UNIVERSITÀ
DI SIENA
1240

Department of
Medical Biotechnologies

Phd Program
Doctorate in Genetics, Oncology and Clinical Medicine (GenOMeC)
XXXVII Cycle (2021-2024)
Coordinator: Prof. Ilaria Meloni

Expanding the Landscape of Breast Cancer-Associated Rare Variants and Combining with Polygenic Risk Score

Tutor:
Dr. Romina D'Aurizio
Institute of Informatics and Telematics (IIT),
National Research Council (CNR)

Supervisors:
Prof. Simone Furini
University of Bologna

Candidate:
Giulia Brunelli
Mat. 118508
Institute of Informatics and Telematics (IIT),
National Research Council (CNR)

Academic Year 2023-2024

Contents

List of Figures

List of Tables

List of Abbreviation

ACMG American College of Medical Genetic

AFR African

AMBER African American Breast Cancer Epidemiology and Risk

AMD Age-related Macular Degeneration

AMR American

ASO Allele-Specific Oligonucleotide

AUC Area Under the Curve

BC Breast Cancer

BCAC Breast Cancer Association Consortium

BD Bipolar Disorder

BhGLM Bayesian hierarchical Generalized Linear Model

BMI Body Mass Index

BOADICEA Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm

$C+T$ Clumping and Thresholding

CAD Coronary Artery Disease

$CD-RV$ Common-Disease/Rare-Variant hypothesis

CFH Complement factor H

$CIMBA$ Consortium of Investigators of Modifiers of BRCA1/2

CNV Copy Number Variation

CSA Central/South Asian

$DBSCAN$ Density-Based Spatial Clustering of Applications with Noise

$DFBBCS$ The Dutch Familial Bilateral Breast Cancer Study

DNA Deoxyribonucleic Acid

DP Read Depth

EAS East Asian

EM Expectation Maximization

$eQTLs$ Expression Quantitative Trait Loci

$ER+$ Estrogen receptor positive

$ER-$ Estrogen receptor negative

ER Estrogen receptor

EUR European

FE Functional Equivalent protocol

$FFPE$ Formalin-Fixed Paraffin-Embedded

FPR False Positive Rate

$GC-HBOC$ German Consortium for Hered- itary Breast and Ovarian Cance

$gDNA$ genomic DNA

$GWAS$ Genome-Wide Association Study

$hDlg$ Homolog of the Drosophila disc Large tumor suppressor Gene

$HGDP$ Human Genome Diversity Project

HT Hypertension

$LASSO$ east Absolute Shrinkage and Selection Operator

LD Linkage Disequilibrium

MAF Minor Allele Frequency

MID Middle Eastern

MR Magnetic Resonance

MSE Mean Squared Error

NGS Next Generation Sequencing

OCE Oceania

$OQFE$ Original Quality Functional Equivalent protocol

OR Odds Ratio

$P+T$ Pruning and Thresholding

PC Principal Component

PCA Principal Component Analysis

PCR Polymerase Chain Reaction

PGS Polygenic Score

PR Progesterone Receptor

PRS Polygenic Risk Score

PTV Protein Truncating Variant

$pVCF$ multi-sample Variant Call Format

QC Quality Control

RA Rheumatoid Arthritis

RAP Research Analysis Platform

RF Random Forest

$RFLPS$ Restriction Fragment Length Polymorphism

RNA Ribonucleic Acid

RR Risk Ratio

RV Rare Variant

SCT Stacked Clumping and Thresholding

$SHAP$ Hapley Additive exPlanations

$SKAT-O$ Optimal Sequence Kernel Association Test

$SKAT$ Sequence Kernel Association Test

SNP Single Nucleotide Polymorphism

SNV Single Nucleotide Variant

$T1D$ Type 1 Diabetes

$T2D$ Type 2 Diabetes

$TNBC$ Triple-Negative Breast Cancer

$UKBB$ UK Biobank

VCF Variant Call Format

VEP Variant Effect Predictor

WES Whole Exome Sequencing

WGS Whole Genome Sequencing

WT Wild Type

$WTCCC$ Wellcome Trust Case Control Consortium

Abstract

Investigating the association at variant-level of rare variants(RVs, MAF < 0.01) with breast cancer (BC) risk in population studies poses challenges due to low statistical power and multiple testing burdens. To increase power, current approaches often aggregate RVs into genetic units, such as genes or gene sets. However, these strategies typically focus on high-penetrance genes and pathways already known to be involved in cancer, limiting their capacity to identify novel contributors. Likewise, most existing methods fail to provide insights into the individual contributions of specific RVs, reducing the interpretability and clinical utility of the findings.

The work described in the present thesis aimed at addressing these two gaps by firstly providing a more systematic and scalable method to comprehensively analyze RV impact to BC and secondly introducing an alternative approach (Bayesian Hierarchical Generalized Linear Model, BhGLM) to investigate the single variant association to BC risk. We trained and test the methods using the UK Biobank (UKBB) cohort (15868 BC cases, 165067 controls). The Burden test assessed the cumulative association of RVs in aggregated genetic units, while BhGLM accounted for complex relationships within the data to identify BC-associated variants. First, we applied the Burden Test to different lists of genes and different RV masks combining Loss Of Function (LoF) and missense to determine the impact of the multiple testing burden on the detection of BC-related RVs and the contribution of different type of variants to BC susceptibility. Second, we exploited the BhGLM to assess

single RV association. We evaluated the quality of the retrieved RVs using the American College of Medical Genetics (ACMG) and ClinVar annotation, and by comparing their impact with the effect of unselected RVs using odds ratio (OR) across different PRS classes. Finally, we built two different RVScores by combining RVs significantly associated to BC by the Burden Tests and the BhGLM model. These scores allowed us to explore the cumulative impact of RVs in BC risk stratification in combination with PRS. The findings were assessed using OR on a distinct test set.

Through the application of the classical Burden test approach we underscored the importance of gene list selection in detecting associations at gene level. We showed how smaller curated lists were more effective at identifying weaker, yet meaningful associations, while larger lists provided a broader view of potential contributors but were less sensitive to subtle signals. Strong associations were consistently observed in well-established BC susceptibility genes like BRCA1, BRCA2, ATM, CHEK2, and PALB2. Notably, we identified two new potential risk genes, ASPRV1 and ADGRA3, that showed a strong relation with BC risk. Weaker associations emerged for further 7 genes (BARD1, MAP3K1, PLCG1, LZTR1, POLD2, DDX1 and NDFUS4), highlighting the need for a balanced approach to gene selection. The RVScore, computed on Burden Test results, showed stable performance across different variant masks, underscoring its robustness as a tool for patient stratification. When calculated using only LoF RVs, the RVScore enabled more precise stratification of BC risk across PRS classes for 2.5% of the population compared to the presence of RVs in high- and moderate-risk genes. Furthermore, when combining LoF and missense RVs, high levels of RVScore yielded higher OR across PRS classes than the sole presence of RVs in high- or moderate-risk genes.

At the same time, the BhGLM approach demonstrated high specificity levels in simulation settings, translating in low false-positive rate (FPR, average ≤ 0.001).

Conversely, sensitivity assumed considerably low values, which highlights an overall conservative trend in the model's classification strategy. When evaluated in a controlled setting of a short list of genes, BhGLM mostly selected pathogenic RVs with higher OR than non the selected ones on the same genes. When extended to the ClinicalExome (5369 genes), we identified a total of 550 LoF RVs, of which 40.2% annotated as Uncertain Significance, and the 24.74% as Pathogenic. Notably around 80% of the annotated Pathogenic RVs are associated with a positive effect size. The comparison the ORs for the selected RVs with the one of unselected RVs across PRS classes reveals their significant contribution to amplifying BC risk.

The comparison between the two approaches revealed notable differences in the number and characteristics of selected variants: BhGLM identified a larger set of RVs on a broader set of genes, likely due to its capacity to model complex relationships.

Nonetheless, high levels of both the RVScores were associated to an increment risk of BC with respect to PRS alone. Furthermore, the here proposed approach based on the a Bayesian hierarchical model, not only introduces a novel methodological framework in the context of BC studies, but enables also the quantification of the collective impact of RVs on BC risk while preserving the capacity to interpret the contribution of individual variants.

However, the reduced discriminatory power at lower BhGLM RVScore levels in the test set suggests that the score's ability to provide finer stratification of cancer risk is influenced by sample size. Similarly, The Burden-derived RVScore showed variability in his distribution between the training and test sets, likely reflecting the smaller size of the test set.

These findings indicated the significant role of rare variants in BC risk and the utility of combining their collective effects into a unified score. Nevertheless, the RVScores need validation with larger external cohorts. In addition, the rarity of the variants poses challenges for RVscores application in small cohorts or individual

patients, limiting their extendibility. Future research should explore novel measures to aggregate the impact of functionally related variants while retaining the ability to analyze the contributions of individual RVs. Finally, this analysis was conducted exclusively on individuals of European ancestry, making it impossible to evaluate its applicability across different populations.

Chapter 1

Introduction

Breast cancer (BC) remains a significant global health concern, being the most commonly diagnosed cancer among women, with approximately 2.3 million new cases reported in 2022 according to World Health Organization[112]. Despite of that, survival rates for breast cancer have improved significantly over the past few decades due to advancements in early detection and treatment modalities. The National Cancer Institute reported a 5-years survival rate of 90.2% for women between 2014 and 2020 [56]. However, survival rates vary significantly by stage at diagnosis; for instance, when BC has metastasized to distant sites, the 5-year survival rate drops to around 31%[56]. The incidence of mammary tumor is influenced by various risk factors, including age, gender, genetic predisposition, and lifestyle choices. Notably, female gender is the strongest risk factor, with about 99% of cases occurring in women[112]. Along with gender, genetic predisposition plays a crucial role in breast cancer risk, particularly for individuals carrying mutations in high-risk genes such as BRCA1 and BRCA2. Mutations in these genes account for approximately 25% of the familial cases of BC and significantly increase the lifetime risk of developing the disease, with estimates suggesting that women with a BRCA1 mutation have up to a 72% chance of being diagnosed before death, while those with a BRCA2

mutation face a lifetime risk of about $55-69\%$ [5]. Throught the years, the increased understanding of BRCA1/2 function and the DNA damage response pathway led to the discovery of other breast cancer susceptibility genes. For eaxample, Douglas et al. [39] estimated that protein-truncating variants in PALB2 confer a relative risks of BC equal to 5%, those on CHEK2 of 3%, on ATM of 2.8%, and of 2.7% on NBN. More limited evidence is available for other genes such as MRE11A, RAD50, MLH1, MSH2, MSH6, MUTYH, MEN1, PPM1D, and PMS2 [39]. Therefore, understanding the genetic predisposition to breast cancer can significantly enhance the survival rate through improved risk stratification, personalized treatment approaches, and proactive management strategies. For instance, genetic testing for mutations in BRCA1 and BRCA2 allows for the identification of individuals at elevated risk, facilitating earlier surveillance and intervention. Moreover, genetic predisposition influences treatment decisions and responses. Research indicates that BRCA mutation carriers often respond better to specific chemotherapies, particularly those targeting DNA damage, such as platinum-based agents. This is particularly relevant for patients with triple-negative breast cancer (TNBC), where BRCA mutations have been associated with improved disease-free survival rates when treated with chemotherapy [93].

In this work we will go through the genetic predisposition to BC focusing on the impact of rare variants (RVs) on the probability to developing the disease by the application of a novel approach. This first chapter will provide an introduction to the genetic background of this study, starting from the data typically used in phenotype-genotype association analysis to the methods used to investigate the relationship between common variants and specific traits. The second chapter will describe the different approaches for RVs association analysis enhancing their advantages and limitations and reporting the state of the art for breast cancer. In the

third chapter we will go through the methods of our study, while in the forth we will explain our findings. In the last chapter we will discuss the innovation of our method and the limitation of our study.

1.1 Genotyping tecnologies

1.1.1 Single nucleotide polymorphism arrays

Compared to many other species, human genetic diversity is relatively lower. The predominant form of genetic variation in humans arises from single nucleotide variants (SNVs), which denote substitutions of a single nucleotide at specific positions within the genome. When such substitutions are present in a sufficiently large fraction of the population (e.g. 1% or more), they are termed single nucleotide polymorphisms (SNPs). Accordingly, a SNP is defined as SNVs with a minor allele frequency (MAF) in the population of at least 1%. SNPs can be located within coding sequences of DNA, non-coding genomic regions, or intergenic regions. Single-nucleotide polymorphisms within coding sequences may not alter the amino acid sequence of the resulting protein due to the redundancy of the genetic code: synonymous SNPs, which preserve the protein sequence. We say non-synonymous SNPs to refer to single nucleotide polymorphisms that modify the amino acid sequence, potentially affecting protein function. Synonymous substitutions, although not changing the amino acid, can still impact protein function. For example, a seemingly silent mutation in the MDR1 gene can slow down translation, altering protein folding and reducing functionality. [23] Non-synonymous substitutions include missense mutations, where a single base change leads to an amino acid alteration, potentially causing disease. An example is the c.1580G-T SNP in the LMNA gene, which results in mandibuloacral dysplasia and progeria syndrome [75]. Nonsense mutations intro-

duce premature stop codons, resulting in truncated and often nonfunctional proteins. For instance, the G542X mutation in the cystic fibrosis transmembrane conductance regulator gene causes cystic fibrosis [99]. SNPs outside coding regions can still affect gene regulation processes such as splicing, transcription factor binding, mRNA degradation, or noncoding RNA sequences. Additionally, non-coding SNPs may regulate gene expression levels as expression quantitative trait loci (eQTLs) and can lead to disease susceptibility. For example, the intronic SNP rs1059060, Ser775Asn in the DNA mismatch repair gene PMS2 is associated with increased sperm DNA damage and the risk of male infertility [42].

Given the important role of SNPs in diseases incidence and evolution, in the past few decades, many advances have been made in the development of technologies that will improve the efficiency, rapidity, and cost effectiveness of genotyping large numbers of individuals for identified SNPs. The first examples of genotyping date back to the 1970s, when early efforts in DNA sequencing laid the groundwork for the development of tools capable of discerning the genetic makeup of organisms. Variations in DNA sequences, identified by bacterial restriction enzymes, lead to cleavage at distinct points, resulting in discrepancies in fragment lengths. In the years between the 1970s and the 1980s, these variations, known as restriction fragment length polymorphisms (RFLPs), were employed for genotyping. [51] Hybridization with allele-specific oligonucleotides (ASO) was initially demonstrated in 1979 [110], showcasing its ability to identify even single-base mismatches. This technique was subsequently employed in 1983 [21] to detect the sickle-cell mutation within the β-globin gene via Southern blot hybridization to human genomic DNA [102]. However, these primitive techniques were laborious and time-consuming, often spanning several days, involving intricate protocols such as gel electrophoresis and hybridization with radioactively labeled probes to visualize the outcomes. Therefore detecting a single base change within the extensive 6×10^9 base pairs of the diploid human

genome represented a significant challenge until the invention of the polymerase chain reaction (PCR) in 1985 [78] that made it possible to design effective assays for genotyping SNPs in elaborate genomes. PCR enabled the amplification of infinite copies of a specific DNA segment, transforming the field of genetic medicine by facilitating DNA comparison, the diagnosis of genetic conditions, and the identification of viruses within human cells [51]. Subsequently, the invention of DNA microarray technologies marked a significant stride towards high-resolution genotyping.

Nowadays microarrays are one of the most widely used techniques for genotyping. They consist of a solid surface that hosts microscopic spots of synthesized DNA. These spots, strategically designed to include segments of DNA that overlap with targeted SNPs, enable genotyping by ensuring that, under appropriate hybridization conditions, target DNA hybridizes only if it is complementary to the DNA in a specific spot [51]. There are several types of genotyping microarray, each employing distinct chemistries. The most common ones on the market are produced by Affymetrix and Illumina. Despite their differences, both brands share several key features. Firstly, their SNP arrays operate on the biochemical principle of binding nucleotide bases to their complementary partners according to the Watson-Crick base pairs. Secondly, the protocols of both technologies entail hybridizing fragmented single-stranded DNA to arrays containing hundreds of thousands of unique nucleotide probe sequences, with each probe designed to bind to a specific DNA subsequence. Specialized equipment measures the signal intensity associated with each probe and its target post-hybridization in both cases. This intensity of the signal depends on the amount of target DNA in the sample and the affinity between the target and the probe. Subsequent processing and analysis of these intensity measures yield inferences about SNP genotype, with both producers boasting genotyping accuracy rates exceeding 99.5% [104].

Nearly two decades ago, Affymetrix pioneered the commercial production of SNP

arrays. The number of SNPs that can be genotyped by Affimetrix microarray technologies has increased dastrically over the years: starting from the first assessment, HuSNP, prototyped by Wang et al. [29] to genotype 1494 SNPs on a chip, it has gradually scaled up to 10 000, 100 000, 500 000 and, finally, to nearly one million SNPs in the latest version [104]. The most recent releases of Affimetrix arrays feature substantial improvements in probe composition aimed at maximizing the limited space available on the array. The Genome-Wide Human SNP Array 6.0, for example, boasts over 906,600 single nucleotide polymorphisms and more than 946,000 probes for detecting copy number variation (CNVs). Examination of each single nucleotide polymorphism on the Human SNP Array 6.0 is performed through three or four replicates for each of the two alleles using six or eight perfect-match probes. In contrast, copy number variations are evaluated with a single probe, specifically targeting regions of the genome devoid of SNPs but potentially polymorphic in terms of CNVs [104]. This version of Affymetrix software rely on Birdsuite, a four-stage analytical framework builded in software for deriving conbined and reciprocally consistent CNV and SNP genotypes [64]. The Birdsuite method, starting from the raw normalized probe intensities, obtains summarized signals using a median polish procedure and expectation–maximization (EM) algorithm resulting in genotypes for each SNP [104].

Regarding Illumina technology, the most popular type of microarray genotyping is Illumina's BeadArray. Like the Affymetrix platform, Illumina's BeadArray has slowly boosted its capacity over time starting from 100 000 SNPs in Human-1 to one of the most recent release HumanHap1M with one million of SNPs [104]. Illumina's BeadArray, consists of bead libraries casually arranged within a substrate of chiseled microwell. Each bead contains several copies of oligonucleotides targeting specific loci, along with a mapping oligonucleotide designed to avoid homology with any sequence of the species under analysis. Through a series of hybridization steps

during chip design, the mapping oligonucleotide helps to identify the precise location of each bead on the array. During the genotyping process, the DNA being analyzed binds to the beads with complementary DNA, halting one base pair before the SNP position. Next, a single-base extension occurs, incorporating one of the four tagged nucleotides. Detection of the incorporated nucleotide emits a signal upon laser excitation, with signal intensity providing information about the allelic ratio at a specific locus. [51] The array data output format consists of a raw measurement for each of the two alleles of each SNP. The raw file of a single HumanHap1M array has approximately one million pairs. One of the computational strengths of the Illumina protocol is the normalization procedure that is performed internally on each individual sample, without requiring multiple arrays. The goal is to produce a pair of raw allele-specific copy measures for each SNP, used for genotype calls by defining a transformed ratio of normalized allelic intensities. Like the current version of the Affymetrix array, HumanHap1M also includes probes for CNVs designed to detect human genetic variations not specifically due to SNPs [104].

1.1.2 Whole exome and whole genome sequencing

The last three years of the 1970s saw the emergence of the first generation of DNA sequencing methods. In the same period in which Allan Maxam and Walter Gilber prototyped their approach involving terminally labeled DNA fragments undergoing base-specific chemical cleavage, followed by separation of the reaction products by gel electrophoresis [76]; Frederick Sanger and colleagues proposed an alternative method that, refined and commercialized, was widely applied in clinical diagnostics. The method developed by Sanger used chain-terminating dideoxynucleotide analogs to induce base-specific termination of primed DNA synthesis [96]. Since then, Sanger's technique has been refined culminating in the sequencing of the first

human genome in 2003 with the Human Genome Project, a monumental undertaking that lasted 13 years and cost about $2.7 billion. In 2008, a human genome was sequenced in just five months at a fraction of the previous cost [30], heralding the ability of "next-generation" sequencing (NGS) technologies. NGS platforms represent a significant advance over Sanger sequencing, featuring massively parallel sequencing of clonally amplified or single DNA molecules spatially segregated in a flow cell. NGS operates through repeated cycles of polymerase-mediated nucleotide extensions or, in some formats, iterative cycles of oligonucleotide ligation. As a result of its massively parallel nature, NGS can produce hundreds of megabases to gigabases of nucleotide sequence output in a single instrument run, depending on the platform [108]. One of the main differences between different NGS platforms is the sequencing method, which can be based on short-reads or long-reads. Short-read protocols typically produce reads of of less than 300 base pairs (bp), while long-read sequencing can provide continuous reads ranging from 10 kbp to many megabases, according to the technology. Analyses utilizing next-generation sequencing encompass both whole-genome sequencing (WGS) and whole-exome sequencing (WES). While whole-genome sequencing is a comprehensive method employed to sequence the entire human genome, WES is a widely used next-generation sequencing method that focuses on the protein-coding regions of the genome (exome) which represents less than 2% of the genome. In a typical individual, WGS can uncover 3 to 4 million variants [20], whereas whole-exome sequencing generally identifies around 20,000 variants within the exome [77].

The decreasing costs of sequencing and the capability to generate an extensive amount of data with modern sequencers render whole-genome sequencing a potent tool for genomics research. Indeed, WGS can be a valuable instrument for capturing both large and small variants that might be missed with targeted approaches, it provides a high-resolution, base-by-base view of the genome and allows to iden-

tify potential causative variants for further follow-up studies of gene expression and regulatory mechanisms; finally, it delivers large volumes of data rapidly to support the assembly of novel genomes and making possible to easily use genetic data to identifying inherited disorders, characterizing oncogenic mutations, and monitoring disease outbursts[52]. WGS analyses find their main application in the diagnosis of rare diseases, especially in cases where conventional methods fail to provide a diagnosis. About 10% of diagnoses of patients with rare diseases were originated by variants in genomic regions that would not have been identified by other approaches, among these a small fraction involved variants coding in regions of low coverage on exome sequencing [19]. Whole-genome sequencing is extensively utilized in cancer research, both to characterize tumors in order to predict treatment responses or choose appropriate therapy, and for detecting somatic driver mutations as well as germline predisposing mutations in tumor suppressor genes. For instance, a study from 2016 revealed 93 mutated genes involved in the genesis of the breast-cancer by analyzing 560 tumor whole-genome sequences [86]. In another paper, K. Nones and colleagues emphasize the role of WGS of paired germline and breast tumor DNA, as well as somatic mutational signatures, in discovering pathogenic germline variants and selecting platinum-based therapy or PARP inhibitors for patients affected by familial breast cancers [87]. However, despite the myriad applications of WGS analysis, when it is necessary to increase throughput, optimize cost per sample, analyze manageable data sets, and maximize data storage, WES provides a more suitable solution. Indeed, whole-exome sequencing is more cost-effective than whole-genome sequencing. Moreover, since over 85% of known disease-causing mutations occur in exons, the exome provides a highly enriched subset of the genome for identifying disease-causing variants, thereby enhancing the efficiency of finding clinically meaningful results [66]. For the aforementioned reasons, WES has found extensive application in research and clinical practice, especially in the understanding

of coding mutations that impact tumor biology. In cancer research, whole-exome sequencing has proved to be a valuable instrument for biomarkers discovery for treatment response. For example, Beltran and colleagues identified a patient with prostate cancer with exceptional response to treatment who harbored a somatic hemizygous deletion of the DNA repair gene FANCA and putative partial loss of function of the second allele through germline missense variant. Subsequent experiments provided the biological rationale behind the evidence by establishing that loss of FANCA function was related to platinum hypersensitivity both in vitro and in patient-derived xenografts[47]. In another study, Dieci et all, applied WES on rare aggressive breast cancer subtypes in order to uncover new potential therapeutic targets. They found that the 30% of samples carry mutations in PYGM, a gene involved in glycogen metabolism, which is dramatically underexpressed in common cancers as compared to normal tissues and that low expression in tumors is correlated with poor relapse-free survival [79]. Whole-exome sequencing can also be used to inform cancer susceptibility by identifying genes that cause an inherited predisposition to different types of cancer [100] [40] [83]. Application of WES in a population-based study in Finland found that a nonsense mutation on the FANCM gene is significantly more frequent among breast cancer patients, particularly those with triple-negative breast cancer [63].

Like the previously mentioned studies, this work aims to identify rare germline variants associated with breast cancer risk. For this purpose, we analyzed WES data from UK Biobank (chapter 3) that were captured using the IDT xGen Exome Research Panel v1.0 with supplemental probes. The samples were sequenced with dual-indexed 75 x 75 bp paired-end reads on the Illumina NovaSeq 6000 platform using S4 flow cells. Figure 1.1 summarizes the Illumina sequencing process[53]; however, this workflow can be extended to any WES pipeline regardless of the used platform, and the sequencing method adopted.

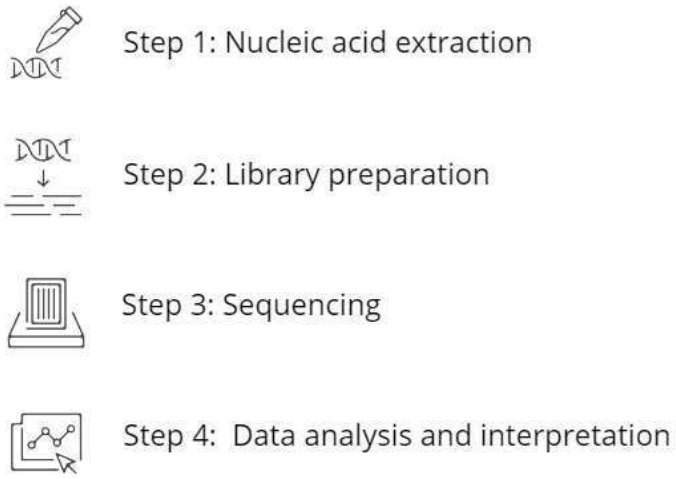


Figure 1.1: **Illumina NGS workflow steps**

Briefly any NGS workflow can be splitted into four steps. A first phase of nucleic acid extraction that involves the acquisition of high-quality genomic DNA (gDNA) from biological samples such as blood, formalin-fixed paraffin-embedded (FFPE) tissue, or saliva. In everyday practice, gDNA is often extracted from peripheral blood leukocyte. After extraction, most NGS workflows require a Quality Control (QC) process. The second step of the WES experiment is the library preparation. This is an important phase for the result of the analysis that implicates preparing sequencing libraries by randomly fragmenting gDNA into smaller molecules and adding specialized adapters to both ends. The DNA fragments then attach to the flow cell through complementary sequences on the adapters, ensuring successful sequencing. At this point, the library is loaded into a flow cell, where the single-stranded fragments selectively hybridize with its inner surface. Each attached fragment then undergoes solid-phase bridge amplification, producing dense clonal clusters with around 1000 copies each. A single flow cell can support millions of these parallel cluster reactions Once the library is ready, the paired-end sequencing can start employing four fluorescently labeled nucleotides to sequence the numerous clusters on the flow cell surface in parallel. These nucleotides act as polymerization terminators, which are

enzymatically cleaved after detection to allow the next round of synthesis. This cycle repeats for 300 or more rounds. Fluorescent detection enhances speed through direct imaging, unlike camera-based methods. After sequencing the forward DNA strand, the reads are washed away, and the process is repeated for the reverse strand. The final step consists in the data analysis and it can began once the reads are aligned to a reference sequence with the aid of bioinformatics tools. At this point the sequencing data can be analyzed using standard or customized approaches[53].
In the next paragraphs and in chapter 2 we will elucidate some of the most common technique used to analyzed the association of genetic variants with various traits and diseases.

1.2 Analysis of genomic data

1.2.1 Genome-wide association studies

In a pioneering investigation published in 2005, researchers led by Robert J. Klein- conducted one of the earliest Genome-Wide Association Studies (GWAS). Their study focused on examining genetic variations associated with age-related macular degeneration (AMD) by scrutinizing the genomic profiles of 96 cases and 50 controls. Through the genotyping of 116,204 single nucleotide polymorphisms (SNPs), the authors identified a significant correlation between AMD and a prevalent intronic variant within the complement factor H (CFH) gene [92]. Since then, over 6,715 research papers have delved into GWAS association studies, leading to the identification of a staggering number of 571,148 common genetic variants - characterized by a minor allele frequency $\geq 5\%$ - associated with various diseases [57].

Essentially, genome-wide association studies aim to elucidate the relationships between genotypes and phenotypes by analyzing variations in allele frequencies of ge-

netic variants among individuals with shared ancestry but distinct phenotypic traits. GWAS can examine copy number variants (CNV)-which involve changes (duplications or deletions) in the number of copies of a particular gene or DNA segment- or sequence alterations in the human genome- where larger structural changes, such as insertions, deletions, or rearrangements of DNA, occur in the genome. However, the most commonly studied type of genetic variation in GWAS is single nucleotide polymorphisms. GWAS typically identify clusters of correlated SNPs that exhibit a statistically significant association with the trait of interest, referred to as genomic risk loci [38]. To identify genetic variants associated with specific traits or diseases, the experimental workflow of a GWAS requires several steps [38]:

- **Study Design Definition**: this first step involves the determination of the sample sizes needed to identify reproducible genome-wide significant associations. Larger sample sizes increase the power to detect associations, especially for traits influenced by many genetic variants with small effects. Besides sample size, study designs for GWAS can vary depending on the nature of the trait being investigated. For dichotomous traits, cases and controls may be included, while quantitative traits may require measurements across the entire study sample. Researchers can also select between population-based and family-based designs.

- **Data Collection**: This step involves collecting DNA and phenotypic data from participants. DNA can be obtained from blood, saliva, or other tissues. It is important to ensure high-quality DNA with sufficient quantity for genotyping. Phenotypic data include disease status, health measures, demographic information (age, sex), and other relevant traits.

- **Genotyping**: Participants' genomes are analyzed using genotyping arrays or sequencing methods to identify genetic variants. Depending on the scope of

the study, two approaches are typically used: Microarray Genotyping when the study focuses on common genetic variants, and Next-Generation Sequencing, used for more comprehensive analysis. These genotyping techniques have been previously described in sections 1.1.1 and 1.1.2.

- **Data processing**: To ensure the accuracy of the genotypic data, stringent QC filters are applied to both the genetic and phenotypic data. Low-Quality SNPs with high rates of missing data, low frequency, or deviations from Hardy-Weinberg equilibrium are filtered out. Similarly, samples with inconsistent or missing phenotypic information, or those with abnormal heterozygosity or ancestry patterns, are removed. Once sample and variant QC step has been performed on GWAS data, variants usually undergo phasing and are imputed. Imputation is a critical step where missing genotypes are statistically inferred using reference panels like the 1000 Genomes Project or TOPMed. This increases the number of SNPs available for the analysis and helps fill in gaps from the genotyping assays.

- **Testing for associations**: this is the core of GWAS. Each genetic variant (e.g., each SNP) is tested for its association with the trait or disease using statistical methods such as Linear Regression for continuous traits or Logistic Regression (applied for binary traits) or Additive, Dominant, Recessive Models that test different modes of inheritance. In this step variables like age, sex, principal components (to account for population stratification), and other relevant factors are included in the model to take into account possible confounders.

- **Accounting for false discovery**: testing millions of SNPs increases the chance of false-positive associations that are avoided using a strict statistical threshold. Since, on average there are approximately one million of indepen-

dent common genetic variants across the human genome, a Bonferroni testing threshold of $P < 5 \times 10^{-8}/P < 5 \times 10^{-7}$ is an appropriate threshold to avoid false positives. However the appropriate threshold might vary depending on the population, on the complexity of the analyzed trait and on the minor allele frequency thresholds for inclusion in the GWAS.

- **Meta-Analysis (Optional)**: In large-scale studies, results from multiple independent cohorts can be combined in a meta-analysis to increase statistical power. This is typically done when single studies do not have large enough sample sizes to detect associations.

- **Replication**: Replication is a crucial step for validating findings. Once genetic associations are identified, they must be tested in an independent cohort to confirm that the results are not due to random chance.

- **GWAS analysis results**: The primary output of a GWAS analysis is a list of P values, effect sizes and their directions as result of the association tests of all the genetic variants with a phenotype of interest. These data provide a first partial view of the association of variants with the phenotype, however further analyses are required to have a more correct interpretation of these results.

- **Post-GWAS analysis**: To interpret GWAS analysis results, further analyses are required such as Fine-Mapping, that aims to narrow down the associated regions to pinpoint the most likely causal variants; or Functional Annotation that annotate significant SNPs with information about gene function, regulatory elements, and expression patterns using databases and tools like ANNOVAR, and VEP (Variant Effect Predictor); or Gene Mapping that identifies which genes are affected by the associated variants using eQTL analyses;

or Pathway and Gene Set Enrichment to determine if associated genes are enriched in specific biological pathways or processes.

As mentioned at the beginning of the paragraph, GWAS represent a powerful approach for uncovering the genetic underpinnings of complex traits and diseases. Landmark studies have demonstrated the capability of this methodology to identify significant genetic variants associated with various health-related conditions, providing valuable insights into disease mechanisms and potential therapeutic targets. Among these works, one of the most important was conducted in 2007: the Wellcome Trust Case Control Consortium (WTCCC) study [28], the largest GWAS study ever conducted at the time of its publication. This research aimed to identifying genetic variants associated with seven common diseases: bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), rheumatoid arthritis (RA), type 1 diabetes (T1D), type 2 diabetes (T2D), and hypertension (HT). The study analyzed 14,000 cases and 3,000 shared controls, leading to the identification of 24 independent association signals across six of the seven diseases, with significant associations at $P < 5 \times 10^{-7}$. The WTCCC GWAS highlighted the effectiveness of a large-scale, shared control approach in uncovering common genetic variants associated with complex diseases, underscoring the importance of sample size, rigorous quality control, and the integration of data across multiple conditions in GWAS. The findings not only confirmed previously known loci but also revealed novel genetic associations, particularly for CD and CAD. Another valuable investigation was published on Nature in 2023 [43]. The study aimed to identify shared genetic risk factors underlying multiple cancers by conducting a pan-cancer and cross-population genome-wide association study. Researchers analyzed genetic data from 13 types of cancer in two large biobank datasets (BioBank Japan and UK Biobank), covering 250,015 East Asians and 377,441 Europeans. The goal was to uncover pleiotropic genetic associ-

ations - i.e. genetic variants influencing multiple cancers - and to better understand the shared heritability and biological pathways contributing to carcinogenesis. The study discovered 10 significant cancer risk variants, with 5 pleiotropic loci associated with multiple cancers. Moreover, a significant genetic correlation was observed between breast and prostate cancer across both East Asian and European populations. Finally this study, trough a large-scale meta-analysis, identified 91 novel loci associated with these cancers, indicating shared heritability and common genetic pathways. This work highlighted the power of pan-cancer and cross-population GWAS in uncovering shared genetic risk factors across diverse cancers.

Breast cancer GWASs

In the preceding sections, we highlighted significant studies that have demonstrated the pivotal role of GWAS in uncovering the genetic basis of complex traits and diseases. Conversely, this section will focus to the major GWASs conducted to elucidate the genetic architecture underlying BC susceptibility, which serves as the central topic of this thesis.

In the last decades, GWAS have found a large application in enhancing knowledge of the genetic structure behind breast cancer predisposition. In a study from 2008, a multi-stage GWAS approach was applied testing over 200,000 SNPs in familial BC cases and controls [41]. Five loci significantly associated with BC risk were identified, including regions containing FGFR2, MAP3K1, TNRC9, LSP1, and an intergenic region at 8q24. Zhang et al. presented the results of a large-scale genome-wide association study that identified 32 new genetic loci associated with BC, 15 of which were subtype-specific (i.e., associated with distinct tumor characteristics like estrogen receptor (ER) status, progesterone receptor (PR), HER2 status, and tumor grade)[50]. In the same year, another work from Fachal an colleagues [12] analyzed over 217,000 participants from the Breast Cancer Association Consortium (BCAC)

and the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA), focusing on 150 genomic regions linked to BC risk. They predicted 191 high-confidence target genes, many involved in cancer-related pathways such as DNA repair, apoptosis, and immune response, among those genes such as FGFR2, MAP3K1, and ESR1 were prominent targets of causal variants. In a study lead by Ahearn, the authors investigated how common genetic variants linked to breast cancer affect the risk of different tumor subtypes [105]. In a most recent study of 2024, the authors selected 240 genetic variants linked to BC through Genome-Wide Association Studies and analyzed their frequencies in various populations founding 11 SNPs that were significantly correlated with mammary tumor incidence [46]. Regardless of the BC subtype considered, some genes were found to be associated with breast cancer risk in most of the previously mentioned GWASs conducted in the last few years. Specifically, SNPs in the gene **FGFR2** have been strongly associated with BC in multiple populations[41][105][50][12], variants in **TOX3** [105][12] and in the regions **8q24** [111][114], **11p15** [48], **16q12** [48] near this gene have been linked with an increased risk of BC, the gene **MAP3K1**, involved in cell signaling pathways that regulate cell proliferation, have been proved to be related to susceptibility of multiple BC subtypes [105][41][50][12]. It has been shown a relationship with BC also for SNPs in genes **CASP8** [41][12], involved in apoptosis, **LSP1** [41][50], a gene encoding an F-actin bundling protein, and **ESR1** [48][12], that encodes an estrogen receptor and ligand-activated transcription factor, and regions **2q35** [11], and **6q25** [17] near ESR1. Finally, **BRCA1**, **BRCA2**, **ATM**, **CHEK2**, and **PALB2** are well known to be risk genes for hereditary BC [41][105][12].

