Universiteit Leiden

Master Computer Science

FAIRification of genetic data and comparison of breast cancer and normal samples within families

Name:	Tushar Mandloi
Student ID:	s2502585
Date:	13/08/2021
Specialisation:	Bioinformatics
1st supervisor:	Dr. Katy J. Wolstencroft, PhD
2nd supervisors:	Prof. Peter Devilee, PhD

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Preface

I want to express my gratitude to my thesis supervisors, Professor Katy Wolstencroft and Peter Devilee, for their invaluable guidance and encouraging mentorship that assisted me to enrich my knowledge and pass-through this challenging yet exciting path. I would also like to express my gratitude to my family and friends who supported, and encouraged me.

Abstract

Breast cancer (BC) is one of the most common forms of cancer seen among females in Western countries. However, many breast cancer cases occur in a minority population at increased risk. Susceptibility to breast cancer is multifactorial. During the last two decades, there has been an increase in studies investigating the genetic risk factors associated with BC. In the late 1990s, a nationwide study started called HEBON with a primary goal to identify genetic variations, estimate cancer risk within families, and develop better treatment methods. However, the HEBON data is poorly connected with publicly available databases and not accessible. Therefore, for allowing a connection to existing non-HEBON datasets, such as ClinVar, LOVD, Gnomad, UniProt, GTex, TCGA, and making it accessible and machine-readable, we propose a FAIR model for FAIRification of the data. In addition, and to demonstrate the power of connecting to public databases, we analyse the association between the genetic profile of HEBON study participants and family history. We performed transcriptome-wide association analysis using the SNP arrays and found 14 genes associated with breast cancer incidence. We found that these 14 genes significantly changed expression levels in cases and played a role in the cellular immune response. Later, we wanted to understand the interaction of our findings with the known BC genes (from Wu et al. and Ferreira et al.) and the implicated processes. We performed enrichment analysis on the network for understanding the involved processes and pathways. Our results found that the genes in the network play a role in chromatin organisation, cellular localisation, hinting at involvement in DNA repair. The network was also enriched in various KEGG pathways involved in cancer. Lastly, to demonstrate the power of FAIR data, we computed polygenic risk scores (PRS) for individuals who are the carrier of BRCA1/2 mutation and non-carrier, followed by correlating the PRS and cancer risk score based on family history.

Contents

Chapter 1

Introduction

Cancer is often termed a "disease of the genes" to emphasise the significance of cataloguing and analysing mutations associated with it. The recent advancements in sequencing technology have underpinned several large-scale projects to compile genomic information related to cancer systematically [1]. For example, the Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov/) focuses on identifying links or genomic mutations associated with cancer have vast clinical applications [2].

Cancer's primary cause is somatic mutations in specific tissues that accumulate over time, although it may be favoured by genetic predisposition, i.e., Germline variants. It has been seen that both rare and common germline variants have been associated with specific diseases [1, 3]. The association of common germline variants with clinical features and disease can be studied through Genome-Wide Association Studies (GWAS). GWAS uses large cohorts of cases to measure and analyse DNA sequence variations to identify the relationship between the disease and mutations across the entire genome [1, 3]. The ultimate aim of GWAS is to predict the risk of an individual to develop the disease using the risk factors and to identify the underpinnings of disease susceptibility for developing better strategies for prevention and treatment.

Breast cancer (BC) is one of the most common forms of cancer seen among females in the Western population. However, a large proportion of breast cancer cases occurs in a minority population at an increased risk [4]. Susceptibility to breast cancer is multifactorial. Many genetic variants and reproductive, lifestyle, and hormonal factors are associated with the risk of incurring the disease [5]. Turnbull et al. represent the frequency-risk profiles for BC based on the clinical characterization by plotting the Allele frequency vs Relative risk [6]. Figure 1.1 demonstrates that the genes BRCA1 and BRCA2 have high penetrance, other BC genes have a moderate penetrance, and risk variants have low penetrance. Colditz et al. showed that an estimated 15% to 30% of cases of breast cancer are heritable. However, the genetic alterations accounting for BC are not fully defined [7]. Mutations in regions of high and moderate penetrance genes which are known with breast cancer susceptibilities, such as BRCA1, BRCA2, PALB2, ATM, and CHEK2, have been identified in 5% of breast cancer cases and about 30%–40% of cases associated with a family history of breast cancer [8, 9, 10, 11, 12]. During the last two decades, there has been an increase in studies investigating the genetic risk factors associated with BC and explaining the missing heritability. GWA studies have identified many genetic loci associated with BC, explaining up to 18% heritability and suggests that BC is a complex, polygenic disease. Studies like The Breast Cancer Association Consortium (BCAC)

investigate the risk of BC in more than 200,000 individuals and found 65 new loci associated with BC [13]. Another study by Milne et al. identified ten more variants associated with BC, specifically with ER-negative disease risk [14]. With all the studies, we can now explain approximately 50% of the familial risk of breast cancer in terms of identified genes and genomic variants. In Figure 1.2, we represent the known familial risk factors as a pie chart.

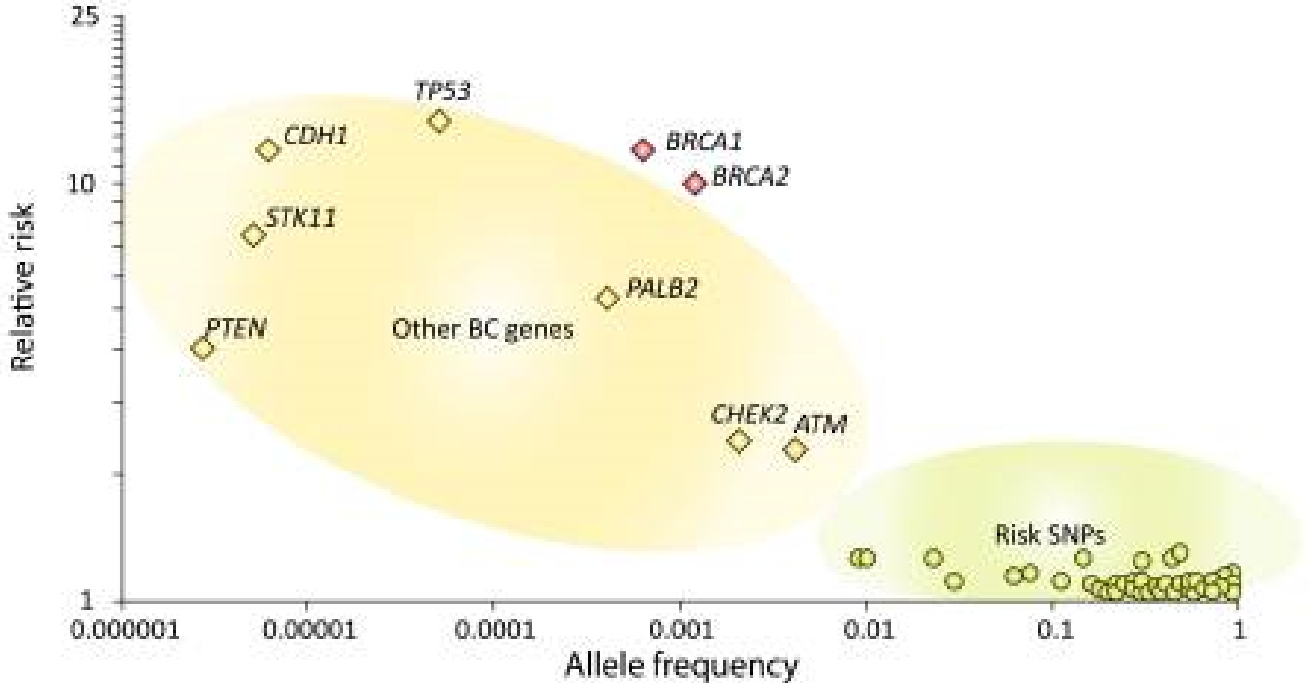


Figure 1.1: Risk penetrance profile for genetic susceptibility factors [6]

Jointly, GWA studies resulted in identifying a total of more than 170 genomic loci with BC risk association. Wu et al. took a different approach in identifying genomic association by using the SNP profile of the individuals to predict their gene expression profile and identified a total of 179 genes with increased risk of breast cancer incidence. Of these, 23 have not been reported previously to have an association [15]. This method of identifying genomic association by predicting the transcriptome from the SNP profiles is known as a Transcriptome-wide association study or TWAS.

In the Netherlands, a nationwide study called HEBON, i.e., **He**reditary **B**reast and **O**varian cancer research **N**etherlands, primarily focussing on the families with breast and ovarian cancer cases started in the late 1990s and is still ongoing. This study's primary goal is to identify genetic variations, estimate the cancer risk within families, and develop better treatment methods. This study consists of various groups and departments from the eight University Medical Centres across the Netherlands and the NKI or the Antoni van Leeuwenhoek institute. Estimation of liability to a disease can be done based on the individual's genotype, i.e., polygenic risk score, or the risk factors (such as hormonal, lifestyle), genetic test results, and family history of an individual. The CanRisk tool [16] estimates the risk based on the risk factors, pedigree information, genetic test results and

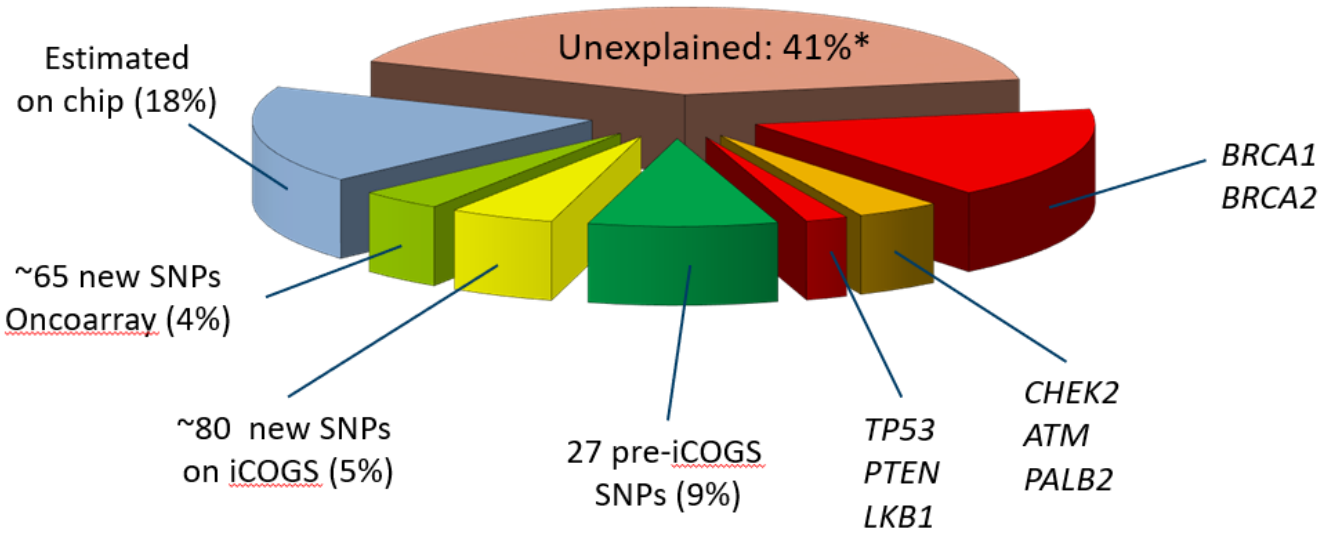


Figure 1.2: The familial risk factors known to be associated with Breast cancer. GWA studies have identified $\sim$ 170 genomic loci, and various genes [8, 9, 10, 11, 12, 13, 14, 15]. The genetic factors that jointly explain $\sim$ 50% of familial risk to BC.

PRS. The CanRisk tool allows researchers and healthcare professionals to carry out multifactorial breast and ovarian cancer risk predictions.

In science, the amount of data generated is astronomical. It requires interoperability for exchanging information between research groups, which leads to the development of FAIR principles [17]. FAIR principles allow tackling the issue of sharing data between researchers by making data Findable, Accessible, Interoperable, and Reusable. The FAIR principles imply the data to be found on the internet, retrievable with or without authentication, integrated with other data sets, and documentation of data generation [18]. Sinaci et al. [19], and Jacobsen et al. [20] presents a workflow for FAIRification of health care data from raw data.

1.1 Problem Statement

The HEBON study has collected various data on participating individuals such as genetic testing information, risk factor, pedigree data, patient data, treatment data, mammographic data. The nine centres collect the data, then Antoni van Leeuwenhoek (AVL) anonymizes it. Various centres hold the data about a specific domain of expertise (such as DNA treatment, risk factors, Imaging). The AVL hold the central record of IDs for mapping the data, the Erasmus Medical Center, Rotterdam manages data on cancer treatment, LUMC manages data on genetic testing.

The development of better treatment approaches requires continually update of the data based on the current community knowledge. For example, new information on genomic variants and their

pathogenicity are published every month. Keeping track of such information can become tedious and can be expedited by following FAIR data standards. FAIR HEBON data will allow researchers to exploit the new knowledge in improving patient treatment, performing a meta-analysis, re-analysis based on updated domain knowledge. Another application of FAIR HEBON data will allow a more streamlined, automated usage of online tools such as the CanRisk tool [5, 16] for risk prediction. Hence, we propose FAIRification of HEBON data in this thesis.

The FAIRification of the data allows easier connection to public databases. As proof of concept, a meta-analysis of genotyped data would identify associated genes with breast cancer. Hence, we propose to perform a TWAS analysis on the genotyped data of HEBON participants. We propose to showcase another application of FAIR HEBON data to compute a risk score based on known genomic variants for Breast Cancer. Lastly, we would compare these polygenic risk scores with risk prediction by the CanRisk tool for each individual.

In this thesis, we answer why FAIRification of HEBON data is important? HEBON data is made FAIR in order to connect it with publicly available databases. To show the application of FAIRified HEBON data, we run the TWAS pipeline to identify the associated genes with BC with help of GTEx data. What are the remaining familial risk factors associated with BC? By performing the TWAS analysis of genotype data. How do the genes found in TWAS analysis interact with known BC genes? By performing network analysis and enrichment of the network. Lastly, we want to see how PRS and predicted risk scores are correlated based on risk factors, genetic test results, and pedigree data.

