence, M., Carter, S. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol 31*, 213–219 (2013).
https://doi.org/10.1038/nbt.2514

Cingolani, P., Platts, A., Wang, leL., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly, 6*(2), 80–92.
https://doi.org/10.4161/fly.19695

Claussnitzer, M., Dankel, S. N., Kim, K. H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puviindran, V., Abdennur, N. A., Liu, J., Svensson, P. A., Hsu, Y. H., Drucker, D. J., Mellgren, G., Hui, C. C., Hauner, H., & Kellis, M. (2015). FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *The New England journal of medicine, 373*(10), 895–907.
https://doi.org/10.1056/NEJMoa1502214

Cleary, J. G., Braithwaite, R., Gaastra, K., Hilbush, B. S., Inglis, S., Irvine, S. A., ... & De La Vega, F. M. (2015). Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. *BioRxiv*, 023754.

Corona, R. I., Seo, J. H., Lin, X., Hazelett, D. J., Reddy, J., Fonseca, M. A. S., Abassi, F., Lin, Y. G., Mhawech-Fauceglia, P. Y., Shah, S. P., Huntsman, D. G., Gusev, A., Karlan, B. Y., Berman, B. P., Freedman, M. L., Gayther, S. A., & Lawrenson, K. (2020). Non-coding somatic mutations converge on the PAX8 pathway in ovarian cancer. *Nature communications, 11*(1), 2020. https://doi.org/10.1038/s41467-020-15951-0

Drake, J. A., Bird, C., Nemesh, J., Thomas, D. J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S. E., Dermitzakis, E. T., & Hirschhorn, J. N. (2006). Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nature genetics, 38*(2), 223–227. https://doi.org/10.1038/ng1710

Dvir, S., Velten, L., Sharon, E., Zeevi, D., Carey, L. B., Weinberger, A., & Segal, E. (2013). Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. *Proceedings of the National Academy of Sciences of the United States of America, 110*(30), E2792–E2801.
https://doi.org/10.1073/pnas.1222534110

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., & Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature, 473*(7345), 43–49.
https://doi.org/10.1038/nature09906

Feng, J., Zhang, Y., Ren, X., Li, D., Fu, H., Liu, C., Zhou, W., Liu, Q., Liu, Q., & Wu, M. (2020). Leucine-rich repeat containing 4 act as an autophagy inhibitor that restores sensitivity of glioblastoma to temozolomide. *Oncogene, 39*(23), 4551–4566. https://doi.org/10.1038/s41388-020-1312-6

Garcia, M., Juhos, S., Larsson, M., Olason, P. I., Martin, M., Eisfeldt, J., DiLorenzo, S., Sandgren, J., Díaz De Ståhl, T., Ewels, P., Wirta, V., Nistér, M., Käller, M., & Nystedt, B. (2020). Sarek: A portable workflow for whole-genome sequencing analysis of germline and somatic variants. *F1000Research, 9*, 63. https://doi.org/10.12688/f1000research.16665.2

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

Giudice, L. C. (2010). Clinical practice. Endometriosis. *The New England Journal of Medicine, 362*(25), 2389-2398. https://doi.org/10.1056/NEJMcp1000274

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., … Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, *446*(7132), 153–158. https://doi.org/10.1038/nature05610

GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature 550*, 204–213 (2017). https://doi.org/10.1038/nature24277

Hermens, M., van Altena, A. M., Nieboer, T. E., Schoot, B. C., van Vliet, H. A. A. M., Siebers, A. G., & Bekkers, R. L. M. (2020). Incidence of endometrioid and clear-cell ovarian cancer in histological proven endometriosis: the ENOCA population-based cohort study. *American journal of obstetrics and gynecology, 223*(1), 107.e1–107.e11. https://doi.org/10.1016/j.ajog.2020.01.041

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America, 106*(23), 9362-9367. https://doi.org/10.1073/pnas.0903103106

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research, 26*(7), 990–999. https://doi.org/10.1101/gr.200535.115

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., & Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nature reviews. Genetics, 17*(2), 93–108. https://doi.org/10.1038/nrg.2015.17

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.), 188*(4184), 107–116. https://doi.org/10.1126/science.1090005

Koboldt D. C. (2020). Best practices for variant calling in clinical sequencing. *Genome medicine, 12*(1), 91. https://doi.org/10.1186/s13073-020-00791-w

Kundaje, A., & Meuleman, W. (2022). Automated sequence-based annotation and interpretation of the human genome. *Nature genetics*, *54*(7), 916–917. https://doi.org/10.1038/s41588-022-01123-x

Latchman D. S. (1993). Transcription factors: an overview*. International journal of experimental pathology, 74*(5), 417–422.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England), 25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, L., Antero, M. F., Zhang, M., Chu, T., Seckin, T., Ayhan, A., Pisanic, T., Wang, T. L., Cope, L., Segars, J., & Shih, I. M. (2021). Mutation and methylation profiles of ectopic and eutopic endometrial tissues. *The Journal of pathology, 255*(4), 387–398. https://doi.org/10.1002/path.5778