Table 1.1 lists the genes retrieved to be associated with BC by GWAS with the respective number of studies where the association have been found.

Despite Genome-Wide Association Studies have revealed many single nucleotide polymorphisms, associated with a wide range of complex traits, most of these SNPs

Gene/Region	Number of GWASs
FGFR2	4
TOX3	2
8q24	2
11p15	1
16q12	1
MAP3K1	4
CASP8	2
LSP1	2
ESR1	2
2q35	1
6q25	1
BRCA1	3
BRCA2	3
ATM	3
CHEK2	3
PALB2	3

Table 1.1: Summary of genes and regions associated with breast cancer risk in GWASs.

exert only a minimal effect and represent a small portion of the truly associated variants, which limits their overall predictive power [101]. In a study of 2008, Douglas F. Easton and Rosalind A. Eeles, conducting GWAS on several common cancers (breast, prostate, colorectal, lung, and melanoma), emphasize the polygenic nature of cancer, where multiple genetic variants with small effects contribute to overall susceptibility. [41] Another work of 2010 shows how the simultaneous combination of 294831 SNPs genotyped on 3925 unrelated individuals was able to explain the 45% of the variance of heritability for human height. [60] This finding suggests that

the genetic architecture of most complex traits is shaped by the combined influence of hundreds or thousands, of small-effect variants. As a result, aggregating multiple SNPs with minimal effects has become a common approach, giving rise to the concept of Polygenic Risk Scores (PRSs) [59].

1.2.2 Polygenic risk score

How we saw in the previous section, GWAS focus on detecting relationships between individual genomic variants, typically single nucleotide polymorphisms, and specific traits or diseases. However, the diagnostic usefulness of individual SNPs discovered by GWAS relies on having substantial effect sizes, which is quite rare. Furthermore, GWAS typically does not evaluate interactions between multiple alleles[59]. Polygenic risk scores, on the other hand, sum the small effects of many SNPs to create a combined score reflecting the overall influence of these variants on a particular phenotype. Although PRS and GWAS originated as distinct concepts, the theoretical foundation of polygenic risk scores is closely tied to GWAS. PRS builds on GWAS findings by combining the effects of multiple genetic variants linked to a particular phenotype. This allows for the calculation of an individual's PRS, which reflects his or her genetic predisposition to a specific phenotype based on the presence of the GWAS selected variants.

Polygenic Risk Score is computed as a weighted sum of risk alleles associated with a specific phenotype using the formula:

$$PRS_{pheno} = \Sigma_{j=1}^{m} X_j \hat{\beta}_j$$

where X_j is the number of risk alleles for the j^th SNP, and $\hat{\beta}_j$ is the effect size derived from GWAS data for that SNP. Finally, m is the number of the variants included in PRS construction which are selected from a discovery cohort [54]. Common

variants (and their effect sizes) are typically chosen from GWAS summary statistics. However, since SNP effects are estimated with uncertainty and not all SNPs influence the trait under study, the use of unadjusted effect size assays of all SNPs could generate poorly estimated PRS with high standard error. Therefore, to improve the predictive power of the PRS, GWAS summary statistics need to undergo a variant selection and/or variant effects modelling. Several methods have been published to address this purpose [54] [101]:

- **Pruning and thresholding (P+T)**: This is the most commonly used method and consists of establishing a p-value threshold for SNP selection and uses informed linkage disequilibrium (LD) pruning to discard SNPs in LD at a given threshold. When the variant with the highest effect size is selected in each LD block, the pruning method is also referred to as 'Clumping' (or C+T). Since the optimal P-value cutoff is not known a priori, PRS can be calculated over a range of thresholds. Subsequentially, the association with the target trait is tested for each P-value upper bound and the prediction is optimised. The C+T method have found a wide application thanks to its straightforward implementation and ease of understanding. In addition, by removing correlated SNPs, it minimizes the risk of overestimating the effect of genetic variants. However the thresholding process may exclude potentially relevant SNPs with smaller effect sizes, for this reason the choice of p-value thresholds can significantly influence the resulting PRS, leading to very different results.

- **Bayesian PGS**: This method, also referred to as 'LDPred', incorporates preliminary knowledge on the distribution of effects and the proportion of causal variants among all SNPs tested. In particular, it requires the definition of a regulatory parameter (ρ), which is an estimate of the genetic variants considered causal. P-value thresholds are varied and multiple 'LDPred risk scores'

are calculated using priors with varying fractions of markers with non-zero effects. Different PGS are calculated for C+T with varying ρ. This method can exceed P+T, particularly with large sample sizes.

- **Stacked clumping and thresholding (SCT)** This is a more novel method that employs machine learning techniques: it integrates clumping and thresholding (C+T) with the Least Absolute Shrinkage and Selection Operator (LASSO) or Ridge regression. SCT utilizes effect sizes and p-values for each SNP to conduct repeated P+T/C+T across a four-dimensional parameter grid, which includes linkage disequilibrium (LD) squared correlation, p-value thresholds, clumping window sizes, and imputation quality.

- **Meta-scoring** is a method that aggregates various polygenic scores related to a specific trait or disease into a single meta-score, known as a metaGRS or metaPGS. This approach operates on the premise that individual polygenic scores may be affected by regression dilution bias to some degree. By combining these scores into a more robust meta-score, the aim is to mitigate this bias. The construction of the meta-score involves creating a linear or non-linear combination of the individual polygenic scores, typically incorporating those for both the primary trait and associated traits.

Once computed, the PRS goes under quality controls and validation against an independent dataset, ideally of biobank-scale and inclusive of diverse ancestries, to assess its predictive power and generalizability. In this phase, the PRS is evaluated to determine its effectiveness in predicting the phenotype using metrics such as sensitivity, specificity, area under the curve (AUC), and mean squared error (MSE) depending on nature of the trait under analysis. As last step, the Polygenic Risk Score is externally validated by calculating it for a separate population or group of subjects that has not been involved in the PRS construction process. This step is

crucial for evaluating its predictive capability, confirming its performance, and reducing the risk of overfitting that may have occurred during the development phase. Figure 1.2 recapitulate all the steps described before.

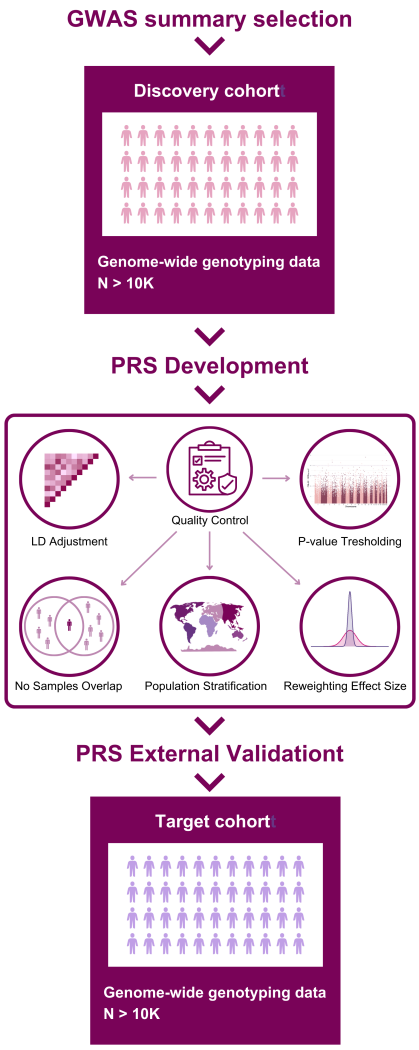


Figure 1.2: **PRS development workflow**

This figure is adapted from [54] and describes the multistage process of PRS development from variant selection to validation

PRS can be used to calculate relative genetic risk for traits or diseases in new cohorts, finding potential applications in various fields. Specifically, it can be used for

screening populations to identify individuals at high risk, allowing for targeted interventions and preventive measures [27]. PRS can also inform treatment decisions by identifying patients who may benefit more from specific therapies and, therefore, it can help in stratifying patients for clinical trials and tailoring management strategies based on genetic predisposition [27][1]. Moreover, for the nature of its definition, PRS predictive accuracy tends to increase with the heritability of a trait, as a larger portion of genetic variation can be captured for risk assessment. On the other hand, predicting traits with low heritability using PRS alone can be more difficult due to the weaker genetic influence. For this reason, by comparing the proportion of variance explained by PRS with the total phenotypic variance observed in a population, we can employ Polygenic Risk Score to explore the heritability of a specific trait, or, viceversa, we gain valuable insights into the genetic factors influencing the characteristic under investigation by evaluating the PRS contribution to its heritability.[27]

Breast cancer PRS

Polygenic Risk Scores for breast cancer have emerged as a pivotal tool in enhancing risk prediction and stratification, even in combination with other risk factors. Key studies, such as the one by Mars et al. [80], have established PRS based on various single nucleotide polymorphisms, demonstrating its ability to improve risk categorization. This paper analyzes data from 122,978 Finnish women from the FinnGen study, which includes 8,401 BC cases. It focuses on two frameshift mutations, c.1592delT in PALB2 and c.1100delC in CHEK2, which are notably enriched in the Finnish population. The authors founds that PRS modifies the risk of breast cancer in women carrying PALB2 and CHEK2 mutations: while PALB2 mutation carriers with a high PRS had an 84% lifetime risk of breast cancer, those with low PRS had a 49% risk. Similarly, subjects with CHEK2 mutations had a lifetime risk

of 59% with a high PRS and only 9% with a low PRS. Morehover, this work showed how PRS also refined the risk estimates for women with a family history of BC, particularly for those with early-onset breast cancer in first-degree relatives. Similar results have been reported by Roberts and colleagues [37] in their review where they demonstrate how PRS substantially improves breast cancer risk prediction when combined with traditional factors like family history and mammographic density and in women carrying high-risk genetic mutations (e.g. BRCA1/2). PRS can also be a potential instrument to predict different breast cancer subtypes, such as estrogen receptor-negative (ER-) breast cancer [37] [8]. Indeed a study conducted by the Breast Cancer Association Consortium [8] defined a 313-SNP PRS that showed a strong association with breast cancer risk, improving prediction for both ER-positive and ER-negative disease. The PRS313 provided an odds ratio (OR) of $1.61(95\%CI 1.57-1.65)$ and an AUC of $0.630(95\%CI 0.628-0.651)$, demonstrating a good discriminatory ability in white European populations. They also showed that women in the highest 1% of the PRS distribution had a 4.37-fold increased risk of ER-positive breast cancer and a 2.78-fold increased risk of ER-negative breast cancer, while those in the lowest 1% had a 0.16-fold risk for ER-positive and a 0.27-fold risk for ER-negative disease. Given that, PRS has potential implications for personalized screening and prevention strategies for BC, however, as PRS becomes more integrated into clinical practice, efforts must be made to increase its accuracy across populations. Indeed while PRS has proven effective in populations of European ancestry, its accuracy diminishes in non-European populations. [6]

1.2.3 Limits of GWAS analysis: the concept of "missing heritability"

Although GWAS and PRS have significantly advanced our understanding of the genetic architecture underlying complex diseases, the role of a large part of the genetic heritage in explaining complex traits remains unclear, even in diseases where large-scale GWAS meta-analyses have been conducted. For instance, in a study involving over 150,000 individuals with type 2 diabetes [18], more than 70 loci were identified with genome-wide significance, yet they only account for approximately 11% of T2D heritability. Similarly, another study[2] involving over 210,000 individuals revealed approximately 70 loci associated with Crohn's disease, but these loci explain only 23% of its heritability [95]. The concept of "missing heritability" [106] describes a common disparity observed between the heritability estimates of complex traits obtained from family-based studies and the variance in traits explained by common SNPs identified through large-scale GWAS endeavors. This discrepancy has prompted the formulation of the common-disease/rare-variant (CD-RV) hypothesis [65]. According to this hypothesis, the genetic basis of common complex traits may be shaped by multiple rare variants affecting one or more genes, often escaping detection by conventional GWAS SNPs [109].

In the next chapter we will take a closer look at the genes involved in breast cancer occurrence, highlighting the impact of rare mutations. Then we will elucidate the different techniques for association analysis of rare variants, and finally, we will examine the different models developed for combining rare and common variants in the study of BC.

Chapter 2

Rare variants association analysis

As outlined in Chapter 1, despite significant advancements in genome-wide association studies, a substantial portion of the genetic basis underlying complex traits remains unresolved, leading to the development of the CD-RV hypothesis [65]. In response, numerous statistical methods have been developed over recent decades to analyze and quantify associations between rare genetic variants and complex diseases. Each approach is characterized by distinct advantages and limitations, making them suitable for specific analytical frameworks and research goals. Rare variants, by nature, exhibit lower linkage disequilibrium compared to common variants, which contributes to a higher burden of multiple testing. Additionally, the low frequency of individuals carrying these variants poses challenges to achieving adequate statistical power. To address these issues, rare variant association studies often aggregate multiple variants within defined genetic units—such as genes, genomic regions, or functional categories—to enhance the ability to detect associations. In the following sections, we will provide a comprehensive examination of the most widely employed statistical tests used to evaluate associations between rare genetic variants and the risk of developing breast cancer. This discussion will delve into the methodologies, underlying principles, and practical applications of these tests, offering insights into

their strengths, limitations, and relevance in breast cancer research.

2.1 Classical approaches for rare variants association tests

In this section, we present an overview of Gene- or Region-Based Aggregation Tests for multiple genetic variants. These methods are designed to assess the cumulative effects of multiple variants within a specific gene or genomic region. By aggregating variants, these approaches enhance statistical power, particularly in scenarios where multiple variants within the group contribute to the risk of a disease or influence a specific trait. Broadly, these methods can be categorized into two primary classes: Burden Tests and Variance-Component Tests (Sequence Kernel Association Test (SKAT) and its optimized version (SKAT-O)). All these tests share a unified baseline statistical framework. Specifically, consider a dataset of n individuals each genotyped at m variant sites and measured for a phenotype y_i, $i \in \{1, \ldots, n\}$, with a mean denoted by μ_i, the tests assume that y_i follows a distribution within the quasi-likelihood family, and the relationship between the phenotype and predictors is modeled using a generalized linear model:

$$h(\mu_i) = \beta_0 + \alpha' \mathbf{X}_i + \beta' \mathbf{G}_i$$

Here, $h(\mu) = \mu$ for continuous phenotypes and $h(\mu) = \text{logit}(\mu)$ for binary phenotypes. The term β_0 represents the intercept, α' is the vector of coefficients for the covariates $\mathbf{X}_i$, and β' corresponds to the coefficients for the allele counts $\mathbf{G}_i$, which can take values of zero, one, or two for each variant allele[72]. The score statistic for the marginal model of a given variant j is mathematically defined as:

$$S_j = \sum_{i=1}^{n} G_{ij}(y_i - \hat{\mu}_i),$$

where $\hat{\mu}_i$ represents the estimated mean of y_i under the null hypothesis $H_0 : \beta = 0$ derived by the null model: $h(\mu_i) = \beta_0 + \alpha' \mathbf{X}_i$. The value of S_j provides an indication of the variant's effect. Specifically, $S_j > 0$ suggests that variant j is associated with an increased risk of the disease, while $S_j < 0$ implies a protective effect of the variant j. The primary distinction between the various tests lies in the methodology used to combine the single-variant into a composite genetic score that serves as the basis for the association test.

Burden Test

Burden tests aggregate information across multiple genetic variants into a single score, which is then tested for association with the phenotype. The genetic score for an individual i is defined as:

$$C_i = \sum_{j=1}^{m} w_j G_{ij},$$

where G_{ij} represents the allele count for variant j in individual i, and w_j is a weight assigned to variant j. This approach is equivalent to assuming $\beta_j = w_j \beta$ in a regression model and testing the null hypothesis $H_0 : \beta = 0$ in the simplified model:

$$h(\mu_i) = \beta_0 + \alpha' \mathbf{X}_i + \beta C_i.$$

The corresponding test statistic is:

$$Q_{\text{burden}} = \left(\sum_{j=1}^{m} w_j S_j \right)^2,$$

In this context, S_j represents the single-variant score statistic, which quantifies the association of individual genetic variants with the trait of interest. To evaluate the statistical significance of the aggregated association pattern, a p-value is derived by comparing the burden test statistic, Q_{burden}, to a chi-square distribution with one degree of freedom. This approach enables the detection of signals that are consistent with the hypothesized genetic model. Burden tests are particularly effective in scenarios where a substantial proportion of the variants within the genetic set contribute causally to the trait and exhibit effects aligned in the same direction. Such alignment amplifies the aggregated signal, enhancing the power of the test to identify associations.

Sequence Kernel Association Test (SKAT)

Variance-component methods, such as the Sequence Kernel Association Test (SKAT), provide a sophisticated framework for evaluating associations between genetic variants and phenotypic traits by modeling the variability in genetic effects rather than aggregating them into a single summary score. These tests specifically assess whether the observed variance in genetic effects across a set of variants deviates significantly from what is expected under the null hypothesis.

Under the null hypothesis, $H_0 : \tau = 0$, SKAT assumes that the effect sizes of individual variants, β_j, are independently distributed according to a normal distribution:

$$\beta_j \sim N(0, \tau w_j^2),$$

where τ represents the variance component, and w_j is a weight assigned to variant j, reflecting its relative contribution based on characteristics such as minor allele frequency or functional annotation.

The test statistic for SKAT is computed as:

$$Q_{\text{SKAT}} = \sum_{j=1}^{m} w_j^2 S_j^2,$$

where S_j denotes the single-variant score statistic, and w_j^2 provides a weighting mechanism that prioritizes variants according to their hypothesized importance. This statistic represents a weighted sum of squared score statistics for individual variants, effectively capturing the collective contribution of multiple genetic effects. To determine statistical significance, the p-value is derived by comparing Q_{SKAT} to a theoretical distribution modeled as a mixture of chi-square distributions, which accounts for the variability in effect sizes across variants.

A key strength of SKAT is its robustness in accommodating a mixture of genetic effects, including both protective and deleterious variants, within the same set. This flexibility makes SKAT particularly well-suited for studying traits influenced by complex genetic architectures.

SKAT-O Test

Omnibus tests, such as the Sequence Kernel Association Test-Optimal (SKAT-O), are designed to combine the strengths of both burden and variance-component methods, offering a robust and adaptable framework for detecting associations with rare genetic variants. These tests are particularly advantageous in situations where the genetic architecture is complex or poorly understood, as they integrate multiple approaches to enhance statistical power. In the case of rare variants, where the effects may be subtle or heterogeneous, SKAT-O provides a flexible method by combining the statistical power of burden and variance-component tests, which are applied to rare variants with low allele frequencies.

The SKAT-O test statistic is a linear combination of the test statistics from

SKAT and the Burden test, expressed as:

$$Q_r = (1 - r)Q_{\text{SKAT}} + rQ_{\text{burden}}, \quad 0 \leq r \leq 1,$$

where r is a weighting parameter that adjusts the contribution of each test statistic. The value of r is adaptively selected to maximize statistical power, depending on the underlying genetic effect of the rare variants. This allows SKAT-O to strike an optimal balance between the burden test's sensitivity to aggregated effects of rare variants and the SKAT test's ability to capture the distribution of genetic effects, whether protective or deleterious, across rare variants. By adaptively adjusting the weighting, SKAT-O is able to accommodate different types of genetic effects and ensure robust performance across various genetic architectures.

To compute the statistical significance of the combined test statistic Q_r, the p-value is obtained through one-dimensional numerical integration. This approach accounts for the complex distribution of the test statistic under the alternative hypothesis, providing a reliable method for obtaining the p-value while maintaining computational efficiency.

The primary advantage of omnibus tests like SKAT-O is their ability to handle the uncertainty of genetic architectures, especially in the case of rare variants where the effects may vary in direction and magnitude. By combining the strengths of both the Burden test and SKAT, SKAT-O provides a powerful tool for genetic association studies focused on rare variants. This flexibility makes SKAT-O particularly effective for investigating complex traits, where rare variants may play a significant yet subtle role in the overall genetic architecture. Despite its flexibility and numerous advantageous characteristics, SKAT-O demonstrates limitations in certain scenarios. Specifically, when the genetic effects of variants within a set are consistently unidirectional—such that all variants exert either uniformly protective or deleteri-

ous effects—the Burden test surpasses SKAT-O in statistical power. This superior performance arises from the Burden test's design, which aggregates the effects of all variants into a single score. In such cases, the cumulative signal is amplified, allowing the Burden test to more effectively detect associations. Conversely, SKAT-O, which balances the strengths of variance-component and Burden methods, may not fully leverage the aggregated effects when unidirectionality dominates, leading to a relative reduction in power under these specific conditions.

In conclusion, the methodologies described provide complementary strategies for rare-variant association testing, each suited to specific genetic scenarios. Burden tests are most effective when causal variants exhibit consistent effects, with all variants contributing either protective or deleterious influences in a uniform direction. This aggregation of effects enhances the statistical power of the test under such conditions. In contrast, variance-component tests excel in cases where the effects of causal variants are heterogeneous, encompassing both protective and deleterious influences. These tests are particularly robust when only a small proportion of variants within the set are causal, ensuring reliable performance despite variability in effect directions. Building on the strengths of these approaches, omnibus tests, such as the Sequence Kernel Association Test-Optimal, provide a flexible and adaptable framework that accommodates a wide range of genetic architectures. By integrating the characteristics of both burden and variance-component methods, SKAT-O is especially advantageous when the underlying genetic architecture driving the observed phenotype is unknown or complex. This adaptability makes SKAT-O a powerful tool for comprehensive rare-variant analysis, offering versatility and robustness in diverse genetic contexts.

2.2 Rare variants association analysis in population studies

In the preceding section, we provided a comprehensive survey of the primary statistical techniques utilized to evaluate the association between rare variants and phenotypic traits. In the subsequent section, we will explore the application of these previously described approaches within the framework of population-based studies. Particular attention will be given to their implementation in assessing the risk of breast cancer and to the integration of RV data with subjects' genotype profiles. As highlighted in Chapter 1, RVs constitute a crucial component of heritability and are indispensable for understanding the underlying mechanisms of disease pathogenesis. Therefore, their identification and characterization are paramount for advancing our comprehension of disease etiology and for fostering the development of innovative therapeutic interventions. The inherent challenge of attaining sufficient statistical power in RVs association studies stems from the low frequency of individuals carrying these variants. This limitation underscores the importance of utilizing large-scale datasets and comprehensive biobanks, such as the UK Biobank, to reach the statistical power required for the effective analysis of rare variant associations. In recent years, numerous studies have leveraged UKBB data, either independently or through integration with data from other cohorts, to investigate the role of rare variants in shaping complex traits, both continuous and discrete, and to explore their polygenic contributions. These investigations frequently employ association analyses that simultaneously evaluate relationships between rare variants and diverse phenotypes. Such an approach is designed to maximize statistical power by aggregating evidence across multiple phenotypic dimensions. Among these studies, the work by Wang et al. [91] represents a pivotal contribution to understanding the relationship between rare protein-coding variants and a broad spectrum of phenotypic

traits. This study utilized exome sequencing data from 269,171 UK Biobank participants of European ancestry and employed gene-based burden test to systematically evaluate associations across 17,361 binary and 1,419 quantitative phenotypes. The analysis identified 1,703 statistically significant gene-phenotype associations for binary traits, with a striking median odds ratio of 12.4, underscoring the substantial effect sizes often attributable to rare variants. Notably, 83% of these associations were not detectable through single-variant association tests, highlighting the unique strength of gene-based collapsing analyses. In another study Weiner and colleagues [36] investigated the contribution of rare coding variants to the heritability of 22 common traits and diseases by analyzing gene-wise burden across exome sequencing data from 394,783 UK Biobank participants. This comprehensive analysis provided critical insights into the genetic architecture of these traits, revealing that rare coding variants implicate a restrict number of large-effect genes. Their findings suggest that the contributions of rare coding variants to heritability and population risk stratification are relatively modest compared to those of common variants. Specifically, the study demonstrated that rare coding variants account for a much smaller percentage of phenotypic variance on average compared to common variants. Most of the burden heritability attributable to rare coding variants arises from ultra-rare loss-of-function variants, stressing the significant role of these variants despite their scarcity. The analysis further revealed a mechanistic convergence between common and rare variants: both types of variants implicate the same cell types, exhibit similar enrichment patterns, and display pleiotropic effects on the same pairs of traits. While there is partial colocalization at individual genes and loci, the extent differs: rare-variant burden heritability is strongly concentrated in significant genes and highly constrained regions, whereas common-variant heritability exhibits a more diffuse, polygenic distribution. These results highlight the complementary roles of common and rare variants in shaping the genetic architecture of complex traits and

diseases. They also emphasize the potential of rare coding variants to illuminate biological pathways involving high-impact genes, while acknowledging their relatively limited contribution to explaining missing heritability and enabling population-level risk prediction. In a similar vein, Backman et al.[61], in their study published on Nature in 2021, conducted an extensive analysis using exome sequencing to investigate the functional consequences of protein-altering variants in a cohort of 454787 participants from the UK Biobank. Their analysis uncovered approximately 12 million coding variants, which were categorized into loss-of-function and deleterious missense variants. These variants were systematically evaluated for their associations with 3994 health-related traits. The authors identified 564 genes exhibiting significant associations with these traits. Notably, rare variant associations were found to be enriched in loci previously implicated in genome-wide association studies; however, a striking 91% of these associations were independent of the common variant signals typically observed in such studies. Finally, the study by Karczewski et al. [15] utilized exome-sequencing data from 454,697 UK Biobank participants to systematically investigate the impact of predicted synonymous, predicted loss-of-function (pLoF), and deleterious missense rare variants across 4,529 phenotypes encompassing both binary and quantitative traits using both single variants and gene-based association technique (Burden and Variance-Component Tests). Their findings revealed that constrained genes—those subject to strong natural selection against pLoFs—exhibited stronger associations with a range of phenotypes, including various cancers. The analysis further demonstrated that RVs classified as pLoF or damaging missense mutations exhibit larger individual effect sizes compared to common variants. This disparity arises from the action of strong negative selection, which suppresses the population frequencies of these deleterious variants. Despite their larger effect sizes, the phenotypic variance captured by rare variants is limited due to their low frequencies in the population. In contrast, common variants,

because their higher frequencies, account for a greater proportion of phenotypic variance, even if their individual effect sizes are smaller. Consequently, for a given effect size, a more common variant contributes more significantly to phenotypic variance and displays a stronger statistical association. Negative selection, however, reduces the prevalence of functional damaging variants, leading to a tendency for variants with substantial effect sizes to be rare. They confirmed these results by partitioned heritability analyses of common variants that highlighted the interplay of opposing forces: while less frequent variants exhibit larger absolute effect sizes, the increase in effect size was insufficient to offset the reduction in variance explained due to their lower frequencies. A similar trend was observed in rare-variant association analyses, where the proportion of variants associated with at least one phenotype increased as their frequency rose. Overall, these results underline the important role of population-based studies in clarifying the contribution of rare variants to the heritability and pathogenesis of a wide range of complex traits and diseases. In the next subsections, we will focus on the association analyses of rare variants with BC risk, providing deeper insights into the key discoveries regarding genes that influence incidence of breast cancer.

2.2.1 Rare variants association analysis for breast cancer

The association between rare variants and the risk of breast cancer has been extensively investigated in recent years, using various classical association tests, including the Burden Test, SKAT, and SKAT-O, across different cohorts. A range of studies have provided important insights into how rare genetic variants contribute to breast cancer susceptibility. A research conducted by Hadda et al. [13] utilized data from the AMBER (African American Breast Cancer Epidemiology and Risk) Consortium, which included 8287 samples, comprising 3629 breast cancer cases and 4658

controls. The primary aim was to investigate rare variants associated with breast cancer risk in African American women. The study employed SKAT-O test, focusing on exome-wide variants with a minor allele frequency $\leq 5\%$. Although no general associations were found with breast cancer risk, two genes, FBXL22 and PDE4D, were associated with an increased risk for estrogen receptor-negative (ER-) breast cancer. Additionally, the SNP rs8100241 at 19p13.11 was identified as associated with a decreased risk of ER- breast cancer. Another significant contribution came from a work of 2021 [4] that analyzed the data from the Breast Cancer Association Consortium (BCAC), which included 60466 cases and 53461 controls. This study sought to evaluate the association of 34 known breast cancer susceptibility genes with rare LoF and missense variants. Using odds ratios and 95% confidence intervals, the authors found significant associations between LoF RVs in genes such as ATM, BRCA1, BRCA2, CHEK2, PALB2, TP53, BARD1, RAD51C, and RAD51D, with a Bayesian false discovery probability (FDP) < 0.05. Moreover, variants in PTEN, NF1, and MSH6 were also associated with breast cancer risk ($p-value < 0.05$). For missense variants, CHEK2, ATM, and BRCA1 were found to be significantly associated ($p-value < 1 \times 10^{-4}$). Similar findings where reported in a paper published on Cancers in 2022 [10], which analyzed data from the German Consortium for Hereditary Breast and Ovarian Cancer (GC-HBOC) and The Dutch Familial Bilateral Breast Cancer Study (DFBBCS). This study, which included 12655 samples (6518 cases and 6137 controls, excluding subjects carrying BRCA1/2 mutations), aimed to identify novel breast cancer susceptibility genes using whole-exome sequencing. For the association analysis, the Burden Test was applied to assess the presence of loss-of-function and missense rare variants across 32 genes known to be associated with breast cancer, in addition to 198 candidate genes. The study found that rare LoF variants in CHEK2, ATM, and PALB2 were significantly associated with breast cancer risk ($p - value < 0.001$), while TP53 showed a weaker association with a p-

value of 0.026. Among the 198 candidate genes, ZFAND1, TYRO3, and TMEM206 showed associations for the presence of rare LoF variants ($p - value < 0.05$), and TMEM161A, ERCC2, and TYRO3 were associated for the presence of rare missense variants ($p - value < 0.05$). Further combining LoF and missense variants, ABCC2, SMARCA2, ZFAND1, TMEM206/PACC1, TMEM161A, and SIPA1L1 emerged as significant with a $p - value < 0.05$. Finally, in a study published on Nature Genetics in 2023 [84] the researchers conducted a meta-analysis combining data from three large cohorts: BRIDGES, PERSPECTIVE, and UK Biobank. This analysis included a total of 26368 breast cancer cases and 217,673 controls. The authors aimed to define the contribution of rare protein-truncating variants (PTVs), rare missense variants, and predicted deleterious rare missense variants to breast cancer risk. The study found that rare PTVs in genes such as ATM, BRCA1, BRCA2, CHEK2, PALB2, and MAP3K1 were significantly associated with breast cancer risk, achieving exome-wide significance ($p - value < 2.5 \times 10^{-6}$). Other genes such as LZTR1, ATR, and BARD1 were also found to have significant associations with a $p - value < 1 \times 10^{-4}$, while CDH1 and RAD51D showed weaker associations ($p - value < 0.01$). In the case of rare missense variants, only CHEK2 reached exome-wide significance, while other genes like SAMHD1, HCN2, CLIC6, and ACTL8 were associated with a $p - value < 1 \times 10^{-4}$. To sum up, the analysis of rare variants has revealed that loss-of-function variants have a particularly strong impact on breast cancer risk, with robust associations observed for RVs in the five key susceptibility genes: BRCA1, BRCA2, ATM, CHEK2, and PALB2. These genes consistently emerge as significant contributors to BC risk across multiple studies. In addition to these primary genes, other genes such as TP53, BARD1, RAD51C, RAD51D, and MAPK1 also exhibit associations with BC risk. Notably, CHEK2 stands out for its strong association with missense variants, observed across several studies. While other genes show associations with varying degrees of strength, the

overall evidence suggests that the risk of BC is most strongly linked to the presence of rare variants in a limited set of genes when the analysis is conduced at the gene level.

2.2.2 Rare and common variants combination for breast cancer risk stratification

As previously discussed, the relationship between rare genetic variants and breast cancer risk has emerged as a focal point of scientific inquiry, with a growing number of studies exploring how these rare variants might be integrated into existing risk prediction models. Considerable evidence suggests that combining polygenic risk scores with rare pathogenic variants from specific high-risk and moderate-risk genes holds promise for improving the accuracy of BC risk assessments. One prominent example of this approach is the BOADICEA model, as described in the study by Antoniou et al. (2019) [3], which integrates rare variants in genes such as BRCA1, BRCA2, PALB2, CHEK2, and ATM, which are known to be associated with high and moderate BC risk. In this model, rare variants in these key susceptibility genes are combined with a polygenic risk score derived from 313 SNPs. Notably, the BOADICEA study conceptualizes a "major" locus composed of six alleles, which includes five of the aforementioned genes, as well as the wild-type (WT) allele, and operates under the assumption of a dominant genetic model of risk. Additionally, the model incorporates non-genetic risk factors, such as mammographic density, family history of BC, and a range of lifestyle factors, including reproductive history, BMI, alcohol consumption, and other hormonal and environmental influences. The results from this study demonstrated that the PRS contributed the most significant factor for stratifying individuals according to their BC risk, followed closely by mammographic density. Importantly, when all risk factors—both genetic and

non-genetic—were considered together, the model was able to provide more precise risk stratification, which emphasizes the value of integrating rare genetic variants into comprehensive, multi-faceted risk assessment tools. A similar methodology was employed in another study published in Genetics in Medicine (2022) [14], which assessed the interactions between polygenic risk, rare pathogenic variants, and family history in a large cohort of 200643 women from the UK Biobank. This study specifically focused on the effects of rare pathogenic variants in high-risk genes (BRCA1 and BRCA2) as well as moderate-risk genes (ATM, CHEK2, PALB2), and their combined impact on BC risk when taken together with the polygenic risk score. The findings revealed that women heterozygous for rare pathogenic variants in the high-risk BRCA1 and BRCA2 genes exhibited a significantly higher risk of BC than women with a high PRS alone. Interestingly, the study also found that women with rare pathogenic variants in moderate-risk genes (ATM, CHEK2, PALB2) and a low PRS had a relatively lower BC risk compared to women with just a high PRS, suggesting that the polygenic risk score may modulate the penetrance of rare variants in both high- and moderate-risk genes. This highlights the potential of the PRS to influence how the genetic risk associated with rare variants is expressed, emphasizing its role in refining risk prediction models. Building upon these findings, a more recent study published in Communications Biology (2024) [62] further explored the integration of PRS and rare pathogenic variants in high-risk (BRCA1, BRCA2) and moderate-risk (ATM, CHEK2, PALB2) genes in order to enhance BC risk stratification. The study employed a clustering algorithm known as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to group genes based on their associated risk levels and used multivariate logistic regression models to calculate the odds ratios for different PRS categories and the presence of rare pathogenic variants. The results of this study strongly reinforced the earlier findings that while PRS alone is a valuable tool for discriminating BC risk, the addition of rare variant information,

particularly from high-risk genes such as BRCA1 and BRCA2, greatly improved the precision of risk stratification across all PRS categories. In particular, women with both a high PRS and rare variants in BRCA1 exhibited the highest risk of BC, with a prevalence rate of 0.67. Furthermore, the presence of rare variants in the moderate-risk genes, even in individuals with a low PRS, was also associated with higher odds of BC compared to those with a high PRS but without rare variants. These results demonstrated the critical role of integrating both genetic factors into the risk prediction process.

Together, these studies provide compelling evidence for the integration of both PRS and rare pathogenic variants into comprehensive BC risk prediction models. While PRS alone remains a powerful tool for stratifying individuals according to their genetic risk, the addition of rare variants in critical susceptibility genes, particularly BRCA1 and BRCA2, significantly enhances the accuracy of risk assessments. This combined approach refines the identification of individuals at the highest genetic risk guiding clinical decision-making and improving patient outcomes.