1.2 Thesis outline

In chapter 2, we will give a brief background about the FAIRification process, GWAS analysis, PRS computation and CanRisk tool. The research in this thesis can be split into two parts, first the FAIRification of HEBON data and second, running the analysis pipeline using the HEBON data. Next, in chapter 3, we present the dataset and the methods used. We discuss all our results in the subsequent chapter 4. Finally, we conclude and describe the future work in chapter 5.

Chapter 2

Background

This chapter describes the underlying concepts and principles to understand FAIRification and the genome-wide association studies process.

2.1 FAIR Principles

The Science builds over prior discoveries, and the progress of science intrinsically depends on the amount of available information. With the digital age, data being produced in the field of science is reaching enormous sizes. With increased data volume, there is a need for scientific data management. In 2016, in the journal Scientific Data, Wilkinson et al. [17] published "FAIR Guiding Principles for scientific data management and stewardship." to provide a framework and need to make the information found, accessed, interoperable, and reusable by the community. The principles primarily emphasise machine-actionability in order to automate the computational processes. The author defined principles for making data FAIR, i.e., findable, accessible, interoperable, and reusable, which can be seen in the following sections.

2.1.1 Findable

1. **Metadata and data are assigned a persistent and globally unique identifier:** A globally unique identifier refers to only one resource globally, and persistence means that this identifier is always used to refer to the same resource [21, 22].

2. **Rich metadata describes data:** This principle allows the resource to be discovered through search or filtering, as a not well-described resource cannot be discovered accurately [21, 22].

3. **Metadata includes the data identifier it describes.**

4. **Indexing of the Metadata and the data in a searchable resource:** Registering and indexing metadata and the data in a searchable resource helps accomplish the first, second, and third principles [21, 22].

2.1.2 Accessible

1. **Metadata and the data can be retrieved using their identifier through a standard communication protocol:** the Communication protocol needs to be universally implemented, open, and free. Moreover, it should allow for authentication and authorisation when needed [21].

2. **Even if the data is not accessible, however, the metadata can still be accessed.**

2.1.3 Interoperable

1. **The metadata and the data use a widely accepted, formal, and accessible language for knowledge representation.**

2. **Use of vocabularies following FAIR principles for metadata and data modelling.**

3. **Metadata and data should include relevant and meaningful references.**

2.1.4 Reusable

1. **Metadata and data are with accurate and relevant attributes:** Metadata and data include a data usage licence, a detailed description of how data was generated and meets community standards.

2.2 FAIRification

Kush et al. [23] show the advantages of using a common data element (CDE) for the FAIRification of data. CDEs are standardised data collection units that try to answer one or more questions with a set of values. CDEs often contain terminology concepts defining the meaning of the data, identifiers for each CDE. CDEs allows the researchers in developing better machine-readable and interoperable semantic models [23]. Zhang et al. [24] showcase semantic modelling of their oncology data for achieving interoperability with the public databases. Zhang proposes an ontology framework for modelling the data using the National Cancer Institute (NCI) Thesaurus [25] and, as proof of concept, perform survival analysis on the data [24].

2.3 Genome-wide association study

The goal of population-based association studies is to identify genomic variants that vary systematically and proportionally between individuals with different disease states and represent the effects of risk-enhancing or protective alleles [26]. *SNPs* are the single base-pair changes occurring in the DNA sequence and have a high frequency in the human genome [27]. Thus, SNPs are considered the modern unit of genetic variation and are used as *genomic markers*. SNPs typically have two alleles, i.e., two commonly occurring base-pair possibilities for an SNP location within a population. SNP frequency is given in terms of the *minor allele frequency* or the frequency of the least common allele [26].

The population geneticists describe the changes in the frequency of genetic variation over time within a population mathematically by Linkage disequilibrium (LD), related to the chromosomal linkage. Chromosomal linkage is where two markers remain physically linked on a chromosome through generations. LD is a property of SNPs, and it is a measure of co-occurrence of two alleles of two different SNPs [3, 26]. LD is commonly represented in D', and r^2, where D' represents the recombination between markers and r^2 is the statistical measure for correlation [28, 29].

Ford et al. showed that the genomic variation in BRCA1 and BRCA2 were contributed towards inherited breast cancer [30]. Over the decades, research studies such as Wu et al., Ferreira et al. analysed genotyped data using transcriptome- and genome-wide association studies to identify more than 88 genes that were not known to be associated with Breast Cancer. GWAS analysis strategies generally include four components: (i) pre-processing the data; (ii) generation of new data; (iii) statistical analysis of the data; and (iv) post-analysis investigations. These investigations' primary goal is to identify and characterise the association among SNPs and measure the disease progression or disease outcomes [31].

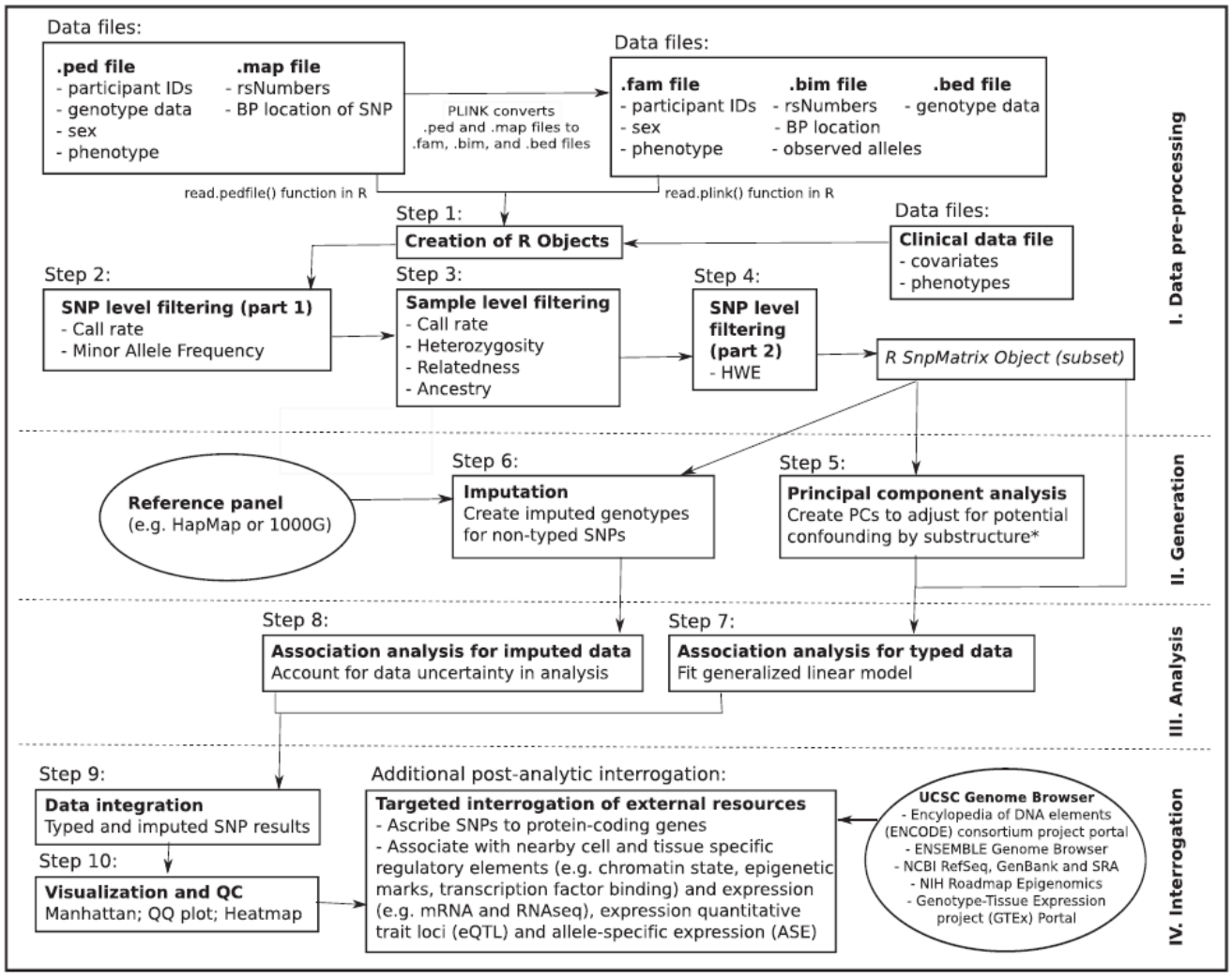


Figure 2.1: GWAS Pipeline as decirbed by Reed et al. [31]

2.3.1 Preprocessing

Pre-processing/Quality control is a vital part of the GWAS pipeline, as the raw genotype data are inherently imperfect. The genotype data can have errors for various reasons, such as poor quality DNA samples, ineffective genotype probes, bad DNA hybridisation, contamination, or mix-ups of the sample. QC involves filtering SNPs and individuals based on missingness, inconsistencies in the sex, minor allele frequency, heterozygosity, relatedness.

SNP-level filtering

SNP-level filtering excludes SNPs with missing information. One of the criteria for filtering is the call rate, i.e., for a SNP, the information of proportion of individuals on the corresponding SNP is available in the study. Thus, for example, a call rate of 95% is used as a filter, that retains SNPs with less than 5% missing data.

The following criteria for filtering are minor allele frequency (MAF). A significant degree of homozygosity for an SNP across participants often results in inadequate power for inferring a significant association between the trait and the SNP. A very small MAF would mean that most individuals will have two copies of the major allele.

Other filtering criteria are the Hardy–Weinberg (dis)equilibrium (HWE) law. HWE assumes an indefinitely large population where no mutation, migration, or selection. Thus, violation of the HWE law would mean that the genotype and the allele frequencies are not constant over generations, and the genotype frequencies are significantly different from expectations. For example, if allele A has the frequency of 0.20 and allele T has the frequency of 0.80, then the expected frequency of genotype AT would be $2 * 0.2 * 0.8 = 0.32$.

In GWA studies, it is assumed that any deviation from the HWE is due to the genotyping errors. The threshold for cases is often less stringent than that of controls. If the case-control status is known, then the HWE law violation indicates genetic association with the trait. The deviations from HWE are measured by performing a goodness-of-fit test between the expected and observed genotypes.

Sample-level filtering

In the next stage of data pre-processing, sample-level filtering is performed for excluding the individuals from the analysis based on sample contamination, missing data, population stratification, and ethnic, gender, or racial ambiguity. Similar to SNP-level filtering, individuals with missing genotype data across the study are excluded based on the call rate. Sample-level filtering also includes filtering based on heterozygosity, i.e., the presence of both the alleles of a given SNP for an individual. The heterozygosity can be filtered using the HWE threshold. Lastly, SNP relatedness, duplicates and population stratification can be dealt with by SNP pruning. This pruning is applied by thresholding the linkage disequilibrium value.

2.3.2 Data generation

After the pre-processing of data, new data is generated before performing the statistical analysis. The genotypes of non-typed SNPs that may have a functional relationship to the results are generated by imputation. Imputation of non-typed SNPs can be performed using a reference haplotypes panel and their LD map, such as Haplotype Reference Consortium (HRC) and 1000 Genomes data. Thus, this provides additional power for the identification of the association. After imputing the non-typed SNPs, as part of quality control, the imputed data with high uncertainty is filtered out. Standard measures of uncertainty are the information content and R^2, and a threshold is often applied to R^2 for filtering. R^2 is the value associated with the linear model regressing the imputation.

2.3.3 Statistical analysis

Generally, association analysis includes regressing each of the SNPs individually for the given trait. A typical GWAS analysis examines each SNP independently for associations with the trait by performing a series of single-locus statistical tests. Commonly, for analysing quantitative traits, a generalised linear model (GLM) and analysis of variance (ANOVA) are used. ANOVA is comparable to linear regression, as it uses a categorical predictor variable. Both methods assume that the traits are normally distributed, and the groups in the study are independent and homoscedastic [3].

These tests measure the deviation from the null hypothesis (i.e., no association between the phenotype and the genotype class) in terms of p-value and effect size. The null hypothesis is rejected 5% of the time, with a small probability of false positives. However, in GWA studies, hundreds of million tests are computed, which needs to be corrected. Correction for multiple testing is often done using the Bonferroni correction, and false discovery rate (FDR)[3]. FDR estimates the proportion of significant results by using a false positive rate (α) of 0.05. In comparison, Bonferroni correction adjusts this alpha to α/k, where k represents the total statistical tests performed [32].

2.3.4 Post-analysis visualization and interrogation

GWA analysis findings can be visualised in various ways. One of them is the Manhattan plot, which allows visualising the significance level of a GWA study by the chromosomal location. Each SNP is plotted based on its chromosomal location and the negative log scaled p-value [33]. SNPs with a smaller p-value will have a higher negative log scaled p-value, meaning by inspecting the plot, one could identify associated SNPs. This identification is often based on the Bonferroni correction used as a threshold in the plots. Lastly, the SNPs which were found to be associated with the trait are mapped to their genomic location using the base pair position and the chromosomal location. This mapping helps in identifying the genes associated with the trait. The SNP-level findings can also be used for predicting the genomic expression levels [15].

Polygenic Risk scores

Another post-analytic interrogation is computing the polygenic risk scores (PRS), a single-valued estimate of an individual developing a disease/trait, which has potential applications in the field of

precision medicines [34]. The PRS is computed as the sum of their genotypes, weighted effect sizes taken from the summary statistics of the GWAS.

$$PRS = \sum_{i=1}^{k} \beta_i N_i \tag{2.1}$$

Igo et al. [35] describe the calculation of a polygenic risk score as the sum of the log of odds ratio, β_i, multiplied by the number of risk alleles, N_i for each locus. The equation 2.1 provides the estimated risk for the disease.

Chapter 3

Materials and Methods

This chapter covers the description of the dataset used and the methods for FAIRification of the data and analysis pipeline.