Li, X., Zhang, Y., Zhao, L., Wang, L., Wu, Z., Mei, Q., Nie, J., Li, X., Li, Y., Fu, X., Wang, X., Meng, Y., & Han, W. (2014). Whole-exome sequencing of endometriosis identifies frequent alterations in genes involved in cell adhesion and chromatin-remodeling complexes. *Human molecular genetics, 23*(22), 6008–6021. https://doi.org/10.1093/hmg/ddu330

Linder, A., Westbom-Fremer, S., Mateoiu, C., Olsson Widjaja, A., Österlund, T., Veerla, S., Ståhlberg, A., Ulfenborg, B., Hedenfalk, I., & Sundfeldt, K. (2024). Genomic alterations in ovarian endometriosis and subsequently diagnosed ovarian carcinoma. *Human reproduction (Oxford, England)*, *39*(5), 1141–1154. https://doi.org/10.1093/humrep/deae043

Luo, J., Chen, L., Huang, X., Xie, J., Zou, C., Pan, M., Mo, J., & Zou, D. (2022). REPS1 as a Potential Biomarker in Alzheimer's Disease and Vascular Dementia. *Frontiers in aging neuroscience*, *14*, 894824. https://doi.org/10.3389/fnagi.2022.894824

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., Wedge, D. C., Fullam, A., Alexandrov, L. B., Tubio, J. M., Stebbings, L., Menzies, A., Widaa, S., Stratton, M. R., Jones, P. H., & Campbell, P. J. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (New York, N.Y.), 348*(6237), 880–886. https://doi.org/10.1126/science.aaa6806

Matsuzaki, S., Canis, M., Vaurs-Barrière, C., Boespflug-Tanguy, O., Dastugue, B., & Mage, G. (2005). DNA microarray analysis of gene expression in eutopic endometrium from patients with deep endometriosis using laser capture microdissection. *Fertility and sterility*, *84*, 1180-1190.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747–753. https://doi.org/10.1038/nature08494

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome biology, 17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

Mortlock, S., Corona, R. I., Kho, P. F., Pharoah, P., Seo, J. H., Freedman, M. L., Gayther, S. A., Siedhoff, M. T., Rogers, P. A. W., Leuchter, R., Walsh, C. S., Cass, I., Karlan, B. Y., Rimel, B. J., Ovarian Cancer Association Consortium, International Endometriosis Genetics Consortium, Montgomery, G. W., Lawrenson, K., & Kar, S. P. (2022). A multi-level investigation of the genetic relationship between endometriosis and ovarian cancer histotypes. *Cell reports. Medicine*, *3*(3), 100542. https://doi.org/10.1016/j.xcrm.2022.100542

Nakagawa, H., & Fujita, M. (2018). Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer science, 109*(3), 513–522. https://doi.org/10.1111/cas.13505

Nassar, L. R., Barber, G. P., Benet-Pagès, A., Casper, J., Clawson, H., Diekhans, M., Fischer, C., Gonzalez, J. N., Hinrichs, A. S., Lee, B. T., Lee, C. M., Muthuraman, P., Nguy, B., Pereira, T., Nejad, P., Perez, G., Raney, B. J., Schmelter, D., Speir, M. L., Wick, B. D., … Kent, W. J. (2023). The UCSC Genome Browser database: 2023 update. *Nucleic acids research*, *51*(D1), D1188–D1195. https://doi.org/10.1093/nar/gkac1072

Pearce, C. L., Templeman, C., Rossing, M. A., Lee, A., Near, A. M., Webb, P. M., Nagle, C. M., Doherty, J. A., Cushing-Haugen, K. L., Wicklund, K. G., Chang-Claude, J., Hein, R., Lurie, G., Wilkens, L. R., Carney, M. E., Goodman, M. T., Moysich, K., Kjaer, S. K., Hogdall, E., Jensen, A., … Ovarian Cancer Association Consortium (2012). Association between endometriosis and risk of histological subtypes of ovarian cancer: a pooled analysis of case-control studies. *The Lancet. Oncology, 13*(4), 385–394. https://doi.org/10.1016/S1470-2045(11)70404-1

Raphael, B. J., Dobson, J. R., Oesper, L., & Vandin, F. (2014). Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. *Genome medicine, 6*(1), 5. https://doi.org/10.1186/gm524

Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics (Oxford, England), 28*(14), 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Shae, D., Baljon, J. J., Wehbe, M., Christov, P. P., Becker, K. W., Kumar, A., Suryadevara, N., Carson, C. S., Palmer, C. R., Knight, F. C., Joyce, S., & Wilson, J. T. (2020). Co-delivery of Peptide Neoantigens and Stimulator of Interferon Genes Agonists Enhances Response to Cancer Vaccines. *ACS nano*, *14*(8), 9904–9916. https://doi.org/10.1021/acsnano.0c02765

Sasidharan, R., & Gerstein, M. (2008). Genomics: protein fossils live on as RNA. *Nature, 453*(7196), 729–731. https://doi.org/10.1038/453729a

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., Tse, T., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic acids research, 50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research, 29*(1), 308–311. https://doi.org/10.1093/nar/29.1.308

Sonnemann, H. M., Pazdrak, B., Antunes, D. A., Roszik, J., & Lizée, G. (2023). Vestigial-like 1 (VGLL1): An ancient co-transcriptional activator linking wing, placenta, and tumor development. *Biochimica et biophysica acta. Reviews on cancer*, *1878*(3), 188892. https://doi.org/10.1016/j.bbcan.2023.188892