2.2.3 Limits of current approaches

The studies outlined in the preceding sections underscore the intricate role of rare variants in conferring breast cancer risk, highlighting both well-established susceptibility genes and emerging candidates. Genes such as BRCA1, BRCA2, ATM, CHEK2, and PALB2 consistently exhibit strong associations with BC risk. At the same time, the discovery of novel genes reinforces the necessity of comprehensive and systematic genomic approaches to fully elucidate the genetic architecture of BC. Importantly, the modulation of rare variant penetrance by polygenic risk scores in both high- and moderate-risk genes indicate the value of integrating these genetic factors into risk prediction frameworks. This integration emphasizes the need for a

robust scoring system that systematically combines rare and common genetic variants while quantifying the strength of association between rare variants and BC risk. Most studies to date have examined the association between PRS and RVs using predefined lists of genes within models that are not readily adaptable to incorporate new genetic discoveries. For instance, the BOADICEA model focuses on a curated set of genes with well-established links to BC, operating under the assumption of a single "major" locus scheme. However, this assumption becomes increasingly problematic as the number of susceptibility genes included expands. Even when more flexible and systematic gene selection approaches are employed, as demonstrated in studies such as [62], the practical implementation of methods that combine PRS with RVs becomes progressively complex with the inclusion of additional genes. Specifically, as more gene clusters are considered, the number of comparisons escalates dramatically, complicating statistical and computational feasibility. Conversely, grouping a growing number of genes into a smaller set of clusters introduces challenges in determining which genes within each cluster drive the observed associations with BC risk. These issues highlight the balance required between expanding genetic inclusivity and maintaining the interpretability and practical application of BC risk models. Finally, conducting breast cancer association analyses at the gene level often fails to provide insight into the specific effects of individual rare variants. In a 2022 paper published in Nature Genetics [90], Dornbos and colleagues introduced a polygenic score integrating both common and rare variants to enhance diabetes diagnosis. Building on this concept, the next chapter explores two distinct approaches for developing a combined score incorporating rare variants and polygenic risk scores to predict breast cancer occurrence. The first approach employs classical gene-based association analysis tests, while the second leverages a Bayesian model to identify RVs associated with BC

Chapter 3

UK Biobank Data

In the first part of this chapter (section 3.1) we will describe the UK Biobank (UKBB) data used for the analysis and list the steps applied for sample quality control and variants selection. On the other hand, the second part (section 3.2) will report the results obtained from this data processing phase.

3.1 UK Biobank

This research has been conducted using the UK Biobank Resource under Application Number 78537. The UK Biobank project stands as a pioneering prospective cohort study, encompassing a vast repository of genetic and phenotypic data gathered from nearly 500,000 individuals residing throughout the United Kingdom. Starting from 2006, participants were recruited from 22 centers located in Scotland, England, and Wales. Spanning an age range of around 40 to 69 years at enrollment, this open-access resource is distinguished by its unprecedented scale and comprehensive scope: each participant contributes to a diverse range of phenotypic and health-related data. This includes biological measurements such as blood pressure, cholesterol levels, and BMI, alongside lifestyle indicators encompassing smoking status, physical

activity levels, and dietary habits. This information is further enriched by data from questionnaires delve into various aspects of participants' lives, capturing details on occupational history, pain perception, cognitive function, digestive health, and mental well-being. Comprehensive data on over 30 key biochemistry markers derived from blood, urine, and saliva samples provide valuable insights into participants' biochemical profiles, while enhancements such as comprehensive magnetic resonance (MR) imaging of the brain, heart, and full body, alongside DEXA scans for bone and joint health, ultrasounds of the carotid arteries, and physical activity patterns monitoring through wrist-worn devices over a 7-day period for 100,000 participants (supplemented by seasonal follow-ups for a subset) provide additional layers of information. Integration with electronic health records spanning mortality, cancer incidences, hospital admissions, and primary care interactions facilitates robust health linkages, enriching the dataset with longitudinal health information. The study benefits from ongoing data updates and meticulous collection of genome-wide genotype data from all 500,000 participants, enabling the exploration of novel genetic associations and the understanding of complex traits and diseases. Cutting-edge technologies such as whole genome sequencing for all participants, whole exome sequencing for a large subset of 470,000 patients, and genotyping encompassing a vast array of genome-wide variants (around 800,000), augmented by imputation to a staggering 90 million variants, further enhance the genetic data available for analysis. Finally, with repeat baseline assessments during the initial imaging phase for 100,000 participants and ongoing sample collections of blood, urine, and saliva, the UK Biobank continues to evolve as a vital resource for advancing biomedical research. UK Biobank data can be accessed through their data showcase, which is organized into data fields. A data field represents the basic unit of information within the UK Biobank, capturing the results of surveys or measurements.

The following sub-sections provide details on the data-fields used for the analysis

in this study.

Clinical Data-Fields

In this study, we extracted data from the following data-fields concerning patients clinical and baseline characteristics:

- **Data-Field 21022: Age at recruitment**: for this field, there are 502,357 data entries, each corresponding to a participant. The unit of measurement is years. This variable is derived from the participant's date of birth (data-field 33) and the date of their initial assessment center visit (data-field 53, instance 0), indicating the participant's age on the day of their initial assessment center visit, rounded to the nearest whole year.

- **Data-Field 31: Sex**: this field contains 502,357 data items, corresponding to 502,357 participants, denoting the participant's gender, where the 273165 Female have been codified as 0 and the 228985 Male as 1. Initially obtained from the central registry upon recruitment, it may have been updated by participants themselves, resulting in a blend of both NHS-recorded and self-reported gender.

- **Data-Field 22001: Genetic Sex**: this field reports the Sex determined by genotyping analysis. There are 488,118 data items, each corresponding to one participant. Participants are indicated with 0 for females and 1 for males. The genetic sex breakdown reveals 264,589 females and 223,332 males within the UKBB cohort. Notably, in 38 cases the genetic sex differs from the values previously calculated by the "Interim 150K" version. This difference results from the re-genotyping of samples deemed to be of inadequate quality by aligning the genetic sex with the information provided in field 31 of the central registry at the time of enrollment, which was updated in case of incorrect

reporting by the participant. In addition, more than 300 cases had a disparity between genetic sex and the self-reported value in field 31.

- **Data-Field 22019: Sex chromosome aneuploidy**: a total of 651 data entries are available, corresponding to 651 participants identified with Sex chromosome aneuploidy markers. These markers indicate samples potentially carrying sex chromosome configurations other than XX or XY. Identification was based on analyzing average log2Ratios for Y and X chromosomes

- **Data-Field 1647: Country of birth (UK/elsewhere)**: This data-field boasts 533,456 data points, representing 501,463 individuals responses to the ACE touchscreen question "Where were you born?". For participants born in the United Kingdom, their place of birth was explored in depth during verbal interviews (Field 129 and Field 130). Whereas, participants born outside the United Kingdom and the Republic of Ireland were subjected to verbal interviews that probed country of birth (Field 20115) and immigration details (Field 3659). Notably, this specific question was removed from the touchscreen protocol as of October 24, 2016.

- **Data-Field 20115: Country of Birth (non-UK origin)**: A total of 40,556 data entries are available. Participants are categorized based on their continent of birth: Africa (12,754), Asia (11,852), Europe (10,155), North America (2,715), Oceania (1,702), and South America (1,378). This field documents participants' countries of birth as provided during interviews. The survey targeted only those who indicated they were born outside the United Kingdom or the Republic of Ireland during the touchscreen questionnaire (Data-Field 1647). However, some participants later reported locations within the United Kingdom or Republic of Ireland, which were recorded accordingly. Responses have been grouped by continent. The coding process occurred post-assessment

visit, utilizing both pre-selected options and free-text entries. Entries that couldn't be mapped onto a specific region have been excluded from the dataset.

- **Data-Field 21000: Ethnic background**: 547,843 available outlines coding the ethnic background of patients: White (517506), Mixed (3191), Asian or Asian British (228), Black or Black British (8356), Chinese (1690), Other ethnic group (4764), Do not know (226) and Prefer not to answer (1778)

- **Data-Field 41270: Diagnoses - ICD10**: This field collects distinct diagnosis codes extracted from participants' medical records, either as primary or secondary diagnoses, according to the International Classification of Disease version 10 (ICD-10) guidelines. The dataset comprises 7,018,114 entries, pertaining to 446,996 participants with 23,415 entries specifically allocated for the diagnosis of Malignant neoplasm of the breast (ICD10 code C50)

- **Data-Field 26220: Standard PRS for breast cancer (BC)**: 485,923 items of data are available, covering 485,923 participants. Polygenic risk scores (PRS) were generated using a Bayesian approach applied to meta-analyzed genome-wide association study (GWAS) summary statistics. Specifically, UKBB Standard PRS for breast cancer was built using summary statistics from external GWAS: the Breast Cancer Association Consortium (BCAC) and CIMBA (Consortium of Investigators of Modifiers of BRCA1/2). SNPs included in the PRS were required to INFO score > 0.8 in UKB; have an INFO score > 0.8 in the GWAS meta-analysis dataset; have an INFO score > 0.7 in other key reference datasets available to Genomics plc; not display large differences in allele frequency between UKB genetically inferred ancestry groups and either Gnomad or 1000 Genomes Project (absolute allele frequency difference between Gnomad and UKB of less than 0.2 in any ancestry group, $p > 1e-12$ and $p > 1e-10$ for Gnomad and 1000 Genomes Project respectively in any

ancestry group); and not display evidence ofmlarge departures from Hardy-Weinberg Equilibrium ($p > 1e - 10$) in any ancestry group. The variants also needed to have a definitive one-to-one mapping between Genome Builds 37 and 38. Indels, the pseudoautosomal regions, and any variants with MAF < 0.05 in the 1000 Genomes Project dataset (for any ancestries used as LD reference panels in the PRS generation step) were excluded from the PRS computation. PRS algorithms were developed through trait-specific meta-analyses using a Bayesian approach, integrating data across multiple ancestries and related traits where applicable. Individual raw PRS values were computed by summing the genome-wide per-variant posterior effect sizes, each weighted by allele dosage. Subsequently, a centering and standardization step was applied to raw PRS aiming to generate a corrected PRS value, ensuring it follows a distribution with approximately zero mean and unit variance for individuals with similar positions in "ancestry space." First, the PRS was centered using the method described by Khera et al.[69], by subtracting the predicted PRS value obtained from a linear regression of the PRS against the first four principal component scores, based on 1000 Genomes Project individuals. Next, each individual's genetic ancestry was inferred. Finally, the centered PRS was standardized by dividing it by the standard deviation of the PRS within the closest matching ancestry group in the 1000 Genomes dataset, resulting in a variance-standardized PRS. To further refine the distributions, PRS values were standardized to have an approximate unit variance within each ancestry group, as inferred from geometric relationships in PC space. The UKBB Standard PRS for breast cancer was developed using only external GWAS data, reducing the risk of overfitting within the UK Biobank dataset. The so calculated PRS was compared to PRS agorithm from The PGS Catalog database (https://www.pgscatalog.org/), excluding PRS algorithms which had used any

UKB GWAS data in the training stage. Further details on the UK Biobank PRSs computation can be found in the reference article [107]

Genomics Data-Fields

Regarding genomic information, we extracted data from the following data-fields:

- **Data-Field 22009: Genetic principal components**: 19,524,720 items of data are available related to the score for each principal component (1-40) covering 488,118 participants. Principla components analysis (PCA) have been computed according to the QC genotype documentation [22]: briefly two rounds of PCA were conducted to address different objectives; the first round involved identifying unrelated samples and computing the top 8 principal components to aid in sample quality control and relatedness inference, while the second round computed the first 40 principal components for comprehensive population structure assessment. The kinship estimation was performed initially to identify unrelated individuals, followed by sample and SNP filtering based on various criteria such as missing rates, MAF, and long-range LD. The genotype data were then processed to compute principal components using fastPCA, and all samples were projected onto the principal components. The outcomes of this analysis were utilized to compute PC-adjusted heterozygosity and to enhance the precision of relatedness inference. After the identification of a cohort comprising high-quality, unrelated samples, a subsequent round of PCA was conducted, computing the initial 40 principal components made available to researchers.

- **Data-Field 22418: Genotype calls**: genotype calling was conducted by Affymetrix utilizing two purpose-designed arrays tailored for closely related purposes. Approximately 50,000 participants underwent genotyping on the

UK BiLEVE Axiom array (Resource 149600), while the remaining roughly 450,000 were genotyped on the UK Biobank Axiom array (Resource 149601). A total of 488,127 items are now available, corresponding to 488,127 participants. This dataset integrates outcomes from both arrays, resulting in 805,426 markers in the released genotype data. Marker positions in the dataset are referenced to GRCh37 coordinates. Notably, genotypes for a fraction of participants (3%) could not be assayed due to inadequate DNA extraction from their blood samples. The genotype data underwent rigorous quality control (QC) procedures. Additionally, the dataset underwent phasing, with approximately 96 million genotypes imputed using computationally efficient methods alongside resources from the Haplotype Reference Consortium and UK10K haplotype resources. Imputation also encompassed classical allelic variations at eleven HLA genes. Information regarding the QC pipeline, including array specifics, as well as crucial genetic characteristics such as population structure and relatedness, are available. Further insights into these analyses and methodologies for deriving additional data such as imputation and haplotypes, are reported in [9]

- **Data-Field 23158: Population level exome OQFE variants, PLINK format - Final exome release:** exomes were acquired using the IDT xGen Exome Research Panel v1.0, including additional probes, targeting 39 Mbp of the human genome encompassing 19,396 genes. Samples were multiplexed and sequenced with dual-indexed 75x75 bp paired-end reads on the Illumina NovaSeq 6000 platform. The initial 50k release sequencing utilized S2 flow cells, while all subsequent samples were sequenced using S4 flow cells. On average, 95.2% of targeted sites in each sample achieved coverage exceeding 20X. Detailed sequencing protocols are provided in the summary manuscript. The

analysis for the UKB final release utilized an updated Functional Equivalence (FE) protocol (original quality functionally equivalent, OQFE protocol) which preserves the original quality scores in the CRAM files and aligns and marks duplicate raw sequencing data (FASTQs) to the full GRCh38 reference in an alt-aware manner. Subsequently, the OQFE CRAMs were used for small variants calling with DeepVariant leading to the production of per-sample gVCFs. These gVCFs were then aggregated and jointly genotyped with GLnexus to produce a single multi-sample VCF (pVCF) for all UKB samples. PLINK [97] [24] files were directly derived from the pVCF. It is important to note that no variant- or sample-level filters were initially applied to the pVCF or PLINK files to ensure broad support for data analysis. The publicly released pVCF, which contains allele-read depths and genotype qualities for all genotypes, serves as the direct output of GLnexus and is used to generate the PLINK files. A total of 469,602 bulk items are available, corresponding to 469,602 participants.

Phenotype data were extracted using the Cohort Browser on the UK Biobank Research Analysis Platform (RAP).

3.1.1 Sample data processing

Given that only 1% of breast cancer cases occur in the male population, and certain genes linked to BC in women are associated with prostate cancer in men [94], we selected the 273294 declared females from the whole set of UK Biobank patients in order to reduce the number of false positive. Moreover, we excluded from the analysis those subjects for which the genetic sex was different from the declared gender and those individuals affected by sex chromosome aneuploidies. We removed also all patients for which the Diagnoses-ICD10 wasn't available. Then, we defined

the binary variable indicating the presence of BC assigning to the case group all the individuals that have been diagnosed with code C50 at least once according to the Diagnoses-ICD10. We did the same considering the ICD10 code of all the other tumors type in order to exclude from the controls all the patients affected by any other kind of cancer. Table 3.1 shows the ICD10 codes indicating the other cancer types.

Additionally, we selected all the women with white ethnic background born in Europe (excluding Finland) according to the data fields 210008, 20115 and 1647. To enhance the quality of the selected samples on the base of their ancestry, we applied a further filter on the declared European women. Specifically, following the idea of [70] we used PLINK2 to project each individual of UKBiobank on the same PCs space of the Human Genome Diversity Project (HGDP) and the 1000 Genomes Project samples according to the pipeline described here [67]. Consequentially, using as covariates the first 10 PCs, we trained a Random Forest (RF) model to assign the HGDP and 1000 Genomes patients to one of the ancestry class: AFR, AMR, CSA, EAS, EUR, MID, and OCE. The RF was implemented using the R package caret version 6.0-94 (R 3.4.0), with 100 trees, nodesize parameter equal to 5 and with 10-fold cross-validation. Finally, for each patients in UKBiobank we predicted the probability of being allocated to each of the mentioned ancestry using the previous trained RF. We assigned each individual to the class with maximum probability. Since some patients had a maximum probability class score below 0.3, we opted to use the hybrid method proposed by [70], aiming to improve the quality of our data. Specifically, we computed the SHAP (SHapley Additive exPlanations) [73] values of the trained RF and select the PCs with the highest impact on the model's prediction of the European ancestry. We used these PCs to run two different clusterization algorithm: K-means[98] and DBSCAN[33]. We used a grid search method to set the algorithms parameters, selecting those that maximized the Average Silhouette

Table 3.1: Other cancer types ICD10 codes

Cancer Type	**ICD10 Codes**
Malignant neoplasms of lip, oral cavity, and pharynx	C00, C01, C02, C03, C04, C05, C06, C07, C08, C09, C10, C11, C12, C13, C14
Malignant neoplasms of digestive organs	C15, C16, C17, C18, C19, C20, C21, C22, C23, C24, C25, C26
Malignant neoplasms of respiratory and intrathoracic organs	C30, C31, C32, C33, C34, C35, C36, C37, C38, C39
Malignant neoplasms of bone and articular cartilage	C40, C41
Melanoma and other malignant neoplasms of the skin	C43, C44
Malignant neoplasms of mesothelial and soft tissue	C45, C46, C47, C48, C49
Malignant neoplasms of female genital organs	C51, C52, C53, C54, C55, C56, C57, C58
Malignant neoplasms of male genital organs	C60, C61, C62, C63
Malignant neoplasms of the urinary tract	C64, C65, C66, C67, C68
Malignant neoplasms of the eye, brain, and other parts of the central nervous system	C69, C70, C71, C72
Malignant neoplasms of thyroid and other endocrine glands	C73, C74, C75
Malignant neoplasms of ill-defined, other secondary, and unspecified sites	C76, C77, C78, C79, C80
Malignant neoplasms of lymphoid, hematopoietic, and related tissue	C81, C82, C83, C84, C85, C86, C87, C88, C89, C90, C91, C92, C93, C94, C95, C96

Index. In particular, for DBSCAN we evaluated values of epsilon, the radius of the epsilon neighborhood, in a sequence of numbers ranging from 0.1 to 1.0, with each successive number increased by 0.1, and we considered $(5, 10, 20, 30, 40, 50)$ minimum points required in the eps neighborhood for core points (parameter minPts). For k-means, istead, we tested from 2 to 10 number of clusters. Then, we applied both the clusterization methods to the data choosing the one providing the highest Average Silhouette Index. The k-means algorithm was the best fit for our data. Therefore, we selected as European Female samples all the women declared European, allocated in the EUR ancestry class by the RF and belonging to the K-means group composed for the majority part by EUR subjects. We divided our final sample in two sets: the 75% of the subjects were used for the association analysis of RVs with BC, while we used the remaining 25% to validate our findings.

3.1.2 Genomic data processing

The steps for the genomic data processing were executed using the app Swiss Army Knife version 4.13.0 on the UK Biobank Research Analysis Platform (RAP)

Exome data

The WES data from the UK Biobank's release of 500,000 participants were processed using the OQFE mapping protocol. Variants were identified with DeepVariant and combined into a multi-sample VCF using GLnexus. Finally, the pVCF files were converted to PLINK format using PLINK 1.9. The multi-sample VCF includes per-genotype metrics such as depth and genotype quality [71]. In order to ensure reliable data for the analysis, we retained only variant sites with at least 90% of the genotypes having read depth (DP)$>$ 10 according to the "90pct10dp" QC filter. Subsequently, we extracted the variants from the whole set of 6700 genes belonging the True Sight

One Sequencing Panel and compute their MAF across the 500,00 individuals in the overall population using PLINK 1.9. We removed all variants with MAF > 0.01 to select just RVs and we excluded all variants with missing call rates exceeding 15%. Next, we selected all the RVs annotated as "LoF" and "Missense" by snpEff effect predictions [25] using the Ensembl v85 gene definitions to determine their functional impact on transcripts and genes. We further annotated and selected missense RVs using the Variant Effect Predictor (VEP) plugin dbNSFP v4 [113] that integrates various annotation sources (among those ClinVar, SIFT, PolyPhen, and CADD). We stratified missense RVs into two groups according to their damaging impact based on the results provided by these annotations. The details of the division in two groups of the missense RVs are illustrated in Figure 3.1.

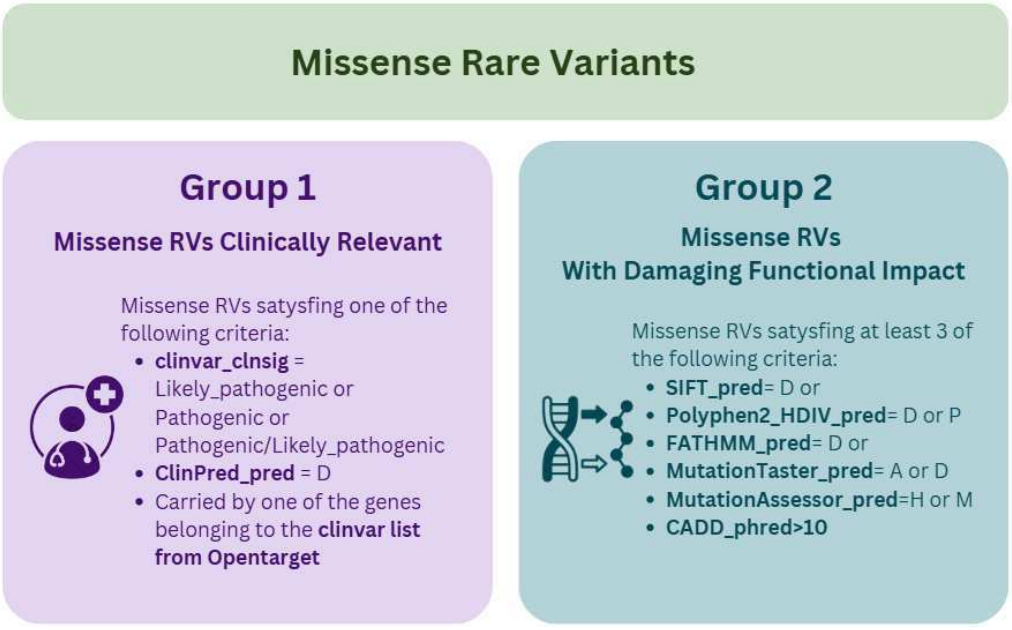


Figure 3.1: **Missense RVs Clustering Criteria**

SNP array data

We used genotype calls data provided by UKBB to estimate the genetic background of each patient when performing the RVs-BC association analysis through classical approaches. The UK-biobank SNP array genotype calls (bulk, genotype calls) are mapped to GRCh37, while the whole-exome sequences are mapped to GRCh38 human reference assembly. We chose to lift the coordinates of genotype calls over to GRCh38 before conducting the rest of the analysis for consistency. Specifically, we used PLINK (1.9 and 2.0) to combine separate chromosome-level genotype files into a single merged dataset and execute a SNP filtering step based on quality criteria as in [39] [3] [74] [44]:

- we removed variants that violate Hardy-Weinberg equilibrium, which could indicate genotyping errors, setting a very low p-value threshold ($1e-50$) to exclude only extreme violations
- we set a minimum allele frequency of 5%, to focus on more common SNPs
- we selected SNPs and samples with less than 15% of missing data, ensuring reliable data for analysis
- we filtered out SNPs with a minor allele count below 20, removing low-confidence variants
- we excluded variants in Linkage Disequilibrium (LD) using the PLINK pruning command with a 1,000 kb window, 50 SNP step size, and an LD threshold of 0.4. This step identifies and removes highly correlated SNPs, leaving only independent variants.

Then, we converted the QC-passed binary PLINK files to VCF format, preparing for coordinate liftover from genome build GRCh37 to GRCh38. The liftover step was

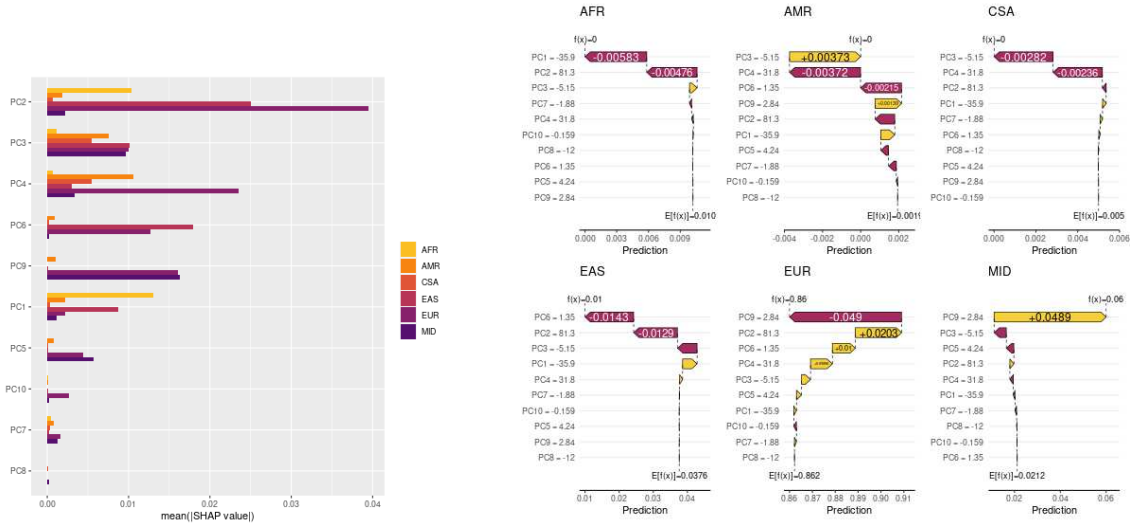
executed using Picard 3.0 [55], a tool from the Broad Institute that allows conversion of genomic coordinates in VCF files using a chain file. We then used bcftools [31] to sort and compress the produced VCF file and PLINK 2.0 to convert it back to PLINK format.

3.2 Results of UK Biobank data

In this section we will go trough the results of the UK Biobank data processing steps previously described.

3.2.1 Sample data processing and description

As mentioned in section 3.1.1, the analysis was executed selecting the 273165 declared females from the whole set of UK Biobank patients. Once we removed subjects for which the genetic sex was different from the declared gender, those individuals affected by sex chromosome aneuploidies and selected samples for which the exome sequence data were available, we obtained a sample size of 254149 women. This number was further reduced to 197191 patients by filtering out those for which the Diagnoses-ICD10 wasn't available, those having prostate cancer and excluding from the controls all the patients affected by any other kind of tumors. Figure A.1 in the Appendix A shows the number of subjects affected by other malignant neoplasms among cases and controls before the filter. From the previously selected dataset, we identified a subset of 184,350 samples consisting of individuals of self-reported white ethnic background born in Europe (excluding Finland). To assign each individual to a specific ancestry, we utilized a Random Forest classifier, as described in section 3.1.1. The performance metrics of the RF model on the test set are summarized in Table A.1, while Figure 3.2 illustrates the SHAP values for the features used in the classification. Figure 3.2, subfigure (a), reveals that principal components 2, 3, 4, 6,



(a) SHAP Importance Plot

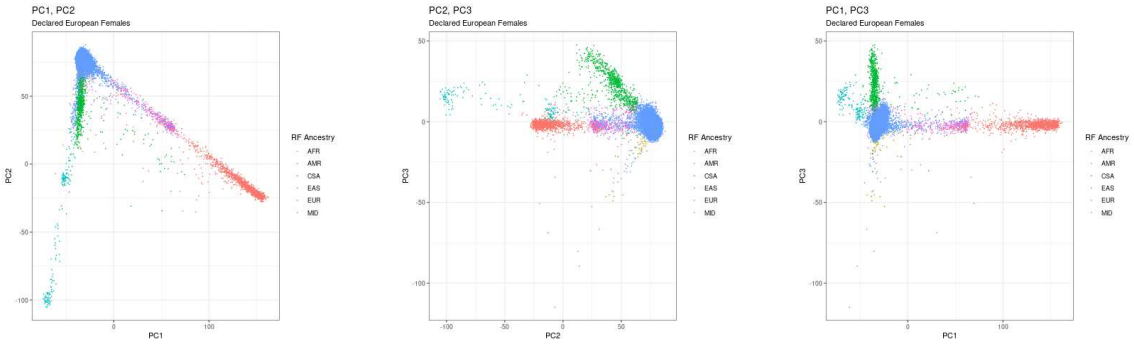
(b) SHAP Waterfall Plot

Figure 3.2: Random forest SHAP values plots

and 9 exert the most significant influence on the model's predictions for European ancestry. Notably, as depicted in Figure A.2 in the Appendix A, a subset of subjects exhibited RF-estimated ancestry probabilities below 0.25. Given that a probability of 0.17 corresponds to random assignment, these individuals can be classified as having "low-confidence" ancestry assignments. For this reason, building on relevant PCs insights, we employed two complementary clustering algorithms—K-means and DBSCAN—to enhance the granularity of classification. The DBSCAN algorithm yielded an average Silhouette index of -0.36381, indicating poor cluster separation, grouping the patients in 4 groups of 183982, 358, 5,and 5 patients respectively; while K-means produced an average Silhouette index of 0.90, suggesting clear and distinct cluster separation and dividing the subjects in 2 groups of 2358 and 181992 individuals. For the final selection, we retained self-declared European female samples that were classified as European (EUR) by the RF model and assigned to the K-means cluster predominantly composed of EUR subjects. This procedure resulted in a refined cohort of 180,935 individuals. Figure 3.3 illustrates the projection of

self-declared European females onto the space defined by the first three principal components. Subfigure (a) depicts the distribution colored by RF-predicted ancestry, while subfigure (b) represents the clustering results from K-means. The figure demonstrates a strong concordance between individuals classified as EUR by the RF model and those assigned to the primary EUR cluster by K-means, validating the reliability of the combined RF and clustering approach in accurately capturing European ancestry within this cohort.

(a) RF predicted ancestries



(b) K-means clusters

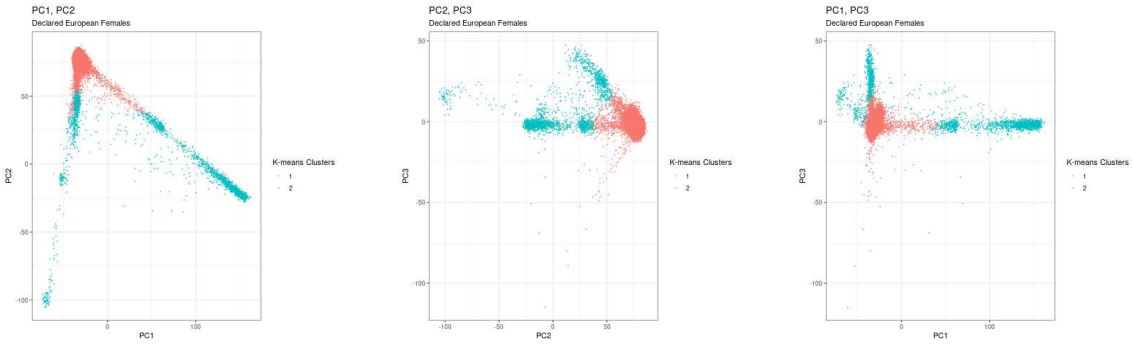


Figure 3.3: Projection of the declared European female subjects in the space of the first 3 PCs

The final population consisted of female subjects aged between 40 and 71 years, with an average age of 56.42 years (median = 58, IQR = 13.78). Age information was unavailable for 358 patients. The ultimate sample included 15,868 cases (8.77%) and

165,067 controls (91.23%), as classified by ICD10 codes. Cases were slightly older than controls, with a median age of 60 years compared to 57 years in the control group ($p - value \leq 2e^{-16}$) as we can see from figure 3.4.

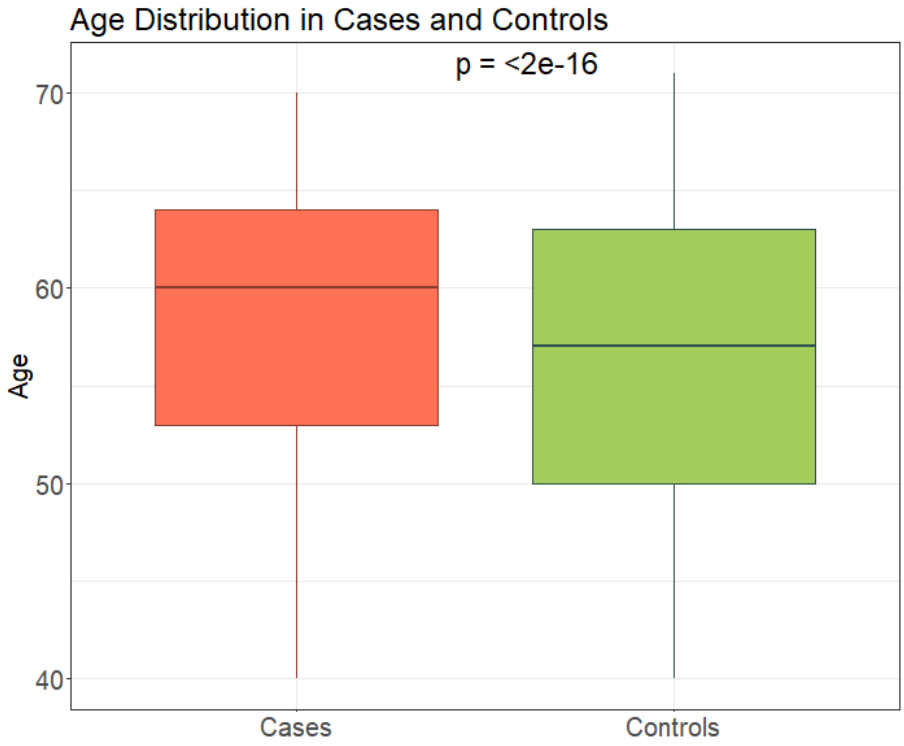


Figure 3.4: Age distribution in cases and controls with Wilcoxon rank-sum test p-value to compare median age

We divided this data into two groups:

- The first group, used for the BC-association analysis, consisted of 135,804 individuals (75% of the total sample), including 11,911 cases (8.77%) and 123,893 controls (91.23%).
- The second group, comprising the remaining 25%, included 45,131 individuals, with 3,957 cases (8.76%) and 41,174 controls (91.24%).

The diagram in figure 3.5 summarize the sample data processing steps described above.

3.2.2 Rare Variants

To enlarge the set of rare variants associated with BC beyond known susceptibility genes, we extended the analysis to the complete set of 6700 genes included in the Clinical Exome,selecting high quality RVs initially annotated as LoF or missense by the snpEff tool. Subsequently, a rigorous filtering process was applied to the missense RVs, retaining only those that were categorized as 'Clinically Relevant' (Group 1) or 'With Damaging Functional Impact' (Group 2) on the based of the annotations derived from the VEP plugin dbNSFP v4, as shown in Figure 3.1. This strategy lead to a final set of 176203 RVs: 82738 LoF (47.0%), 63100 missense RVs belonging to Group 1 (35.8%), and 30365 missense RVs in the Group 2 (17.2%) that was subsequently tested.

Although RVs classified as missense in the Group 2 account for only 17% of the selected variants, they are the most frequent as can be noticed in Figure 3.6. Missense RVs in Group 2 have a median MAF of 3.01×10^{-4}, which is considerably higher than the median MAF of missense RVs in Group 1 (3.79×10^{-5}, pairwise-comparison Wilcoxon test $p_{\text{adj}} \leq 2 \times 10^{-16}$), and higher also than the median MAF of LoF RVs, which is 2.42×10^{-5} (pairwise-comparison Wilcoxon test $p_{\text{adj}} \leq 2 \times 10^{-16}$). This difference in allele frequencies translates into a disparity in the number of patients carrying these variants in the selected subpopulation. Specifically, the median number of patients harboring LoF RVs is 1 ($IQR = 2$). For missense RVs in Group 1, the median number of patients is 3 ($IQR = 6$), while for missense RVs in Group 2, the median number of patients increases to 7 ($IQR = 26$). These observations are consistent with the biological interpretation of the impact of the three groups of RVs under investigation. Specifically, LoF variants, which typically result in a complete and often irreversible disruption of gene function, are frequently associated with severe phenotypic consequences and effects. Given their potential to disrupt

essential biological processes, it is expected that LoF variants would be relatively rare in the population, as they are subject to strong negative selection pressure. Similarly, missense variants classified within Group 1, which are expected to have a high likelihood of contributing to clinically significant conditions, show a relative moderate frequency in the population. Lastly, the missense variants in Group 2 exhibit a wide IQR in the number of patients carrying them, suggesting a high degree of variability in their distribution across the population. The damaging impact and prevalence of each category of RVs were further evident in the distribution of RV counts among patients. Specifically, the median number of RVs per patient was 2 for LoF RVs ($IQR = 3$), 4 for missense RVs in Group 1 ($IQR = 3$), and 14 for missense RVs belonging to Group 2 ($IQR = 5$)

In addition, for some of the analysis of the presented work, we tested the RV belonging to restricted gene lists known to be related with BC from clinics and GWAS:

- **ClinGen**: this list contains 20 genes and was extracted from the file downloaded from [26]. This file provides a summary of the Gene-Disease Validity curation completed by ClinGen's GCEPs. The genes are classified on the base of supportive evidences in 6 classes of Clinical Validity for the disease. We considered all the genes related to breast cancer except for those classified as "Refuted" (Refuted: evidence refuting the initial reported evidence for the role of the gene in the specified disease has been reported and significantly outweighs any evidence supporting the role);

- **Genturis**: This list of 229 genes was curated by ERN GENTURIS - European Reference Network (ERN) for all patients with one of the rare genetic tumour risk syndromes (genturis) [45];

- **Harmonizome 3.0**: includes 595 genes associated with breast cancer in

GWAS and other genetic association datasets from the GWASdb SNP-Disease Associations database[7];

- **Open Targets**: 427 genes downloaded from the breast neoplasm association page of the Open Target Platform [89] [88] and filtered for association score ≥ 0.5.