3.1 Dataset

In $5-10\%$ of breast cancer incidences, heredity is the cause, and other factors play a role in the other $90-95\%$, along with the genetic predisposition. In the mid-90s, researchers found BRCA1 and BRCA2 to be related to breast and ovarian cancer. Women with a mutation in either of the two genes have significantly increased breast and ovarian cancer risk. Men and women can pass on the BRCA gene mutation to their daughters and their sons [36].

The dataset used in this project is from the HEBON study, a nationwide survey of families with breast and ovarian cancer incidence, to facilitate researchers and these families. Individuals from families with breast or ovarian cancer incidences and have at least one member genetically tested can participate. Each tested individual receives an invitation letter and is grouped based on their response to participation in the study, does not want to participate, and does not respond. HEBON collects various information from the participating individuals, such as the age of cancer incidence, lifestyle factors (i.e., alcohol consumption, smoking status), genetic mutations, family history, and medical records (such as MRI images, mammograms).

Analysis	**Instances**	**Data source**
FAIRification	54,890 individuals	HEBON DNA data
TWAS analysis	2155 cases; 1778 controls	cases: HEBON; controls: BCAC
Risk score	2132 carrier cases and control 2155 non carrier cases 1778 non-carrier controls	BRCA1/2 carrier: CIMBA; non-carrier cases: HEBON; non-carrier control: BCAC

Table 3.1: Data distribution for the data used in this study.

In this project, we FAIRified the DNA determinations or testing information and genomic variant

data and analysed the genotype and pedigree data of the participants. The DNA determinations data is represented using a relational database using ten tables containing information of 54,890 individuals. The DNA determinations data consist of genetic testing results, variants found after testing, various studies the individual was part of, and testing centre details. The genetic testing data consist of mutation in either of the eight genes (BRCA1, BRCA2, RAD51C, RAD51D, BRIP1, PALB2, CHEK2 and ATM). For our TWAS analysis, we use the genotype data, i.e., SNP array is represented using a PLINK file [37] containing a total of 2323 individuals who do not carry a BRCA1/BRCA2 mutation. Lastly, for comparison between the PRS and the CanRisk tool prediction, we used the genotype data of individuals with a mutation in BRCA1/BRCA2 from the CIMBA study, pedigree information and diagnostic test information.

3.2 FAIRification

As shown by Jacobsen et al. [20] (seen in figure 3.1), the FAIRification can be split into three phases: pre-FAIRification, FAIRification and post-FAIRification. The first step of the pre-FAIRification phase is identifying the FAIRification objective, which here is to make the HEBON data interoperable and findable. The next step in this phase is analysing the HEBON data and metadata by checking the data representation, data description and existing FAIR features. Before moving to the FAIRification process, the data is preprocessed, such as mapping the variant to their genomic location, anonymising individual identifiers while following the FAIR guiding principles.

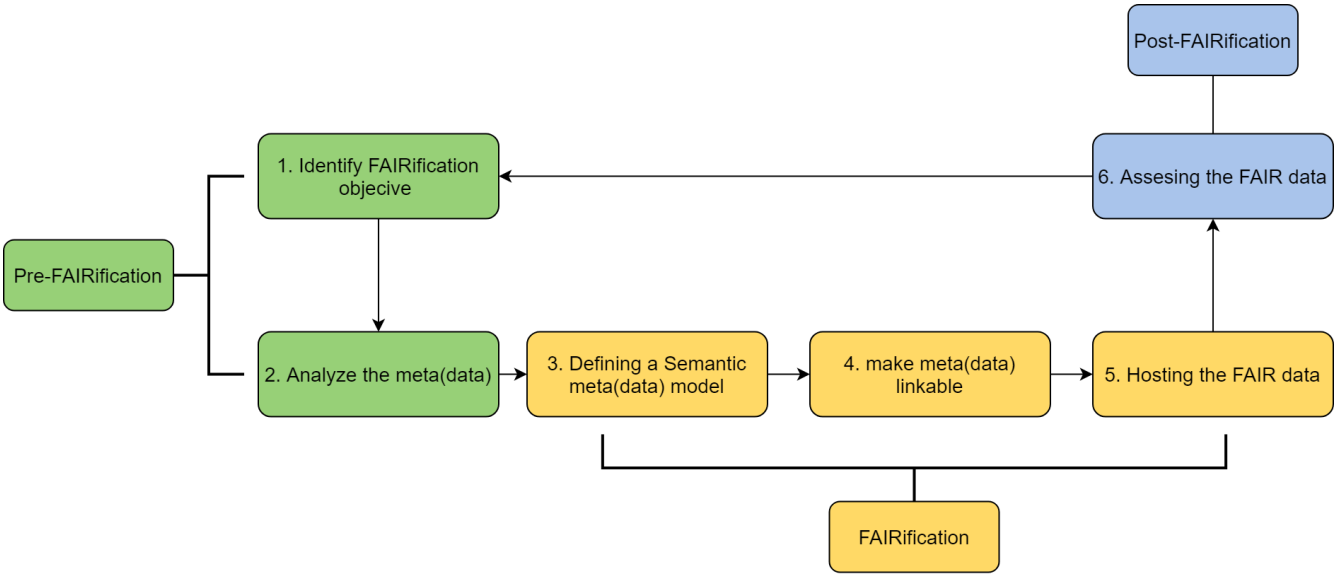


Figure 3.1: FAIRification workflow as decribed by Jacobsen et al. [20]. The workflow can be divided into 3 parts, pre-FAIRification, FAIRification,, and post-FAIRificaition.

FAIRification of data requires defining a semantic model, which depicts the relation between the available data. Semantic models are often considered as a template for transforming the data into an interoperable and machine-readable format. It describes the data, formally depicting the relationship between data elements while using common vocabularies for annotation and interoperability with

other available resources. We referred to the common data element model of EJP-rare diseases. We used Resource Description Framework (RDF), Resource Description Framework Schema (RDFS), and Web Ontology Language (OWL) for our semantic model. RDF, RDFS and OWL are the common representation languages used in Semantic modelling.

3.3 TWAS pipeline

We used SNP arrays of 2155 cases from the HEBON study and 1778 controls from the BCAC study [13] for our TWAS analysis. All participants in our study are females who come from the Netherlands and do not carry a BRCA1/BRCA2 mutation.

3.3.1 Pre-processing the genotype data

A crucial step for analysing population-wide genomic data is pre-processing the data by performing quality control (QC). We use PLINK [37], a command-line program that allows the users to perform a wide range of genetic analyses for handling the genotyped data. The first step of QC is to deal with SNPs and individuals' missingness, and we excluded the SNPs and individuals with a call rate below the threshold of $> 95\%$. The next step in the QC is handling the SNPs with low MAF and deviation from the Hardy-Weinberg equilibrium. We excluded the SNPs below the MAF threshold of 0.05 and the Hardy-Weinberg equilibrium p-value below e^{-6} for controls and e^{-10} for cases. The next step is handling the heterozygosity and the relatedness of samples, discrepancies related to sex. We performed QC with strict thresholds, similar to the authors [15].

3.3.2 Imputation

Next, we phased the genotype data using Eagle [38], which estimates the haplotype phase using a phased reference panel, improving the imputation quality. We imputed the phased genotype data using Minimac3 [39] to the Haplotype Reference Consortium reference panel [40] with the genome built GRCh37. Minimac3 is a computationally efficient and lower memory software for genotype imputation [39]. We included only the SNPs with an imputation quality of ≥ 0.8, a MAF of ≥ 0.05, and a call rate $\geq 98\%$ for the association study.

3.3.3 TWAS analysis

We predict the gene expression of each individual based on the included SNPs using the PrediXcan algorithm [41]. PrediXcan uses an elastic net with alpha = 0.5 as recommended by Gamazon et al. We use weights of the pre-trained model on GTEx V7 data of Breast tissue from European individuals. We applied PrediXcan using these weights on the genotype data; first, it predicts the gene expression and then runs association tests for a trait. The genes with a prediction score (R^2) ≥ 0.09 and a p-value below the Bonferroni-correlated threshold, $\leq 5.82 * 10^{-6}$, were associated with breast cancer.

3.3.4 Network analysis

After predicting the gene expression and identifying the associated genes, we wanted to understand the interaction of our findings with the known genes and the implicated processes. We used the STRING [42] database for identifying the interaction between associated genes and known genes associated with breast cancer [36] using network analysis. The network is in the form of a graph $G = (V, E)$. The vertices, V, represent genes and the edges, E, exhibits interaction based on curated-literature interactions from STRING [42]. Next, we perform network enrichment for understanding the relationships between genes in the network. The enrichment of network was done using gene ontology (GO) terms to investigate enriched GO biological processes.

3.3.5 PRS Calculation

We compute the PRS for the individuals with a BRCA1/BRCA2 mutation using the formula given by Igo et al [35] and in Equation 2.1 based on the 313-breast cancer-associated variants.

Chapter 4

Results and Discussion

In this chapter, we report our findings of the FAIRification and association analysis of the HEBON data. First, the results of the FAIRification process are discussed where the proposed semantic model is reported, followed by the FAIRness of the data. Next, the results of the association and network analysis are discussed in depth. The HEBON data is patient sensitive, therefore has restricted access and is stored at the LUMC. The code for this study can be found at GitHub: Mandloi2309/HEBON-analysis.

4.1 FAIRification

We followed the FAIRification workflow, and our objective for HEBON FAIRification was to make the data interoperable and findable with the publicly available databases and tools. Before FAIRification, the data was represented using a relational database and lacked persistent URIs. The information on variants is represented in a VCF, which was standardised according to the EVA standards. Next step, after analysing the data, we defined a semantic model for representing the data. We used existing ontologies like Semanticscience Integrated Ontology (SIO)[43], Genotype Ontology (GENO)[44], SNOMED Clinical Terms (SNOMEDCT)[45], Sequence Types and Features Ontology (SO)[46], National Cancer Institute Thesaurus (NCIT)[25], and DCMI Metadata Terms (DC) for describing the data.

The semantic model can be seen in Figure A.1 and in figures 4.1, 4.2, 4.3, and4.4. In figure 4.1, the information about the individual is represented, such as their HEBON identifier, the study they were part of and the date and time of the entry. Each HEBON participant is assigned a unique identifier, i.e., HEBON number, and we use this as a URI for each individual in our model. Most participants of the HEBON study has also taken part in various other studies (like CIMBA, BRIDGES, IBCCS), and we use SIO to define this relationship between the study and the participant. We created 13 study instances of 'NCIT: study' for defining the relationship of the HEBON individuals and the studies.

In figure 4.2, the information pertaining the variant detected is represented. The variant information includes HGVS notation of the variant, its effect on DNA and protein molecules, the genomic location of the variant, the reference sequence used for detecting the variant, and the gene where

the variant is found. We used ClinVar IDs as a URI and Wiki ontology for HGVS notation for each variant. Sequence alteration caused by the variants is represented using instances of the sequence alteration class of GENO. Effect on protein is represented using SO and SNOMEDCT vocabularies. Each individual in the HEBON study was tested at one of the 9 test centres, and we represent the

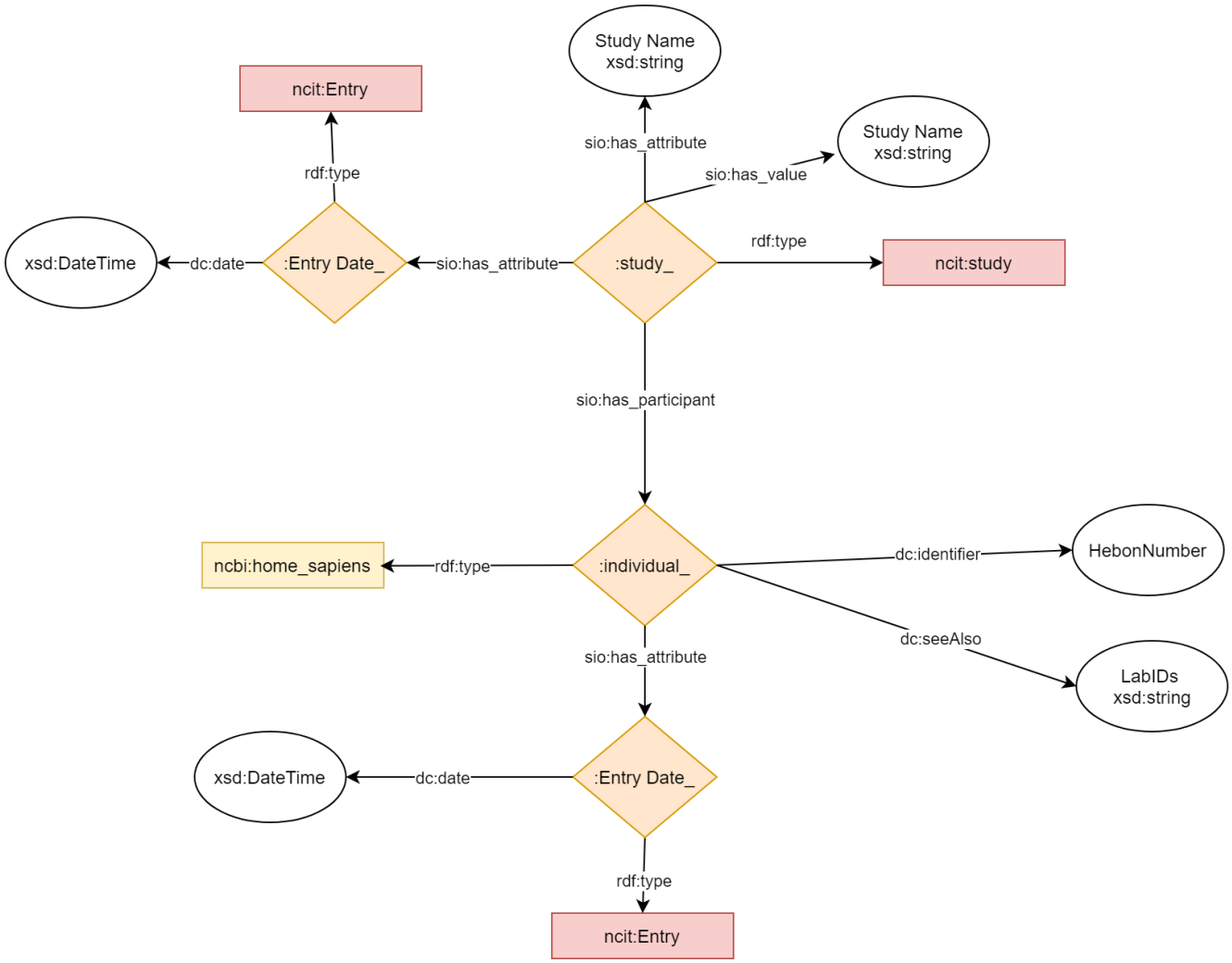


Figure 4.1: Semantic model representing the general information and detailed studies regarding the individuals part of the FAIRification process. Ontologies such as SIO, NCBI, NCIT were used in representing the data.