Steinberg, M. H., & Sebastiani, P. (2012). Genetic modifiers of sickle cell disease. *American journal of hematology, 87*(8), 795–803. https://doi.org/10.1002/ajh.23232

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome*. Nature, 458*(7239), 719–724. https://doi.org/10.1038/nature07943

Suda, K., Nakaoka, H., Yoshihara, K., Ishiguro, T., Tamura, R., Mori, Y., Yamawaki, K., Adachi, S., Takahashi, T., Kase, H., Tanaka, K., Yamamoto, T., Motoyama, T., Inoue, I., & Enomoto, T. (2018). Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Reports, 24*(7), 1777-1789. https://doi.org/10.1016/j.celrep.2018.07.037

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics, 43*(1110), 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Ward, K., Chettier, R., Farrington, P., & Albertsen, H. (2012). Sipping from the firehose: complete genome sequencing of endometriosis patients. *Fertility and Sterility*, *98*(3), S68.

Ward, L. D., & Kellis, M. (2012). Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology, 30*(11), 1095-1106. https://doi.org/10.1038/nbt.2422

Wang, Y., Nicholes, K., & Shih, I. M. (2020). The Origin and Pathogenesis of Endometriosis*. Annual review of pathology, 15*, 71–95. https://doi.org/10.1146/annurev-pathmechdis-012419-032654

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research, 7*, 1338. https://doi.org/10.12688/f1000research.15931.2

Wittkopp, P., Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet 13*, 59–69 (2012). https://doi.org/10.1038/nrg3095

Wu, R. C., Wang, P., Lin, S. F., Zhang, M., Song, Q., Chu, T., ... & Wang, T. L. (2019). Genomic landscape and evolutionary trajectories of ovarian cancer precursor lesions. The Journal of pathology, 248(1), 41-50.

World Medical Association. (2013). *Declaration of Helsinki*. Declaration of Helsinki – WMA – The World Medical Association

Xu C. (2018). A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and structural biotechnology journal, 16*, 15–24. https://doi.org/10.1016/j.csbj.2018.01.003

Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., Franci, C., Cheung, T. K., Fritsche, J., Weinschenk, T., Modrusan, Z., Mellman, I., Lill, J. R., & Delamarre, L. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature, 515*(7528), 572–576. https://doi.org/10.1038/nature14001

Yachida, N., Yoshihara, K., Yamaguchi, M., Suda, K., Tamura, R., & Enomoto, T. (2021). How Does Endometriosis Lead to Ovarian Cancer? The Molecular Mechanism of Endometriosis-Associated

Ovarian Cancer Development. *Cancers (Basel), 13*(6), 1439. https://doi.org/10.3390/cancers13061439

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods, 12*(10), 931–934. https://doi.org/10.1038/nmeth.3547

Zhou, J., Park, C. Y., Theesfeld, C. L., Wong, A., Yuan, Y., Sheckel, Fak, J., Yao, K., Tajima, Y., Packer, A., & Darnell, R. (2019). Whole-genome deep-learning analysis identifies the contribution of noncoding mutations to autism risk. *Nature Genetics, 51*, 973-980. https://doi.org/10.1038/s41588-019-0420-0

Appendix I

Table1. Variants in exonic regions that were filtered out in this study.

Sequence region
5_prime_UTR_premature_start_codon_gain_variant
conservative_inframe_deletion
conservative_inframe_insertion
disruptive_inframe_deletion
disruptive_inframe_insertion
frameshift_variant
frameshift_variant&splice_region_variant
frameshift_variant&stop_gained
intragenic_variant
missense_variant
missense_variant&splice_region_variant
non_coding_transcript_exon_variant
start_lost
stop_gained
stop_gained&splice_region_variant
stop_lost
stop_retained_variant
synonymous_variant
splice_region_variant&synonymous_variant
splice_acceptor_variant&missense_variant&splice_region_variant&intron_variant

Appendix II

Scripts used for extraction of eQTLs and to run de Sei framework were created by Chen et al., (2022) and can be found at:

https://github.com/FunctionLab/sei-manuscript/tree/main/eQTL_effect_sizes

https://github.com/FunctionLab/sei-framework

All the scripts to reproduce the plots of this report can be found at:

https://github.com/manuelrujano/modig_thesis

Appendix III

Table 1. MultiQC general statistics report of genome coverage per sample

Cohort 1			Cohort 2 – lane 1			Cohort 2 – lane 2		
Sample Name	**≥ 10X**	**≥ 30X**	**Sample Name**	**≥ 10X**	**≥ 30X**	**Sample Name**	**≥ 10X**	**≥ 30X**
1E.md	4.00%	3.00%	**2A.md**	2.00%	2.00%	**2C.md**	3.00%	2.00%
1E.recal	4.00%	3.00%	**2A.recal**	2.00%	2.00%	**2C.recal**	3.00%	2.00%
1N.md	4.00%	3.00%	**2B.md**	3.00%	2.00%	**3A.md**	2.00%	1.00%
1N.recal	4.00%	3.00%	**2B.recal**	3.00%	2.00%	**3A.recal**	2.00%	1.00%
1O.md	3.00%	2.00%	**2C.md**	3.00%	2.00%	**3B.md**	3.00%	2.00%
1O.recal	3.00%	2.00%	**2C.recal**	3.00%	2.00%	**3B.recal**	3.00%	2.00%
2E.md	3.00%	2.00%	**3A.md**	2.00%	1.00%	**3C.md**	3.00%	2.00%
2E.recal	3.00%	2.00%	**3A.recal**	2.00%	1.00%	**3C.recal**	3.00%	2.00%
2N.md	3.00%	1.00%	**3B.md**	3.00%	2.00%	**6B.md**	3.00%	2.00%
2N.recal	3.00%	1.00%	**3B.recal**	3.00%	2.00%	**6B.recal**	3.00%	2.00%
2O.md	3.00%	2.00%	**3C.md**	3.00%	2.00%	**6C.md**	4.00%	2.00%
2O.recal	3.00%	2.00%	**3C.recal**	3.00%	2.00%	**6C.recal**	4.00%	2.00%
3E.md	2.00%		**6B.md**	3.00%	2.00%	**8B.md**	3.00%	1.00%
3E.recal	2.00%		**6B.recal**	3.00%	2.00%	**8B.recal**	3.00%	1.00%
3N.md	4.00%	2.00%	**6C.md**	4.00%	2.00%	**8C.md**	3.00%	2.00%
3N.recal	4.00%	2.00%	**6C.recal**	4.00%	2.00%	**8C.recal**	3.00%	2.00%
3O.md	4.00%	2.00%	**8B.md**	3.00%	1.00%	**9A.md**	2.00%	1.00%
3O.recal	4.00%	2.00%	**8B.recal**	3.00%	1.00%	**9A.recal**	2.00%	1.00%
4E.md	2.00%	1.00%	**8C.md**	3.00%	2.00%	**9B.md**	3.00%	2.00%
4E.recal	2.00%	1.00%	**8C.recal**	3.00%	2.00%	**9B.recal**	3.00%	2.00%
4N.md	3.00%	2.00%	**9A.md**	2.00%	1.00%	**9C.md**	3.00%	2.00%
4N.recal	3.00%	2.00%	**9A.recal**	2.00%	1.00%	**9C.recal**	3.00%	2.00%
4O.md	3.00%	2.00%	**9B.md**	3.00%	2.00%			
4O.recal	3.00%	2.00%	**9B.recal**	3.00%	2.00%			
5N.md	3.00%	2.00%	**9C.md**	3.00%	2.00%			
5N.recal	3.00%	2.00%	**9C.recal**	3.00%	2.00%			
5O.md	1.00%							
5O.recal	1.00%							
6E.md	3.00%	2.00%						
6E.recal	3.00%	2.00%						
6N.md	3.00%	2.00%						
6N.recal	3.00%	2.00%						
6O.md	3.00%	2.00%						
6O.recal	3.00%	2.00%						

≥ 10X: sequence coverage of 10X, ≥ 30X: sequence coverage of 30X. Letters assigned: E, O, N, A, B, and C are the initial sample processing names and may differ from Table 1 of the report.

Table 2. General statistics report the percentage of duplication, number of reads, and GC content

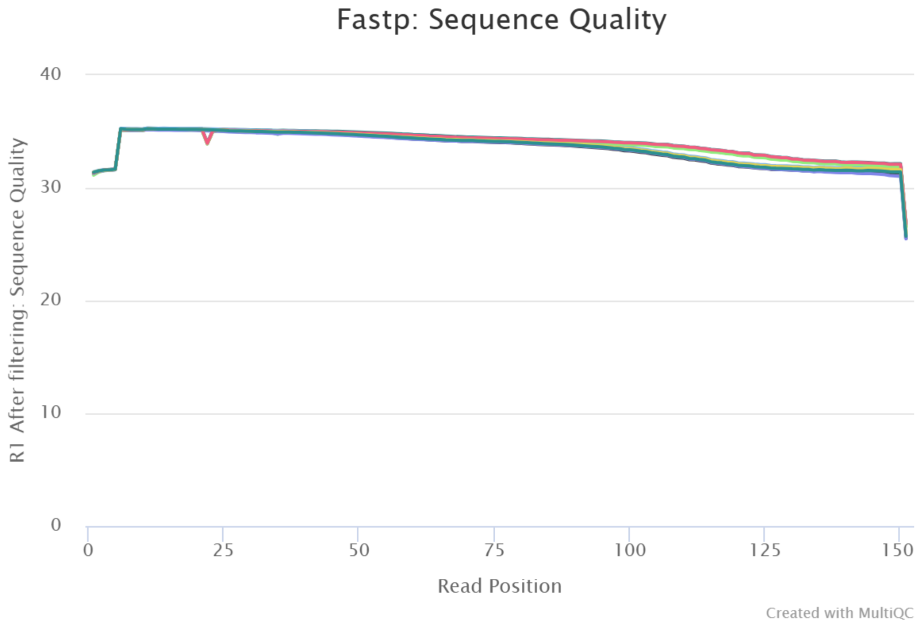
Sample Name	% Duplication	M Reads After Filtering	GC content	Sample Name	% Duplication	M Reads After Filtering	GC content
4O-lane_1	8.20%	139.6	53.80%	**2A-lane_1**	25.70%	125.9	50.10%
1N-lane_1	2.30%	130.4	49.30%	**9A-lane_1**	13.50%	71.6	51.80%
1E-lane_1	2.50%	128.6	51.70%	**2B-lane_1**	12.70%	98	51.90%
2E-lane_1	9.60%	93.3	52.80%	**3A-lane_1**	9.90%	59.7	49.20%
6O-lane_1	6.60%	93.3	52.10%	**9B-lane_1**	8.50%	67.5	52.90%
3N-lane_1	1.70%	83.9	49.70%	**6B-lane_1**	8.10%	60.9	53.70%
3O-lane_1	2.70%	81.2	49.10%	**3B-lane_1**	7.20%	61.3	49.90%
6N-lane_1	2.10%	79.7	50.00%	**9C-lane_1**	6.40%	88.2	51.90%
6E-lane_1	3.30%	71.2	50.30%	**6C-lane_1**	5.30%	104.7	52.40%
4E-lane_1	6.20%	67.2	54.40%	**8C-lane_1**	5.20%	87.2	52.20%
5N-lane_1	1.90%	63.1	49.90%	**2C-lane_1**	5.00%	95.3	52.40%
2O-lane_1	3.20%	61.8	52.40%	**8B-lane_1**	4.80%	49	52.90%
1O-lane_1	2.90%	59.2	51.20%	**3C-lane_1**	4.00%	66	52.70%
2N-lane_1	6.90%	48.9	51.10%	**9A-lane_2**	14.10%	73.3	51.80%
4N-lane_1	1.50%	47.8	50.50%	**3A-lane_2**	10.40%	61.1	49.20%
5O-lane_1	11.70%	37	52.90%	**9B-lane_2**	8.90%	69	52.90%
3E-lane_1	11.40%	36.6	54.10%	**6B-lane_2**	8.50%	62.1	53.70%
5E-lane_1	10.60%	13.2	53.50%	**3B-lane_2**	7.70%	62.6	49.90%
Min	**1.50%**		**49.10%**	**9C-lane_2**	6.80%	90	51.90%
Median	**3.25%**		**51.45%**	**6C-lane_2**	5.70%	106.9	52.40%