. These selected variants were then categorized into distinct 'MASKs' for subsequent analysis. The Table 3.2 and Figure 3.7 show the distribution of variants within each group across the investigated gene lists. Additionally, Table 3.3 provides a detailed breakdown of the variant distribution across each MASK, with respect to the different gene lists utilized in this study. Interesting, genes belonging to the ClinGen and Genturis sets, are those less characterized for the presence of missense RVs in Group 2.

	LoF	**Missense_Group1**	**Missense_Group2**
Harmonizome	4939 (42.5%)	4633 (39.9%)	2040 (17.6%)
ClinGen	597 (31.9%)	1258 (67.3%)	15 (0.802%)
Genturis	3057 (25.4%)	7964 (66.3%)	999 (8.31%)
Open Targets	3601 (24.3%)	9265 (62.6%)	1926 (13.0%)

Table 3.2: RVs in each group by genes lists

Table 3.3: Number of variants in each MASK for each gene list

MASK	Description	Gene List	Variants
M1	LoF RVs	ClinGen	597
		Genturis	3057
		Harmonizome	4939
		Open Targets	3601
		Clinical Exome	82738
M2	Missense RVs, Group1	ClinGen	1258
		Genturis	7964
		Harmonizome	4633
		Open Targets	9265
		Clinical Exome	63100
M3	Missense RVs, Group2	ClinGen	15
		Genturis	999
		Harmonizome	6979
		Open Targets	1926
		Clinical Exome	30365
M4	Missense RVs, Group1 and 2	ClinGen	1273
		Genturis	8963
		Harmonizome	6673
		Open Targets	11191
		Clinical Exome	93465
M5	LoF and Missense RVs, Group1	ClinGen	1855
		Genturis	11021
		Harmonizome	9572
		Open Targets	12866
		Clinical Exome	145838
M6	LoF and Missense RVs, Group2	ClinGen	612
		Genturis	4056
		Harmonizome	14018
		Open Targets	5527
		Clinical Exome	113103
M7	LoF and Missense RVs (Groups 1 and 2)	ClinGen	1870
		Genturis	12020
		Harmonizome	11612
		Open Targets	14792
		Clinical Exome	176203

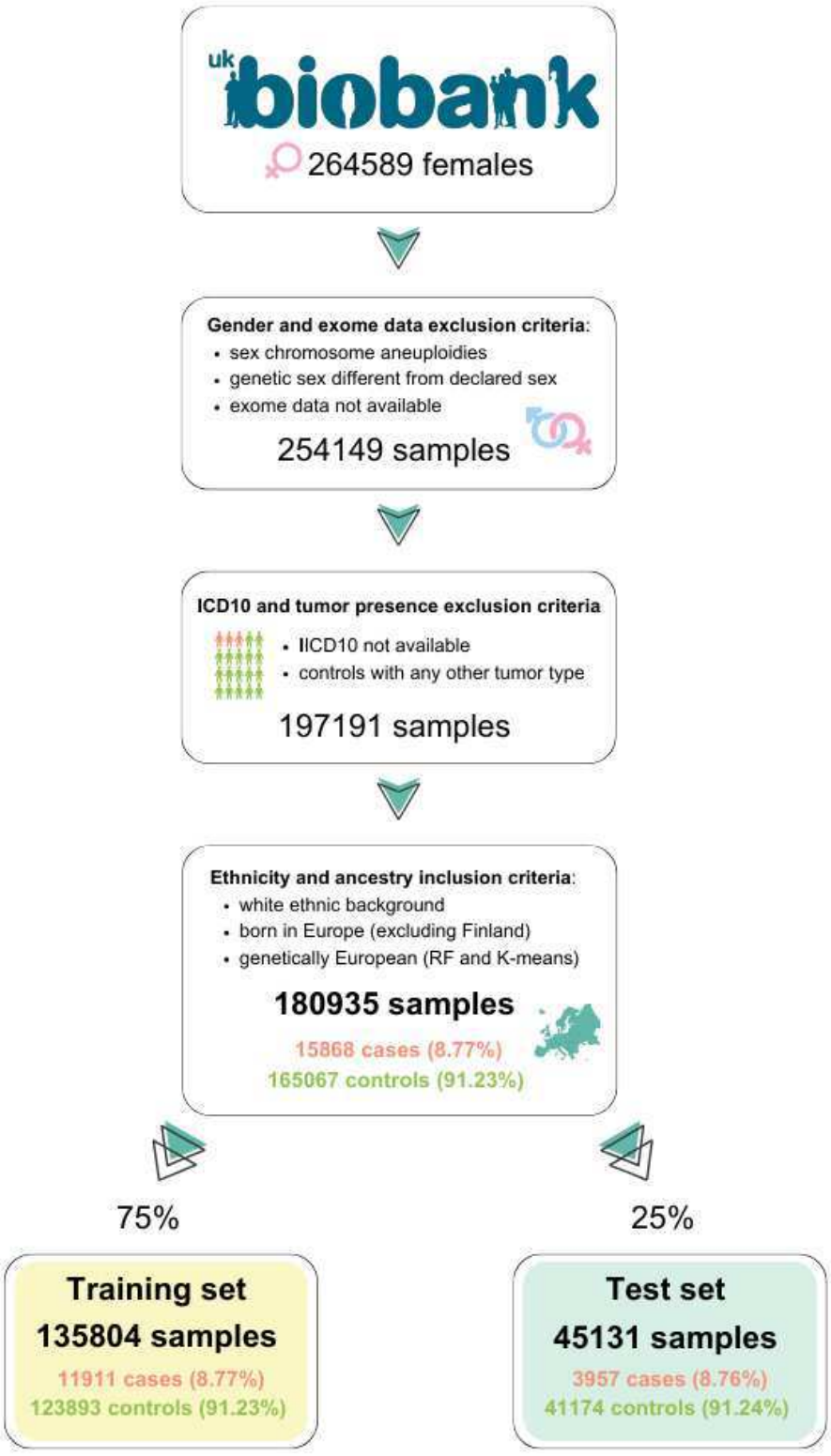


Figure 3.5: Sample data processing steps

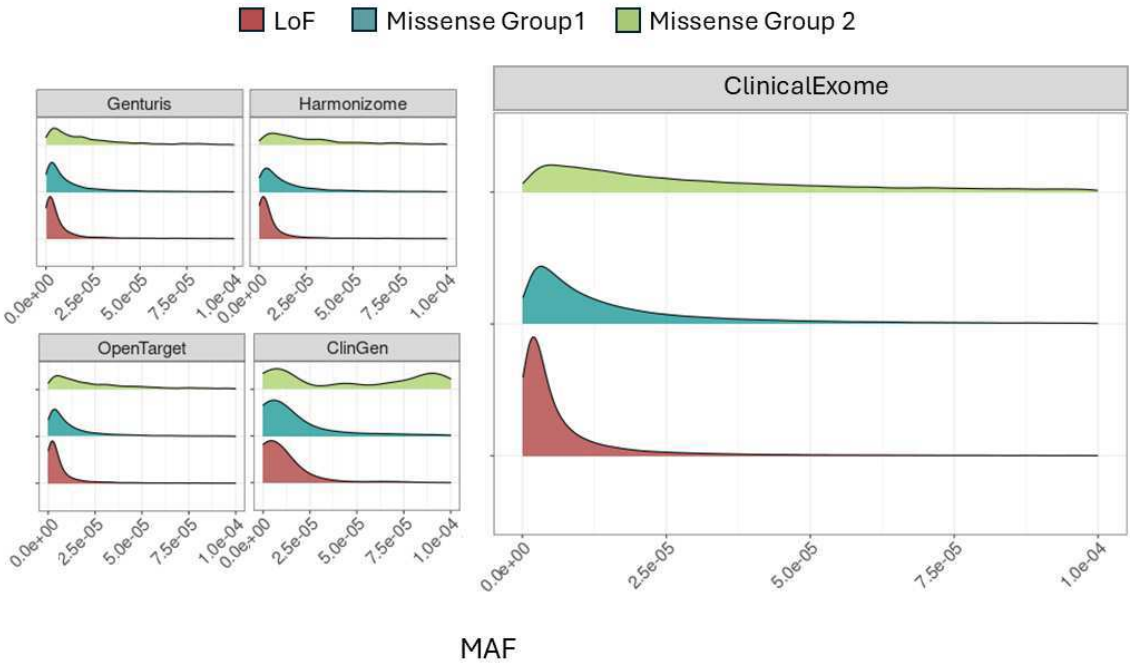


Figure 3.6: MAF distribution by RVs group by gene lists

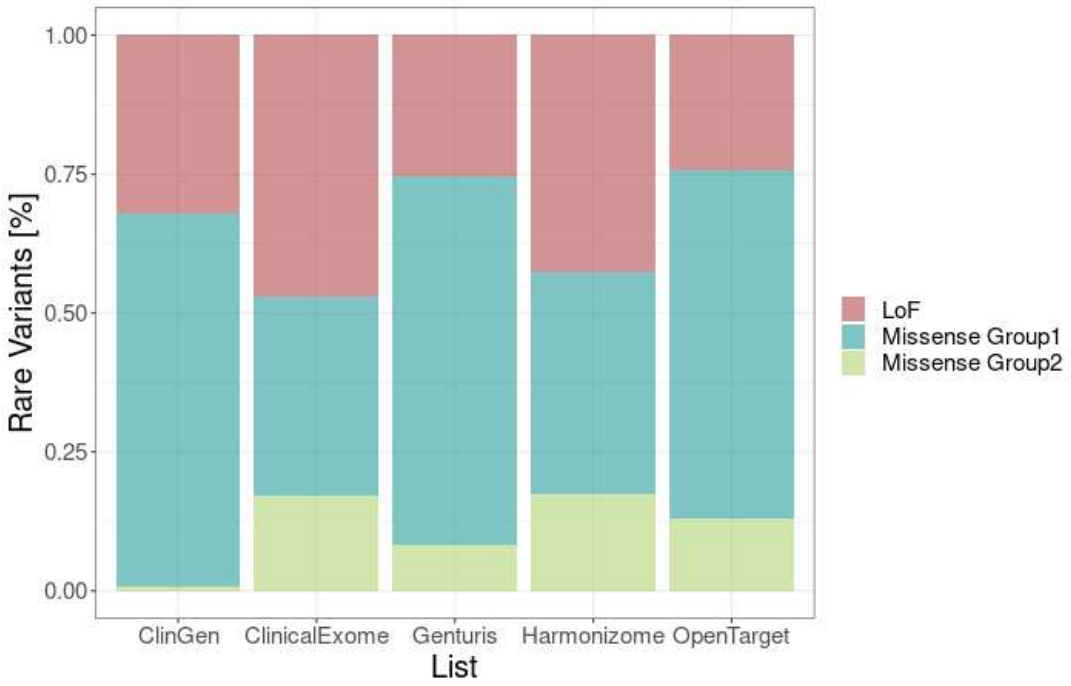


Figure 3.7: Percentage of RVs in each group by genes lists

Chapter 4

Rare variants-breast cancer association analysis using classical approaches

This chapter reports the methods applied for the BC-RVs association analysis using classical approaches (section 4.1) and the results obtained from the application of these technique (section 4.2).

4.1 Methods for rare variants-breast cancer association analysis using classical approaches

In Chapter 3.2.2, we explored a variety of methodologies for analyzing the associations between rare variants and various traits, with a particular focus on models designed to evaluate the influence of RVs on breast cancer risk. Recognizing that each approach offers distinct advantages and limitations, we adopted a comprehensive strategy, examining a number of different methods to provide a systematic frame-

work for understanding the genetic architecture of RVs in breast cancer. Specifically, we applied both Burden and Variance component tests (SKAT, Sequence Kernel Association Test, and SKAT-O, Sequence Kernel Association Test - Optimal) to evaluate the direction of the RVs effect on BC in UKBB cohort. We then aggregated the results from the Burden test into a combined score, which was further integrated with Polygenic Risk Scores to assess the additional contribution of RVs to BC risk (section 4.1.3). To reduce the multiple testing burden and enhance the power for detecting gene associations, we chose to conduct the analysis considering rare variants belonging to four different carefully curated lists of genes known to be related with BC from clinics and GWAS and that have been described in section 5.3.4: **ClinGen**, **Genturis**, **Harmonizome 3.0**, and **Open Targets**.

As shown in Figure 4.1, there is limited overlap among the gene lists, with the exception of the ClinGen panel, which shares 60% of its genes with Genturis and Open Targets lists. Considering this, with the aim to broaden the investigation of the involvement of rare variants in a enlarged list of gens to achieve a more comprehensive overview of the genetic of breast cancer, we extended our analysis to the entire set of 6,700 genes from the **Clinical Exome** (TrueSight One Sequencing Panel v2). Although this choice may reduce statistical power and increase the risk of undetected associations due to a larger multiple testing burden.

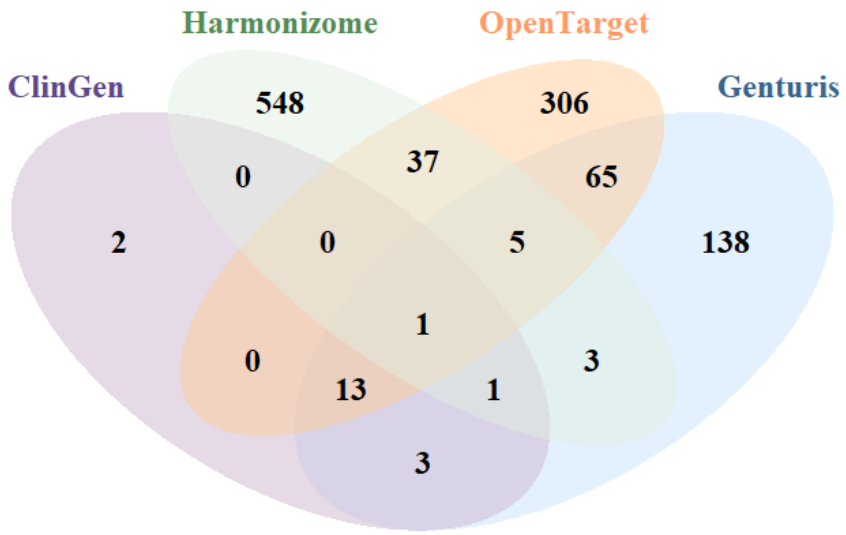


Figure 4.1: **Gene Overlap Between Lists**

For each list of genes, included the CLinical Exome, we defined seven distinct masks aggregating RVs according to VEP annotation as explained in section 3.1.2. The details of each mask are outlined below:

- **Mask 1 (M1)**: includes only loss-of-function (LoF) RVs.
- **Mask 2 (M2)**: consists of missense RVs classified as clinically relevant (Group 1).
- **Mask 3 (M3)**: includes missense RVs with a damaging functional impact (Group 2).
- **Mask 4 (M4)**: combines M2 and M3.
- **Mask 5 (M5)**: combines M1 and M2.
- **Mask 6 (M6)**: combines M1 and M3.
- **Mask 7 (M7)**: aggregates all LoF RVs with all missense RVs i.e. M1, M2 and M3.

In the following subsections, we will provide a comprehensive overview of the implementation of the rare variants-breast cancer association test using the Regenie software [58] (section 4.1.1 and 4.1.1). In section 4.1.2, we will described the definition of the score based on the combined effects of RVs across different genes. This RVs-based score is designed to quantify the cumulative influence of rare variants, allowing us to assess their collective impact on breast cancer risk. Finally, we will discuss how this score is integrated with PRS to provide a deeper understanding of RVs contributions to BC susceptibility (section 4.1.3).

4.1.1 Burden and variance component tests implementation

We utilized Regenie, a whole genome regression modelling of large genome-wide association studies, through the Swiss Army Knife application on the UKBB Research Analysis Platform to perform gene-level association analyses [58]. Regenie evaluates the cumulative effect of multiple genetic variants within predefined genomic regions, typically genes, on a target trait—breast cancer, in this instance. This gene-level approach is particularly powerful for capturing the collective impact of rare variants within a gene, an effect that may be undetectable using single-variant analyses. Regenie's analytical workflow is organized into two main steps. In Step 1, a null model is built to adjust for both covariates and individual genetic backgrounds. This initial phase generates predictions that account for population structure and systematic effects related to covariates, providing a set of "baseline" values. These values are later used as a reference for identifying true associations between genetic variants and the trait. By constructing this null model, Regenie helps control for population stratification and other potential confounders, thereby enhancing the reliability of downstream association testing. In Step 2, Regenie applies the predictions from the null model to conduct the primary association tests, assessing the relationship between variant sets and breast cancer. This step leverages the foundation established in Step 1, testing each gene-level variant set while maintaining control over confounding effects related to population structure and covariates.

Regenie: Step 1

Given that our phenotype of interest is binary (e.g., presence or absence of breast cancer), we specified a binary trait model in Step 1 of the Regenie analysis. In this initial phase, we trained the null model on high-quality genotype data, filtered according to the criteria outlined in Section 3.1.1. To account for population struc-

ture and mitigate potential confounding factors, we included age and the first ten principal components (PCs) of genetic ancestry as covariates. Regenie employs a ridge regression approach, which partitions the data into blocks for efficient processing. In our setup, we configured these blocks to contain 1,000 single nucleotide polymorphisms (SNPs) each, enabling the model to handle data in more manageable segments. To enhance the model's generalizability, training was performed using leave-one-out cross-validation, allowing for more robust performance across different subsets of the data. Additionally, we allocated 16 CPU threads to parallelize computation, significantly accelerating the analysis. Step 1 was performed a single time across the entire dataset, independently of the specific gene lists under consideration, and utilizing all high-quality genotype calls.

Regenie: Step 2

For each of the previously described gene lists, as well as for the complete set of genes in the Clinical Exome, we executed Regenie Step 2 three separate times: once for the Burden test, once for the SKAT test, and once for the SKAT-O test. For each run, we specified the relevant test by adjusting the corresponding parameter within the Regenie configuration. The association testing was conducted on a chromosome-by-chromosome basis, iterating over all chromosomes within a looped workflow to streamline the analysis. This procedure was performed for each of the variant masks defined in the previous section to evaluate gene-level associations between RVs within each mask and the risk of developing breast cancer. By running the analysis separately for each variants mask, we aimed to capture the distinct contributions of different types of RVs to breast cancer susceptibility, allowing for a more nuanced understanding of variant effects at the gene level. Again, to adjust for population structure and reduce potential confounding variables, we incorporated age and the first ten PCs of genetic ancestry as covariates in our analysis. These

covariates were included alongside the null model's predictions generated in Step 1, allowing for a comprehensive adjustment that enhances the accuracy of our association tests by controlling for both demographic and genetic components. Given that the control group in the UK Biobank dataset is approximately ten times larger than the group of breast cancer cases, we opted to use Firth regression with a saddlepoint approximation [32]. This method improves the precision of p-values, particularly for rare variants, where conventional tests may lack accuracy due to the imbalance in case-control ratios. In this step as well, we enabled multi-threaded execution to boost computational efficiency and expedite the workflow.

4.1.2 RVs-based score definition using Burden Test results

To quantitatively evaluate the cumulative influence of rare variants and to assess their collective impact on the risk of developing breast cancer, we constructed a rare variants-based score for each of the specified gene lists and the associated masks. This score was build by leveraging the effect sizes obtained from the burden test, which allows us to capture and integrate the contributions of multiple rare variants across the relevant genomic regions. Specifically, for each list of genes and each mask, the score based on rare variants was computed using the following formula:

$$Score_{RV_i} = \sum_g \beta_g \sum_{v \in g} \frac{d_{iv}}{\mathbb{E}(d_v)}$$

where:

- $Score_{RV_i}$: represents the rare variant-based score for individual i.
- β_g: is the estimated effect size of gene g (taking into account the RVs belonging to the considered mask).
- d_{iv}: represents the allelic dosage for variant v in subject i.

- $\mathbb{E}(d_v)$: is the expected allele frequency of variant v belonging to gene g.

The inner summation $\sum_{v \in g}$ represents the aggregate allelic dosage d_{iv} of all variants v within the gene g, weighted for their expected frequency $\mathbb{E}(d_v)$. The expected frequency represents the MAF across all populations retrieved using gnomAD (v4.1.0). The outer summation $\sum_g$ aggregates these scores across all genes, weighted by the effect size β_g for each gene, accounting for eventually multiple gene:variant effects.

We computed the RVScore, using the genes selected by Burden Test for BC-association on the Clinical Exome and within specific gene masks ($M1$, and $M7$), after correcting $p - values$ by the Benjamini-Hochberg procedure. The Clinical Exome and these two masks were chosen to provide a comprehensive assessment of the impact of different RVs on breast cancer susceptibility.

4.1.3 Combination of RVs-based score and PRS

To evaluate the cumulative impact of RVs across various variant masks in conjunction with a polygenic risk score, our initial step involved selecting the BC-PRS most suitable for our dataset. We decided to compare the UK-Biobank Standard PRS for breast cancer (the polygenic risk scores described above and proven to be more powerful than almost all previously released PRSs in individuals with European and non-European ancestries [35]) with the genome-wide score derived using PRS-CS [103] proven to outperform the 313 SNP [16] score and the LDpred score [82] by Mars et al.(2020) [81]. To evaluate which PRS had the strongest association with breast cancer, each PRS was individually standardized to have a mean of zero and a standard deviation of one, enabling the calculation of odds ratios per standard deviation. Logistic regression models, adjusted for age, were employed to compute the OR for each PRS. The PRS with the highest OR, indicating the most robust correlation with BC, was selected for subsequent analysis.

We investigated the association between the RVScore, computed across the entire Clinical Exome and within specific gene masks ($M1$, and $M7$; denoted as $RVScore_M$, where $M \in \{M1, M7\}$, with breast cancer risk in conjunction with polygenic risk scores (PRS). The Clinical Exome and these two masks were chosen to provide a comprehensive assessment of the impact of different rare variants on breast cancer susceptibility. To achieve this, individuals were stratified into three PRS categories based on tertile distributions: *High* PRS values corresponded to the upper tertile ($> 70\%$), *Medium* PRS values spanned the inter-tertile range ($30\% - 70\%$), and *Low* PRS values were below the 30th percentile. Similarly, the RVScore was categorized into three groups: *Null RVScore* for individuals with an RVScore of zero, *Low RVScore* for values below the median computed excluding the *Null RVScore* group, and *High RVScore* for values above the median. For each selected mask, we also assessed the RVScore's behavior in comparison to the presence of RVs within high-risk (BRCA1/2) and moderate-risk (ATM, CHEK2, PALB2) genes as was done by [62]. The logistic regression model were constructed to estimate odds ratios (ORs), using the *Intermediate* PRS group without any RVs and/or a *Null RVScore* as the reference category. This model was fitted on the training set and on the test set (this latter subset was not used in the initial RV-breast cancer association analysis). By comparing the significance and the OR of $RVScore_M$ when included alongside PRS, we aimed to isolate and evaluate the cumulative impact of rare variants belonging to different variant masks on breast cancer susceptibility.

Up to this point, we have outlined a methodological framework for systematically developing a score that incorporates rare variants and combines it with polygenic risk scores to predict breast cancer occurrence, utilizing classical gene-based association analysis tests. In the following sections, we will introduce a more innovative approach for consistently analyzing the impact of rare variants on breast cancer risk, enabling the interpretation of associations at the single-variant level.

4.2 Results of rare variants-breast cancer association analysis using classical approaches

In this section, we report the results of our gene-level association analysis aimed at uncovering robust relationships between rare variants and breast cancer. To ensure the reliability of these findings, we first determined the most appropriate statistical approach for our dataset by employing two widely recognized statistical frameworks: the Burden Test and Variance Component Tests. Each method presents distinct advantages and limitations based on the underlying characteristics of the RV-BC relationships within each gene, thereby offering insights into the direction of the RVs association effect. At the same time, by comparing different lists of genes, we aimed to asses the impact of the multiple testing burden on the detection of RVs truly related to breast cancer. The size of the evaluated gene lists plays a significant role in shaping the outcomes of the genetic association studies. In general, smaller gene lists tend to be less affected by the impact of multiple testing corrections, allowing for the identification of genes with weaker, but still potentially meaningful, associations. In contrast, larger gene lists, though more susceptible to the effects of multiple testing, are expected to identify a few genes, but with stronger associations, potentially leading to more robust findings. Finally, we aimed to evaluate the contribution of LoF and missense RVs to BC risk by performing our analysis combining the variants in different MASKs. As we saw in chapter 2, the current literature on BC genetics highlights the significant role of LoF variants in BC susceptibility, particularly in genes such as BRCA1, BRCA2, ATM, CHEK2, and PALB2. Large cohort studies [4] [84], have identified additional genes such as TP53, BARD1, RAD51C, RAD51D and MAPK1 playing a role in BC predisposition. In addition to LoF variants, missense mutations, especially in genes like CHEK2, are also frequently associated with BC risk [4] [84]. Finally, to combine and evaluate

the cumulative influence of RVs on BC we built a RVScore.

4.2.1 Burden, SKAT and SKATO test results and multiple testing burning impact

Regarding the test outcomes, the SKAT and SKAT-O tests detected CHEK2 as a significant gene associated with BC, with $p_{adj} \leq 0.05$, particularly when using genes from the lists Open Targets, ClinGen, and Genturis in MASKS M5 and M7. The burden test, by comparison, identified a larger set of genes associated with BC across different MASKs and gene panels. This initial finding suggests that, within each gene, the majority of the evaluated rare variants exert a consistent directional effect on breast cancer susceptibility. In the light of that we decided to proceed in our study building on Burden Test results.

The analysis of the genetic lists revealed that, despite their initial size, only a subset of genes presents at least a RV in the UKBB cohort and, thus, was ultimately tested for association with breast cancer . Specifically, 48.6% of the genes in Harmonizome3 carried rare RVs that were included in the association analysis, the 73% in Open Targets, in Genturis and in the Clinical Exome were tested the 87.3% and the 88.3% of the genes respectively, while the whole batch of ClinGen genes was evaluated (Table 4.1). Across all gene lists, regardless of the applied MASK, approximately 10% (645 genes) of the genes in the Clinical Exome were found to be potentially associated with BC, exhibiting a $p - value \leq 0.05$. This proportion was similar to that observed for Open Targets (11.8%, 37 genes), but lower than the fractions of genes showing weak associations in Harmonizome3 (13.1%, 38 genes), Genturis (15.5%, 31 genes), and ClinGen (30%, 6 genes). Upon applying the Benjamini-Hochberg procedure to correct for multiple testing, however, the number of genes identified as associated with BC significantly diminished, with only a small

subset of genes maintaining significant associations. Notably, the Top 5 gene were consistently retrieved across all the gene lists, underscoring their robust association with BC.

List	**Genes with** $p-value \leq 0.05$ **(%)****	**Genes with** $p_{adj} \leq 0.05$ **(%)****	**Total Number Of Genes Tested (%)***	**Total Number Of Considered Genes**
ClinGen	6 (30.0%)	6 (30.0%)	20 (100%)	20
Genturis	31 (15.5%)	9 (4.5%)	200 (87.3%)	229
Open Targets	37 (11.8%)	9 (2.88%)	313 (73.3%)	427
Harmonizome3	38 (13.1%)	1 (0.346%)	289 (48.6%)	595
ClinicalExome	645 (10.9%)	7 (0.118%)	5914 (88.3%)	6700

** computed on the total number of genes,*
*** computed on the number of tested genes.*
The p_{adj} have been computed applying the Benjamini-Hochberg method

Table 4.1: Summary of Burden Test analysis results.

These results highlighted the critical role of gene list size in shaping the outcomes of genetic association studies. The smaller curated gene lists are less susceptible to the burden of multiple testing, facilitating the identification of weaker genetic associations. However, larger gene lists, like the ClinicalExome, while offering a more comprehensive analysis, tend to reduce the power to detect these weaker associations due to the increased number of comparisons. This is a trade-off inherent in the use of larger gene sets, which can uncover stronger associations but at the cost of a greater risk for false positives. These results became even more clear when looking at the details of the MASKs analysis results.

Figure 4.2 shows the genes with association $p_{adj} \leq 0.05$ for the presence of RVs in each MASK for all the list. In particular it can be noticed that:

- For LoF variants (MASK M1), the key genes ATM, BRCA1, BRCA2, CHEK2, and PALB2, were consistently associated with BC across all gene lists were present. BARD1 (evaluated in Clingen, Genturis, Open Targets and ClinicalExome), while showing significant association in the analysis performed

on ClinGen and Genturis lists, lost significance when considering the larger gene lists like Open Targets and ClinicalExome. Other genes, such as DDX11 (observed in Genturis, and in ClinicalExome), MAP3K1 and PLCG1 (both present in Open Targets, and in ClinicalExome), showed a similar variability in their levels of association, losing significance when considered in the larger lists due to the increased multiple testing burden. Finally ASPRV1 e ADGRA3 belonging just to the clinical exome where selected as associated to BC risk for the presence of LoF RVs.

- In the analysis of missense variants (MASKS M2, M3, M4), CHEK2 emerged as a strong candidate, with significant associations observed across multiple groups and in all the list when present. Specifically, it was identified as associated to BC for carrying missense RVs in Groups 1 and 2 (MASK M4), and in Group 1 alone (MASK M2), suggesting a potentially greater impact of missense RVs in Group 1. NDUFS4, which was present in Open Targets and ClinicalExome, showed a significant association with BC for Group 2 missense RVs (MASK M3), but lost significance when considering the full ClinicalExome.

- When both LoF and missense variants were considered together (M5, M6, M7), significant shifts in the strength of associations were observed. Genes like PLCG1 and ADGRA3, which were significant for LoF variants alone, lost significance when both LoF and missense variants were considered together across all MASKs. Genes such as BRCA2 and BARD1 lost significant associations when considering LoF in combination with missense RVs in Group 1 (M5, M7). MAP3K1 remained significant in Open Targets when evaluated for the combination of LoF RVs with missense in Group 1 (M5) but lost significance when LoF and missense RVs in Group 2 were combined (M6, M7).

NDUFS4, initially significant in Open Targets for missense variants (M4), lost significance when LoF variants were introduced in all the MASKs M5, M6, and M7. Finally, both LZTR1 (present in Genturis and ClinicalExome) and POLD2 (belonging to Open Targets and ClinicalExome) have been selected as associated to BC when analyzed in shortest lists for the presence of both LoF e missense RVs in Group 1 (M6), but lost significance when considering the full ClinicalExome.

In conclusion, we confirmed several well-established associations, such as LoF variants in BRCA1, BRCA2, ATM, CHEK2, and PALB2, as well as missense mutations in CHEK2. Furthermore, we revealed new potential strong candidates like ASPRV1 - a gene coding for a Protease responsible for filaggrin processing, essential for the maintenance of a proper epidermis organization- and ADGRA3 - gene encoding a member of the G protein-coupled receptor superfamily, a membrane protein which may play a role in tumor angiogenesis through its interaction with the human homolog of the Drosophila disc large tumor suppressor gene (hDlg). At the same time we identified more weakly associated genes such as BARD1, MAP3K1, LZTR1, NDUFS4, POLD2, PLCG1, and DDX1. To ensure the robustness of our results, we decided to construct the RVScore using the genes selected from the ClinicalExome for $p_{adj} \leq 0.05$. As the most extensive gene list evaluated, the Clinical Exome is more affected by the multiple testing burden, making it a stringent filter for identifying associations. Consequently, genes found to be associated with BC using this list are those with the strongest and most robust associations. Moreover, we decided to pursue in the RVScore definition focusing on MASKs M1 (just LoF RVs) and M7 (LoF RVs combined with all the missense in Group 1 and 2). The use of MASK M1 ensures a focus on LoF variants, which are widely recognized for their strong functional impact and relevance to BC. This MASK consistently identified well-

established BC risk genes, such as ATM, BRCA1, BRCA2, CHEK2, and PALB2, across all gene lists. These results highlighted the validity of prioritizing LoF variants as key contributors to BC susceptibility. Additionally, novel candidates such as ASPRV1 and ADGRA3 were identified in the Clinical Exome analysis, further emphasizing the utility of this MASK in uncovering previously unrecognized associations. In contrast, MASK M7 combines LoF and missense RVs from both Group 1 and Group 2, allowing for a more comprehensive analysis of variant types. This MASK captures genes with potentially synergistic effects between LoF and missense variants, such as CHEK2, which demonstrated consistent associations across multiple groups and gene lists. By focusing on the Clinical Exome and these two masks, we leveraged the rigorous filtering effect of the extensive gene list to ensure that only the strongest associations are considered. Building the RVScore exploiting the results obtained by considering the RVs in masks M1 and M7 on the genes from the Clinical Exome, aligns with our objective of identifying robust genetic associations with BC while maintaining the capacity to explore diverse variant types.

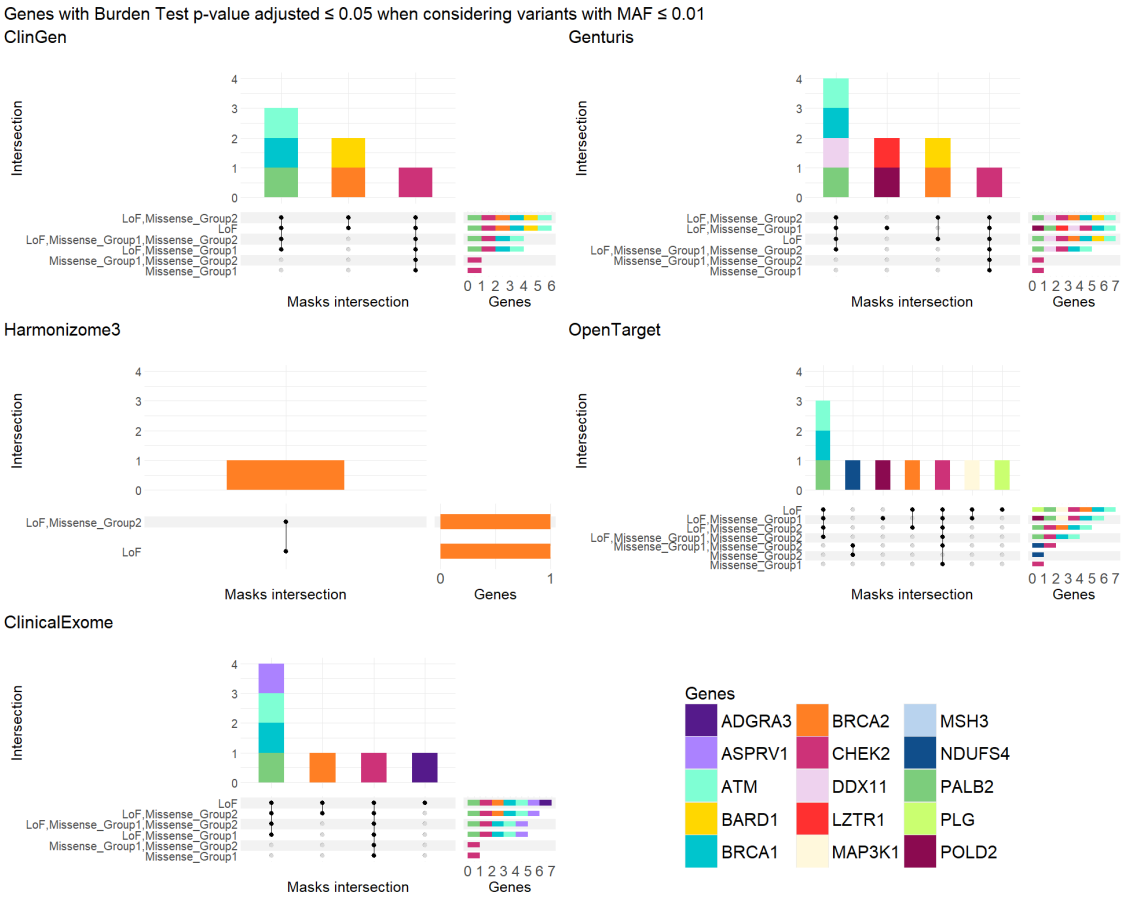


Figure 4.2: Burden association test results for RVs (MAF ≤ 0.01) across different gene lists.