test centre information using an instance in figure 4.3. The test type information is represented using instances of NCIT genetic testing. The instances have either partial screening, mutation carrier screening, genetic screening and are undetermined as its value. The test results are represented using clinical test results from NCIT, and each instance is described using vocabularies from SNOMEDCT and SIO. The genetic tests identified the variants for each individual, which are represented as an instance of modification class from GENO. Test results were divided into two categories, general test results and variant test results. The general test results provide information if a known variant was found during the test. The variant test result provides detailed information on the variant, such as heterozygous or homozygous. Lastly, in the HEBON study, each variant is classified using

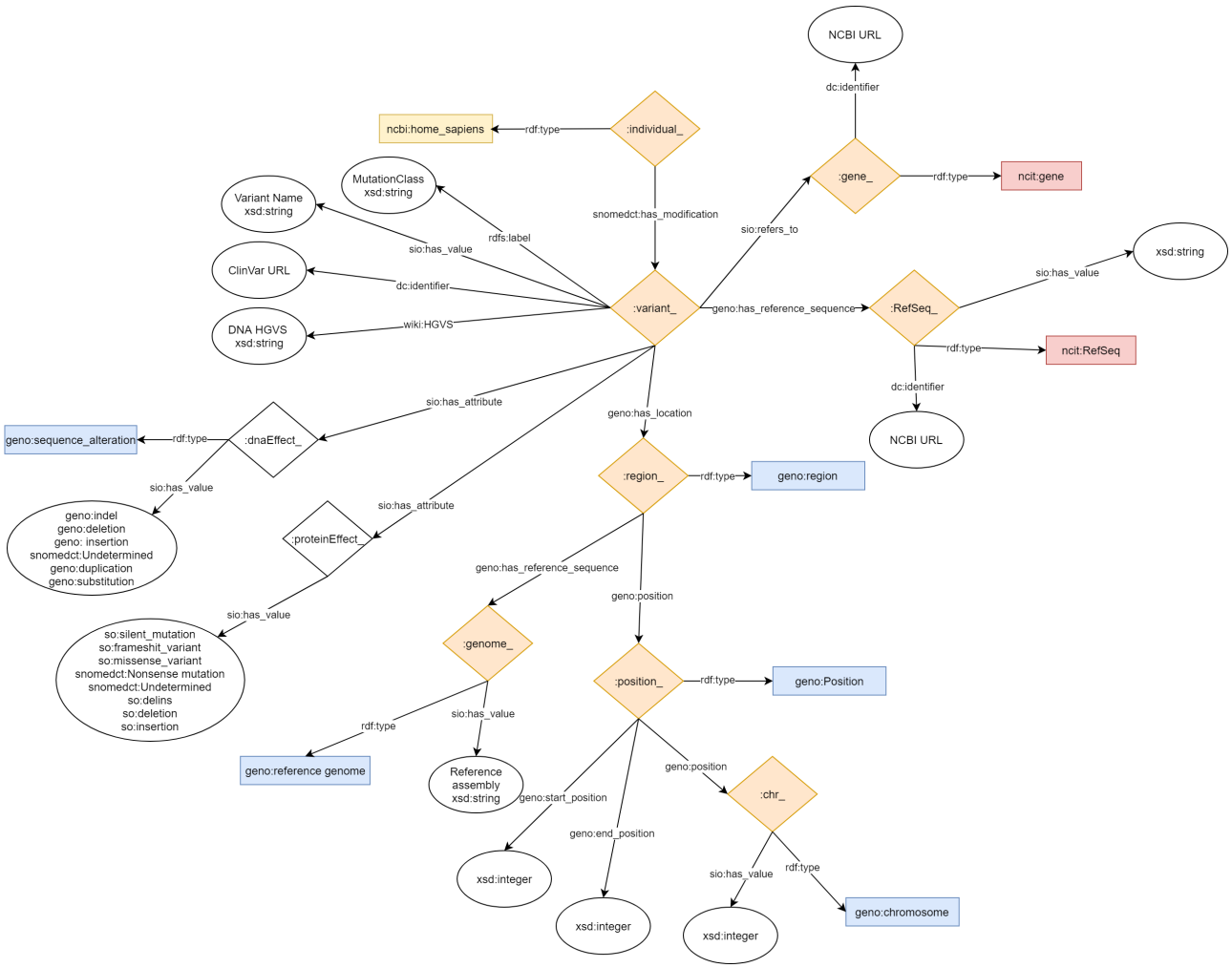


Figure 4.2: The model represents the variant information such as HGVS notation of the variant, its effect on DNA and protein molecules, the genomic location of the variant, the reference sequence used for detecting the variant, and the gene where the variant is found. We used ontologies such as GENO, NCIT, SIO, SNOMEDCT for model building.

three classifications; gold classification, HEBON classification, and IARC classification and is represented in the figure 4.4. The gold classification describes whether the individual is a carrier or non-carrier of a pathogenic mutation for each gene and the number of pathogenic mutations. It is based on the joint analysis of available genetic data for an individual and represents a carrier status summary score. The International Agency for Research on Cancer (IARC) categorised the variants based on their pathogenicity to humans [47]. Lastly, the HEBON classification is the carrier classification based on original reports from the diagnostic lab with a variant description. HEBON classification and gold classification, in principle, should be identical. However, based on available data, gold classification will evolve and be different over time. Each classification was represented using instances of NCIT classification class and relationship from SNOMEDCT.

The FAIRification of HEBON data for linking it with heterogeneous datasets has demonstrated the usefulness of using semantic data integration to resolve schematic, syntactic and semantic

heterogeneities across various data sources. The use of ontologies and common data elements (CDEs) facilitates data integration in various ways:

1. A shared, controlled vocabulary helps in standardising the data elements. It makes it easy to understand the data for humans and computers.

2. Modelling the data and defining semantic relationships between data elements explicitly states the assumptions based on domain and data.

3. FAIR represents the data in a formal and machine-readable language.

Figure 4.3: The model represents the testing information such as test centre details, type of test done for variant detection, and the result of the tests. We used ontologies such as GENO, NCIT, SIO, SNOMEDCT for model building.

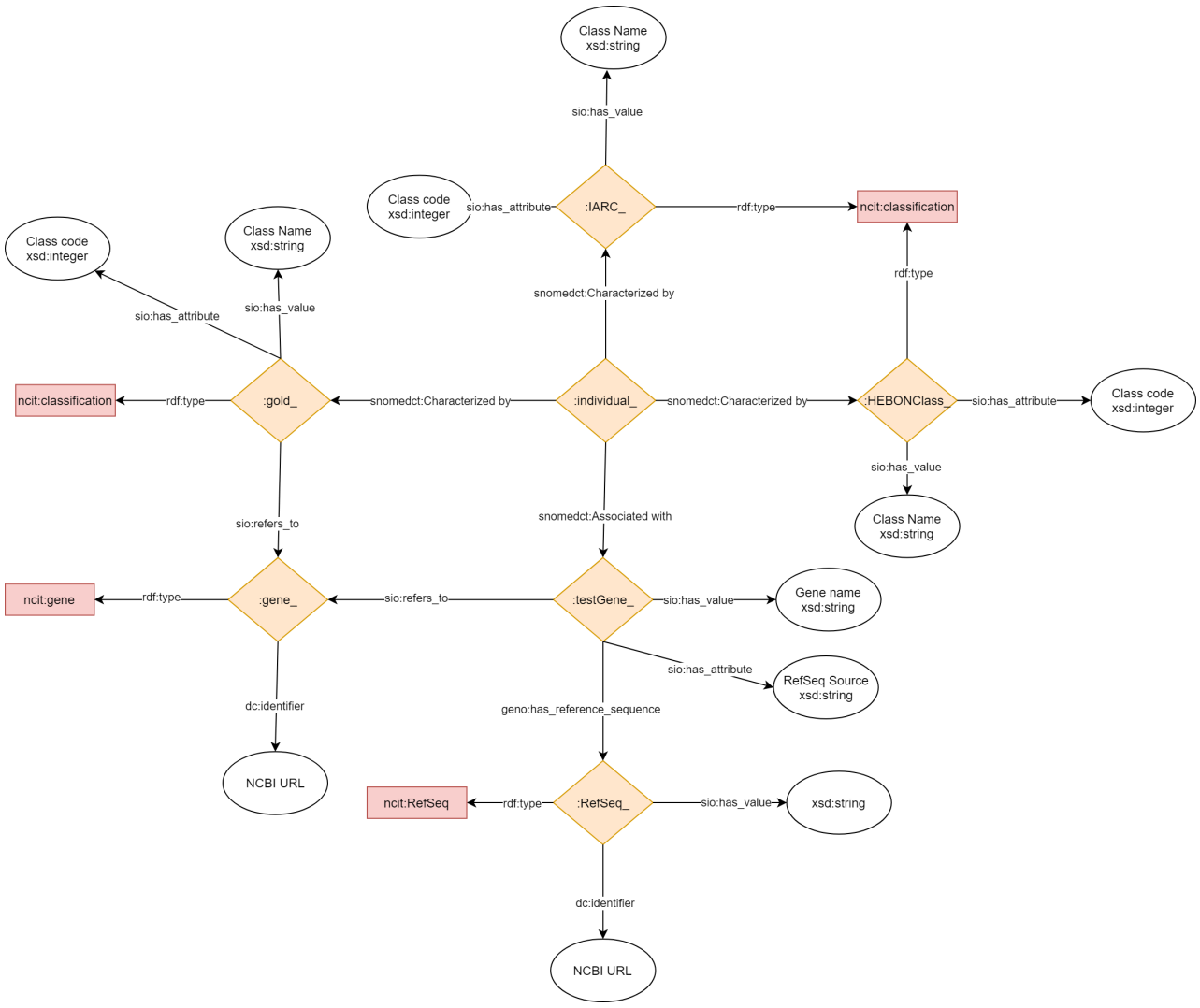


Figure 4.4: The model represents the three classifications used in HEBON study such as gold, IARC, and HEBON classification test. The gold classification describes whether the individual is a carrier or non-carrier of a pathogenic mutation based on the available genetic data and represents a carrier status summary score. The International Agency for Research on Cancer (IARC) categorised the variants based on their pathogenicity to humans [47]. Lastly, the HEBON classification is the carrier classification based on original reports from the diagnostic lab with a variant description. We used ontologies such as GENO, NCIT, SIO, SNOMEDCT for model building.

4.2 Association analysis

To assess the association between genes in breast cancer, we analysed the genotypes of participating individuals in the HEBON study. We performed the association analysis on SNPs between cases and controls, and we can see the association analysis results in the Manhattan plot in Figure 4.5. The Manhattan plot represents the p-values of SNPs and their genomic location. The x-axis represents the chromosome position, and the y-axis represents the $-log_{10}$ of the p-value. The genetic variants have significant p-values and rise on the Manhattan plot. SNPs above the red line have a

p-value less than the Bonferroni threshold and are considered to be associated with the trait. The Figure4.5 shows SNPs above the red line accumulating on specific locations, such as chromosome 6 and chromosome 10. We plotted 7371964 SNPs using the Manhattan plot, and 1596 SNPs were below the Bonferroni threshold. Out of the 1596 SNPs, 1455 SNPs overlap with the known SNPs associated with BC.