Max	**11.70%**		**54.40%**	**8C-lane_2**	5.50%	89	52.30%
				2C-lane_2	5.30%	97.2	52.40%
				8B-lane_2	5.20%	50.1	52.90%
				3C-lane_2	4.30%	67.1	52.70%
				Min	**4.00%**		**49.20%**
				Median	**13.80%**		**52.35%**
				Max	**25.70%**		**53.70%**

Appendix IV

Figure 1.

A)



B)

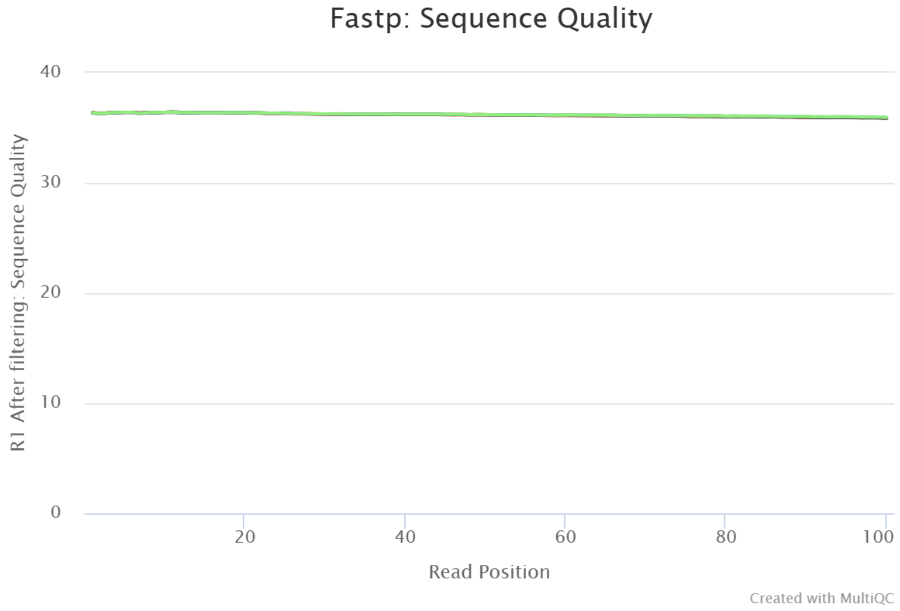


Figure 2.