This figure consists of five individual upset plots visualizing for each gene list the intersections of Burden Test significant genes across different RVs masks. Each upset plot consists of a top part bar chart indicating the significant genes belonging to the intersection, and a right side bar chart indicating the genes with $p_{adj} \leq 0.05$ for each mask. Specifically, in **ClinGen** BRCA1, ATM and PALB2 are significant when considering the combination of LoF and missense, and LoF alone; BARD1 and BRCA2 are significant in the LoF and missense Group 2 and LoF alone masks; CHEK2 is found in all masks. In **Genturis** list ATM, BRCA1, DDX11, and PALB2 are shared by all the masks considering the combination of LoF and missense, and LoF alone; POLD2 and LZTR1 are significant for the LoF and missense Group 1 mask; BARD1 and BRCA2 are significant in the LoF and missense Group 2 and LoF alone masks; CHEK2 is significant in all masks. **Harmonizome3** displays just BRCA2 found in LoF + missense Group2 and LoF alone mask. In **Open Targets**, ATM, BRCA1, and PALB2 are shared by all the masks considering the combination of LoF and missense, and LoF alone; NDUFS4 is significant when considering missense in Group 2 and missense in Group 1 and 2 together; POLD2 when considerinf LoF and missense Group 1, BRCA2 is significant in the LoF and missense Group 2 and LoF alone masks, CHEK2 is significant in all the masks with the exception of the missense Group 2 one; MAP3K1 is significant in the LoF and missense Group 1 and LoF alone masks; PLG is significant when considering just LoF. Finally, in the **Clinical Exome**, ATM, BRCA1, ASPRV1, and PALB2 are shared by all the masks considering the combination of LoF and missense, and LoF alone; BRCA2 is significant in the LoF and missense Group 2 and LoF alone masks; CHEK2 is significant in all the masks, and ADGRA3 is significant when considering just LoF.

4.2.2 RVs-based score and PRS

To investigate the relationship between the RVScore and breast cancer risk in conjunction with the Polygenic Risk Score (PRS), we calculated the RVScore by combining significantly associated genes obtained using the Clinical Exome list with selected gene masks (M1 and M7). Both PRS and RVScore were then categorized into three groups, as detailed in Section 4.1.3. We then analyzed the performance of the RVScore by comparing it with the condition of the solely presence of RVs in high-risk genes (BRCA1/2) and moderate-risk genes (ATM, CHEK2, PALB2) across PRS categories, using odds ratios as the metric. We conducted odds ratio analyses on both the Training and Test datasets to evaluate the scalability of the proposed score. This section provides a comprehensive description of the computed RVScore and its validation when integrated with PRS.

The RVScore was constructed using 336 RVs for M7 and 310 for M1 (Table 4.2). The median MAF of RVs included in the computation of the $RVScore_{M7}$ was statistically significantly higher, with an increase of 1.1×10^{-6}, compared to the mediam MAF of the RVs used for the $RVScore_{M1}$ (Wilcoxon test $p-value < 0.001$).

MASK	N. RVs used in the score	Mean MAF [95%CI]	Median MAF (IQR)	Wilcoxon test p-value (H1:M7>M1)
M7	336	3e-05 [1e-05, 4e-05]	4.3e-06 (1.06e-05)	< 0.001
M1	310	2e-05 [0, 3e-05]	3.2e-06 (7.2e-06)	

Table 4.2: RVs included in the RVScore by MASK

Tables from 4.3 to 4.6 report the distribution measures of the RVScore defined in each mask in Training and Test set across the whole population and in cases and controls. As expected, the proportion of individuals with an RVScore $\neq 0$ was considerably lower in the Test set compared to the Training set (1.09% vs. 12.43% for M1 and 1.92% vs. 20.06% for M7). Consequently, the average RVScore was higher in the Training set than in the Test set. Furthermore, the score derived from RVs in M1 was higher than that based on RVs in M7 (Table 4.3), likely due to the inherent definition of the score and the higher frequency of RVs included in M7 compared to the LoF RVs in M1. Overall, the RVScore was zero for more than 50% of the population, regardless of the mask or dataset analyzed (median RVScore = 0, IQR: 0).

An examination of the RVScore distribution in cases and controls, reveals that the percentage of cases with $RVScore \neq 0$ is generally nearly double that observed in controls (Table 4.5). This results in a higher average RVScore in cases compared to controls, regardless of the mask or dataset explored (Table 4.5). However, when looking at the subpopulation of subjects with $RVScore \neq 0$, the significance of the differences is lost on the Test set (Table 4.6), may be influenced by the high prevalence of subjects with $RVScore = 0$. Indeed, given the smaller size of the Test set, we acknowledged the likelihood that certain RVs might not be represented among its individuals. It can be also noticed that on the Test set the median of the $RVScore_{M1}$ of the Cases is the same of the overall population, but it shows an higher IQR, suggesting greater variability in $RVScore_{M1}$ among cases (Table 4.4 and Table 4.6). For $RVScore_{M7}$ both Cases and Controls report the same median value of the overall population with a similar IQR, indicating uniformity in score distribution across groups for this mask in the Test set (Table 4.4 and Table 4.6). In the light of that, in order to evaluate the ability of the defined score in combination with PRS, for each MASK, we separated our score in 3 classes: "Null RVScore"

for $RVScore = 0$, and "Low" and "High" RVScore for values respectively below and above the median computed on the subpopulation of subjects not in the first class. Using the median of the RVScore distribution within the subpopulation of $RVScore \neq 0$ ensures that the classification is not overly influenced by extreme values allowing to capture the nuances of score variation among those who exhibit variability and avoiding diluting the analysis with the high proportion of null scores.

Set	MASK	Mean [95%CI]	N. subjects	N. subjects with RVScore $\neq 0$ (%)
Test Set	M1	522.78 [410.25, 635.31]	45022	493 (1.09%)
Test Set	M7	429.8 [345.51, 514.09]	45022	865 (1.92%)
Training Set	M1	1607.11 [1408.53, 1805.7]	135555	1685 (12.43%)
Training Set	M7	1107.81 [965.1, 1250.52]	135555	2720 (20.06%)

Table 4.3: RVScore distribution including patients with null RVScore

Set	MASK	Mean [95%CI]	Median (IQR)
Test Set	M1	47741.8 [38349.6, 57133.99]	3657.43 (37586.99)
Test Set	M7	22370.4 [18236.85, 26503.96]	2696.56 (12603.42)
Training Set	M1	129289.14 [114533.9, 144044.38]	6371.21 (83222.93)
Training Set	M7	55209.35 [48399.01, 62019.69]	3303.07 (21845.64)

Table 4.4: RVScore distribution excluding patients with null RVScore

Set	MASK	BC Satus	Mean [95%CI]	N. subjects	N. subjects with RVScore $\neq 0$ (%)	T-test p-value (H1:>)
Test Set	M1	Cases	1583.38 [825.26, 2341.51]	3948	110 (2.78%)	≤ 0.001
Test Set	M1	Controls	420.84 [321.37, 520.31]	41074	383 (0.93%)	
Test Set	M7	Cases	1041.75 [514.02, 1569.48]	3948	154 (3.90%)	0.01
Test Set	M7	Controls	370.98 [293.78, 448.18]	41074	711 (1.73%)	
Training Set	M1	Cases	4965.46 [3764.79, 6166.12]	11896	381 (3.20%)	≤ 0.001
Training Set	M1	Controls	1284.04 [1099.62, 1468.46]	123659	1304 (1.05%)	
Training Set	M7	Cases	3169.29 [2260.09, 4078.5]	11896	459 (3.85%)	≤ 0.001
Training Set	M7	Controls	909.5 [779.85, 1039.15]	123659	2261 (1.82%)	

Table 4.5: RVScore distribution in cases and controls, including patients with null RVScore

Set	MASK	BC Satus	Mean [95%CI]	Median (IQR)	Wilcoxon test p-value (H1:>)	T-test p-value (H1:>)
Test Set	M1	Cases	56829.07 [31603.65, 82054.49]	3657.43 (41555.77)	0.34	0.20
Test Set	M1	Controls	45131.87 [35447.28, 54816.47]	2307.64 (35995.58)		
Test Set	M7	Cases	26706.63 [13784.69, 39628.56]	2696.56 (12454.34)	0.61	0.22
Test Set	M7	Controls	21431.19 [17250.83, 25611.55]	2696.56 (12455.58)		
Training Set	M1	Cases	155036.92 [120776.95, 189296.9]	16203.57 (127016.88)	< 0.001	0.03
Training Set	M1	Controls	121766.21 [105553.91, 137978.5]	3657.43 (64382.69)		
Training Set	M7	Cases	82139.26 [59733.36, 104545.17]	3657.43 (32425.13)	0.04	0.00
Training Set	M7	Controls	49742.38 [42947.24, 56537.52]	2696.56 (18333.39)		

Table 4.6: RVScore distribution in cases and controls, excluding patients with null RVScore

As a preliminary step, before the RVScore validation phase, we assessed which PRS was most suitable for our dataset. Specifically, we compared the PRS-CS [103] with the UKBB Standard PRS for BC (named BC-PRS hereafter)[107]. Both PRS were individually standardized, and their respective odds ratios per standard deviation were calculated using logistic regression models adjusted for age. In the cohort under investigation in this work, the OR of the BC-PRS was 1.85 (95% CI: 1.83–1.87), compared to the OR of 1.73 (95% CI: 1.71–1.75) calculated using PRS-CS 4.7. Based on these results, the BC-PRS was chosen in downstream analysis.

	OR (95% CI)*
PRS-CS	1.73 (1.71–1.75)
Standard PRS BC (UKBB)	1.85 (1.83–1.87)

** adjusted with age, OR p-value ≤ 0.001. OR computed on the overall population (Training and Test Sets)*

Table 4.7: OR of BC-PRS (Standard UK Biobank PRS for BC) and of PRS-CS

We, therefore divided BC-PRS in three classes, to evaluate the performance of the RVScore across PRS categories by comparing its association with the simply presence of RVs in high and moderate-risk gene. In Figure 4.3 and in the Tables from 4.8 to 4.11 are reported the results of this analysis. The RVScore effectively stratified genetic risk across PRS classes, as evidenced by progressively increasing ORs from Low to High PRS categories when compared with the Null RVScore class.

The consistency of these trends across masks and datasets underscores the RVScore's capacity to refine risk stratification individuals with differing genetic burdens. Furthermore, in the Training set, individuals in the RVScore High subpopulation exhibited higher ORs within each PRS class compared to those in the RVScore Low group. These differences are still visible in the Test set, particularly in M7 in the High PRS class, but less pronounced. When compared with the presence of RVs on moderate risk genes, it can be noticed that, on both Training an Test set, the High RVScore class consistently showed higher ORs. On the other hand, the Low RVScore captured a similar gradient of increasing risk as moderate-risk genes. In M1, the presence of RVs on high-risk genes conferred the higher risk of BC, specially in combination with highest PRS values where it reach an OR = 12.06 (95% CI: 9.96-14.61) on the Training set and OR = 8.25 (95% CI: 5.6-12.16) on the Test set (Table 4.8 and 4.9). Conversely, when considering M7 the presence of the analyzed RVs on the high risk genes had less storng impact, while the RVScore demonstrated greater utility in distinguishing risks, capturing intermediate risk levels where high-risk gene presence alone may not suffice.Specifically, in this mask, high levels of RVScore yielded higher ORs across PRS classes than the presence of RVs in high- and moderate-risk genes.

In conclusion, the RVScore provided a scalable approach to genetic risk assessment, complementing single-gene metrics and PRS. While high-risk and moderate-risk genes contribute significantly to risk prediction, the RVScore captures the combined effect of multiple RVs across a combination of genes. This allows it to provide additional insight into risk profiles, particularly in cases where the presence of RVs on high or moderate-risk gene alone does not fully explain the observed risk. Its ability to differentiate intermediate-risk categories suggests that the RVScore may offer a broader perspective on genetic susceptibility.

PRS_classes	covariate	OR (95%CI)	P_Value	n_samples
Low	No Rv/Null RVScore	0.52 (0.51-0.53)	< 0.001	44636
Intermediate	No Rv/Null RVScore	1 (0.98-1.02)	1	44599
High	No Rv/Null RVScore	1.97 (1.94-2)	< 0.001	44587
Low	RVs on moderate risk genes	1.35 (1.22-1.5)	< 0.001	397
Intermediate	RVs on moderate risk genes	2.62 (2.36-2.89)	< 0.001	433
High	RVs on moderate risk genes	5.17 (4.67-5.72)	< 0.001	458
Low	RVs on high risk genes	3.13 (2.59-3.79)	< 0.001	119
Intermediate	RVs on high risk genes	6.1 (5.04-7.39)	< 0.001	128
High	RVs on high risk genes	12.06 (9.96-14.61)	< 0.001	109
Low	RVScore Low	1.37 (1.21-1.55)	< 0.001	263
Intermediate	RVScore Low	2.67 (2.36-3.02)	< 0.001	282
High	RVScore Low	5.29 (4.67-5.98)	< 0.001	312
Low	RVScore High	2.12 (1.87-2.4)	< 0.001	270
Intermediate	RVScore High	4.12 (3.64-4.67)	< 0.001	287
High	RVScore High	8.17 (7.21-9.26)	< 0.001	269

Table 4.8: Training Set, MASK M1

PRS_classes	covariate	OR (95%CI)	P_Value	n_samples
Low	No Rv/Null RVScore	0.54 (0.52-0.56)	< 0.001	14842
Intermediate	No Rv/Null RVScore	1 (0.97-1.03)	1	14840
High	No Rv/Null RVScore	2.06 (2-2.12)	< 0.001	14829
Low	RVs on moderate risk genes	1.6 (1.33-1.92)	< 0.001	135
Intermediate	RVs on moderate risk genes	2.99 (2.5-3.59)	< 0.001	130
High	RVs on moderate risk genes	6.15 (5.13-7.37)	< 0.001	135
Low	RVs on high risk genes	2.15 (1.46-3.18)	< 0.001	24
Intermediate	RVs on high risk genes	4.01 (2.72-5.91)	< 0.001	27
High	RVs on high risk genes	8.25 (5.6-12.16)	< 0.001	34
Low	RVScore Low	1.61 (1.3-2)	< 0.001	86
Intermediate	RVScore Low	3.02 (2.43-3.74)	< 0.001	97
High	RVScore Low	6.2 (5-7.68)	< 0.001	98
Low	RVScore High	1.72 (1.34-2.2)	< 0.001	74
Intermediate	RVScore High	3.21 (2.51-4.11)	< 0.001	64
High	RVScore High	6.59 (5.15-8.44)	< 0.001	74

Table 4.9: Test Set, MASK M1

PRS_classes	covariate	OR (95%CI)	P_Value	n_samples
Low	No Rv/Null RVScore	0.52 (0.51-0.53)	< 0.001	40971
Intermediate	No Rv/Null RVScore	1 (0.98-1.02)	1	40904
High	No Rv/Null RVScore	1.97 (1.94-2)	< 0.001	40849
Low	RVs on moderate risk genes	1.06 (0.99-1.14)	0.11	833
Intermediate	RVs on moderate risk genes	2.05 (1.91-2.21)	< 0.001	833
High	RVs on moderate risk genes	4.05 (3.77-4.36)	< 0.001	935
Low	RVs on high risk genes	0.58 (0.55-0.61)	< 0.001	3417
Intermediate	RVs on high risk genes	1.12 (1.08-1.17)	< 0.001	3497
High	RVs on high risk genes	2.21 (2.12-2.3)	< 0.001	3443
Low	RVScore Low	0.98 (0.89-1.08)	0.7	444
Intermediate	RVScore Low	1.9 (1.72-2.09)	< 0.001	434
High	RVScore Low	3.75 (3.4-4.13)	< 0.001	537
Low	RVScore High	1.29 (1.17-1.43)	< 0.001	428
Intermediate	RVScore High	2.51 (2.26-2.77)	< 0.001	436
High	RVScore High	4.95 (4.47-5.48)	< 0.001	441

Table 4.10: Training Set, MASK M7

PRS_classes	covariate	OR (95%CI)	P_Value	n_samples
Low	No Rv/Null RVScore	0.54 (0.52-0.56)	< 0.001	13628
Intermediate	No Rv/Null RVScore	1 (0.97-1.03)	1	13603
High	No Rv/Null RVScore	2.06 (2-2.12)	< 0.001	13564
Low	RVs on moderate risk genes	1.24 (1.09-1.41)	0	279
Intermediate	RVs on moderate risk genes	2.32 (2.04-2.64)	< 0.001	266
High	RVs on moderate risk genes	4.77 (4.19-5.42)	< 0.001	286
Low	RVs on high risk genes	0.59 (0.55-0.64)	< 0.001	1105
Intermediate	RVs on high risk genes	1.1 (1.03-1.19)	0.01	1153
High	RVs on high risk genes	2.27 (2.11-2.44)	< 0.001	1175
Low	RVScore Low	1.2 (1.01-1.43)	0.04	145
Intermediate	RVScore Low	2.24 (1.88-2.67)	< 0.001	142
High	RVScore Low	4.6 (3.87-5.48)	< 0.001	159
Low	RVScore High	1.32 (1.1-1.57)	0	145
Intermediate	RVScore High	2.46 (2.06-2.94)	< 0.001	135
High	RVScore High	5.06 (4.23-6.04)	< 0.001	139

Table 4.11: Test Set, MASK M7

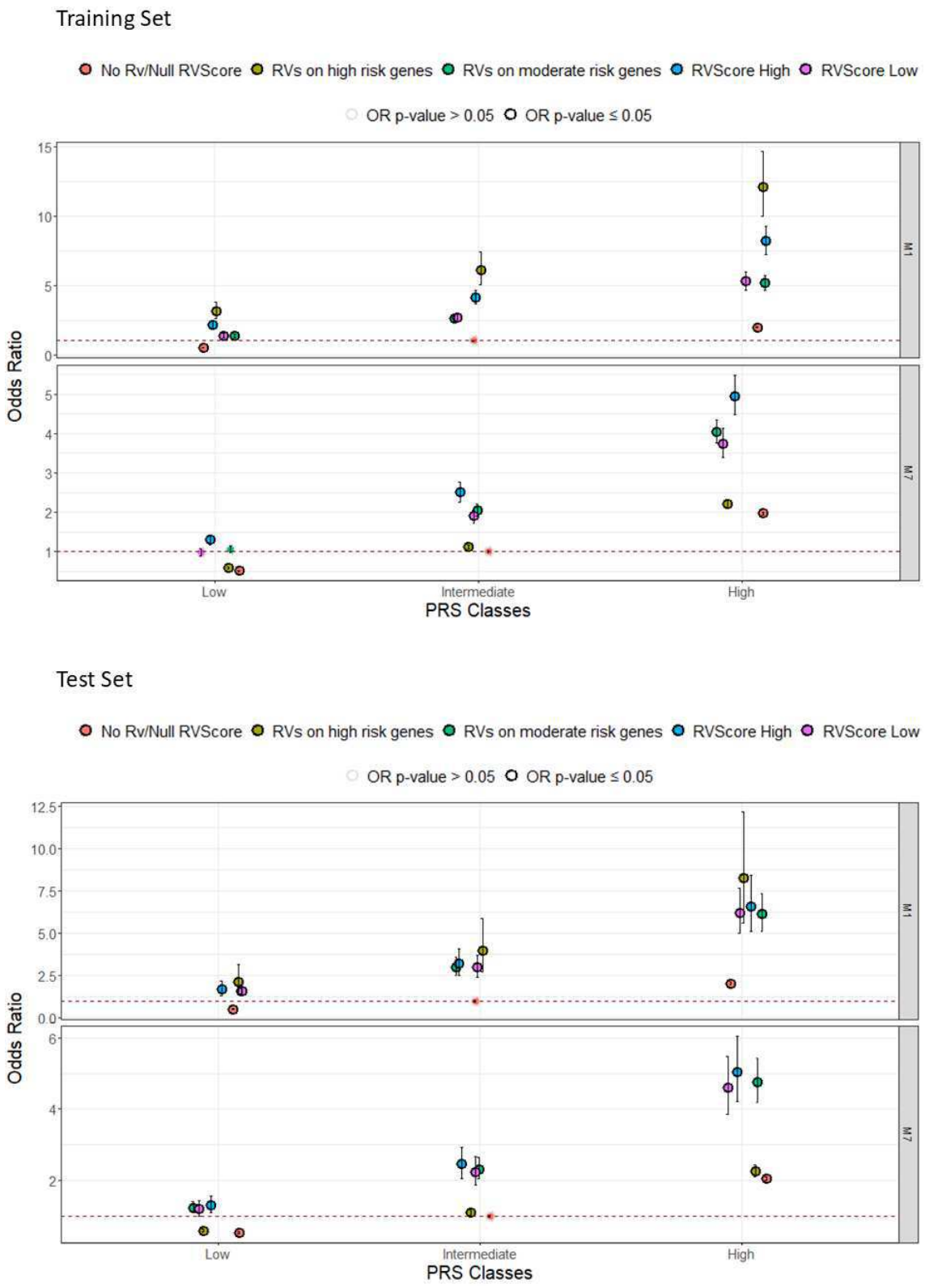


Figure 4.3: RVScore OR across PRS categories

Chapter 5

Rare variants-breast cancer association analysis using BhGLM

5.1 Bayesian hierarchical generalized linear model (BhGLM)

As discussed in Chapter 2, conducting rare variant association analyses at the gene level can boosts statistical power, allowing us to detect associations between rare variants and phenotypic traits that might remain undetected in single-variant tests. However, aggregating variants into genetic units, such as genes, can reduce interpretability. To address this limitation, we sought to exploit an alternative approach: the Bayesian hierarchical generalized linear model (BhGLM) [49]. This method leverages prior information through a spike-and-slab prior on the beta parameters, enabling association analysis at the level of individual RV. By retaining information at the single-variant level, this approach is expected to provide a more interpretable assessment of each variant's contribution to BC risk. The model is implemented in the R package BhGLM [85]. In classical Generalized Linear Model (GLM) analysis,

the parameters are estimated by maximizing the likelihood function

$$p(y|X\beta,\phi) = \prod_{i=1}^{n} p(y_i|X_i\beta,\phi)$$

where y represents the observed values of the response variable, and X denotes the $n \times p$ predictors matrix. Here, β is a $1 \times p$ vector of predictors coefficients and ϕ includes any additional parameters, such as the dispersion coefficient. Differently, Bayesian analysis relies on estimating coefficients using the posterior distribution, which is proportional to the product of the likelihood and prior distributions of parameters:

$$p(\beta,\phi|y,X) \propto p(y|X\beta,\phi)p(\beta,\phi).$$

The package BhGLM uses a uniform distribution as prior for the dispersion parameter ϕ and the intercept β_0 and allows to choose between four different prior distributions for the coefficients array β. Specifically, denoting with π the BC risk, the relationship between RVs and breast cancer probability was defined for each patient i as:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \Sigma_{j=1}^{p}\beta_j x_{ij}$$

The β_j represent the effects size of the p rare variants under analysis. For each coefficient with assumed a spike-slab mixture Student-t prior:

$$\beta_j \sim t_\nu(0,(1-\gamma_j)s_0 + \gamma_j s_1)$$

Where s_1 and s_0 are, respectively, big and small values employed to model informative and non-informative predictors[85]. More specifically, the distribution determined by s_0 is centered around zero with a small standard deviation, making it appropriate for modeling the near-zero effect sizes of rare variants unrelated to

breast cancer. In contrast, a prior shaped by the parameter s_1 exhibits a larger variance, allowing for a wider spread of values, which makes it more suitable for modeling the effect sizes of rare variants associated with breast cancer. Figure 5.1 shows the shape of the distribution regulated by the two parameters.

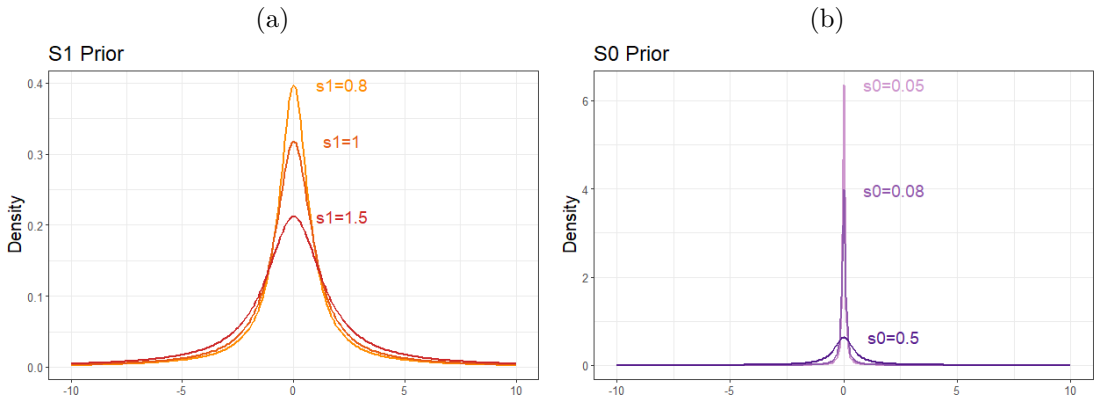


Figure 5.1: s_1 (a) and s_0 (b) distribution densities

Finally, γ_j is the indicator parameter. In the spike-and-slab mixture prior framework, predictors can be grouped by assuming that the indicator parameters for predictors within group g follow a Bernoulli distribution defined as:

$$\gamma_j|\theta_g \sim Bin(\gamma_j|1,\theta_g) = \theta_g^{\gamma_j}(1-\theta_g)^{1-\gamma_j}$$

For the group-specific probability θ_g, a uniform prior is assigned, represented as $\theta_g \sim Beta(1, b_g)$. The parameter b determine the value of θ and, consequently, the degree of shrinkage applied to the coefficients. As illustrated in Figure 5.2, smaller values of b result in a higher probability of θ being close to 1, thereby increasing the likelihood of $\gamma = 1$, which corresponds to a prior distribution regulated by s_1. Conversely, larger values of b increase the probability of θ being near 0, as does γ, leading to a prior distribution governed by s_0.

To summarize, the BhGLM model leverages the Bayesian framework to enhance

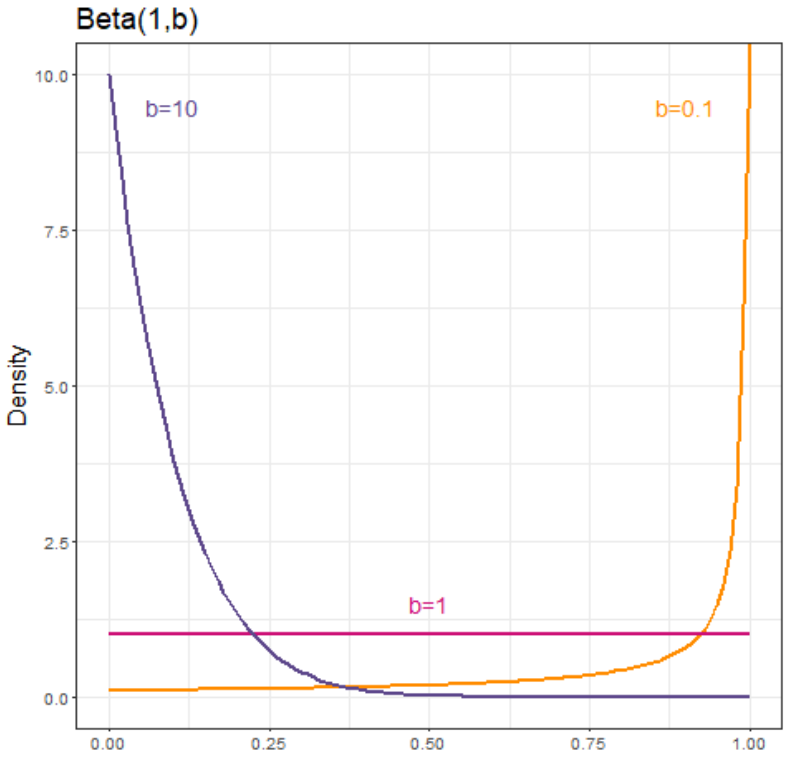


Figure 5.2: Beta density varying the parameter b

statistical power, thereby facilitating the detection of associations between rare variants and breast cancer. Furthermore, the spike-and-slab priors exhibit heavy tails and a peak at zero, enabling reduced shrinkage for relevant predictors and stronger shrinkage for irrelevant ones. This characteristic ensures robust inferences, particularly in large-scale datasets. In addition, the priors not only apply varying degrees of shrinkage across coefficients but also integrate group structures among predictors into the analysis [85].

5.2 Methods for rare variants-breast cancer association analysis using BhGLM

The previous section provided a detailed explanation of the BhGLM methodology. Here, it's presented how we adapted this approach for analyzing the association between RVs and BC in the UKBB cohort. As a first step, two distinct simulation scenarios were implemented to assess the model's capability to detect RVs associated with the trait of interest and to identify the parameter combination that maximizes accuracy in capturing RV-phenotype associations. Subsequently, the model was validated in a controlled setting using a restricted subset of genes. Finally, we extended the BhGLM to the full set of genes belonging to the Clinical Exome assay. In the following sections, a detailed description of each step is provided.

5.2.1 BhGLM validation and parameter selection in simulation settings

Following the idea of [34], we established two distinct simulation frameworks designed to evaluate the effectiveness of BhGLM in detecting RVs associated with the phenotype across different scenarios and to identify the optimal set of param-

eters that maximize detection accuracy. In each simulation setting, the model's performance was assessed by evaluating its sensitivity, specificity, false positive rate (FPR), and balanced accuracy. These metrics provided a comprehensive view of the BhGLM's ability to correctly identify rare variants truly associated with the phenotype. To further examine the precision of BhGLM in estimating the effect sizes of RVs, we calculated the logarithm of the mean squared error (MSE) for the estimated beta coefficients. This measure is defined as follows:

$$log(MSE) = log(\Sigma_{j=1}^{p}(\hat{\beta}_j - \beta_j))^2$$

where p represents the total number of RVs included in the model. The $log(MSE)$ provides insight into the correctness of the model's effect size estimation by quantifying the average squared deviation of the estimated betas from the true values, with lower values indicating better accuracy in beta estimation. In both of the settings a binary phenotype was assumed (eg. presence or absence of a disease) and used 1000 loss-of-function rare variants (with minor allele frequency ≤ 0.01) randomly selected form the 82738 LoF RVs of the ClinicalExome, stratified into three MAF classes:

- **Extremely rare variants**: 333 RVs with MAF in the lower 30% quantile.
- **Moderately rare variants**: 333 RVs with MAF in the middle 40% quantile.
- **Weakly rare variants**: 334 RVs with MAF in the upper 30% quantile.

Finally, we assumed a spike-slab mixture Student-t prior for the β parameters with the parameter indicating the degree of freedom $\nu = 1$

Simulation Setting 1

In Simulation Setting 1, our aim was to assess the sensitivity of the Bayesian hierarchical generalized linear model in response to differing proportions of BC-associated variants and varying parameters S0 and S1. In this framework we assumed that there was no correlation between the selected RVs and assigned the effect sizes (βs) to variants in proportion to their MAF using the following formula:

$$\beta = \begin{cases} \text{uniform}(-\ln(100), \ln(100)) \cdot |\ln(\text{MAF})| & \text{if the variant is associated with trait,} \\ 0 & \text{otherwise.} \end{cases}$$

The effects where chosen randomly to be either positive or negative for variants considered associated with the binary trait of interest. Variants not associated with the phenotype were assigned a beta of zero. The probability of the binary trait was calculated for each subject as the inverse-logit of the linear combination of the assigned betas and the genotype value of each variant. To control for baseline risk, an intercept term of zero was applied, setting the trait baseline probability to 0.5 for each individual. We tested the model under various combinations of phenotype-associated variants proportions $(0.01, 0.02, 0.05, 0.1, 0.2, 0.3)$ and parameter settings (S0 $\in \{0.03, 0.05, 0.1, 0.5, 0.8\}$ and S1 $\in \{1, 1.5\}$). We set the parameter $b = 1$. For each setting of proportions and parameters, we ran the model 10 times to evaluate BhGLM consistency and accuracy in detecting RV-phenotype associations.

Simulation Setting 2

In Simulation Setting 2, the objective was to evaluate the dependence of the BhGLM on group-specific prior information, as well as on the values of S0 and S1. In this setting we designed the simulation framework dividing the 1000 randomly selected rare variants into five groups with specific assumptions on their relation with the

phenotype:

- strongly phenotype related variants: two groups consist of variants that increase or decrease disease odds by a 10-fold factor ($\beta = \pm log(10)$)
- weakly phenotype related variants: two groups consist of variants that increase or decrease disease odds by a 5-fold factor ($\beta = \pm log(5)$)
- one group comprises variants unrelated to the phenotype ($\beta = 0$).

Again for each subject, the probability of the binary trait was computed as the inverse-logit of a linear combination of model coefficients (βs) and the variant dosage. The intercept was set to 0, providing a baseline disease probability of 0.5 for each individual. In Section 5.1, we detailed how the parameter b controls the mixing between the priors defined by S0 and S1 and described how this parameter can be adjusted for distinct groups of variants. In the present setting, we conveyed information about the group structure of the rare variants to the model by varying the parameter b. Specifically, the model was ran under three conditions:

- **Correct Group Information**: we set $b = 0.001$ for variants in the group strongly related to the phenotype, $b = 1$ for variants weakly related to the phenotype, and $b = 100$ for variants unrelated to the phenotype. This setup allowed the model to incorporate group-specific priors that reflect the varying degrees of association with the phenotype.
- **Incorrect Group Information**: We next evaluated the model's response to false group information by intentionally misassigning the parameter values. In this scenario, $b = 100$ was set for variants strongly associated with the phenotype, $b = 0.001$ for weakly related variants, and $b = 1$ for phenotype-unrelated variants. This configuration enabled us to observe the effect of providing misleading information regarding the cluster structure of the RVs.

- **No Group Information**: Finally, we assessed the model without any group-specific information by setting $b = 1$ uniformly for all RV groups. This condition represented a scenario in which the model received no prior knowledge about the group structure of the variants.

These three conditions allowed us to explore how the BhGLM responds to varying levels of accuracy in group prior information, examining its sensitivity to the parameter b as it pertains to variant clusters with differing relationships to the trait of interest. We tested the model under various combinations of the parameter S0 $\in \{0.03, 0.05, 0.1, 0.5, 0.8\}$ and S1 $\in \{1, 1.5\}$. For each parameter combination, the BhGLM was run 10 times. In each iteration, a different subset of variants was randomly assigned to each of the five groups.

5.2.2 Application to real data in controlled settings

Following validation of the model on simulated data, we applied it to real data within a controlled experimental framework. This setting was designed to allow us to systematically assess the quality and relevance of the selected rare variants while maintaining control over the evaluation parameters. Specifically, the BhGLM was implemented on a targeted selection of genes, grouped into three categories based on their established relationship with BC:

- **Top 5 BC-Related Genes**: This group included five well-characterized BC-related genes (ATM, BRCA1, BRCA2, CHEK2, and PALB2).

- **ClinGen Genes**: This set consisted of 15 ClinGen cancer-related genes from which those in the Top 5 BC genes were excluded. These genes are recognized by the ClinGen resource as having strong evidence of a relationship with cancer, providing a broader context for exploring RV linked to BC risk.

- **Non-ClinGen Genes**: This category comprised 20 genes that were not part of the ClinGen cancer list. This group was included to evaluate associations in genes with less established cancer relevance.

Table 5.1 shows the genes in each category. We hypothesized that the model would

Table 5.1: List of Genes by Category

Gene Categories and Associated Genes	
Top 5 BC-Related Genes	ATM, BRCA1, BRCA2, CHEK2, PALB2
ClinGen Genes	BARD1, CDH1, CHEK1, EPCAM, GEN1, MCPH1, MLH1, MSH6, MUTYH, PIK3CA, PMS2, RAD51C, RAD51D, RECQL, SLX4
Non-ClinGen Genes	UGT1A7, PTGDR, CCDC151, LRBA, QARS, RHBDF2, FLVCR1, RELN, C15orf41, ADGRG1, UTRN, PNKD, BTRC, MAPRE2, SCN4A, LRRC6, GFI1B, PODXL, COL4A3, MAP3K1

select a higher number of RVs in the first two gene categories, which are more directly linked to BC, and fewer RVs in the Non-ClinGen group. To further assess the significance of the selected RVs, we used ClinVar and ACMG annotations.

For each selected variant, we determined the minimum number of carriers required to achieve statistical significance ($\alpha = 0.05$) with a power of 0.8, based on the formula outlined by Kelsey et al for Risk Ratio (RR) [68]. This calculated threshold was then compared to the observed number of carriers for each variant. For RVs with a sufficient number of carriers, the RR was estimated to assess the strength of their association with breast cancer risk. Subsequently, the model-selected RVs were integrated with PRS to quantify their impact on BC risk, independent of the underlying genetic background. To achieve this, we computed the OR stratified by defined PRS categories. Specifically, we categorized individuals based on PRS tertiles distribution: "High" PRS values were those above the 66th percentile, "Medium"

encompassed PRS values between the 33th and 66th percentiles, and "Low" PRS values fell below the 33th percentile, as it was done for the evaluation of the Burden Test RVScore. We compared the OR for the presence of the RVs selected with the OR of the presence of RVs not selected by the model for each class of PRS.

The BhGLM was applied exclusively to LoF RVs within the three gene categories, removing RVs carried by just one subject. RV presence was coded as a binary variable (1 indicating the presence of the RV in the patient and 0 indicating its absence). Additionally, we incorporated patient age and the first ten genetic principal components (calculated according to UK Biobank Data-Field 220099) as covariates. To ensure uniformity in the data range, both age and the principal components were scaled to the $[0, 1]$ interval using min-max normalization. To address class imbalance between cases and controls, we applied class weights to the model, helping to balance the influence of minority and majority classes on model training. The model was trained on 75% of the data designated for RVs association analysis. We kept all rare variants that were present in at least one patient of the training set. The β parameters in the BhGLM were modeled using a spike-slab mixture Student-t prior with degrees of freedom set at $\nu = 1$. We set $S0 = 0.5$ and $S1 = 1$ that is the optimal parameter combination according to the simulation settings. Below we present a comprehensive overview of the BhGLM configuration, detailing the specific parameters setting.