The single SNPs often do not have sufficient power to be detected in the GWA study; however, we know that multiple SNPs can influence gene expression by working together. Thus, it is possible that multiple SNPs that individually did not associate with the disease might affect a single gene in the same way and whose increased (or decreased) expression could be associated with the disease. Hence, we wanted to see the gene expression of the participants, and we used PrediXcan for predicting the expression and run association tests to a trait, i.e., breast cancer, in the cohort. We predicted expression for 4628 genes by using 98143 out of 143827 SNPs in the model. Our association analysis found 14 genes associated with breast cancer with a predictive score of more

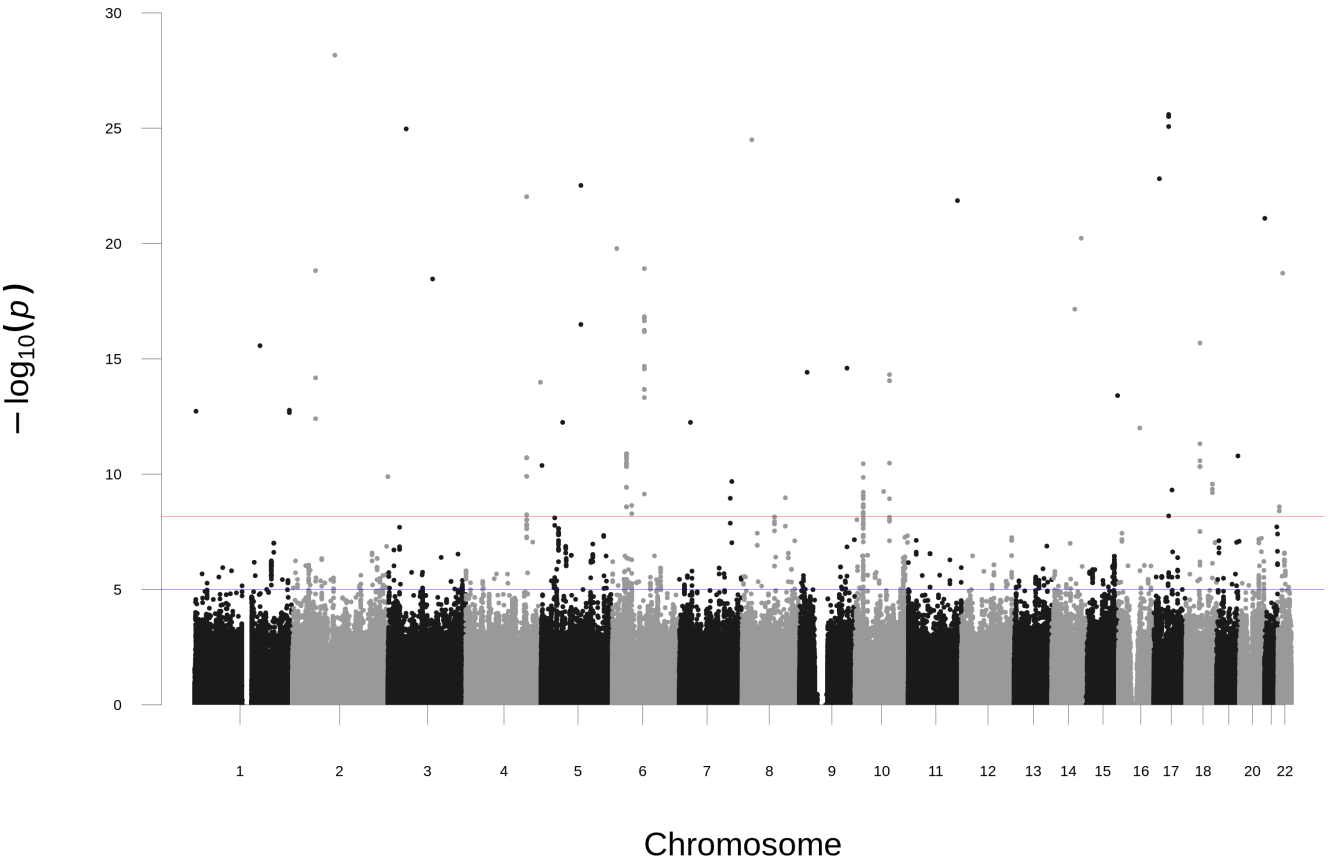
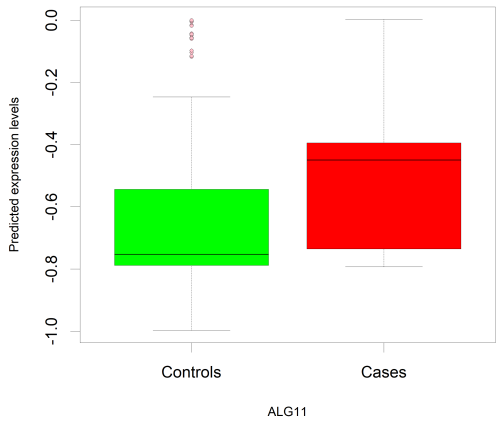
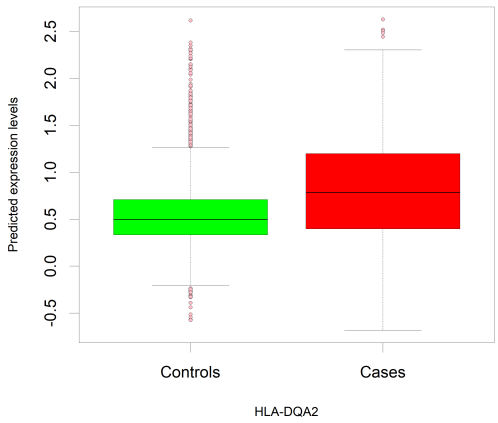


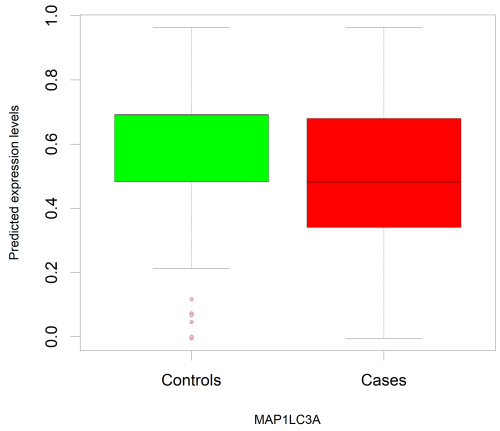
Figure 4.5: Manhattan plot of genome-wide association analysis results between 2155 cases from the HEBON study and 1778 controls from the BCAC study [13]. The figure depicts the level of statistical significance (y-axis), measured using the negative log of the p-value for each SNP, arranged by the chromosomal location on the x-axis. A grey or black dot indicates each typed SNP. Imputation was performed using the HRC reference panel [40]. The red line represents the Bonferroni level of significance ($p \leq 5 * 10^{-8}$), and the blue line represents the False Discovery Rate.



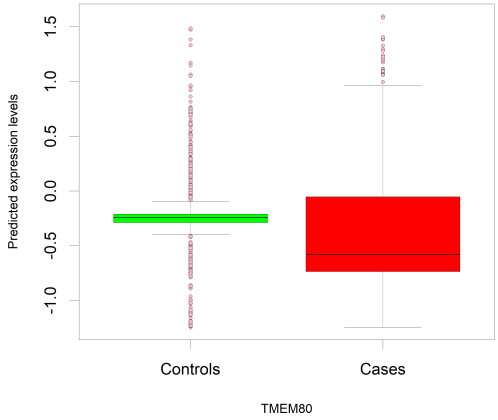
(a) ALG11



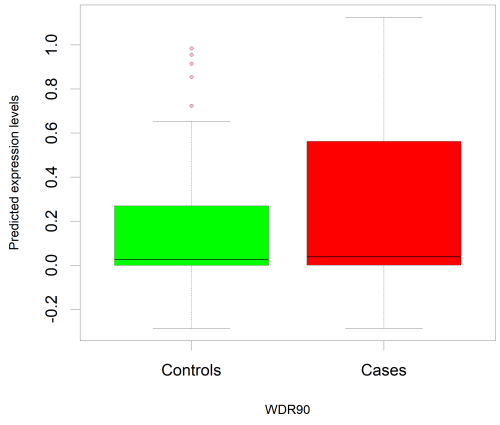
(b) HLA-DQA2



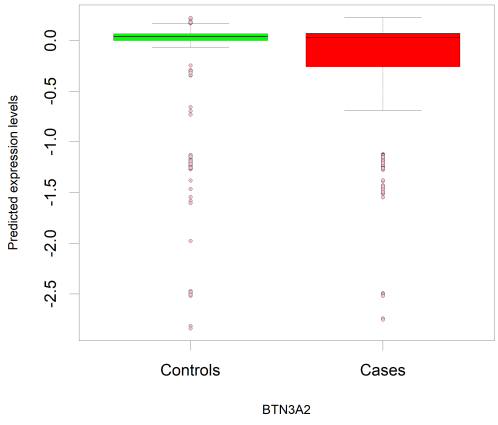
(c) MAP1LC3A



(d) TMEM80



(e) WDR90



(f) BTN3A2

Figure 4.6: Predicted gene expression levels in 2155 cases from the HEBON study and 1778 controls from the BCAC study.

than 0.9 and p-value below the Bonferroni threshold (See Appendix B Table B.1). We represented the expression of the genes using a boxplot (can be seen in Figure 4.6 and Appendix C. Out of 14 genes, ALG11, PSMG1, C16ORF13, HLA-DQA2 were upregulated in cases relative to controls. MAP1LC3A, TMEM80 and BTN3A2 were downregulated in cases relative to controls, and the rest genes had ambiguous expression relative to controls. The expression of BTN3A2 is seen to be suppressed in many tumor cell lines. BTN3A2 encodes the protein involved in the adaptive immune response. CASP8 plays a role in cell apoptosis, PSMG1 plays a role in proteasome assembly, HLA-DQA2 plays a role in peptide folding, and most genes in immune response and tumour suppression.

Our TWAS analysis of 2155 cases and 1778 controls predicted gene expression of ~ 4600 genes and found 14 gene found to be associated with BC. In our findings, few genes play a role in peptide folding, chromatin assembly, and most genes play a role in the immune response. BC genes are known to play a role in DNA damage but not so much in immune response, which can be due to the small sample size, and missed gene signals as we were only able to predict the expression of a quarter of genes. Out of our findings three genes CASP8, BTN3A2, and TRIM4 overlap with the BC associated findings of Wu et al. and Ferrire et al. and literature such as Cox et al. [48] and Cai [49] has shown that these genes associated with BC and regulate DNA repair and T-cell receptor interaction respectively. One of the reason for such a small overlap could be due to the fact that the HEBON participants used in this study have never been part of any GWAS studies as of now. Another reason could be the sample size, previous studies used a larger sample size compared to the one used in this study.

4.3 Network analysis

We performed network analysis for identifying the biological interaction of our findings (i.e. 14 genes) with the genes from Wu et al. [15] and Ferreira et al. [36] (i.e. a total of 192 genes found to be associated with BC). We analysed 204 genes (combined from our findings and the two studies) using the STRING database and then expanded to their first neighbours for identifying all possible interactions. We used a cutoff of 0.7 for the interaction between the genes and expanded to 275 genes (can be seen in figure D.1), and then we clustered the network using Markov Cluster Algorithm based on the interactions (can be seen in figure 4.7). We can see that our findings interact with the known set of genes to be associated with breast cancer.

The summary statistics of the network is represented in the table 4.1. The average number of neighbours in the network is 29.578, and the network density of 0.92. In figure 4.7 we represent the interaction network where the size of the node depends on the connectivity of each gene. The green coloured genes are the ones found in our analysis, and purple coloured genes (i.e. CASP8, TRIM4, BTN3A2) are overlapping genes. In the table D.1 we represent the network statistics for each node. The Histone genes have the highest number of connections in the network, i.e. 32, betweenness and closeness centrality and the clustering coefficient (≥ 0.9) and can be seen as a cluster at the top left in the figure.

Figure 4.7: STRING interaction network. We analysed 204 genes (combined from our findings, Wu et al. [15] and Ferreira et al. [36]) using the STRING database and then expanded to their first neighbours for identifying all possible interactions. We used a cutoff of 0.7 for the interaction between the genes. The network then expanded to 275 genes, and then we clustered the network based on their interactions. The green coloured genes are the ones found in our analysis, and purple coloured genes (i.e. CASP8, TRIM4, BTN3A2) are overlapping genes.

Summary statistics	
Number of nodes	275
Number of edges	1071
Avg. number of neighbours	29.578
Network diameter	2
Network radius	1
Connected components	73
Network density	0.92
Network heterogeneity	0.12
Network centralization	0.08
Clustering coefficient	0.95

Table 4.1: Summary statistics of the STRING network.

4.4 Enrichment analysis

We performed enrichment analysis for exploring the biological function of genes associated with breast cancer. Figure 4.8 represents the enriched network, and we found 158 genes to be involved in various biological processes and enriched in a network based on gene ontology (GO) terms and KEGG pathways (see table 4.2). The top 8 enriched GO processes were organelle organisation, a cellular component organisation, cellular localisation, symbiont process, chromatin organisation, interspecies interaction between organisms, antigen processing and presentation of exogenous peptide antigen via MHC class II, and intracellular transport. The enriched processes such as chromatin organisation and cellular localisation, which can be essential in DNA repair. The top 8 enriched KEGG pathways were viral carcinogenesis, systemic lupus erythematosus, alcoholism, Huntington's disease, pathways in cancer, MicroRNAs in cancer, Colorectal cancer, platinum drug resistance. These enriched pathways show that the genes are involved in pathways in cancer.

Few genes such as RBL2, TRIM4, CASP8 out of our findings (i.e. 14 genes from TWAS analysis) are enriched in processes such as chromosome organization, cellular component organization etc.and the remaining genes were interacting with the enriched genes. In the figure 4.8 and table E.1, the cluster 1 with Histone genes is enriched in the organelle and chromosome organization, and cellular localization. The cluster 2 with CASP8 and RBL2 is enriched in organelle and chromosome organization, and intra-cellular transport. The cluster 3 with HLA-DQA2 is enriched in antigen processing, and intra-cellular transport. The cluster 4 (with TRIM4) is enriched in chromosome organization and regulation of mitotic cell cycle. The processes enriched are involved in DNA repair mechanism, aligning with the literature that the BC genes are often involved in regulating DNA repair. We can see our findings are also involved in these processes showing an association.