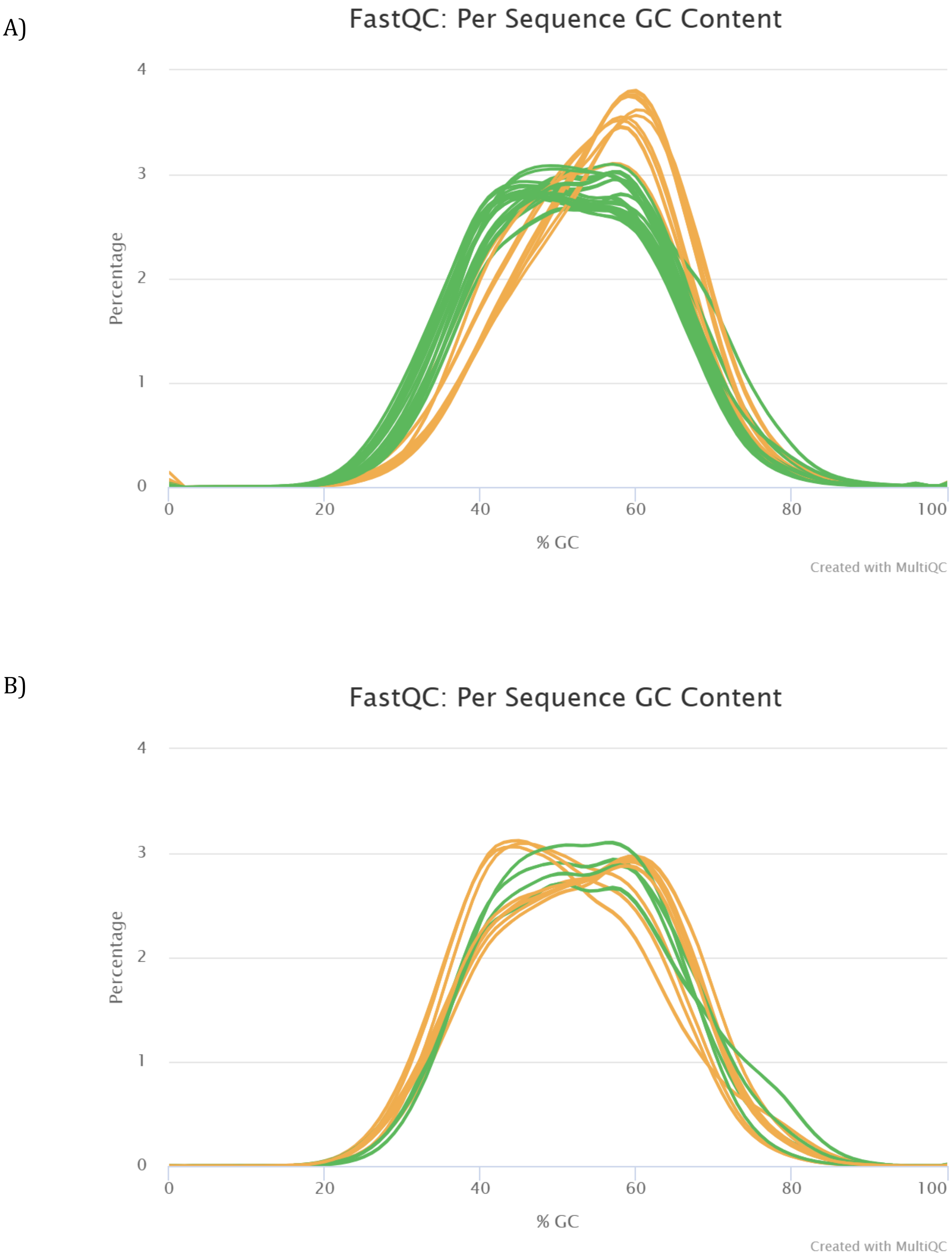
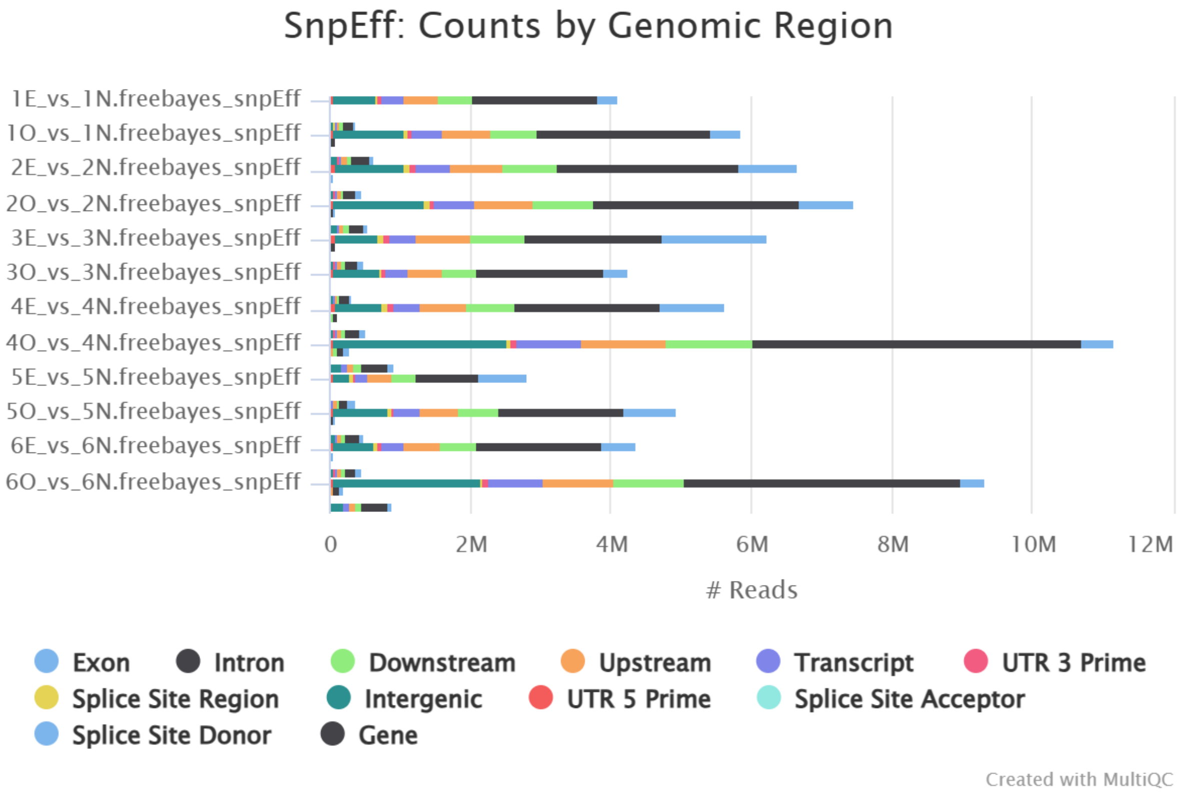