In this scenario the following numbers of RVs were selected in each gene group:

- **Top 5 BC-Related Genes**: 125 LoF RVs.
- **ClinGen Genes**: 141 LoF RVs
- **Non-ClinGen Genes**: 141 LoF RVs

We incorporated prior information into the model to reflect the expected strength

of association between rare variants and breast cancer risk. This was achieved by setting the parameter b at varying levels based on gene category and variant frequency, thereby guiding the model in its assessment of variant relevance. Specifically:

- $b = 0.001$ was set for variants within the **Top 5 BC-Related Genes**, which are strongly associated with BC.
- For variants within the **ClinGen genes**, b was set to 0.01, indicating a slightly weaker but still meaningful association.
- For **non-ClinGen genes**, the parameter b was adjusted according to the minor allele frequency to account for the potential effect size differences: ultra-rare variants with $MAF \leq 0.0001$ were assigned $b = 0.1$, as they are presumed to have a higher damaging effect, while variants with $MAF \in (0.0001, 0.01]$ were assigned $b = 1$. Just 4 LoF variants in non-ClinGen genes were found with $MAF \in (0.0001, 0.01]$

Additionally, we set $b = 1$ for covariates representing age and the first 10 genetic principal components. This framework allowed the model to prioritize variants in a biologically meaningful way based on gene relevance and variant rarity.

5.2.3 Extension to the whole Clinical Exome

Subsequently, we enforced the BhGLM to the whole set of genes belonging to the Clinical Exome, in order to further enhance the statistically BC-associated RVs. As it was done for the controlled setting, we applied BhGLM exclusively to loss-of-function RVs carried by more than one patient in the training set.Following the previous scheme, RV presence was coded again as a binary variable and we included patient age and the first ten genetic principal components scaled with min-max normalization as covariates in the model. Again a weighted model was used to

control for class imbalance, and we trained the BhGLM on 75% of the data. Finally, we assigned a spike-slab mixture Student-t prior with degrees of freedom set at $\nu = 1$ to the β parameters and set $S0 = 0.5$ and $S1 = 1$, which represents the optimal parameter combination derived from the simulation settings. For computational reason we couldn't run the model including all the RVs simultaneously. To overcome to this problem, we implemented a resampling strategy in which 20 RVs subsets of equal size were analyzed using the BhGLM model.Given that the strength of association for each RV is estimated by BhGLM in the context of other variants within each subset, the resampling scheme was further repeated 10 times. This iterative process allowed us to explore multiple combinations of RVs, ensuring a more comprehensive assessment of the associations while accounting for possible interactions and dependencies among variants. We ultimately selected all the LoF RVs with a BhGLM association p-value ≤ 0.05 in at least 5 out of 10 resamplings. Using these as covariates, along with age and the first 10 principal components, we re-ran the BhGLM analysis. As we did before, for each resampling and for the last application of the BhGLM, prior information was incorporated into the model to reflect the expected strength of association between rare variants and breast cancer risk by setting the parameter b to different values. Specifically, the variants were divided in 3 subgroups:

- **RVs on clinically relevant genes**: all the LoF RVs belonging to the genes in the curated lists ClinGen and Genturis, described in section 4.1, were assigned to this group. For these variants we set $b = 0.001$, reflecting a strong expected association

- **RVs in genes related to BC by GWAS**: this category was composed by all the LoF RVs in the genes from Open Targets and Harmonizome 3.0. We assigned $b = 0.01$ describing a moderate effect on breast cancer risk

- **RVs on genes of unknown significance**: RVs in this category receive $b = 1$, reflecting a decreasing supposed association strength

By setting b values for each group of genes and RVs we enabled the model to incorporate prior biological knowledge to prioritize variants with stronger evidence of disease association. Again, we set $b = 1$ for covariates representing age and the first 10 genetic principal components.

Finally, the p-values of the RVs identified by the final model were adjusted for multiple testing, and those with $p_{adj} \leq 0.05$ were selected to construct the RVScore.

5.2.4 RVs-based score definition using the results of the BhGLM application to the whole Cllinical Exome

The final step of our study involved constructing an RVScore to evaluate the collective impact of the RVs identified by BhGLM on breast cancer risk and to validate our findings using the sub-cohort excluded from the RV-BC association analysis. To achieve this, we selected all RVs significantly associated with BC in the final BhGLM implementation and defined the RVScore for each patient i using the following formula:

$$Score_{RV_i} = \sum_v \frac{\beta_v d_{iv}}{se(\beta_v)}$$

where:

- β_v represents the estimated effect size of the v-th RV on the phenotype, derived from the BhGLM model
- $se(\beta_v)$ is the standard error of the estimated effect size β_v, reflecting the uncertainty in the estimate
- finally, d_{iv} corresponds to the allelic dosage for variant v in subject i

In simulation settings, it is acknowledged that the estimated β_v may not accurately represent the true magnitude of the association between the RVs and the phenotype due to statistical noise, model assumptions, or sample size limitations. To address this issue, we chose to scale each RV's contribution to the score using the ratio $\frac{\beta_v}{se(\beta_v)}$. This approach places greater emphasis on RVs with more reliable associations by accounting for the precision of the effect size estimates. Specifically, the ratio $\frac{\beta_v}{se(\beta_v)}$ can be interpreted as a signal-to-noise measure of the effect size, thereby incorporating the uncertainty associated with the estimation process. By doing so, our scoring methodology reflects not only the magnitude of the estimated association but also its statistical robustness, providing a more nuanced representation of the potential impact of RVs on BC risk. Following the approach used for the validation of the RVScore built on the Burden Test results, we stratified the BC-PRS into three categories based on tertile distributions, and categorized RVScore into the three groups: Null RVScore, Low RVScore for values below the median computed excluding the Null RVScore group, and High RVScore for values above the median. A logistic regression model on both the training and test set, was fitted to estimate ORs using the Intermediate PRS group with Null RVScore as the reference category.

Through the application of the BhGLM for RVs selection and coefficients estimation, we aimed to develop a combined score able to capture the collective impact of RVs on the risk of BC, while preserving the ability to dissect and interpret the individual contribution of each RV to BC susceptibility.

5.3 Results of rare variants-breast cancer association analysis using BhGLM

To evaluate the effectiveness of BhGLM in detecting RVs associated with the phenotype and to identify the optimal set of parameters that maximize detection accuracy, two different simulation settings were implemented as previously described (section 5.3.1). Following, the model was applied to real data within a controlled experimental framework designed to allow us to systematically assess the quality and relevance of the selected RVs (section 5.3.2). Subsequently, we enforced the BhGLM to the whole set of genes belonging to the Clinical Exome, in order to further enhance the statistically BC-associated RVs (section 5.3.3). As final step, a RVScore was built to evaluate the collective impact of the RVs identified by BhGLM on BC risk (section 5.3.4).

5.3.1 BhGLM in the simulation settings

This section describes the outcomes of the two simulation setting built to evaluate the BhGLM and select the model parameters. Specifically, the first one aimed at assessing the susceptibility of the BhGLM in response to differing proportions of BC-associated RVs, while the second one had the goal to establishing the dependence of the BhGLM performances on group-specific prior information. We selected the combination of parameters providing the best performances results according to the evaluation measures reported in section 5.2.1.

Simulation Setting 1

Focusing on the impact of the percentage of phenotype-associated rare variants on BhGLM performance, figure 5.3 demonstrates that the specificity consistently

remained close to 1, regardless of the proportion considered. Consequently, the FPR exhibited average values below 0.001, highlighting the model's strong ability to effectively minimize false positives. Conversely Sensitivity assumed considerably lower values, emphasizing the model's tendency to maintain a more stringent approach, which reduce the detection of both true and false positives. Consequence of that can be also the observed slight, albeit non-significant, increment in sensitivity as the proportion of BC-associated RVs decreased. Specifically, as the number of phenotype-associated RVs increases, the likelihood of these RVs being classified as "unrelated" by BhGLM also rises. This behavior explains also the observed trends in balanced accuracy and log10MSE: balanced accuracy exhibited a slight decrease as the proportion of phenotype-associated RVs increased, while log10MSE showed an upward trend. The latter is likely due to the model estimating the regression coefficients (β) of false negative RVs as zero, despite their actual non-zero associations with the phenotype.

There was not too much difference between the BhGLM performance varying the parameters in this setting. The combination $S0 = 0.5$, and $S1 = 1$ (Figure 5.4) was the one providing the best Balanced Accuracy for almost all the proportions of trait-associated RVs (average Balanced Accuracy among proportions = 0.65).

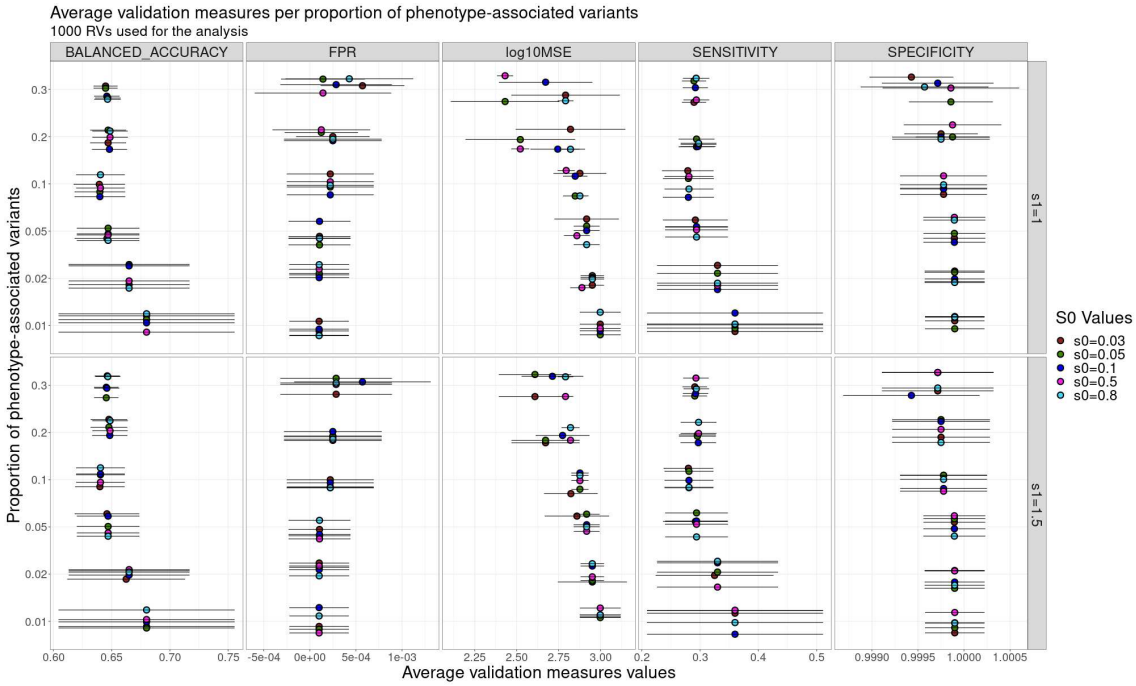


Figure 5.3: Simulation Setting 1 BhGLM evaluation measure varying the parameters combination

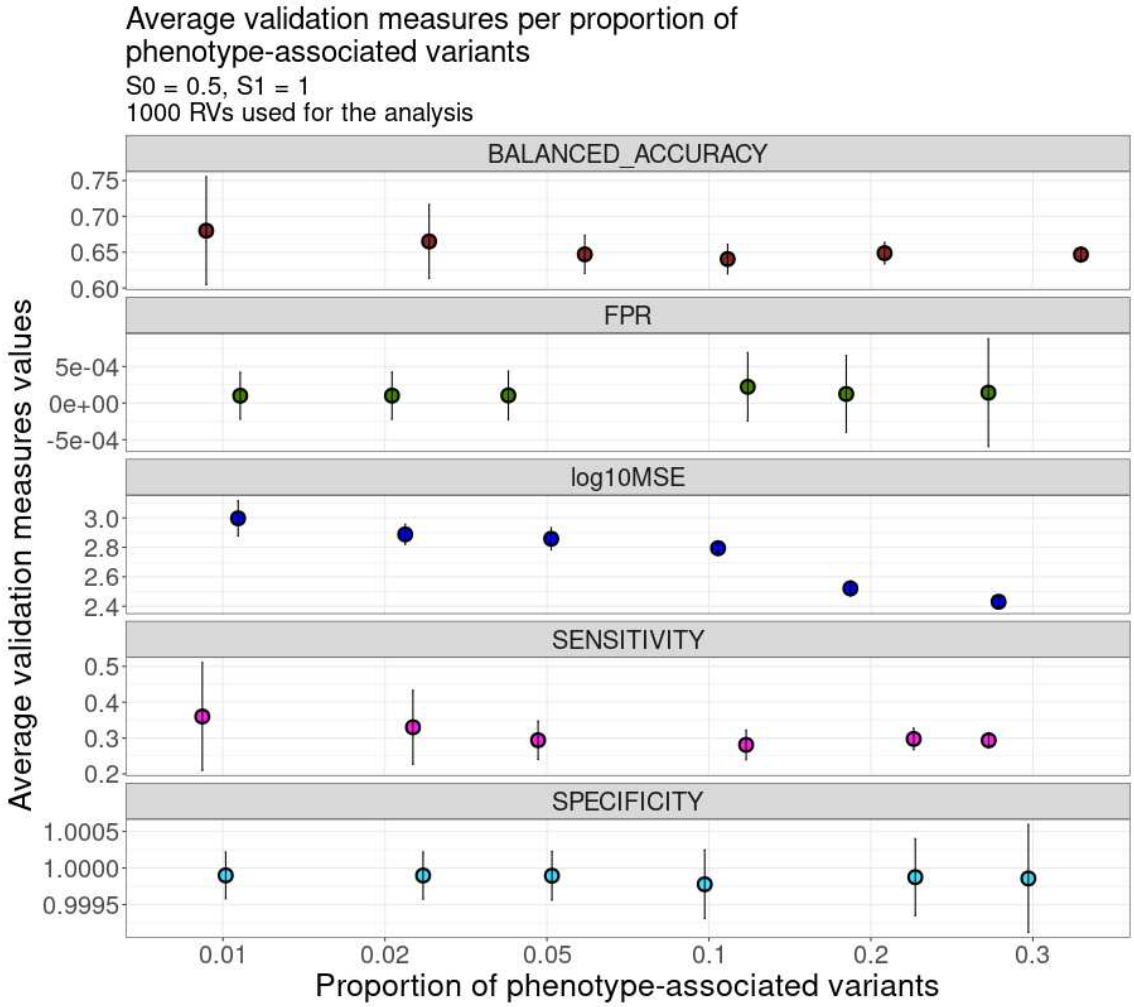


Figure 5.4: Simulation Setting 1 BhGLM evaluation measure when $S0 = 0.5$ and $S1 = 1$

Simulation Setting 2

The results shown in Figure 5.6 demonstrate that incorporating prior knowledge about potential group structures in the data enhances the model's performance. While the trend of improvement is evident, the overall stability of the results highlights the Bayesian approach effectively leverages data-driven selection processes, even in the absence of/wrong group prior information. Notably, the model exhibits again a pronounced tendency to minimize false positives, resulting in consistently high specificity and correspondingly low sensitivity(Figure 5.5). The balanced accuracy reached an average value of 0.66 when $S0 = 0.5$ and $S1 = 1$, with group prior information provided. The parameter combination of $S0 = 0.5$ and $S1 = 1$ yielded the most consistent and favorable evaluation metrics also in this setting, indicating it as the optimal couple of parameters for achieving reliable model performance across the tested conditions.

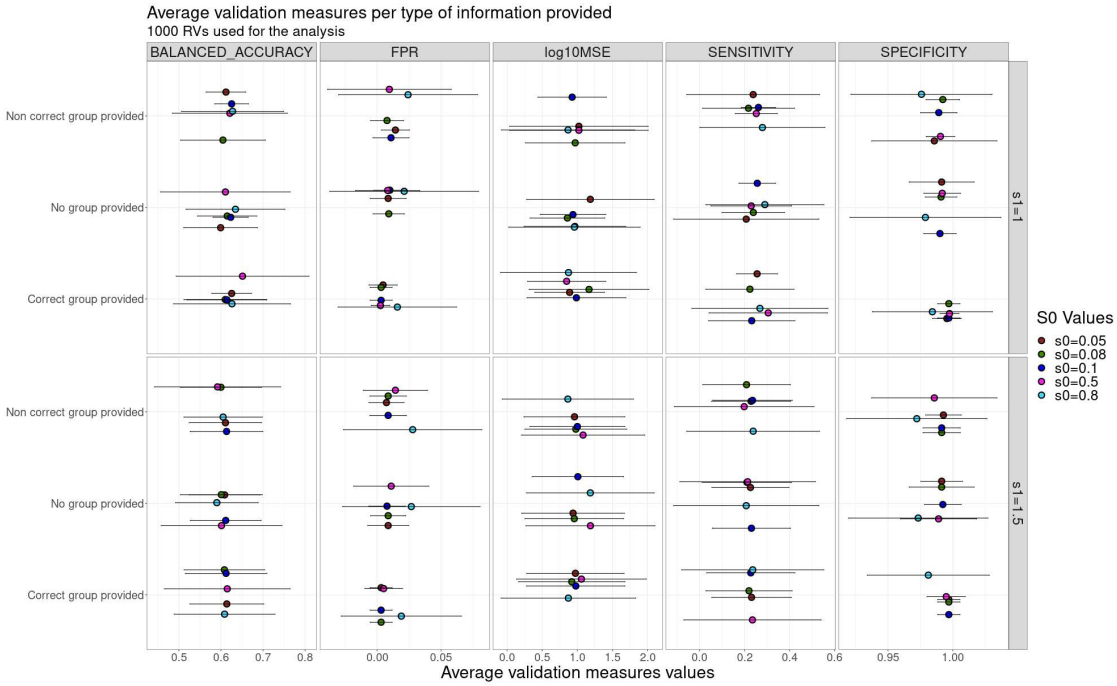


Figure 5.5: Simulation Setting 2 BhGLM evaluation measure varying the parameters combination

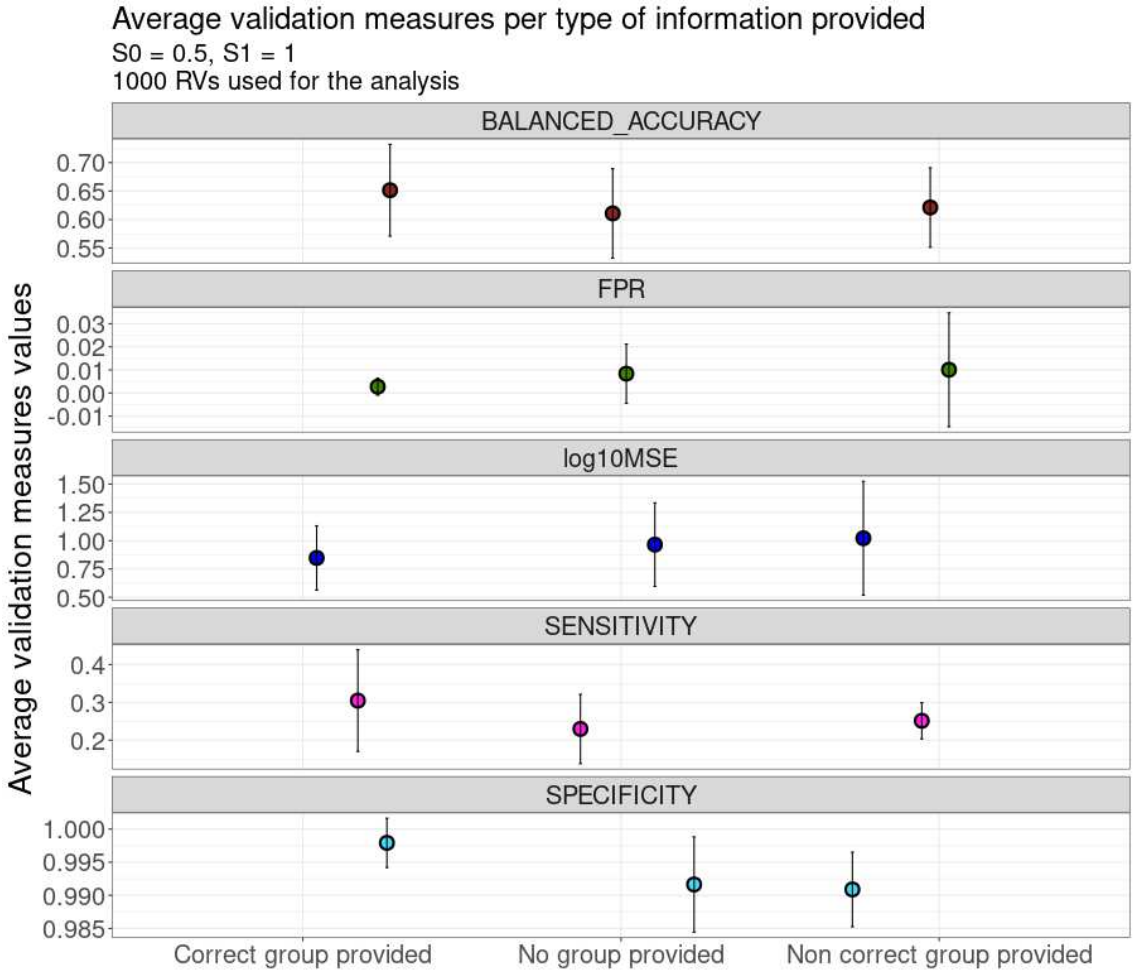


Figure 5.6: Simulation Setting 1 BhGLM evaluation measure when $S0 = 0.5$ and $S1 = 1$

5.3.2 Application of BhGLM in the controlled setting

In this section we report the result of the application of BhGLm on the controlled setting designed to allow us to systematically assess the quality and relevance of the selected RVs. With this goal, the BhGLM was fitted on a targeted selection of genes, grouped into three categories based on their established relationship with breast cancer as explained in section 5.2.2, expecting an higher number of RVs to be retrieved in the more BC-related gene categories. To further assess the significance of the selected RVs, we estimate the RR for the selected RVs with a sufficient number of carriers and used ClinVar and ACMG annotations. Finally we compared the OR for the presence of the RVs selected with the OR of the presence of RVs not selected by the model stratified by defined PRS classes to evaluate their impact to breast cancer risk, independent of the underlying genetic background.

As shown in Figure 5.7, among the 407 LoF rare variants included as covariates in the model (125 in the Top 5 BC-related genes and 141 across both ClinGen and non-ClinGen genes), 46 were identified as significantly associated with BC ($p - value \leq 0.05$). Specifically, 26.4% of RVs in the Top 5 BC-related genes were determined to be BC-associated, compared to 7.8% in ClinGen genes and only 1.42% in non-ClinGen genes, aligning with our initial hypothesis. Among the top five genes, BRCA1 exhibited the highest proportion of BC-associated RVs, with 7 out of 14 variants (50%) identified by the model. BRCA2 ranked second, with 30.96% of its RVs deemed significant by BhGLM. In contrast, PALB2, ATM, and CHEK2 displayed lower proportions of BC-associated RVs (Figure 5.8). These findings are consistent with the established roles of BRCA1/2 as high-risk genes and PALB2, ATM, and CHEK2 as moderate-risk genes. Within ClinGen genes, MUTYH, which is implicated in telomere packaging and base excision repair pathways, and SLX4, involved in DNA repair pathways similar to BRCA1 and BRCA2, demonstrated

proportions and numbers of BC-associated RVs comparable to those observed in moderate-risk genes (Figure 5.8). For non-ClinGen genes, two BC-associated RVs were identified in UTRN, a gene involved in actin-binding pathways. These variants represented 10.53% of the analyzed RVs for this gene (Figure 5.8).

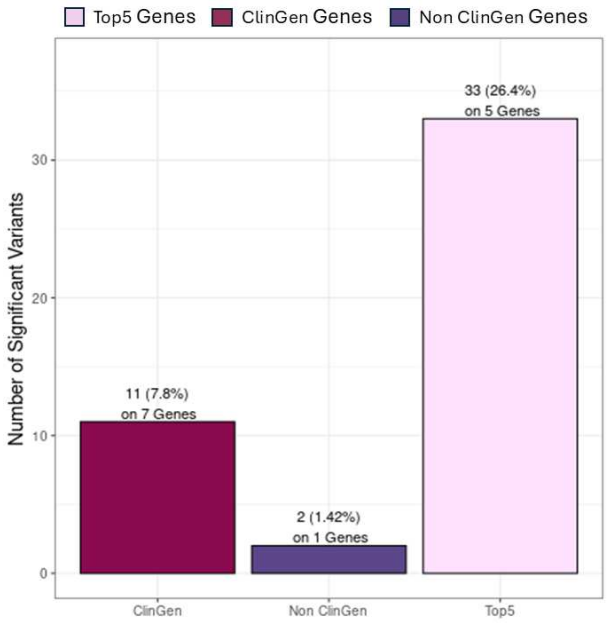


Figure 5.7: Number of BhGLM selected RVs per group of genes

Among the selected RVs, only 25 were observed in a sufficient number of patients to allow for evaluation using RR analysis (Figure 5.9). Nearly all of these RVs were located within the Top 5 BC-related genes, with the exception of a single variant in BARD1, which is annotated by ClinVar as "Pathogenic/Likely pathogenic" for hereditary breast and ovarian cancer syndrome, familial breast cancer, and hereditary cancer-predisposing syndrome. Notably, the RR analysis identified a significant association ($p - value \leq 0.05$) for all these variants. The selected RVs were subsequently evaluated using ACMG and, when unavailable, ClinVar annotations. Remarkably, the majority of the selected RVs were classified as "Pathogenic." The sole exception was a single RV in GEN1, which was annotated as a "Variant of Un-

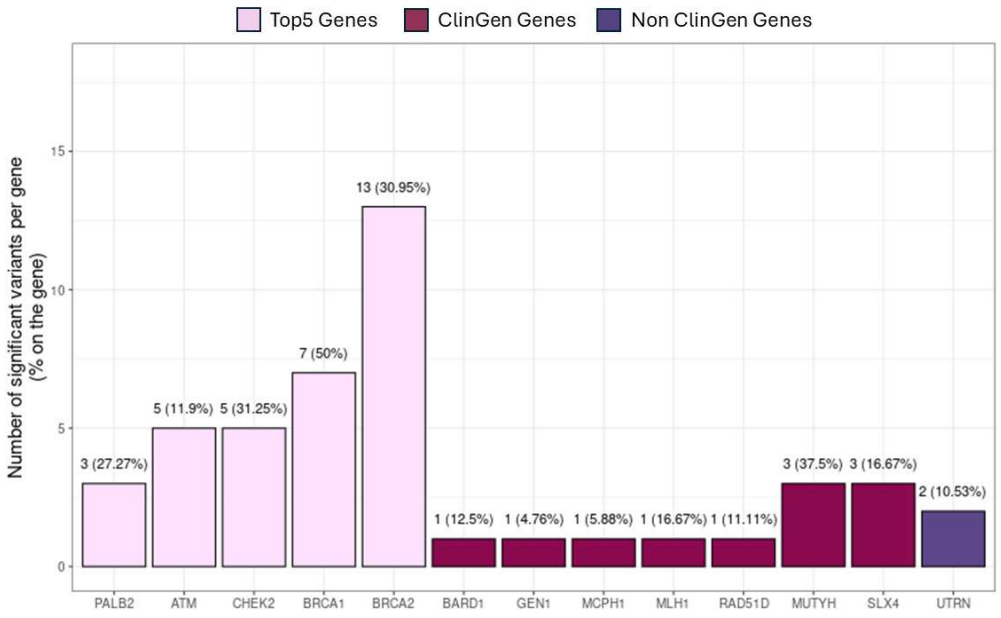


Figure 5.8: Number of BhGLM selected RVs per gene and group of genes

certain Significance" (Figure 5.10). Of the 21 RVs that could not be assessed using RR analysis, only 5 were left without annotation: the 3 RVs of SLX4 and the 2 on UTRN.

Finally, to evaluate the impact of the selected RVs to breast cancer risk, independent of the underlying genetic background, we compared the OR for the presence of the RVs selected with the OR of the presence of RVs not selected by the model in the same genes across defined BC-PRS classes.

Overall, the presence of RVs selected by the BhGLM model within the analyzed genes substantially increased the risk of BC compared to unselected RVs, particularly in individuals with high PRS (OR = 7.06, 95% CI: 6.37–7.83). However, even among individuals with low PRS values, the presence of BhGLM-selected RVs was associated with a higher risk of BC (OR = 1.86, 95% CI: 1.68–2.06) than the intermediate PRS group (OR = 1.00, 95% CI: 0.98–1.02) and was comparable to the risk observed in the high PRS group alone (OR = 1.97, 95% CI: 1.94–2.00). Notably,

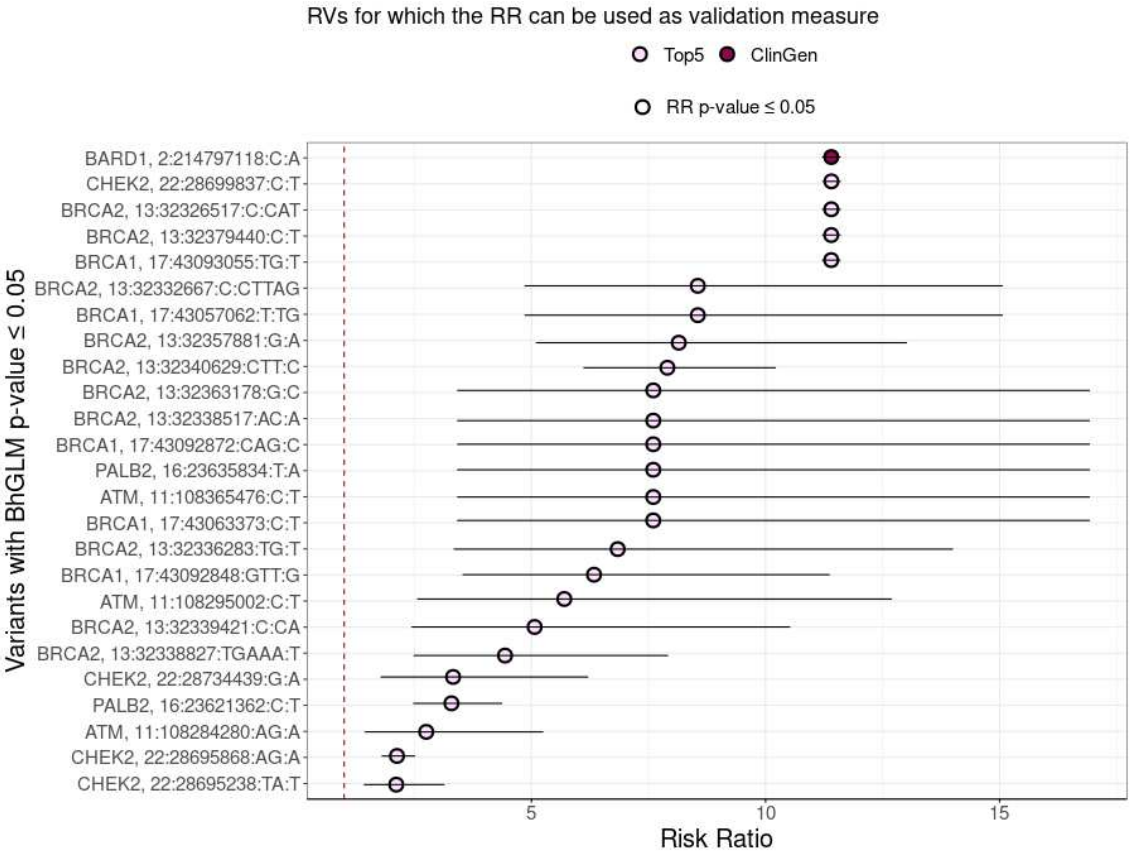


Figure 5.9: BhGLM selected RVs valuable using RR

the presence of unselected RVs within the same genes followed a risk stratification pattern consistent with PRS alone, without an additional increase in BC risk (Table 5.2, Figure 5.11).

When analyzing individual genes, the top five BC-related genes, along with MUTYH and SLX4, exhibited an elevated BC risk associated with the presence of selected RVs in individuals with high PRS values compared to PRS alone. Furthermore, BhGLM-selected RVs in ATM, BRCA1, BRCA2, and CHEK2 showed a significant increase in (ORs compared to unselected RVs in these genes, regardless of PRS class. For PALB2, MUTYH, and SLX4, although an increase in ORs was observed with BhGLM-selected RVs, the magnitude of this increment was less pronounced, potentially indicating false negatives in RV selection for these genes

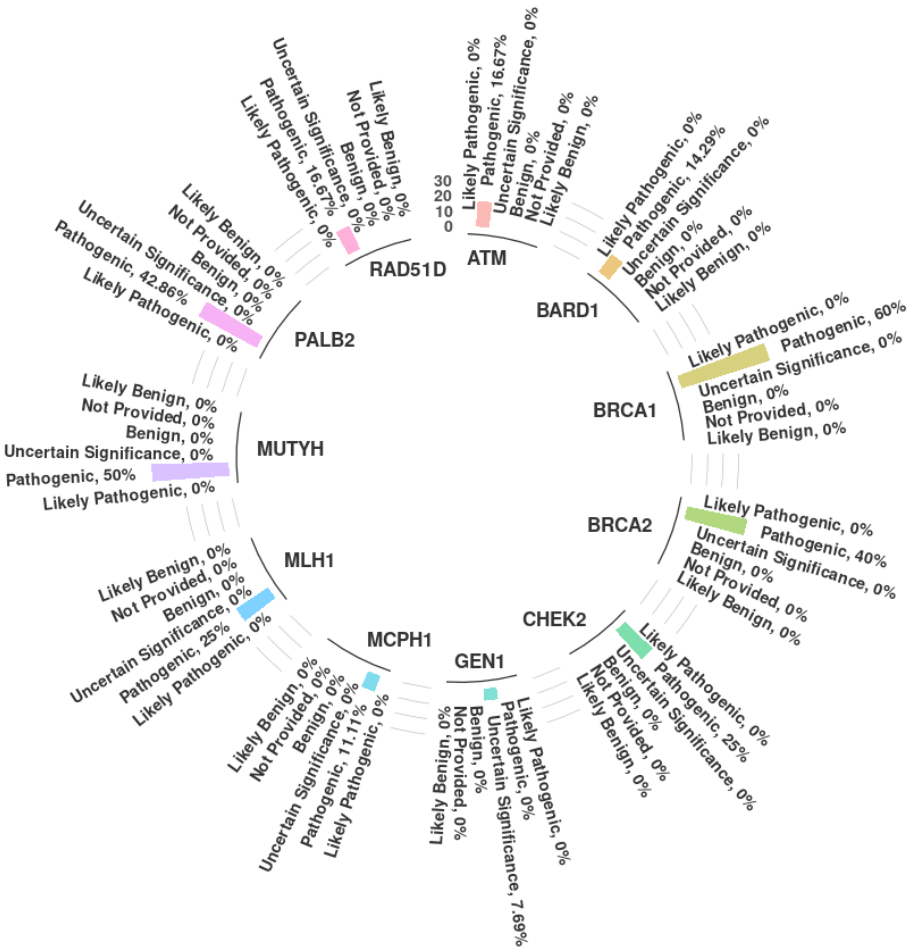


Figure 5.10: Annotation of BhGLM selected RVs

(Figure 5.12).