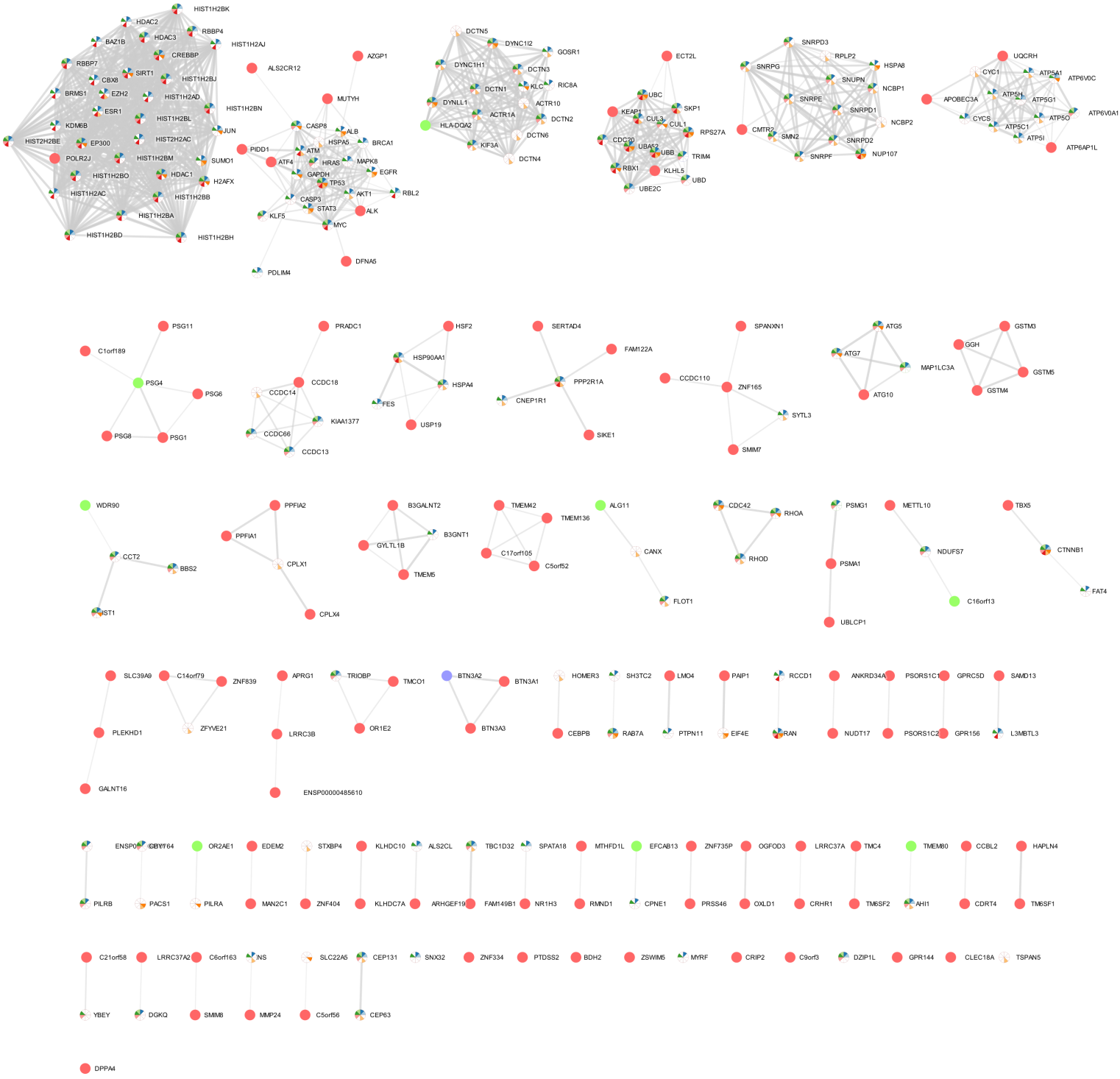


Figure 4.8: Network enrichment of the STRING network based on GO terms. We found 158 genes to be involved in various biological processes and enriched in a network based on gene ontology (GO) terms (see table 4.2).

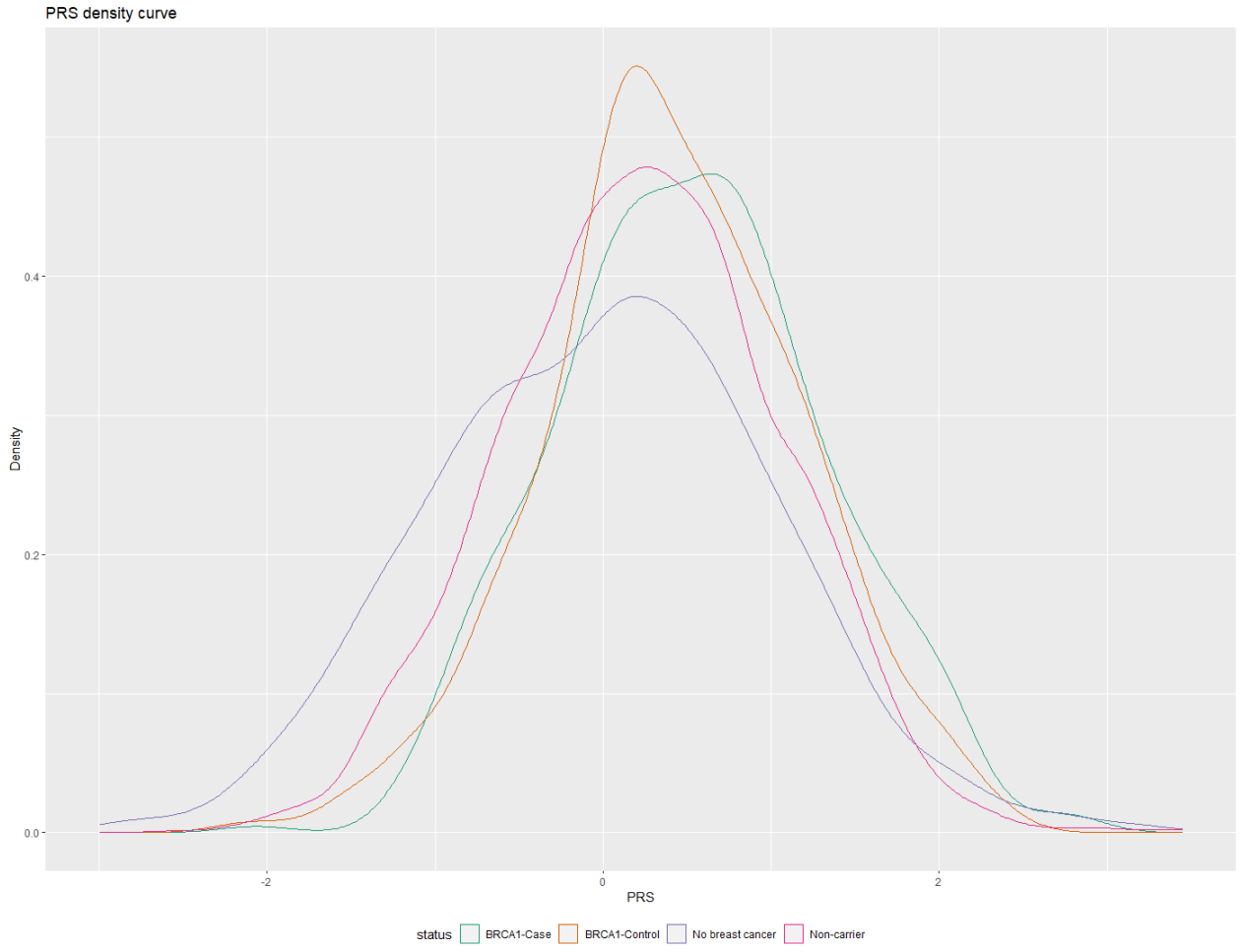
background genes	genes	category	description	FDR value	p-value	term name
3131	103	GO Process	organelle organization	6.00E-14	1.37E-17	GO.0006996
5163	137	GO Process	cellular component organization	4.04E-13	1.84E-16	GO.0016043
2180	74	GO Process	cellular localization	4.31E-10	6.87E-13	GO.0051641
650	37	GO Process	symbiont process	7.99E-10	1.56E-12	GO.0044403
683	38	GO Process	chromatin organization	7.99E-10	1.45E-12	GO.0006325
724	39	GO Process	interspecies interaction between organisms	8.11E-10	1.84E-12	GO.0044419
96	16	GO Process	antigen processing and presentation of exogenous peptide antigen via MHC class II	1.39E-09	3.47E-12	GO.0019886
1390	55	GO Process	intracellular transport	1.58E-09	5.04E-12	GO.0046907
183	27	KEGG Pathways	Viral carcinogenesis	3.59E-16	1.58E-18	hsa05203
94	19	KEGG Pathways	Systemic lupus erythematosus	1.58E-13	1.39E-15	hsa05322
142	21	KEGG Pathways	Alcoholism	8.38E-13	1.11E-14	hsa05034
193	20	KEGG Pathways	Huntington's disease	1.08E-09	1.90E-11	hsa05016
515	28	KEGG Pathways	Pathways in cancer	1.17E-07	2.58E-09	hsa05200
149	14	KEGG Pathways	MicroRNAs in cancer	2.51E-06	6.64E-08	hsa05206
85	11	KEGG Pathways	Colorectal cancer	3.00E-06	9.24E-08	hsa05210
70	10	KEGG Pathways	Platinum drug resistance	4.42E-06	1.56E-07	hsa01524

Table 4.2: Top eight enrichment based on GO terms and KEGG pathways.

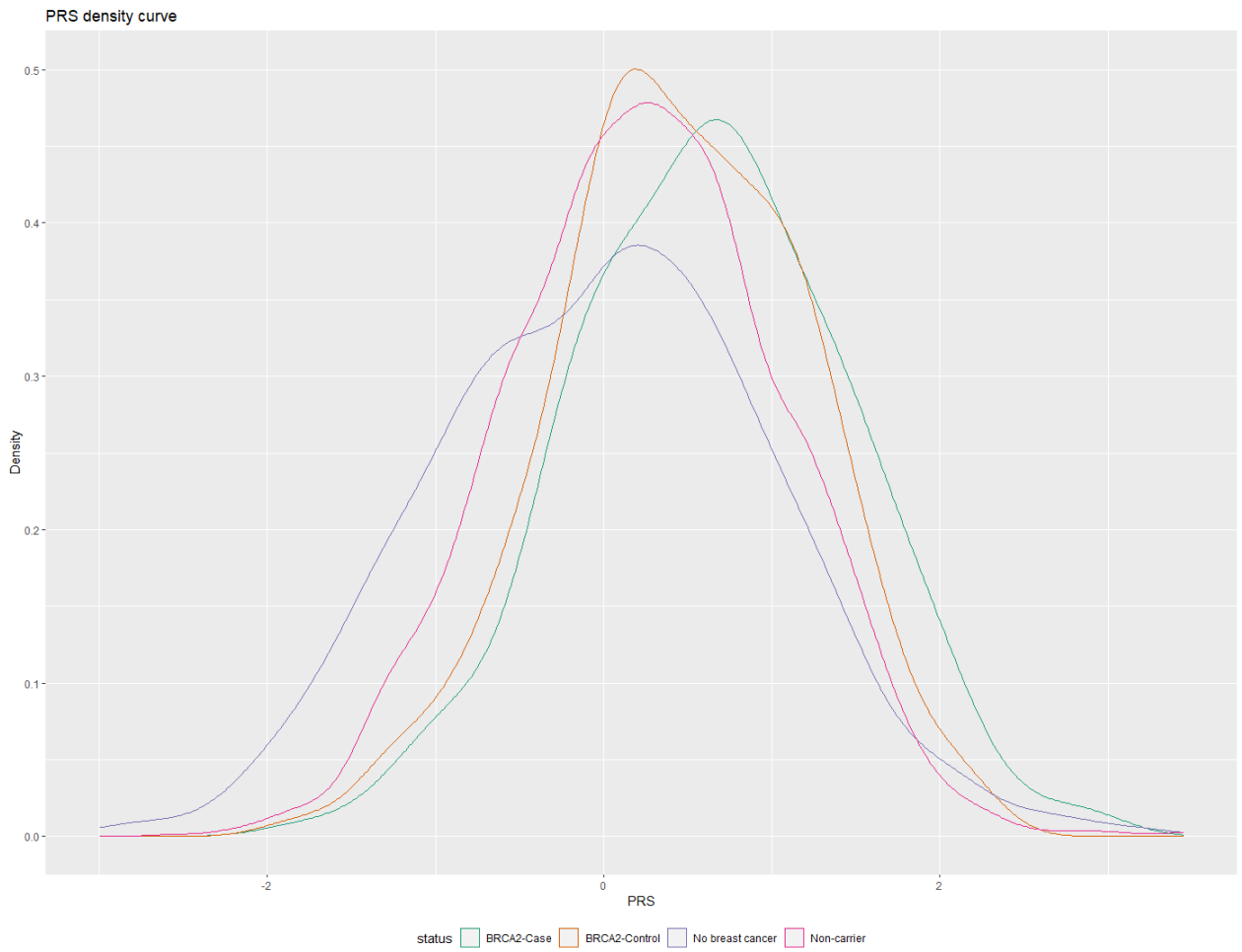
4.5 Polygenic risk score

We analysed 2132 individuals with the BRCA1/2 mutation, 2155 without a BRCA1/2 mutation and 1778 controls for computing the polygenic risk score based on 313 -breast cancer-associated variants. We standardised the PRS scores by computing z-scores using the 1778 BCAC control individuals.

The PRS density curve can be seen in the figure 4.9 and the statistics can be seen in table 4.3, and we can see that the PRS distribution for the non-carrier controls and controls with BRCA1/2 mutation have a similar mean PRS. On the other hand, the cases that are BRCA1/2 carriers have a higher mean than the non-carrier cases. The distribution shows the BRCA1/2 mutation is involved in cancer risk as we see a right shift in the cases with a BRCA1/2 mutation compared to non-carriers. Based on the curve between the controls with and without the mutation, we can say that it does not mean that the individuals will develop cancer; it might increase the chances.



(a) PRS curve of BRCA1



(b) PRS curve of BRCA2

Figure 4.9: PRS density curve between 2132 individuals with the BRCA1/2 mutation, 2155 without a BRCA1/2 mutation and 1778 controls. On the x-axis is the standardised PRS scores and y-axis represent the density. PRS in a population is known to follow a normal distribution.

Category	**Mean**	**Standard deviation**
No breast caner (non-carrier controls)	1.40	0.61
Non-carrier cases (HEBON-cases)	1.72	0.58
BRCA1 carrier - controls	2.04	0.64
BRCA1 carrier - cases	2.42	0.56
BRCA2 carrier - controls	2.14	0.62
BRCA2 carrier - cases	2.64	0.58

Table 4.3: Mean and standard deviation of PRS distribution based on the categories.

Lastly, to see if the family history and genetic test result correlate with the polygenic risk score, we used CanRisk Tool to predict cancer risk based on the family history and genetic test. CanRisk

tool takes pedigree data along with risk factors such as Menarche, BMI, height etc and returns a risk score. We can see in the figure 4.10 that the PRS and risk based on family history and genetic tests are very weakly correlated with a Pearson coefficient of -0.009. CanRisk tool predicts the scores based on the rare variants found to be associated with breast and ovarian cancer whereas PRS is computed based on the common variants. A weak correlation shows that PRS is not entirely dependent on the family history but more on the genetic variations. This shows that both the variables are independent of each other hence can be combined in computing an overall risk score.