A)
FastQC: Per Sequence GC Content
Percentage
4
3
2
1
0
0
20
40
60
80
100
% GC
Created with MultiQC
B)
FastQC: Per Sequence GC Content
Percentage
4
3
2
1
0
0
20
40
60
80
100
% GC
Created with MultiQC

Figure 3. Counts by genomic regions reports:

A)



B)

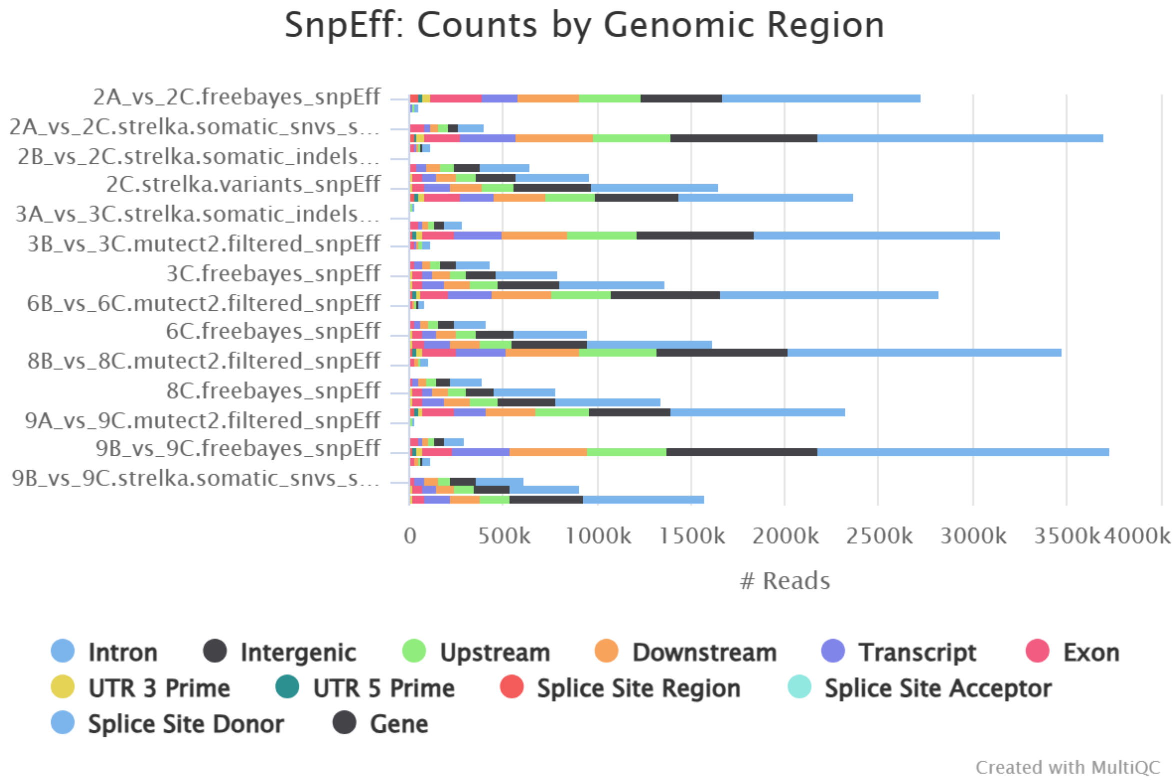


Table 1. Common mutations within OE and OC specimens. Frequency: variants in >3 samples per group.

Chr	**Pos**	**Ref**	**Alt**	**Gene**	**Type**	**Sequence class**	**VAF (Range)**
VAF: Median of the variant allele frequencies							
19	50568204	G	A	LRRC4B	Upstream Gene Variant	E1 Stem cell	0.61 (0.29-0.86)
6	138987956	C	A	REPS1	5' UTR Variant	E4 Multi tissue	0.31 (0.17-0.51)
9	4490549	A	G	SLC1A1	5' UTR Variant	E4 Multi tissue	0.87 (0.73-1)
22	30047228	A	T	CNN2P1	Upstream Gene Variant	E11 T cell	0.19 (0.16-0.25)
9	64408025	GT	TG	ANKRD20A4P	Intron Variant	TF1 NANOG FOXA1	0.22 (0.13-0.35)
2	132318793	T	C	ZNF806	Intron Variant	E12 Erythroblast like	0.37 (0.31-0.42)