PRS_classes	covariate	OR (95%CI)	P_Value	n_samples
Low	No Rv	0.52 (0.51-0.53)	<0.001	44,878
Intermediate	No Rv	1 (0.98-1.02)	1	44,877
High	No Rv	1.97 (1.94-2)	<0.001	44,832
Low	selected RVs	1.86 (1.68-2.06)	<0.001	369
Intermediate	selected RVs	3.58 (3.23-3.97)	<0.001	391
High	selected RVs	7.06 (6.37-7.83)	<0.001	445
Low	unselected RVs	0.55 (0.51-0.6)	<0.001	775
Intermediate	unselected RVs	1.06 (0.97-1.16)	0.17	756
High	unselected RVs	2.09 (1.92-2.28)	<0.001	749

Table 5.2: OR of BhGLM selected RVs by PRS classes

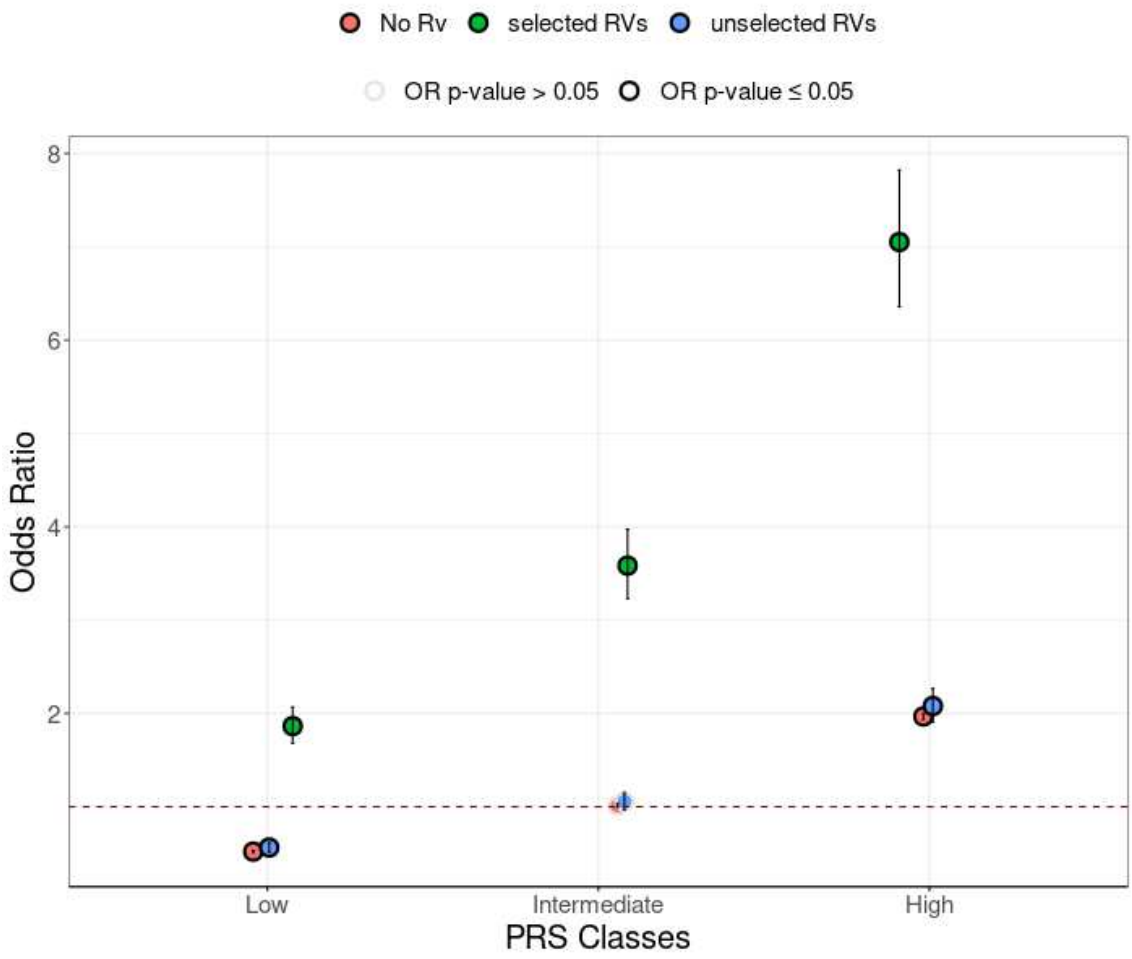


Figure 5.11: OR of BhGLM selected RVs in Controlled Setting by PRS classes

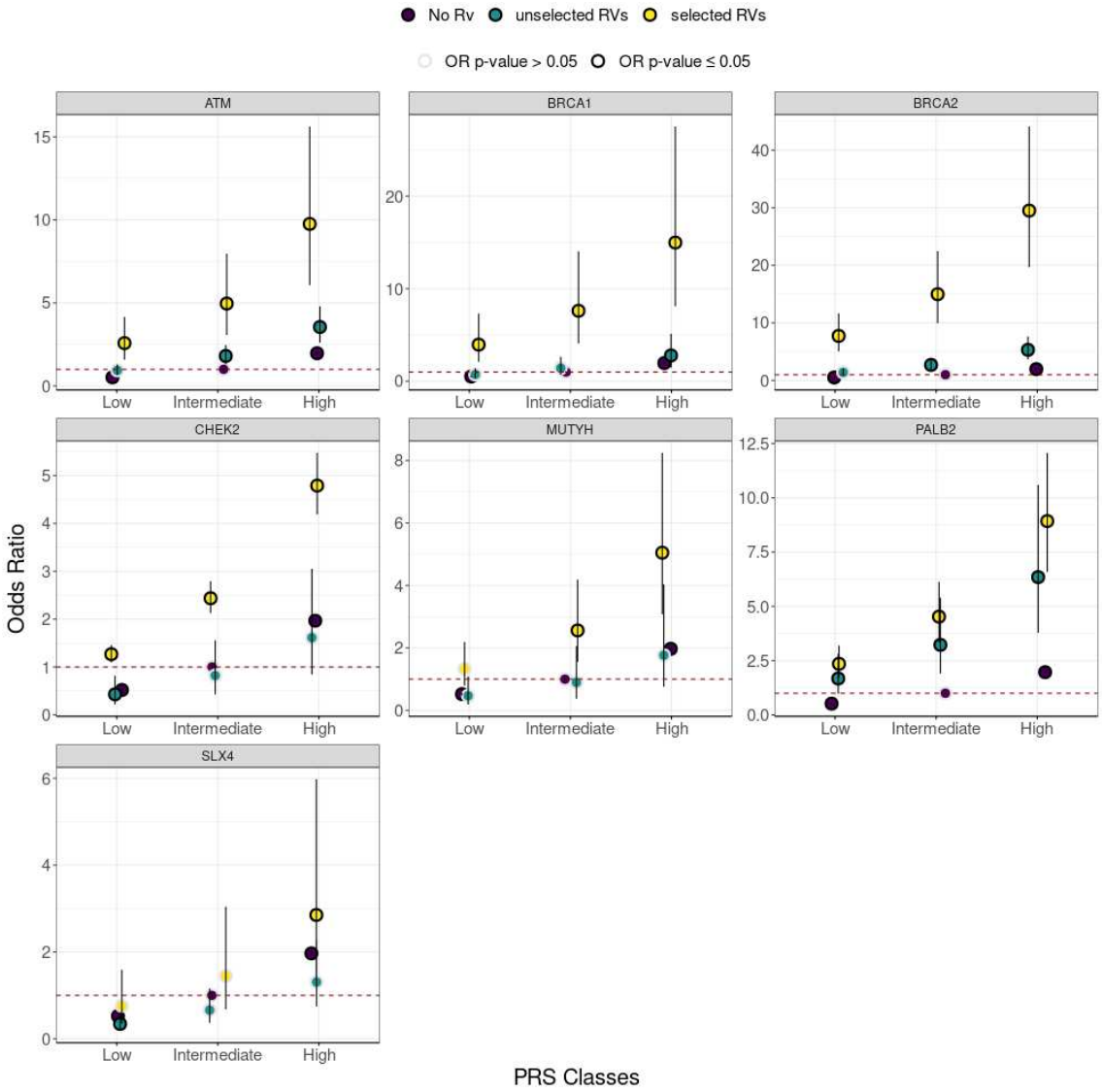


Figure 5.12: OR of BhGLM selected RVs by PRS classes in those genes were the impact of selected RVs was higher

5.3.3 Application of BhGLM to the whole Clinical Exome

Subsequently, we enforced the BhGLM to the whole set of genes belonging to the Clinical Exome, in order to further enhance the statistically BC-associated RVs and choose the RVs to combine in the construction of the RVScore. Again, the BhGLM retrieved RVs were evaluated by using ACMG and ClinVar annotation and by comparing the OR for the presence of the RVs selected with the OR for the presence of RVs not picked by the model across BC-PRS classes.

Among the 34121 RVs in the Clinical Exome present in multiple individuals in the UK Biobank subsample used for the association analysis, the resampling strategy led to the selection of 1657 RVs across 1248 genes. These variants, along with age and the first 10 principal components were included as covariates in the final implementation of the BhGLM. This analysis initially identified 1016 RVs with $p-value \leq 0.05$, which was reduced to 550 RVs after correcting for multiple testing. These 550 RVs were distributed across 494 genes, with only 9% of these genes harboring more than one selected RV. Specifically, 39 genes contained two RVs, four had three, one (CHEK2) had four, and one (BRCA2) had seven. BRCA1 and PALB2 each contributed two RVs, while ATM had only one. Notably, 50% of these 550 RVs exhibited a minor allele frequency below 6.81e-05, with an interquartile range of 0.00013 (Table 5.3).

N. RVs with $p_{adj} \leq 0.05$	Mean MAF [95%CI]	Median MAF (IQR)
550	0.00026 [0.00021, 0.00032]	6.81e-05 (0.0001315)

Table 5.3: MAF of the selected RVs

We annotated these variant using ACMG and ClinVar annotations as we did before for the controlled setting. The majority of the identified rare variants were classified as Variants of Uncertain Significance (40.2%, 221 RVs), while 24.74% (136

RVs) were annotated as Pathogenic. A smaller proportion of RVs were categorized as Likely Pathogenic (1.09%, 6 RVs), Benign (0.5%, 3 RVs), or Likely Benign (1.27%, 7 RVs). For 177 (32.1%) RVs the annotation wasn't available (5.13). Notably, 79.41% of the annotated Pathogenic RVs were associated with a positive effect size ($\beta > 0$), indicating their potential deleterious impact on BC. Similarly, all variants classified as Likely Pathogenic exhibited positive coefficients, further supporting their potential contribution to increased risk or severity. In contrast, all Benign variants were associated with a negative effect size ($\beta < 0$), suggesting a protective or neutral influence. Among the Likely Benign variants, the majority (5 out of 7) also exhibited negative effect sizes, aligning with their classification as less likely to contribute to adverse phenotypic outcomes (Fugure 5.14).

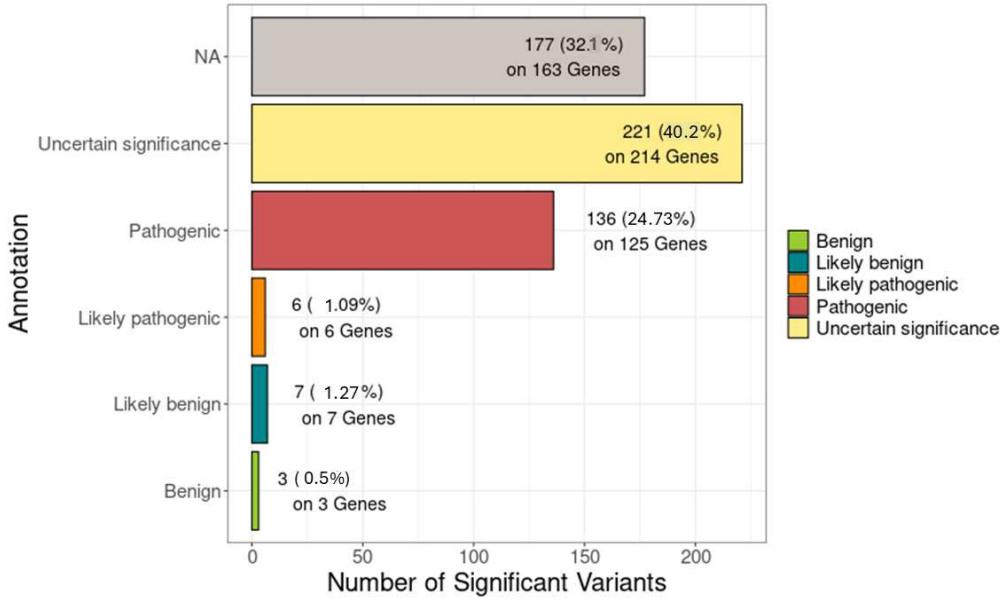


Figure 5.13: Annotation of BhGLM ClinicalExome selected RVs

Examining the ORs for selected rare variants across PRS classes reveals their significant contribution to amplifying risk. Notably, the inclusion of selected RVs results in the largest relative increase in the Intermediate PRS class, where the OR

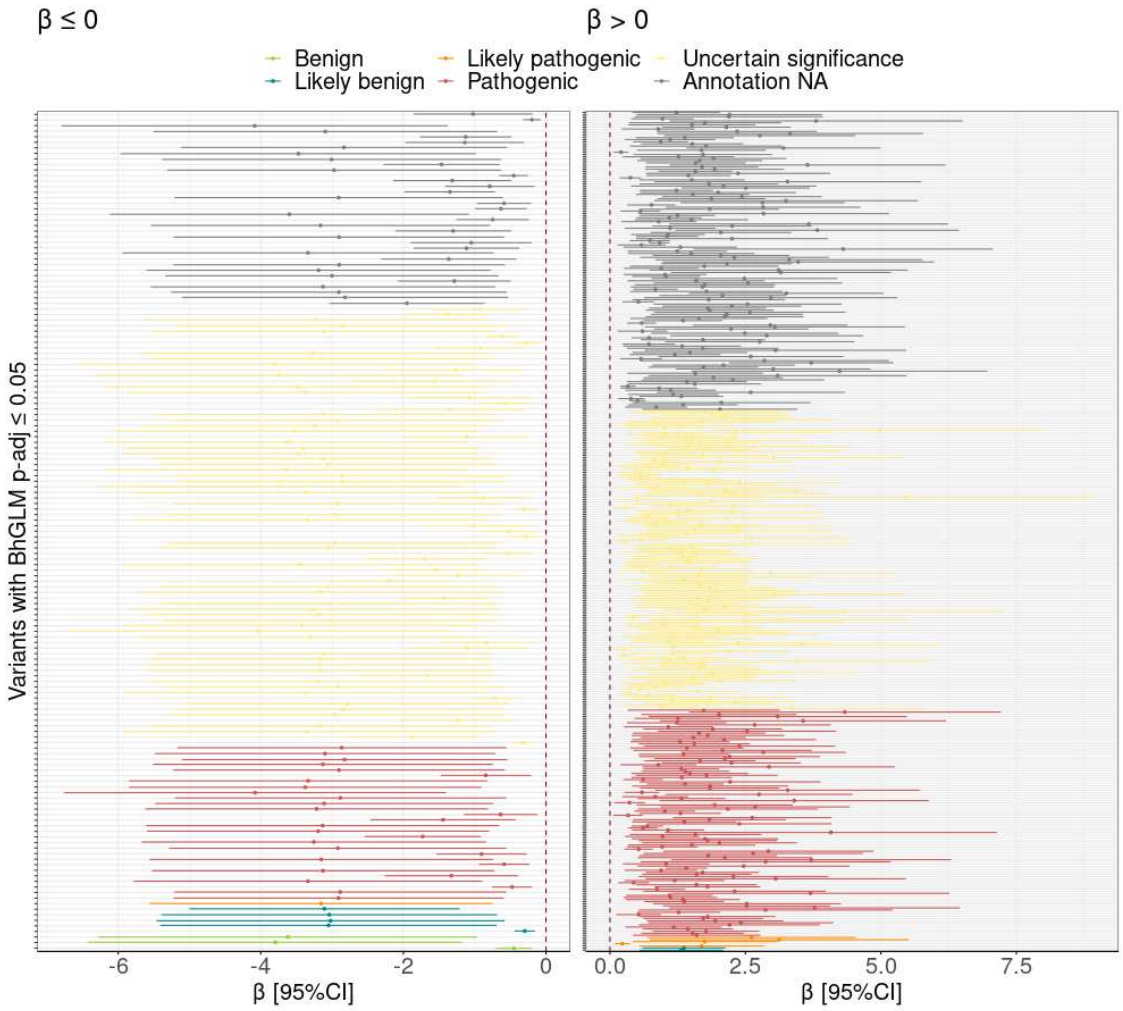


Figure 5.14: Estimated BhGLM beta coefficient by annotation

shifts from the baseline of 1.00 (95% CI: 0.98–1.02) to 1.31 (95% CI: 1.28–1.35). A similar pattern is observed in the High PRS group, with the OR increasing from 1.97 (95% CI: 1.94–2.00) to 2.59 (95% CI: 2.53–2.66). In contrast, unselected RVs exhibit a minimal impact on risk across all PRS classes, with ORs slightly lower than those of PRS alone. This suggests that unselected RVs likely include a mix of benign and less impactful variants, which dilute their overall contribution to disease risk (Table 5.4 and 5.15.

The observed substantial effect of selected RVs on risk amplification highlights

PRS_classes	covariate	OR (95%CI)	P_Value	n_samples
Low	No Rv	0.52 (0.51-0.53)	< 0.001	997
Intermediate	No Rv	1 (0.98-1.02)	1	931
High	No Rv	1.97 (1.94-2)	< 0.001	947
Low	selected RVs	0.68 (0.66-0.7)	< 0.001	11144
Intermediate	selected RVs	1.31 (1.28-1.35)	< 0.001	11207
High	selected RVs	2.59 (2.53-2.66)	< 0.001	11218
Low	unselected RVs	0.49 (0.48-0.5)	< 0.001	44717
Intermediate	unselected RVs	0.94 (0.92-0.95)	< 0.001	44770
High	unselected RVs	1.84 (1.81-1.87)	< 0.001	44747

Table 5.4: OR of BhGLM ClinicalExome selected RVs by PRS classes

the efficacy of the BhGLM in identifying RVs that are both statistically and biologically relevant. The ability of the selected RVs in refining risk stratification across PRS Classes, underscores their quality and their alignment with the underlying disease architecture. We used these RVs to construct the RVScore, which evaluation will be described in the next section.

5.3.4 BhGLM RVs-based score and PRS

The last step of these study was building on BhGLM results to develop a combined score able to capture the collective impact of RVs on the risk of BC, while preserving the ability to interpret the individual contribution of each RV to BC susceptibility. To evaluate the association between RVScore classes with BC across differnt levels of PRS, we computed the OR using logistic regression as explained in section 5.2.4. We conducted ORs analyses on both the Training and Test datasets to evaluate the scalability of our score. We acknowledged the same limitation about the Test set

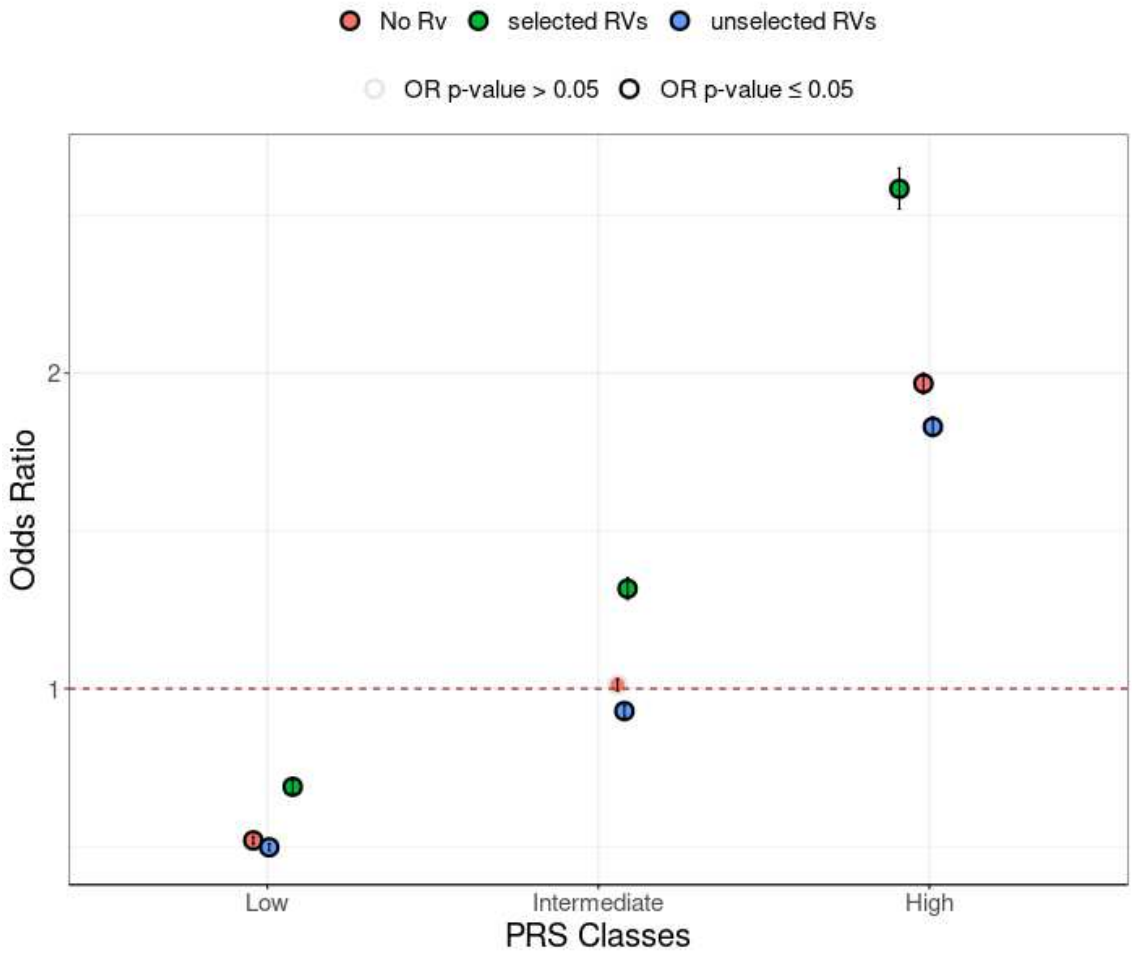


Figure 5.15: OR of BhGLM ClinicalExome selected RVs by PRS classes

sample size pointed out in the evaluation of the previously developed RVScore.

The RVScore was developed using the 550 RVs identified through the application of the BhGLM. Compared to the RVScore constructed based on Burden Test results, the inclusion of a larger number of RVs resulted in approximately 24% of individuals having an RVScore $\neq 0$ in both the Training and Test datasets. On average, the RVScore was comparable between the two datasets, with both showing a median and interquartile range (IQR) of 0 when considering the subjects with null RVSCore (Table 5.5). These balanced distribution between the Training and the Test set is preserved also removing those individuals for which the RVScore was equal to 0

(Table 5.6). The similar distribution of RVScore values across the Training and Test sets enhances the reliability of the model and facilitates the interpretation of our findings.

Set	Mean [95%CI]	N. subjects	N. subjects with RVScore $\neq 0$ (%)
Test Set	0.10 [-3.78, 4]	45131	10764 (23.85%)
Training Set	0.11 [-3.76, 3.98]	135804	32691 (24.07%)

Table 5.5: RVScore distribution including patients with null RVScore

Set	Mean [95%CI]	Median (IQR)
Test Set	0.45 [0.37, 0.53]	2.43 (6.46)
Training Set	0.47 [0.43, 0.51]	2.47 (6.46)

Table 5.6: RVScore distribution excluding patients with null RVScore

When looking at the differences distributions of the RVScore between Cases and Controls in both the datasets, it can be observed that the proportion of subjects with RVScore $\neq 0$ was the same as in the overall population for both cases and controls in the Test Set. In the Training set, while the Cases showed a similar percentage of subjects with null RVScore of the one of the overall population, the Controls reported an higher fraction of individuals with RVScore $\neq 0$ (Table 5.7). The median and the IQR where still equal to zero in both the datasets and independently from the considered population. On average, the RVScore was significantly elevated among cases compared to controls across both datasets, with the difference being more pronounced in the Training set (Table 4.5). This trend persisted even after excluding individuals with $RVScore = 0$. Specifically, in the Training set, cases exhibited a median RVScore of 2.99 (IQR: 1.44), markedly higher than the median of 0.8 (IQR: 6.35) observed in controls ($Wilcoxon; p < 0.001$). The notably wider interquartile range in the control group suggests greater variability in the RVScore distribution,

likely indicative of a heterogeneous mix of protective and neutral variants diluting the overall score. Similarly, in the Test set, the median RVScore remained higher among cases (2.52, IQR: 4.63) compared to controls (1.42, IQR: 6.46, $Wilcoxon; p < 0.001$) (Table 4.6). These preliminary observations suggest that elevated RVScores may reflect an increased burden of risk-associated rare variants in cases, whereas lower or negative RVScores are more likely to be associated with protective or neutral variants that are more represented in controls.

Set	BC Satus	Mean [95%CI]	N. subjects	N. subjects with RVScore $\neq 0$ (%)	T-test p-value (H1:>)
Test Set	Cases	0.22 [0.15, 0.28]	3957	965 (24.38%)	< 0.001
Test Set	Controls	0.1 [0.08, 0.12]	41174	9799 (23.79%)	
Training Set	Cases	0.67 [0.63, 0.71]	11911	3450 (34.52%)	< 0.001
Training Set	Controls	0.06 [0.05, 0.07]	123893	2924 (23.60%)	

Table 5.7: BhGLM RVScore distribution in cases and controls, including patients with null RVScore

Set	BC Satus	Mean [95%CI]	Median (IQR)	Wilcoxon test p-value (H1:>)	T-test p-value (H1:>)
Test Set	Cases	0.89 [0.61, 1.16]	2.52 (4.63)	≤ 0.001	< 0.001
Test Set	Controls	0.41 [0.33, 0.49]	1.42 (6.46)		
Training Set	Cases	2.31 [2.19, 2.43]	2.99 (1.44)	≤ 0.001	< 0.001
Training Set	Controls	0.25 [0.21, 0.3]	0.8 (6.35)		

Table 5.8: BhGLM RVScore distribution in cases and controls, excluding patients with null RVScore

On both the datasets, High values of RVScore were associated with an higher risk of developing BC indipendently across all the PRS classes. In particular on the Training set the OR of the High RVScore class (OR = 1.99, 95%CI:(1.9-2.07)) at intermediate level of PRS was comparabele to High values of PRS alone (OR = 1.97, 95%CI:(1.94-2)) (Table 5.9 and Figure 5.16). While on the Training Set Low levels of RVScore were able to provide a finer stratification of the population in terms of BC risk if compared to just the PRS, on the Test set the ORs of the Low RVScore group were comparable to the PRS performance across all classes, highlight the need

for further replication in independent, well-powered cohorts.

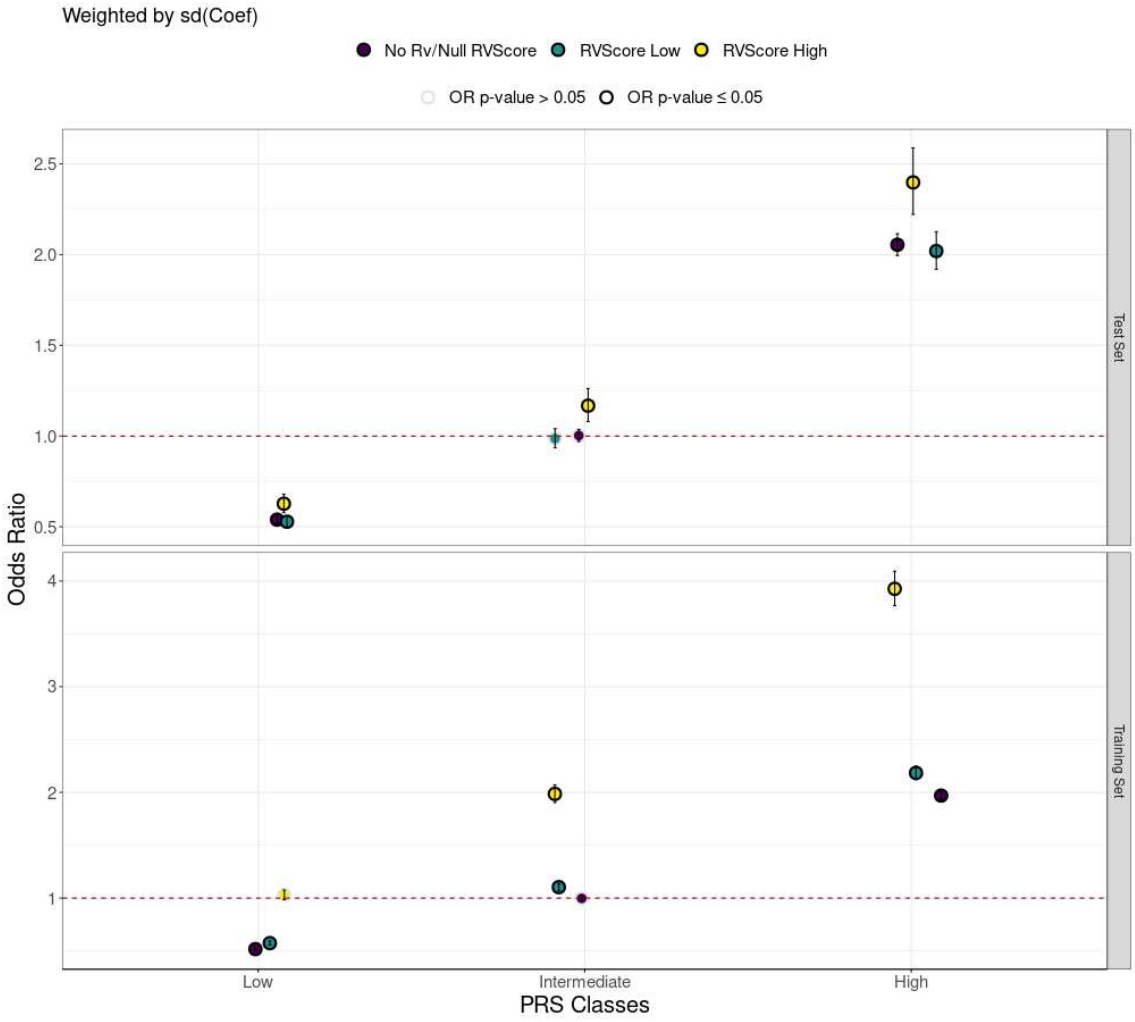


Figure 5.16: BhGL RVScore OR across PRS classes

In conclusion, laverging the BhGLM results, we built a RVScore that effectively captured the burden of risk-associated RVs, with higher scores consistently observed in cases compared to controls across both Training and Test datasets. Moreover, while capturing the collective impact of RVs on the risk of BC, this defined score preserved the ability to dissect and interpret the individual contribution of each RV to BC susceptibility. Nonetheless, the less robust replication of some findings in the Test, emphasizing the need of further studies to refine the score, improve its

PRS_classes	covariate	OR (95%CI)	P_Value	n_samples
Low	No Rv/Null RVScore	0.52 (0.51-0.53)	< 0.001	34349
High	No Rv/Null RVScore	1.97 (1.94-2)	< 0.001	34281
Intermediate	No Rv/Null RVScore	1 (0.98-1.02)	1	34274
Low	RVScore Low	0.57 (0.55-0.59)	< 0.001	7998
High	RVScore Low	2.18 (2.12-2.25)	< 0.001	7822
Intermediate	RVScore Low	1.1 (1.07-1.14)	< 0.001	7944
Low	RVScore High	1.03 (0.98-1.07)	0.21	2829
High	RVScore High	3.93 (3.77-4.1)	< 0.001	3073
Intermediate	RVScore High	1.99 (1.9-2.07)	< 0.001	2958

Table 5.9: BhGL RVScore OR across PRS classes, Training Set

PRS_classes	covariate	OR (95%CI)	P_Value	n_samples
High	No Rv/Null RVScore	2.06 (2-2.12)	< 0.001	11435
Low	No Rv/Null RVScore	0.54 (0.52-0.56)	< 0.001	11429
Intermediate	No Rv/Null RVScore	1 (0.97-1.03)	1	11422
High	RVScore Low	2.02 (1.92-2.12)	< 0.001	2605
Low	RVScore Low	0.53 (0.5-0.56)	< 0.001	2631
Intermediate	RVScore Low	0.98 (0.93-1.04)	0.52	2607
High	RVScore High	2.4 (2.22-2.59)	< 0.001	966
Low	RVScore High	0.63 (0.58-0.68)	< 0.001	947
Intermediate	RVScore High	1.17 (1.08-1.26)	< 0.001	977

Table 5.10: BhGL RVScore OR across PRS classes, Test Set

generalizability, and validate its utility in diverse populations.

Chapter 6

Discussion

This study analyzed the UK Biobank dataset to investigate the role of rare variants in breast cancer susceptibility, employing both classical methods and an innovative advanced statistical Bayesian model (BhGLM) to derive a comprehensive understanding of RV contributions. The Burden test assessed the cumulative association of RVs in aggregated genetic units, while the BhGLM accounted for complex relationships within the data to identify BC-associated variants. The under investigation UK Biobank cohort encompasses only European women (15868 BC cases, 165067 controls).

We applied Burden Test to different lists of genes to investigate the impact of the multiple testing burden on the detection of RVs truly related to breast cancer. Moreover, we aimed to evaluate the contribution of LoF and missense RVs to BC risk by performing our analysis combining the variants in different masks.

To assess BhGLM performances and indentify the combination of parameters maximizing the detection accuracy, we applied the model in two simulated settings and in a controlled setting. Finally the BhGLM was extended to the whole ClinicalExome including 5369 genes. Singleton RVs were excluded aiming to mitigate potential instability associated with these extremely infrequent variants. In both

the controlled setting and the ClinicalExome analysis, we evaluated the quality of the retrieved RVs using ACMG and ClinVar annotation, and by comparing their impact with the effect of unselected RVs using OR across different PRS classes. Finally, the variants selected by both the Burden Tests or the BhGLM model were combined in two different RVScores to explore the cumulative impact of RVs on BC risk stratification in combination with PRS. We assessed our findings using OR on a Test set cohort (25% of the full set) excluded from the previuos BC-RVs association analysis.

Using the Burden test, we highlighted the impact of gene list selection on the ability to detect associations. Indeed, smaller curated lists, such as ClinGen, were more effective at identifying weaker, yet meaningful associations, owing to reduced multiple testing burdens. Conversely, larger lists, such as the Clinical Exome, offered a more exhaustive overview of potential contributors but were less sensitive to subtle signals. Strong associations were consistently observed in well-established BC susceptibility genes like BRCA1, BRCA2, ATM, CHEK2, and PALB2, while weaker associations emerged for genes such as BARD1, MAP3K1, PLCG1, LZTR1, POLD2, DDX1 and NDFUS4 across different RVs masks, emphasizing the need for a balanced approach to gene selection that prioritizes both sensitivity and comprehensiveness. With Burden Test and extending the investigation to ClinicalExome genes, we identified new potential candidate genes as strongly associated to BC: ASPRV1 and ADGRA3.

In simulation settings, BhGLM demonstrated high specificity levels, translating in low false-positive rate (FPR, average ≤ 0.001). On the other hand, sensitivity assumed considerably lower values, highlighting the model's conservative classification strategy, which is better suited for identifying the most robust associations. In controlled setting, BhGLM retrieved mostly RVs annotated as pathogenic that provided a much higher OR than the unselected RVs on the same genes especially at high

PRS values (OR = 7.06 (95%CI:(6.37-7.83)) vs OR = 2.09 (95%CI:(1.92-2.28))). When extended to the ClinicalExome, 550 heterogeneous LoF RVs were selected: 40.2% RVs of Uncertain Significance, 24.74% Pathogenic, 1.09% Likely Pathogenic, 0.5% Benign, 1.27% Likely Benign, while for the 32.1% of these RVs the annotation wasn't available. Notably, around the 80% of the annotated Pathogenic RVs were associated with a positive effect size. Comparing the ORs for the selected RVs with the one of unselected RVs across PRS classes reveals their significant contribution to amplifying risk especially in the Intermediate PRS class, where the OR shifts from the baseline of 1.00 (95% CI: 0.98–1.02) to 1.31 (95% CI: 1.28–1.35). In contrast, unselected RVs exhibit a minimal impact on risk across all PRS classes, with ORs slightly lower than those of PRS alone.

The comparison between the Burden test and BhGLM revealed notable differences in the number and characteristics of selected variants, shaped by the methodological approaches and by the exclusion of singleton RVs in the BhGLM analysis. BhGLM identified a larger set of variants than the Burden Test, likely due to its ability to capture complex relationships within the data. Furthermore, the variants selected by BhGLM exhibited a higher median MAF compared to those identified by the Burden test. Specifically, BhGLM favored variants with a median MAF of 4.3×10^{-6} (IQR: $1.06 \times 10{-}5$), while the Burden test selected variants with a median MAF of 3.2×10^{-6} (IQR: 7.2×10^{-6}). This difference reflects the exclusion of RVs carried by just one patient in the BhGLM analysis, focusing instead on variants with slightly higher frequencies that may contribute more consistently to BC risk. By excluding these singleton RVs in the BhGLM analysis, we aimed to mitigate potential biases and instability associated with these variants, which may lack statistical power in smaller datasets. The inclusion of moderately rare variants likely contributed to the improved scalability and robustness of the BhGLM-based RVScore, as evidenced by its consistent distributions across training and test sets.

In contrast, the Burden test's inclusion of ultra-rare variants, while potentially capturing high-impact signals, may have introduced greater variability and reduced generalizability.

When focusing on Burden Test RVScore built combining LoF and missense RVs, we noticed that elevated Burden Test RVScore levels were associated with higher ORs across PRS classes compared to the presence of RVs in high- and moderate-risk genes alone.

High levels of the BhGLM RVScore increased the risk of BC across all the PRS classes. The maximum increment was observed at intermediate class of PRS, where the high RVScore reached an OR comparable to the one of the high PRS category alone ((OR = 1.99, 95%CI:(1.9-2.07) vs (OR = 1.97, 95%CI:(1.94-2)). Furthermore, this defined RVscore, enable us to interpret the contributions of each RV selected by the association analysis methods. When comparing the RVScores derived from the Burden test and BhGLM, notable differences emerged. The Burden-derived RVScore, built on fewer variants, showed variability between the training and test sets, reflecting limited scalability to smaller populations. In contrast, the BhGLM-based RVScore exhibited similar distributions in training and test sets, underlining a major ability to generalization.