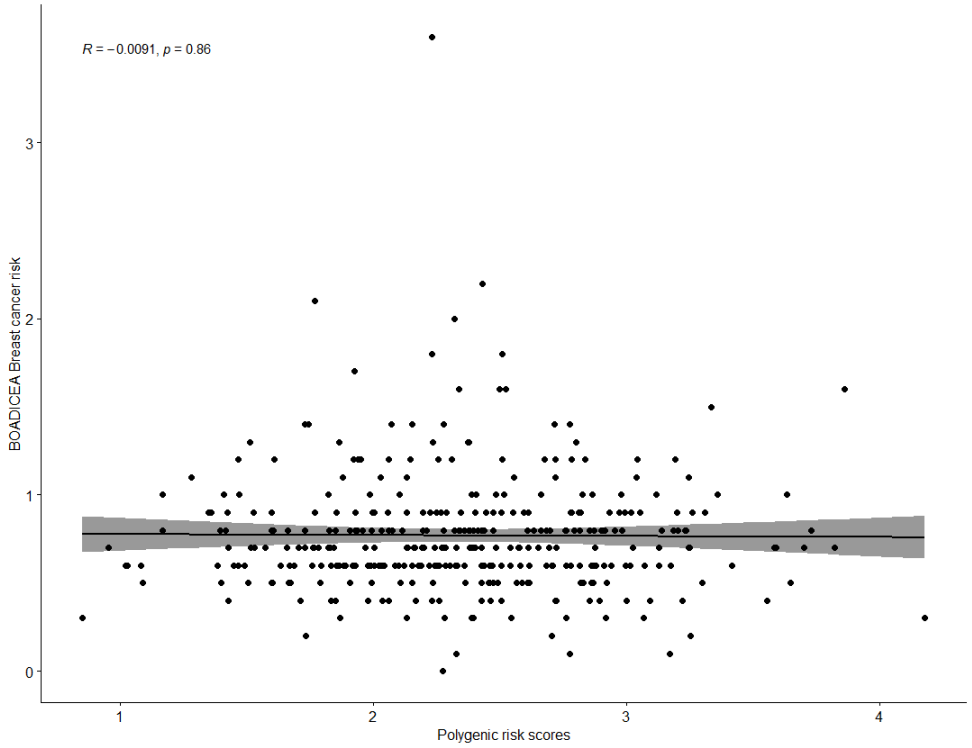


Figure 4.10: Scatter plot representing the Pearson correlation between the PRS and CanRisk predicted cancer risk score.

Chapter 5

Conclusion and future work

In this work, our objective was FAIRification of HEBON data to make it interoperable with publicly available datasets and tools. We worked on the DNA determination data from the HEBON study. We followed the workflow presented by Jacobsen et al. and FAIR principles for the FAIRification of HEBON DNA data. In our semantic model of the data, we made use of various existing ontologies and vocabularies.

Post-FAIRification, as proof of concept, we analysed the genotype data using the TWAS pipeline for association analysis. We performed TWAS pipeline on the genotype data of individuals who did not have a BRCA1/BRCA2 mutation and found 14 genes below the Bonferroni threshold associated with breast cancer. We used the 14 genes and the findings of Wu et al. and Ferreira et al. for network analysis. In network analysis, we used the STRING database for identifying the possible interactions. We found that our findings interact with the known set of genes with high network connectivity. Moreover, we wanted to see the role of these genes in various biological processes. Therefore, we performed enrichment analysis on the network and found that 158 genes were involved in various biological processes based on GO terms organelle organisation, cellular localisation, chromatin organisation, and immune response. The top enriched KEGG pathways were viral carcinogenesis, alcoholism, Huntington's disease, pathways in cancer, MicroRNAs in cancer, Colorectal cancer, platinum drug resistance.

Moreover, we computed polygenic risk scores for individuals who were carriers of BRCA1 and BRCA2 mutation vs the non-carrier. We found from the PRS distribution that the risk scores of individuals with BRCA1/2 mutation and non-carriers in controls have similar means. However, cases with the BRCA1/2 mutation had a much higher mean compared to the non-carries cases. Lastly, we looked at the correlation between the risk prediction based on the risk factors and PRS using the CanRisk tool and found that they are very weakly correlated. This shows that the PRS is majorly dependent on the genetic profile.

The most plausible extension of this project would be to host the FAIRified data using a FAIR data point and perform a meta-analysis using a bigger sample size. Studies like this could help in identifying and understanding the risk factors involved in BC development and treatment. The detected genes could have a possible role in immune therapy in cancer and help develop personalized treatment methods.

Bibliography

[1] Miguel Vazquez, Victor de la Torre, and Alfonso Valencia. Chapter 14: Cancer genome analysis. *PLoS Computational Biology*, 8(12), Dec 2012.

[2] Guy Haskin Fernald, Emidio Capriotti, Roxana Daneshjou, Konrad J. Karczewski, and Russ B. Altman. Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13):1741–1748, May 2011.

[3] William S. Bush and Jason H. Moore. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12), Dec 2012.

[4] Montserrat Garcia-Closas, Necdet Burak Gunsoy, and Nilanjan Chatterjee. Combined Associations of Genetic and Environmental Risk Factors: Implications for Prevention of Breast Cancer. *JNCI: Journal of the National Cancer Institute*, 106(11), Nov 2014.

[5] Andrew Lee, Nasim Mavaddat, Amber N. Wilcox, Alex P. Cunningham, Tim Carver, Simon Hartley, Chantal Babb de Villiers, Angel Izquierdo, Jacques Simard, Marjanka K. Schmidt, Fiona M. Walter, Nilanjan Chatterjee, Montserrat Garcia-Closas, Marc Tischkowitz, Paul Pharoah, Douglas F. Easton, and Antonis C. Antoniou. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genetics in Medicine*, 21:1708–1718, Aug 2019.

[6] Clare Turnbull, Amit Sud, and Richard S. Houlston. Cancer genetics, precision prevention and a call to action. *Nature Genetics*, 50(9):1212–1218, August 2018.

[7] Graham A Colditz, Walter C Willett, David J Hunter, Meir J Stampfer, JoAnn E Manson, Charles H Hennekens, Bernard A Rosner, and Frank E Speizer. Family history, age, and risk of breast cancer: prospective data from the nurses' health study. *Jama*, 270(3):338–343, 1993.

[8] Yoshio Miki, Jeff Swensen, Donna Shattuck-Eidens, P Andrew Futreal, Keith Harshman, Sean Tavtigian, Qingyun Liu, Charles Cochran, L Michelle Bennett, Wei Ding, et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182):66–71, 1994.

[9] Richard Wooster, Graham Bignell, Jonathan Lancaster, Sally Swift, Sheila Seal, Jonathan Mangion, Nadine Collins, Simon Gregory, Curtis Gumbs, Gos Micklem, et al. Identification of the breast cancer susceptibility gene brca2. *Nature*, 378(6559):789–792, 1995.

[10] Nazneen Rahman, Sheila Seal, Deborah Thompson, Patrick Kelly, Anthony Renwick, Anna Elliott, Sarah Reid, Katarina Spanova, Rita Barfoot, Tasnim Chagtai, et al. Palb2, which

encodes a brca2-interacting protein, is a breast cancer susceptibility gene. *Nature genetics*, 39(2):165–167, 2007.

[11] Anthony Renwick, Deborah Thompson, Sheila Seal, Patrick Kelly, Tasnim Chagtai, Munaza Ahmed, Bernard North, Hiran Jayatilake, Rita Barfoot, Katarina Spanova, et al. Atm mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature genetics*, 38(8):873–875, 2006.

[12] Hanne Meijers-Heijboer, Ans Van den Ouweland, Jan Klijn, Marijke Wasielewski, Anja de Snoo, Rogier Oldenburg, Antoinette Hollestelle, Mark Houben, Ellen Crepin, Monique van Veghel-Plandsoen, et al. Low-penetrance susceptibility to breast cancer due to chek2* 1100delc in noncarriers of brca1 or brca2 mutations. *Nature genetics*, 31(1), 2002.

[13] K. Michailidou, S. Lindström, J. Dennis, and J. Beesley. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, 11 2017.

[14] R. L. Milne, K. B. Kuchenbaecker, and K. Michailidou. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nature Genetics*, 49(12):1767–1778, Dec 2017.

[15] Lang Wu, Wei Shi, Jirong Long, and Wei Zheng. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature genetics*, 50(7):968–978, Jul 2018.

[16] Tim Carver, Simon Hartley, Andrew Lee, Alex P. Cunningham, Stephanie Archer, Chantal Babb de Villiers, Jonathan Roberts, Rod Ruston, Fiona M. Walter, Marc Tischkowitz, Douglas F. Easton, and Antonis C. Antoniou. CanRisk tool—a web interface for the prediction of breast and ovarian cancer risk and the likelihood of carrying genetic pathogenic variants. *Cancer Epidemiology Biomarkers & Prevention*, 30(3):469–473, December 2020.

[17] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), March 2016.

[18] Charles Vesteghem, Rasmus Froberg Brøndum, Mads Sønderkær, Mia Sommer, Alexander Schmitz, Julie Støve Bødker, Karen Dybkær, Tarec Christoffer El-Galaly, and Martin Bøgsted. Implementing the FAIR data principles in precision oncology: review of supporting initiatives. *Briefings in Bioinformatics*, 21(3):936–945, June 2019.

[19] A. Anil Sinaci, Francisco J. Núñez-Benjumea, Mert Gencturk, Malte-Levin Jauer, Thomas Deserno, Catherine Chronaki, Giorgio Cangioli, Carlos Cavero-Barca, Juan M. Rodríguez-Pérez, Manuel M. Pérez-Pérez, Gokce B. Laleci Erturkmen, Tony Hernández-Pérez, Eva Méndez-Rodríguez, and Carlos L. Parra-Calderón. From raw data to FAIR data: The FAIRification workflow for health research. *Methods of Information in Medicine*, 59(S 01):e21–e32, June 2020.

[20] Annika Jacobsen, Rajaram Kaliyaperumal, Luiz Olavo Bonino da Silva Santos, Barend Mons, Erik Schultes, Marco Roos, and Mark Thompson. A generic workflow for the data FAIRification process. *Data Intelligence*, 2(1-2):56–65, January 2020.

[21] Annika Jacobsen, Ricardo deMiranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T. Evelo, Carole Goble, Giancarlo Guizzardi, Karsten Kryger Hansen, Ali Hasnain, Kristina Hettne, Jaap Heringa, Rob W.W. Hooft, Melanie Imming, Keith G. Jeffery, Rajaram Kaliyaperumal, Martijn G. Kersloot, Christine R. Kirkpatrick, Tobias Kuhn, Ignasi Labastida, Barbara Magagna, Peter McQuilton, Natalie Meyers, Annalisa Montesanti, Mirjam van Reisen, Philippe Rocca-Serra, Robert Pergl, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Juliane Schneider, George Strawn, Mark Thompson, Andra Waagmeester, Tobias Weigel, Mark D. Wilkinson, Egon L. Willighagen, Peter Wittenburg, Marco Roos, Barend Mons, and Erik Schultes. FAIR principles: Interpretations and implementation considerations. *Data Intelligence*, 2(1-2):10–29, January 2020.

[22] Mark Thompson, Kees Burger, Rajaram Kaliyaperumal, Marco Roos, and Luiz Olavo Bonino da Silva Santos. Making FAIR easy with FAIR tools: From creolization to convergence. *Data Intelligence*, 2(1-2):87–95, January 2020.

[23] R.D. Kush, D. Warzel, M.A. Kush, A. Sherman, E.A. Navarro, R. Fitzmartin, F. Pétavy, J. Galvez, L.B. Becnel, F.L. Zhou, N. Harmon, B. Jauregui, T. Jackson, and L. Hudson. FAIR data sharing: The roles of common data elements and harmonization. *Journal of Biomedical Informatics*, 107:103421, July 2020.

[24] Hansi Zhang, Yi Guo, Qian Li, Thomas J. George, Elizabeth Shenkman, François Modave, and Jiang Bian. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Medical Informatics and Decision Making*, 18(S2), July 2018.

[25] Jennifer Golbeck, Gilberto Fragoso, Frank Hartel, Jim Hendler, Jim Oberthaler, and Bijan Parsia. The national cancer institute's thesaurus and ontology. *SSRN Electronic Journal*, 2003.

[26] A. T. Marees, H. de Kluiver, S. Stringer, F. Vorspan, E. Curis, C. Marie-Claire, and E. M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):e1608, 06 2018.

[27] Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, Oct 2010.

[28] International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, Oct 2005.

[29] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322, Sep 1995.

[30] D. Ford, D.F. Easton, M. Stratton, S. Narod, D. Goldgar, P. Devilee, D.T. Bishop, B. Weber, G. Lenoir, J. Chang-Claude, H. Sobol, M.D. Teare, J. Struewing, A. Arason, S. Scherneck, J. Peto, T.R. Rebbeck, P. Tonin, S. Neuhausen, R. Barkardottir, J. Eyfjord, H. Lynch, B.A.J. Ponder, S.A. Gayther, J.M. Birch, A. Lindblom, D. Stoppa-Lyonnet, Y. Bignon, A. Borg, U. Hamann, N. Haites, R.J. Scott, C.M. Maugard, H. Vasen, S. Seitz, L.A. Cannon-Albright, A. Schofield, and M. Zelada-Hedman. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *The American Journal of Human Genetics*, 62(3):676–689, March 1998.

[31] E. Reed, S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes. A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, 34(28):3769–3792, Dec 2015.

[32] Y. Hochberg and Y. Benjamini. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7):811–818, Jul 1990.

[33] G. Gibson. Hints of hidden heritability in GWAS. *Nature Genetics*, 42(7):558–560, Jul 2010.

[34] F. Dudbridge. Power and predictive accuracy of polygenic risk scores. *PLoS Genetics*, 9(3):e1003348, Mar 2013.