9	133732414	A	G	SARDH	Intron Variant	L4 Low signal	0.31 (0.2-0.46)
7	24710448	T	G	GSDME	Intron Variant	E6 Weak epithelial	0.33 (0.21-0.48)
10	90919271	A	C	ANKRD1	Splice Region Variant & Intron Variant	TN2 Transcription	0.23 (0.18-0.38)
X	136550909	A	T	VGLL1	Intron Variant	TN1 Transcription	0.31 (0.19-0.41)
9	69377795	A	G	FAM189A2	Intron Variant	E11 T cell	0.28 (0.23-0.32)
5	179800826	A	G	MIR1229	Upstream Gene Variant	PC2 Weak Polycomb	0.35 (0.34-0.46)
3	184035950	T	C	HTR3D	Intron Variant	L4 Low signal	0.19 (0.13-0.22)
21	36938761	T	C	DPRXP5	Upstream Gene Variant	TN4 Transcription	0.2 (0.13-0.21)
20	61998083	C	G	TAF4	Intron Variant	TN2 Transcription	0.3 (0.27-0.36)
2	71431517	C	T	ZNF638	Intron Variant	TN2 Transcription	0.21 (0.16-0.35)
19	38289242	G	A	ENSG00000286037	Downstream Gene Variant	TN2 Transcription	0.26 (0.11-0.36)
16	21036665	CT	AG	DNAH3	Intron Variant	TF3 FOXA1 AR ESR1	0.43 (0.28-0.86)
19	12643981	G	A	RPL10P16	Upstream Gene Variant	TN4 Transcription	0.19 (0.13-0.2)
15	41480540	T	C	RTF1	Intron Variant	PC1 Polycomb Heterochromatin	0.27 (0.24-0.32)
15	31048008	G	A	TRPM1	Intron Variant	L4 Low signal	0.19 (0.15-0.33)
16	2449953	C	T	MIR6767	Downstream Gene Variant	TN4 Transcription	0.14 (0.09-0.21)
11	72280676	A	G	ENSG00000213365	Intron Variant	HET3 Heterochromatin	0.18 (0.09-0.26)
1	37888205	C	G	SNORA63	Downstream Gene Variant	E11 T cell	0.24 (0.22-0.32)
1	36043421	T	G	AGO3	Intron Variant	TN1 Transcription	0.25 (0.14-0.52)

1	36043427	C	G	AGO3	Intron Variant	E9 Liver Intestine	0.24 (0.17-0.55)
1	209429262	T	C	MIR205	Upstream Gene Variant	L4 Low signal	0.16 (0.13-0.17)
2	233170662	A	G	INPP5D	Intron Variant	TN4 Transcription	0.24 (0.19-0.32)
1	172387417	A	G	DNM3	Intron Variant	L3 Low signal	0.21 (0.19-0.3)
2	232527505	A	G	PRSS56	Downstream Gene Variant	PC2 Weak Polycomb	0.22 (0.09-0.31)
2	86885009	A	G	ENSG00000213605	Downstream Gene Variant	L4 Low signal	0.22 (0.2-0.3)
19	42670318	A	G	CEACAMP5	Downstream Gene Variant	HET3 Heterochromatin	0.21 (0.12-0.3)
1	23798576	A	G	LYPLA2	Downstream Gene Variant	TN1 Transcription	0.2 (0.15-0.28)
17	7342526	A	G	ACAP1	Intron Variant	TN4 Transcription	0.14 (0.1-0.22)
1	206897905	A	C	IL24	Intron Variant	L5 Low signal	0.19 (0.1-0.24)
19	2194438	C	T	DOT1L	Intron Variant	TN4 Transcription	0.19 (0.18-0.29)
17	59275100	T	C	GDPD1	3' UTR Variant	L4 Low signal	0.13 (0.09-0.16)
14	73109547	T	A	RBM25	Intron Variant	TN4 Transcription	0.28 (0.13-0.33)
19	54974107	A	G	NLRP2	Intron Variant	L5 Low signal	0.22 (0.17-0.28)
17	7350876	T	C	TMEM95	Upstream Gene Variant	L5 Low signal	0.2 (0.18-0.23)
5	181055522	T	C	BTNL9	Intron Variant	PC3 Polycomb	0.18 (0.12-0.22)
5	119071721	T	C	DMXL1-DT	Upstream Gene Variant	E4 Multi tissue	0.37 (0.31-0.45)
9	133810112	A	G	VAV2	Intron Variant	PC2 Weak Polycomb	0.25 (0.13-0.4)
5	884122	C	G	BRD9	Intron Variant	TN1 Transcription	0.19 (0.17-0.38)
3	195729589	G	T	MUC20	Intron Variant	TN1 Transcription	0.18 (0.17-0.18)

3	38374585	A	G	XYLB	Intron Variant	E5 B cell like	0.21 (0.18-0.24)
1	154970567	C	T	SHC1	5' UTR Variant	E2 Multi tissue	0.41 (0.3-0.46)
5	119071716	G	A	DMXL1-DT	Upstream Gene Variant	CTCF CTCF Cohesin	0.34 (0.21-0.45)
10	384502	C	T	DIP2C	Intron Variant	TN1 Transcription	0.13 (0.12-0.14)
9	133810120	G	C	VAV2	Intron Variant	PC2 Weak Polycomb	0.25 (0.14-0.36)