Overall, high levels of both the RVScores were associated to an increase risk of BC with respect to PRS alone with an higher ORs values observed for higher levels of PRS. Specifically, on the Test set in the High PRS class the OR of high Burden Test RVScore was 6.59 ((95% CI:5.15-8.44)) for M1 and 5.06 (95% CI:(4.23-6.04)) for M7, while the one of the BhGLM RVScore was 2.40 ((95% CI:(2.22-2.59)) against the one equal to 2.06 (95% CI:(2-2.12)) reached by the PRS alone. The higher ORs reached by the RVScore built on Burden Tests selection may reflect the inclusion of ultra-rare variants that potentially captured high-impact signals at the cost of generalisability. Nevertheless, the loose of the discriminatory power at lower

BhGLM RVScore levels in the test set suggest that the score's ability to effectively provide finer stratification of cancer risk at low values can be influenced by sample size. These findings emphasize the need for larger validation cohorts to confirm the utility of the RVScore in broader contexts.

Several studies have explored the association between rare variants and BC risk using classical association tests, including the Burden Test, SKAT, and SKAT-O. Notably, studies using the AMBER Consortium, BCAC, GC-HBOC, and large meta-analyses combining BRIDGES, PERSPECTIVE, and UK Biobank have confirmed strong associations for LoF variants in ATM, BRCA1, BRCA2, CHEK2, PALB2, and MAP3K1. Additional associations were observed for LZTR1, ATR, and BARD1 [13][4][10][84]. Beyond gene-level associations, integrating RVs with PRS has been a focus of risk prediction models. The 2022 study of Hassanin E et al [14], using UK Biobank data, assessed interactions between PRS and RVs, finding that PRS modulates RV penetrance, especially for moderate-risk genes. The BOADICEA model [3] combined RVs in high- and moderate-risk genes with a PRS of 313 SNPs, incorporating non-genetic risk factors for improved stratification. Finally, the 2024 study by Schwarzerova et al. [59] further explored the integration of PRS and rare pathogenic variants in genes clusterized in high-risk (BRCA1, BRCA2) and moderate-risk (ATM, CHEK2, PALB2) to enhance BC risk stratification. Despite the importance of these studies in providing convincing evidence for the integration of both PRS and rare pathogenic variants into comprehensive BC risk prediction models, most of them have examined the association between PRS and RVs using predefined lists of genes or applying models that are not readily adaptable to incorporate new genetic discoveries (see BOADICEA [3]). Even when more flexible and systematic gene selection approaches are employed, such as in [62], the practical implementation of methods that combine PRS with RVs becomes progressively

complex with the inclusion of additional genes. Finally, all of these works aggregate RVs into genetic units without offering insights into the role of specific variants. In contrast, this study presents a systematic methodological framework for developing a comprehensive risk score that integrates rare variants with PRS to predict breast cancer occurrence. It employs classical gene-based association tests alongside an advanced analytical approach that allows to consistently assess the impact of RVs on BC risk while enabling the interpretation of associations at the single-variant level. In conclusion, our work underscores the significant contribution of rare variants on BC risk and the utility of combining their collective effects into a unified score. The proposed approaches enhanced patient stratification, empowering a more nuanced understanding of genetic influence to disease. Moreover the BhGLM, not only introduces a novel methodological setting not previously explored in the context of BC studies, but allows also the quantification of the collective impact of RVs on BC risk while preserving the capacity to interpret the contribution of individual variants. This dual capability addresses a critical limitation in current approaches to RV analysis. Despite these advances, this work has limitations that warrant consideration. The rarity of the variants underlying the RVScore poses challenges for its application to individual patients or small cohorts, limiting its extendibility and emphasizing the necessity for further validation in a larger and more adequately powered cohort than the Test set utilized in this study. Moreover, future research could explore for novel measures to aggregate the impact of functionally related variants, while preserving the capability to analyze the impact of the single RV, offering a potentially more scalable measure of genetic risk. Additionally, the analysis was conducted exclusively on individuals of European ancestry. On top of that, the UK Biobank cohort is characterized by limited genetic diversity [61]. This constraint further emphasize the need to validate the RVScore in external cohorts, particularly those representing diverse ancestries, to ensure its scalability and evaluate its

applicability across populations. Expanding the scope of the study to include non-European individuals would not only improve the generalizability of the findings but also address disparities in genetic research.

These limitations notwithstanding, the present study highlights the critical role of RVs in BC risk and lays the groundwork for refining genetic risk models that integrate rare and common variants to enhance predictive accuracy.

Appendix A

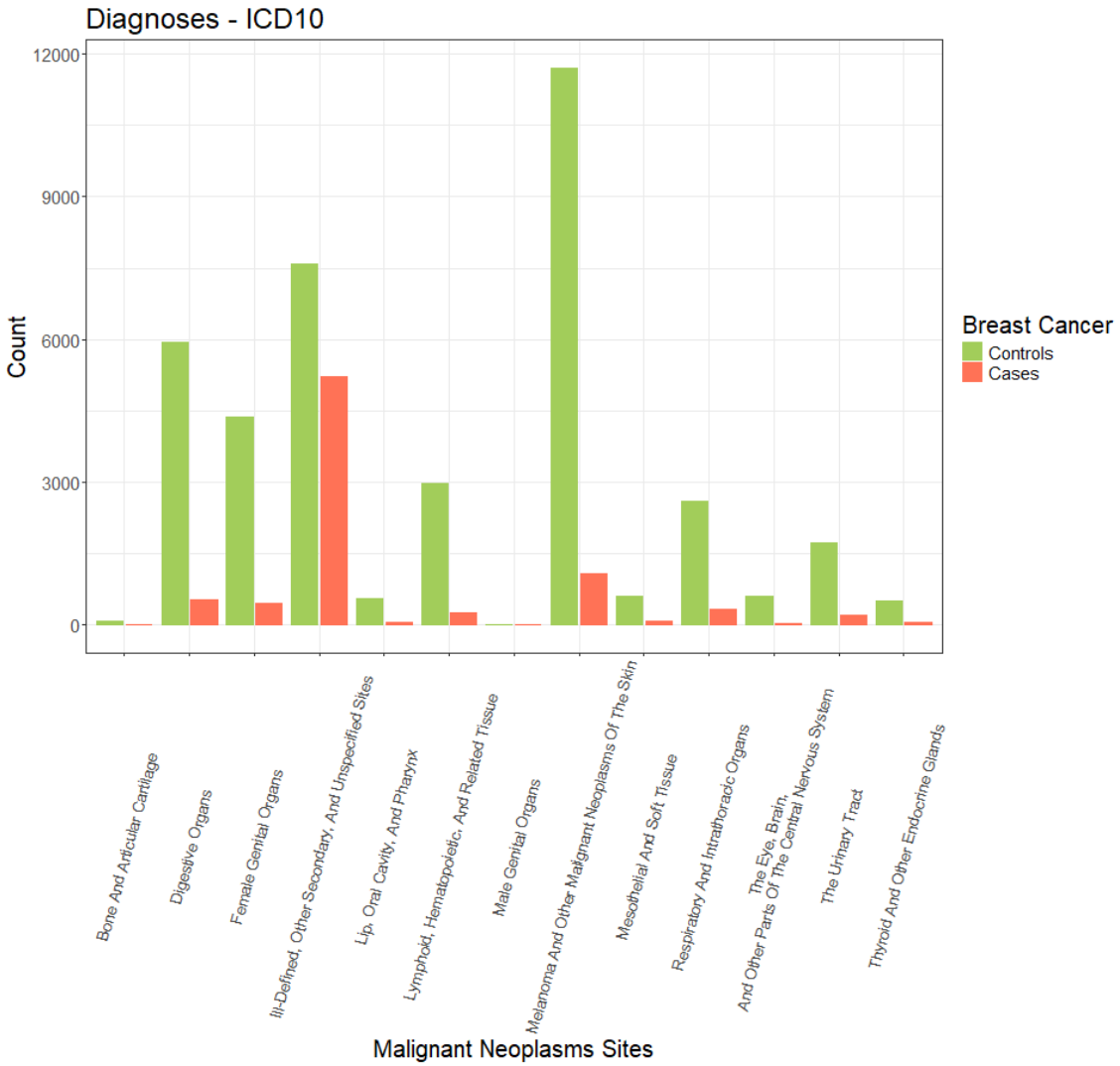


Figure A.1: Number of cases and controls affected by other cancer types

Table A.1: Random Forest Performance Metrics

Metric	**Value**
Accuracy	0.9933333
Kappa	0.9371855
F1	0.8517160
Sensitivity	0.8642169
Specificity	0.9949534
Positive Predictive Value	0.8568909
Negative Predictive Value	0.9949483
Precision	0.8568909
Recall	0.8642169
Detection Rate	0.1419048
Balanced Accuracy	0.9295852

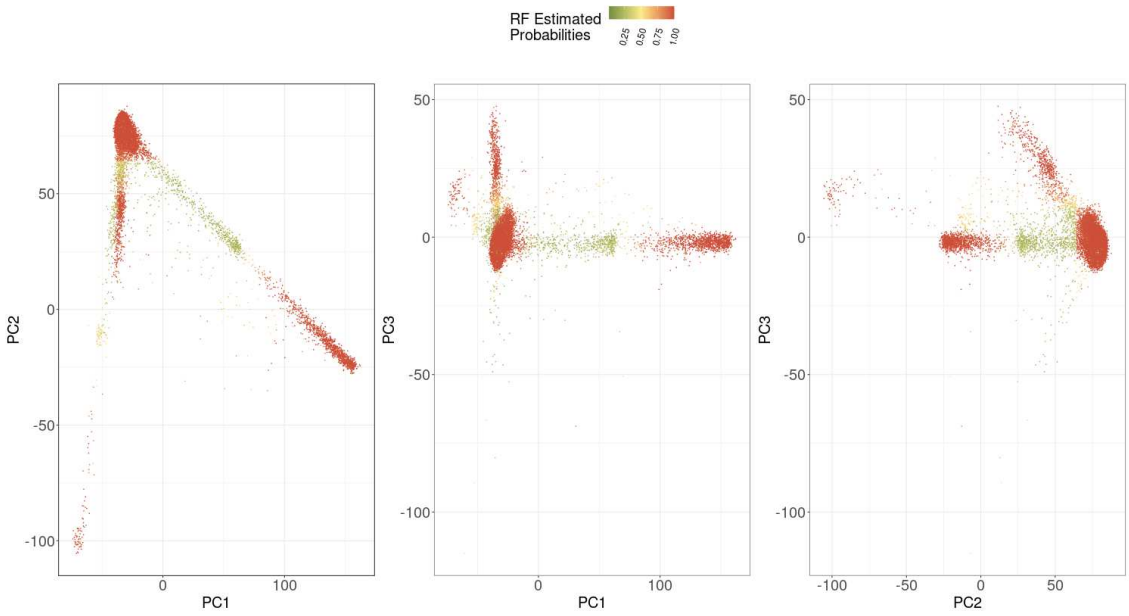


Figure A.2: RF classes estimate probabilities

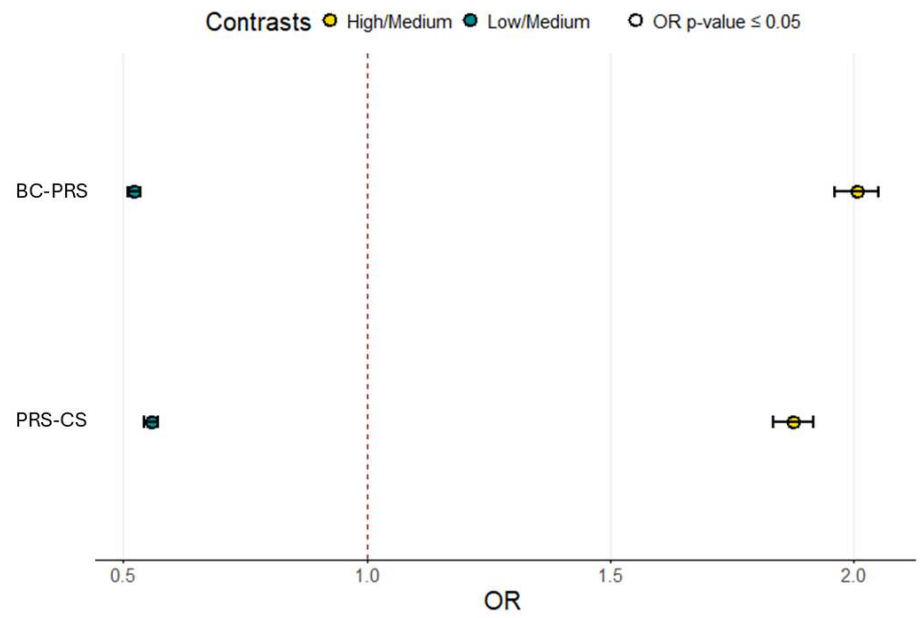


Figure A.3: OR UKBB BC-PRS and PRS-CS by classes

PRS	**PRS Classes Contrasts**	**Number of Subjects Per Classes**	**OR (95% CI)***
PRS-CS	Low/Medium	60205/60205	0.56 (0.54–0.57)*
	High/Medium	60205/60205	1.88 (1.84–1.91)*
UKBB BC-PRS	Low/Medium	60183/60182	0.52 (0.51–0.53)*
	High/Medium	60182/60182	2.00 (1.96–2.05)*

* adjusted with age, OR *p-value* ≤ 0.001. OR computed on the overall population (Training and Test Sets).

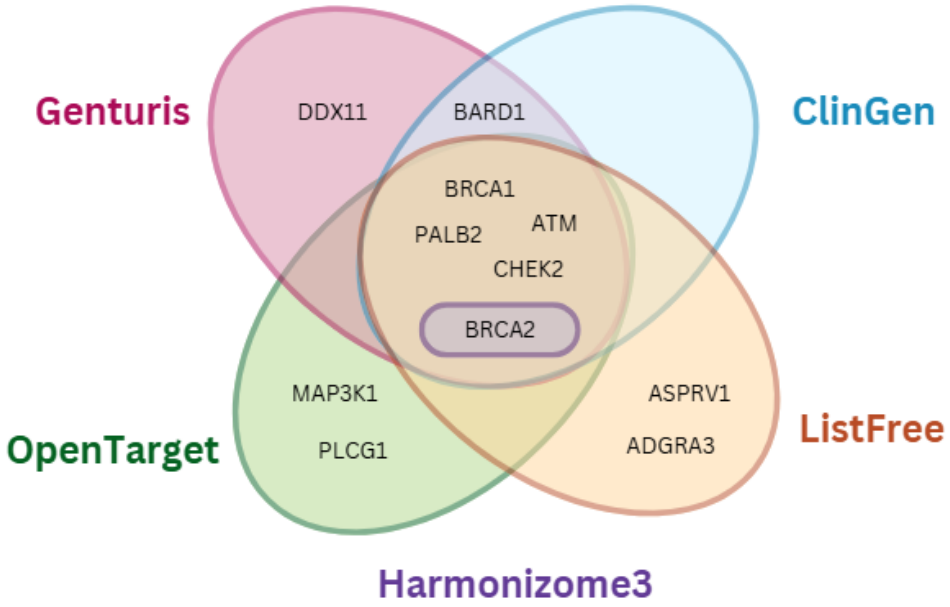


Figure A.4: Genes with Burden association test $p_{adj} \leq 0.05$ for the presence of LoF RVs (M1) with MAF ≤ 0.01 among the different gene lists. Note that in Harmonizome, among the genes considered in the picture, was present just BRCA2.

Bibliography

[1] Abu-El-Haija A et al. "The clinical application of polygenic risk scores: A points to consider statement of the American College of Medical Genetics and Genomics (ACMG)". In: *Genetics in Medicine* (2023). DOI: 10.1016/j.gim.2023.100803.

[2] Franke A et al. "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci". In: *Nat. Genet* (2010). DOI: 10.1038/ng.717.

[3] Lee A et al. "BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors." In: *Genet Med* (2019). DOI: 10.1038/s41436-018-0406-9.

[4] Breast Cancer Association Consortium et al. "Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women." In: *The New England journal of medicine* (2021). DOI: 10.1056/NEJMoa1913948.

[5] Daly MB et al. "NCCN Guidelines® Insights: Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic, Version 2.2024". In: *J Natl Compr Canc Netw* (2023). DOI: 10.6004/jnccn.2023.0051.

[6] Lakeman IMM et al. "Clinical applicability of the Polygenic Risk Score for breast cancer risk prediction in familial cases". In: *J Med Genet* (2023). DOI: 10.1136/jmg-2022-108502.

[7] Li MJ et al. "GWASdb: a database for human genetic variants identified by genome-wide association studies". In: *Nucleic Acids Res* (2012). DOI: 10.1093/nar/gkr1182.

[8] Mavaddat N et al. "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes". In: *Am J Hum Genet* (2019). DOI: 10.1016/j.ajhg.2018.11.002.

[9] Bycroft et al. "The UK Biobank resource with deep phenotyping and genomic data." In: *Nature* (2018). DOI: https://doi.org/10.1038/s41586-018-0579-z.

[10] Dumont M et al. "Uncovering the Contribution of Moderate-Penetrance Susceptibility Genes to Breast Cancer by Whole-Exome Sequencing and Targeted Enrichment Sequencing of Candidate Genes in Women of European Ancestry." In: *Cancers* (2022). DOI: 10.3390/cancers14143363.

[11] Easton DF et al. "Genome-wide association study identifies novel breast cancer susceptibility loci." In: *Nature* (2007). DOI: 10.1038/nature05887.

[12] Fachal L et al. "Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes." In: *Nat Genet* (2020). DOI: 10.1038/s41588-019-0537-1.

[13] Haddad SA et al. "An exome-wide analysis of low frequency and rare variants in relation to risk of breast cancer in African American Women: the AMBER Consortium." In: *Carcinogenesis* (2016). DOI: 10.1093/carcin/bgw067.

[14] Hassanin E et al. "Breast and prostate cancer risk: The interplay of polygenic risk, rare pathogenic germline variants, and family history." In: *Genetics in medicine* (2022). DOI: 10.1016/j.gim.2021.11.009.

[15] Karczewski Konrad J et al. "Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes". In: *Cell genomics* (2022). DOI: doi:10.1016/j.xgen.2022.100168.

[16] Mavaddat N et al. "Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes." In: *American journal of human genetics* (2019). DOI: 10.1016/j.ajhg.2018.11.002.

[17] Stacey SN et al. "Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer." In: *Nat Genet* (2007). DOI: 10.1038/ng2064.

[18] Morris AP et al. "Wellcome Trust Case Control Consortium. Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators. Genetic Investigation of ANthropometric Traits (GIANT) Consortium. Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium. South Asian Type 2 Diabetes (SAT2D) Consortium. DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes." In: *Nat. Genet.* (2012). DOI: 10.1038/ng.2383.

[19] Jespersen AS et al. Bagger FO Borgwardt L. "Whole genome sequencing in clinical practice." In: *BMC Med Genomics.* (2024). DOI: https://doi.org/10.1186/s12920-024-01795-w.

[20] L. Biesecker. "Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project." In: *Genet. Med.* (2012). DOI: https://doi.org/10.1038/gim.2011.78.

[21] Conner BJ et al. "Detection of sickle cell beta S-globin allele by hybridization with synthetic oligonucleotides." In: *Proc Natl Acad Sci U S A.* (1983). DOI: 10.1073/pnas.80.1.278.

[22] Clare Bycroft et al. "Genome-wide genetic data on 500,000 UK Biobank participants." In: *Nature* (2017). DOI: https://doi.org/10.1101/166298.

[23] Kimchi-Sarfaty C et al. "Silent polymorphism in the MDR1 gene changes substrate specificity". In: *Science.* (2007). DOI: 10.1126/science.1135308.

[24] Christopher CC et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *GigaScience* (2015). DOI: https://doi.org/10.1186/s13742-015-0047-8.

[25] P. Cingolani et al. "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". In: *Fly* 6.2 (2012), pp. 80–92.

[26] *ClinGen Genes List.* 2024. URL: https://search.clinicalgenome.org/kb/downloads#section_gene-disease-validity.

[27] Lewis CM and Vassos E. "Polygenic risk scores: from research tools to clinical instruments." In: *Genome Med* (2020). DOI: https://doi.org/10.1186/s13073-020-00742-5.

[28] The Wellcome Trust Case Control Consortium. "Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." In: *Nature* (2007). DOI: 10.1038/nature05911.

[29] Wang D.G. et al. "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome". In: *Science.* (1998). DOI: 10.1126/science.280.5366.1077.

[30] Wheeler DA et al. "The complete genome of an individual by massively parallel DNA sequencing." In: *Nature* (2008). DOI: 10.1038/nature06884.

[31] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *GigaScience* (2021). DOI: 10.1093/gigascience/giab008. URL: https://doi.org/10.1093/gigascience/giab008.

[32] Firth David. "Bias reduction of maximum likelihood estimates". In: *Biometrika* (1993). DOI: https://doi.org/10.1093/biomet/80.1.27.

[33] Dingsheng Deng. "DBSCAN Clustering Algorithm Based on Density". In: *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*. 2020, pp. 949–953. DOI: 10.1109/IFEEA51475.2020.00199.

[34] Liu DJ and Suzanne ML. "A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions". In: *PLoS genetics* (2010). DOI: 10.1371/journal.pgen.1001156.

[35] Thompson DJ et al. "A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release." In: *PLoS One* (2024). DOI: 10.1371/journal.pone.0307270.

[36] Weiner DJ, Nadig A, and Jagadeesh KA et al. "Polygenic architecture of rare coding variation across 394,783 exomes". In: *Nature* (2023). DOI: https://doi.org/10.1038/s41586-022-05684-z.

[37] Roberts E, Howell S, and Evans DG. "Polygenic risk scores and breast cancer risk prediction". In: *Breast* (2023). DOI: 10.1016/j.breast.2023.01.003.

[38] Uffelmann E et al. "Genome-wide association studies". In: *Nat Rev Methods Primers* (2021). DOI: https://doi.org/10.1038/s43586-021-00056-9.

[39] Douglas F et al. Easton. "Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk". In: *The New England journal of medicine* (2015). DOI: 10.1056/NEJMsr1501341.

[40] Clara Esteban-Jurado et al. "Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer". In: *Genetics in Medicine* (2015). DOI: https://doi.org/10.1038/gim.2014.89.

[41] Douglas FE and Rosalind AE. "Genome-wide association studies in cancer." In: *Human Molecular Genetics* (2008). DOI: https://doi.org/10.1093/hmg/ddn287.

[42] Ji G et al. "Common variants in mismatch repair genes associated with increased risk of sperm DNA damage and male infertility". In: *BMC Med.* (2012). DOI: 10.1186/1741-7015-10-49.

[43] Sato G, Shirai Y, and Namba S et al. "Pan-cancer and cross-population genome-wide association studies dissect shared genetic backgrounds underlying carcinogenesis." In: *Nat Commun* (2023). DOI: https://doi.org/10.1038/s41467-023-39136-7.

[44] Sato G, Shirai Y, and Namba S et al. "Pan-cancer and cross-population genome-wide association studies dissect shared genetic backgrounds underlying carcinogenesis." In: *Nat Commun* (2023). DOI: https://doi.org/10.1038/s41467-023-39136-7.

[45] *Genturis Genes List*. 2024. URL: https://www.genturis.eu/l=eng/home.html.

[46] Da Costa Nunes GG et al. "Genomic Variants and Worldwide Epidemiology of Breast Cancer: A Genome-Wide Association Studies Correlation Analysis". In: *Genes (Basel)* (2024). DOI: 10.3390/genes15020145.

[47] Beltran H, Eng K, and Mosquera JM et al. "Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response." In: *JAMA Oncol* (2015). DOI: 10.1001/jamaoncol.2015.1313.

[48] Chen H, Fan S, and Stone J et al. "Genome-wide and transcriptome-wide association studies of mammographic density phenotypes reveal novel loci." In: *Breast Cancer Res* (2022). DOI: https://doi.org/10.1186/s13058-022-01524-0.

[49] Mallick H and Yi N. "Hierarchical Models for Genetic Association Studies". In: *J Biomet Biostat* (2013). DOI: 10.4172/2155-6180.1000e124.

[50] Zhang H, Ahearn T.U, and Lecarpentier J et al. "Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses". In: *Nat Genet* (2020). DOI: https://doi.org/10.1038/s41588-020-0609-2.

[51] Kockum I, Huang J, and Stridh P. "Overview of Genotyping Technologies and Methods." In: *Protoc.* (2023). DOI: 10.1002/cpz1.727..

[52] Illumina. *A high-resolution view of the entire genome.* URL: https://emea.illumina.com/techniques/sequencing/dna-sequencing/whole-genome-sequencing.html.

[53] Illumina. *NGS workflow steps.* URL: https://emea.illumina.com/science/technology/next-generation-sequencing/beginners/ngs-workflow.html.

[54] Illumina. *Polygenic risk scores. Using research to better understand the heritable risk of disease.* URL: https://emea.illumina.com/areas-of-interest/complex-disease-genomics/polygenic-risk-scores.html.

[55] Broad Institute. *Picard Toolkit*. 2019. URL: https://broadinstitute.github.io/picard/.

[56] National Cancer Institute. *Cancer Stat Facts: Female Breast Cancer*. URL: https://seer.cancer.gov/statfacts/html/breast.html.

[57] MacArthur J et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." In: *Nucleic Acids Res.* (2017). DOI: 10.1093/nar/gkw1133.

[58] Mbatchou J, Barnard L, and Backman J et al. "Computationally efficient whole-genome regression for quantitative and binary traits". In: *Nat Genet* (2021). DOI: https://doi.org/10.1038/s41588-021-00870-7.

[59] Schwarzerova J et al. "A perspective on genetic and polygenic risk scores—advances and limitations and overview of associated tools." In: *Briefings in Bioinformatics* (2024). DOI: https://doi.org/10.1093/bib/bbae240.

[60] Yang J et al. "Common SNPs explain a large proportion of the heritability for human height." In: *Nat Genet* (2010). DOI: doi:10.1038/ng.608.

[61] Backman JD, Li AH, and Marcketta A. et al. "Exome sequencing and analysis of 454,787 UK Biobank participants". In: *Nature* (2021). DOI: https://doi.org/10.1038/s41586-021-04103-z.

[62] Kang JH, Lee Y, and Kim DJ et al. "Polygenic risk and rare variant gene clustering enhance cancer risk stratification for breast and prostate cancers." In: *Commun Biol* (2024). DOI: https://doi.org/10.1038/s42003-024-06995-9.

[63] Kiiski JI et al. "Exome sequencing identifies FANCM as a susceptibility gene for triple-negative breast cancer." In: *Proc Natl Acad Sci U S A.* (2014). DOI: 10.1073/pnas.1407909111.

[64] Korn JM et al. "Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs." In: *Nat Genet.* (2008). DOI: 10.1038/ng.237.

[65] Frazer KA et al. "Human genetic variation and its contribution to complex traits." In: *Nat Rev Genet.* (2009). DOI: 10.1038/nrg2554.

[66] Johansen Taber KA, Dickinson BD, and Wilson M. "The promise and challenges of next-generation genome sequencing for clinical care." In: *JAMA Intern Med.* (2014). DOI: 10.1001/jamainternmed.2013.12048.

[67] Masahiro Kanai. *PCA projection*. 2021. URL: https://github.com/covid19-hg/pca_projection/blob/master/README.md.

[68] Evans AS Kelsey JL Whittemore AS and Thompson WD. *Methods of sampling and estimation of sample size.* Methods in Observational Epidemiology, Oxford University Press, 1996.

[69] Amit V et al. Khera. "Whole-Genome Sequencing to Characterize Monogenic and Polygenic Contributions in Patients Hospitalized With Early-Onset Myocardial Infarction". In: *Circulation* (2019). DOI: 10.1161/CIRCULATIONAHA.118.035658.

[70] Karczewski KJ et al. "Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes". In: *Cell Genom* (2022). DOI: 10.1016/j.xgen.2022.100168.

[71] Olga Krasheninina et al. "Open-source mapping and variant calling for large-scale NGS data from original base-quality scores". In: *bioRxiv* (2020). DOI: https://doi.org/10.1101/2020.12.15.356360.

[72] Seunggeung Lee et al. "Rare-Variant Association Analysis: Study Designs and Statistical Tests". In: *The American Journal of Human Genetics* (2014). DOI: 10.1016/j.ajhg.2014.06.009.

[73] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[74] Esai Selvan M et al. "Germline rare deleterious variant load alters cancer risk, age of onset and tumor characteristics." In: *NPJ Precis Oncol* (2023). DOI: 10.1038/s41698-023-00354-3.

[75] Al-Haggar M et al. "A novel homozygous p.Arg527Leu LMNA mutation in two unrelated Egyptian families causes overlapping mandibuloacral dysplasia and progeria syndrome". In: *Eur J Hum Genet.* (2012). DOI: 10.1038/ejhg.2012.77.

[76] Gilbert W. Maxam AM. "A new method for sequencing DNA." In: *Proc Natl Acad Sci U S A.* (1977). DOI: 10.1073/pnas.74.2.560.

[77] Bamshad MJ et al. "Exome sequencing as a tool for Mendelian disease gene discovery." In: *Nat Rev Genet.* (2011). DOI: 10.1038/nrg3031.

[78] K. Mullis et al. "Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction." In: *Cold Spring Harb Symp Quant Biol.* (1986). DOI: https://doi.org/10.1101/sqb.1986.051.01.032.

[79] Dieci MV, Smutná V, and Scott V et al. "Whole exome sequencing of rare aggressive breast cancer histologies." In: *Breast Cancer Res* (2016). DOI: https://doi.org/10.1007/s10549-016-3718-y.

[80] Mars N, Widén E, and Kerminen S et al. "The role of polygenic risk and susceptibility genes in breast cancer over the course of life". In: *Nat Commun* (2020). DOI: https://doi.org/10.1038/s41467-020-19966-5.

[81] Mars N, Widén E, and Kerminen S et al. "The role of polygenic risk and susceptibility genes in breast cancer over the course of life." In: *Nat Commun* (2020). DOI: https://doi.org/10.1038/s41467-020-19966-5.

[82] Mars N, Koskela JT, and Ripatti P et al. "Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers." In: *Nat Med* (2020). DOI: https://doi.org/10.1038/s41591-020-0800-0.

[83] Wilcox N, Dumont M, and González-Neira A et al. "Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk". In: *Nat Genet* (2023). DOI: https://doi.org/10.1038/s41588-023-01466-z.

[84] Wilcox N, Dumont M, and González-Neira A et al. "Exome sequencing identifies breast cancer susceptibility genes and defines the contribution of coding variants to breast cancer risk." In: *Nat Genet* (2023). DOI: https://doi.org/10.1038/s41588-023-01466-z.

[85] Yi N et al. "BhGLM: Bayesian hierarchical GLMs and survival models, with applications to genomics and epidemiology". In: *Bioinformatics* 35.8 (2019), pp. 1419–1421. DOI: 10.1093/bioinformatics/bty803.

[86] Staaf J et al. Nik-Zainal S Davies H. "Landscape of somatic mutations in 560 breast cancer whole-genome sequences." In: *Nature.* (2016). DOI: https://doi.org/10.1038/nature17676.

[87] K. Nones et al. "Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers." In: *Annals of Oncology* (2019). DOI: https://doi.org/10.1093/annonc/mdz132.

[88] David Ochoa et al. "The next-generation Open Targets Platform: reimagined, redesigned, rebuilt". In: *Nucleic Acids Research* (2022). DOI: https://doi.org/10.1093/nar/gkac1046. URL: 10.1093/nar/gkac1046.

[89] *OpenTarget Breast Melanoma*. 2024. URL: https://platform.opentargets.org/disease/EFO_0003869/associations.

[90] Dornbos P, Koesterer R, and Ruttenburg A et al. "A combined polygenic score of 21,293 rare and 22 common variants improves diabetes diagnosis based on hemoglobin A1C levels". In: *Nat Genet* (2022). DOI: https://doi.org/10.1038/s41588-022-01200-1.

[91] Wang Q, Dhindsa RS, and Carss K et al. "Rare variant contribution to human disease in 281,104 UK Biobank exomes". In: *Nature* (2021). DOI: https://doi.org/10.1038/s41586-021-03855-y.

[92] Klein RJ et al. "Complement factor H polymorphism in age- related macular degeneration." In: *Science* (2005). DOI: 10.1126/science.1109557.

[93] De Talhouet S, Peron J, and Vuilleumier A et al. "Clinical outcome of breast cancer in carriers of BRCA1 and BRCA2 mutations according to molecular subtypes". In: *Sci Rep 10* (2020). DOI: https://doi.org/10.1038/s41598-020-63759-1.

[94] Shah S et al. "BRCA Mutations in Prostate Cancer: Assessment, Implications and Treatment Considerations". In: *International Journal of Molecular Science* (2021). DOI: 10.1093/gigascience/giab008. URL: https://doi.org/10.3390/ijms222312628.

[95] Lee S. et al. "Rare-variant association analysis: study designs and statistical tests." In: *Am J Hum Genet.* (2014). DOI: 10.1016/j.ajhg.2014.06.009.

[96] Coulson AR. Sanger F Nicklen S. "DNA sequencing with chain-terminating inhibitors." In: *Proc Natl Acad Sci U S A.* (1977). DOI: 10.1073/pnas.74.12.5463.

[97] Christopher Chang Shaun Purcell. *PLINK 2.0.* URL: www.cog-genomics.org/plink/2.0/.

[98] Kristina P. Sinaga and Miin-Shen Yang. "Unsupervised K-Means Clustering Algorithm". In: *IEEE Access* (2020). DOI: 10.1109/ACCESS.2020.2988796.

[99] Cordovado SK et al. "CFTR mutation analysis and haplotype associations in CF patients." In: *Mol Genet Metab.* (2012). DOI: 10.1016/j.ymgme.2011.10.013.

[100] Anna P. Sokolenko et al. "Identification of novel hereditary cancer genes by whole exome sequencing". In: *Cancer Letters* (2015). DOI: https://doi.org/10.1016/j.canlet.2015.09.014.

[101] Choi SW, Mak TS, and O'Reilly PF. "Tutorial: a guide to performing polygenic risk score analyses." In: *Nat Protoc* (2020). DOI: 10.1038/s41596-020-0353-1.

[102] AC. Syvänen. "Accessing genetic variation: genotyping single nucleotide polymorphisms." In: *AC Nat Rev Genet.* (2001). DOI: https://doi.org/10.1038/35103535.

[103] Ge T, Chen CY, and Ni Y et al. "Polygenic prediction via Bayesian regression and continuous shrinkage priors." In: *Nat Commun* (2019). DOI: https://doi.org/10.1038/s41467-019-09718-5.

[104] LaFramboise T. "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". In: *Nucleic Acids Res.* (2009). DOI: 10.1093/nar/gkp552.

[105] Ahearn T.U, Zhang H, and Michailidou K et al. "Common variants in breast cancer risk loci predispose to distinct tumor subtypes." In: *Breast Cancer Res* (2022). DOI: https://doi.org/10.1186/s13058-021-01484-x.

[106] Manolio TA et al. "Finding the missing heritability of complex diseases." In: *Nature.* (2009). DOI: 10.1038/nature08494.

[107] Deborah J. Thompson et al. "UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits". In: *medRxiv* (2022). DOI: 10.1101/2022.06.16.22276246.

[108] Durtschi JD. Voelkerding KV Dames SA. "Next-generation sequencing: from basic research to diagnostics." In: *Clin Chem.* (2009). DOI: 10.1373/clinchem.2008.112789.

[109] Chen W., Coombes BJ., and Larson NB. "Recent advances and challenges of rare variant association analysis in the biobank sequencing era." In: *Front Genet.* (2022). DOI: 10.3389/fgene.2022.1014947.

[110] R. B. et al. Wallace. "Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch." In: *Nucleic Acids Res.* (1979). DOI: 10.1093/nar/6.11.3543.

[111] Gong WF et al. "Single nucleotide polymorphism 8q24 rs13281615 and risk of breast cancer: meta-analysis of more than 100,000 cases." In: *PLoS One* (2013). DOI: 10.1371/journal.pone.0060108.

[112] WHO. *Breast cancer.* URL: https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

[113] Liu X et al. "dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs". In: *Genome Med* (2020). DOI: 10.1186/s13073-020-00803-9.

[114] Zhang Y et al. "Association between 8q24 (rs13281615 and rs6983267) polymorphism and breast cancer susceptibility: a meta-analysis involving 117,355 subjects". In: *Oncotarget* (2016). DOI: https://doi.org/10.18632/oncotarget.12009.

Acknowledgements

I would first of all like to thank my tutor, Romina D'Aurizio, for the great support and infinite patience shown during this period. My group mates: Elia, Valeria, Danilo, Barbara, Maurizio, Mohamad and Francesco because, despite the distance, I could not have asked for better companions for this journey. Kristina, for being a guide in the intriguing world of university bureaucracy. Federico and my parents, for the emotional support during this years.

Giulia Brunelli Firmato il 05.03.2025 17:10:28