[35] Robert P. Igo, Tyler G. Kinzy, and Jessica N. Cooke Bailey. Genetic risk scores. *Current Protocols in Human Genetics*, 104(1), November 2019.

[36] Manuel A. Ferreira, , Eric R. Gamazon, Fares Al-Ejeh, and Kristiina Aittomäki. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nature Communications*, 10(1), April 2019.

[37] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A.R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I.W. de Bakker, Mark J. Daly, and Pak C. Sham. PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, September 2007.

[38] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, Richard Durbin, and Alkes L Price. Reference-based phasing using the haplotype reference consortium panel. *Nature Genetics*, 48(11):1443–1448, October 2016.

[39] Sayantan Das, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, Emily Y Chew, Shawn Levy, Matt McGue, David Schlessinger, Dwight Stambolian, Po-Ru Loh, William G Iacono, Anand Swaroop, Laura J Scott, Francesco Cucca,

Florian Kronenberg, Michael Boehnke, Gonçalo R Abecasis, and Christian Fuchsberger. Next-generation genotype imputation service and methods. *Nature Genetics*, 48(10):1284–1287, August 2016.

[40] A reference panel of 64, 976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279–1283, August 2016.

[41] Eric R Gamazon, , Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, August 2015.

[42] Andrea Franceschini, Damian Szklarczyk, Sune Frankild, Michael Kuhn, Milan Simonovic, Alexander Roth, Jianyi Lin, Pablo Minguez, Peer Bork, Christian Von Mering, et al. String v9. 1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, 41(D1):D808–D815, 2012.

[43] Michel Dumontier, Christopher JO Baker, Joachim Baran, Alison Callahan, Leonid Chepelev, José Cruz-Toledo, Nicholas R Del Rio, Geraint Duck, Laura I Furlong, Nichealla Keath, Dana Klassen, James P McCusker, Núria Queralt-Rosinach, Matthias Samwald, Natalia Villanueva-Rosales, Mark D Wilkinson, and Robert Hoehndorf. The semanticscience integrated ontology (SIO) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, 5(1):14, 2014.

[44] Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, November 2014.

[45] Shaker El-Sappagh, Francesco Franda, Farman Ali, and Kyung-Sup Kwak. SNOMED CT standard ontology based on the ontology for general medical science. *BMC Medical Informatics and Decision Making*, 18(1), August 2018.

[46] Karen Eilbeck, Suzanna E Lewis, Christopher J Mungall, Mark Yandell, Lincoln Stein, Richard Durbin, and Michael Ashburner. The sequence ontology: a tool for the unification of genome annotations. *Genome Biology*, 6(5):R44, 2005.

[47] Sharon E. Plon, Diana M. Eccles, Douglas Easton, William D. Foulkes, Maurizio Genuardi, Marc S. Greenblatt, Frans B.L. Hogervorst, Nicoline Hoogerbrugge, Amanda B. Spurdle, and Sean V. Tavtigian and. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Human Mutation*, 29(11):1282–1291, November 2008.

[48] Angela Cox, , Alison M Dunning, Montserrat Garcia-Closas, Sabapathy Balasubramanian, Malcolm W R Reed, Karen A Pooley, Serena Scollen, Caroline Baynes, Bruce A J Ponder, Stephen Chanock, Jolanta Lissowska, Louise Brinton, Beata Peplonska, Melissa C Southey, John L Hopper, Margaret R E McCredie, Graham G Giles, Olivia Fletcher, Nichola Johnson, Isabel dos Santos Silva, Lorna Gibson, Stig E Bojesen, Børge G Nordestgaard, Christen K Axelsson, Diana Torres, Ute Hamann, Christina Justenhoven, Hiltrud Brauch, Jenny Chang-Claude, Silke Kropp, Angela Risch, Shan Wang-Gohrke, Peter Schürmann, Natalia Bogdanova, Thilo Dörk,

Rainer Fagerholm, Kirsimari Aaltonen, Carl Blomqvist, Heli Nevanlinna, Sheila Seal, Anthony Renwick, Michael R Stratton, Nazneen Rahman, Suleeporn Sangrajrang, David Hughes, Fabrice Odefrey, Paul Brennan, Amanda B Spurdle, Georgia Chenevix-Trench, Jonathan Beesley, Arto Mannermaa, Jaana Hartikainen, Vesa Kataja, Veli-Matti Kosma, Fergus J Couch, Janet E Olson, Ellen L Goode, Annegien Broeks, Marjanka K Schmidt, Frans B L Hogervorst, Laura J Van't Veer, Daehee Kang, Keun-Young Yoo, Dong-Young Noh, Sei-Hyun Ahn, Sara Wedrén, Per Hall, Yen-Ling Low, Jianjun Liu, Roger L Milne, Gloria Ribas, Anna Gonzalez-Neira, Javier Benitez, Alice J Sigurdson, Denise L Stredrick, Bruce H Alexander, Jeffery P Struewing, Paul D P Pharoah, and Douglas F Easton and. A common coding variant in CASP8 is associated with breast cancer risk. *Nature Genetics*, 39(3):352–358, February 2007.

[49] Peian Cai, Zhenhui Lu, Jianjun Wu, Xiong Qin, Zetao Wang, Zhi Zhang, Li Zheng, and Jinmin Zhao. BTN3a2 serves as a prognostic marker and favors immune infiltration in triple-negative breast cancer. *Journal of Cellular Biochemistry*, 121(3):2643–2654, November 2019.

Appendix A

Semantic model

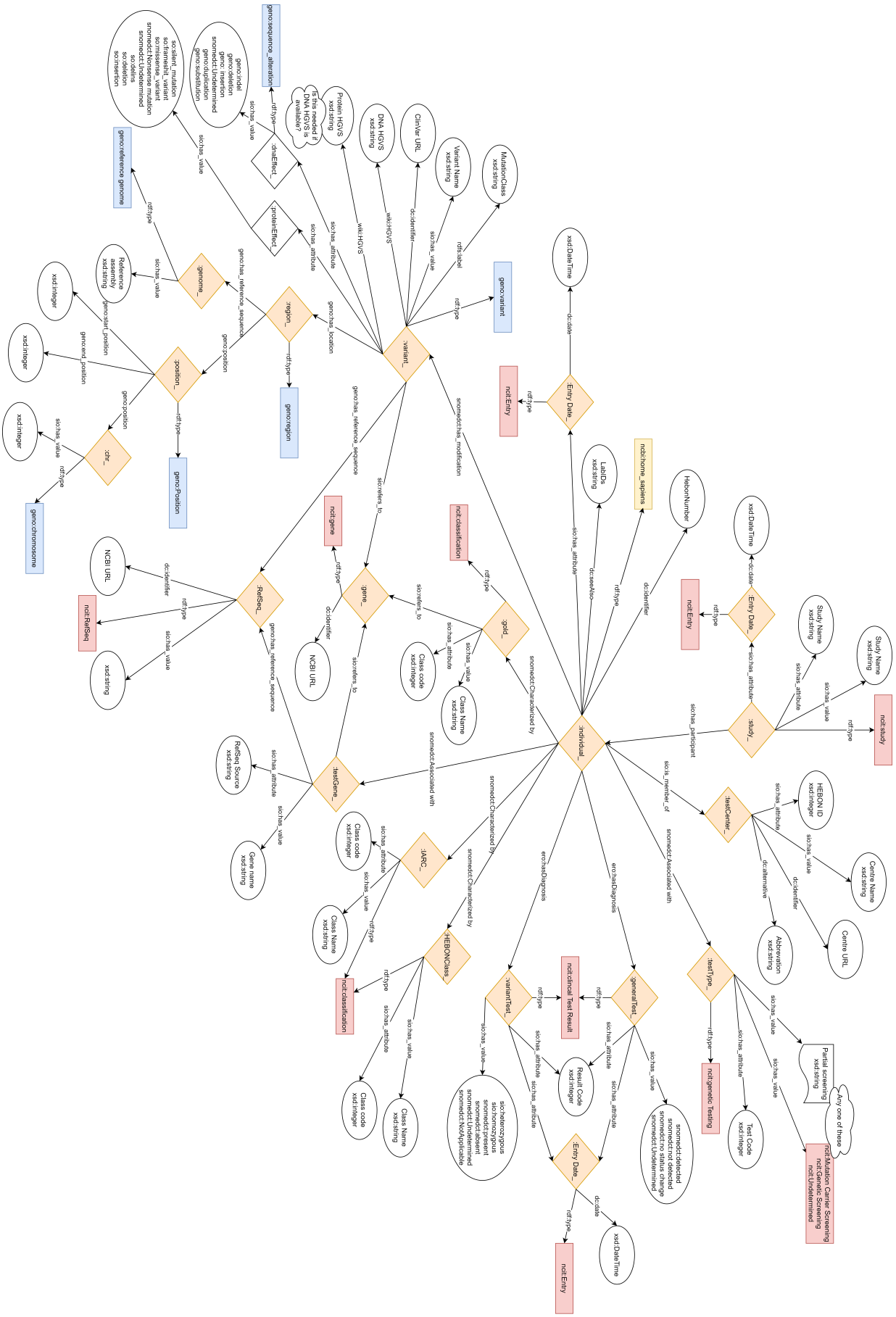


Figure A.1: Semantic model representing the HEBON-DNA data

Appendix B

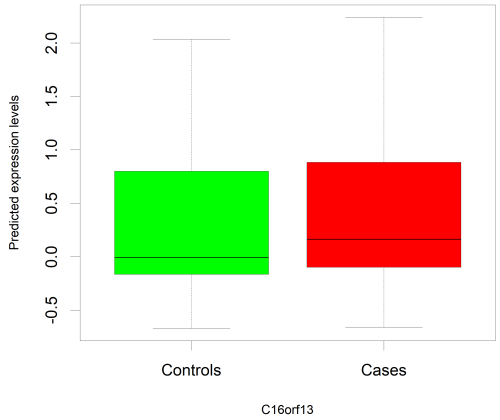
Associated genes

gene_name	**zscore**	**effect_size**	**pvalue**	**pred_r2**	**pred_pval**	pred_qval	n_snps_used	n_snps_ in_model	CHR
HLA-DQA2	4.74081	2.495364	2.13E-06	0.508558	9.89E-30	5.95E-28	25	64	6
WDR90	3.947492	0.620137	7.90E-05	0.122929	1.14E-06	3.36E-06	13	30	16
TRIM4	3.696812	0.277719	0.000218	0.440706	1.29E-24	3.97E-23	43	53	7
MAP1LC3A	-3.54137	-0.47661	0.000398	0.158839	2.30E-08	9.26E-08	22	24	20
C16orf13	3.440602	0.603253	0.00058	0.437531	2.15E-24	6.30E-23	18	53	16
PSG4	3.418645	1.822694	0.000629	0.305343	5.07E-16	5.58E-15	4	35	19
TMEM80	-3.34024	-0.27934	0.000837	0.428118	9.77E-24	2.79E-22	24	45	11
CASP8	3.287565	0.404776	0.001011	0.232872	4.59E-12	3.21E-11	14	21	2
BTN3A2	-3.28176	-0.20309	0.001032	0.417555	5.18E-23	1.28E-21	95	98	6
OR2AE1	-3.12497	-0.51987	0.001778	0.210671	6.37E-11	3.82E-10	6	18	7
PSMG1	3.057589	0.376174	0.002231	0.358447	3.51E-19	5.54E-18	25	31	21
ALG11	-2.99074	-0.43916	0.002783	0.197151	3.06E-10	1.67E-09	10	12	13
RBL2	2.97243	0.259059	0.002955	0.33747	6.65E-18	8.98E-17	27	31	16
EFCAB13	2.872187	0.4434	0.004076	0.196594	3.26E-10	1.77E-09	17	37	17

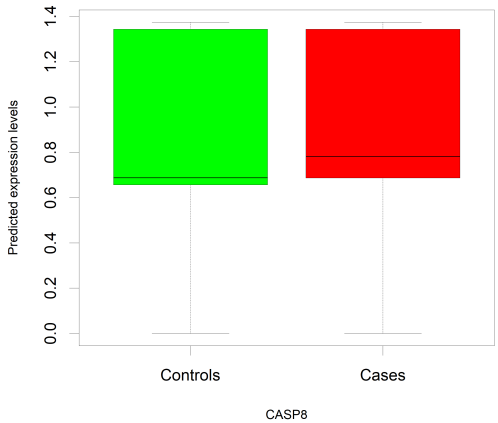
Table B.1: Genes identified to be associated with BC after predicting the gene expression using PrediXcan. $Pred\_R^2$ represents the prediction quality, *n snps used* represents the SNPs from the samples that matched with the SNPs in model.

Appendix C

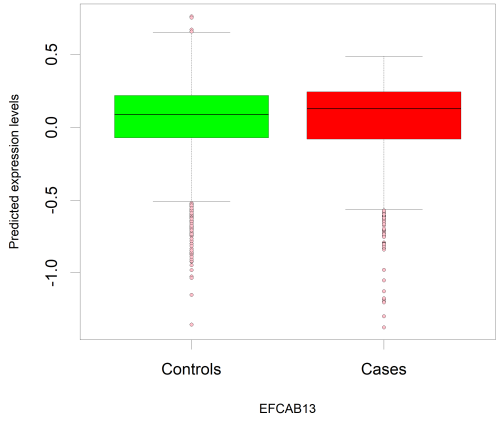
Predicted expression levels



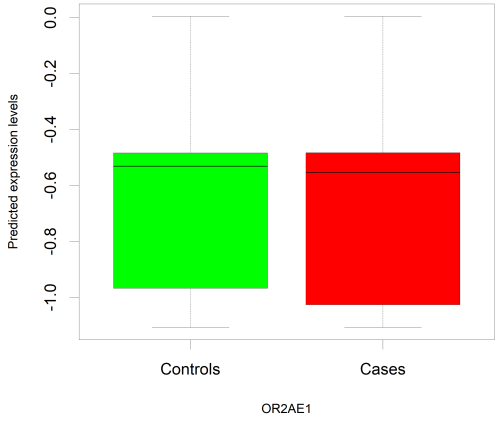
(a) C16orf13



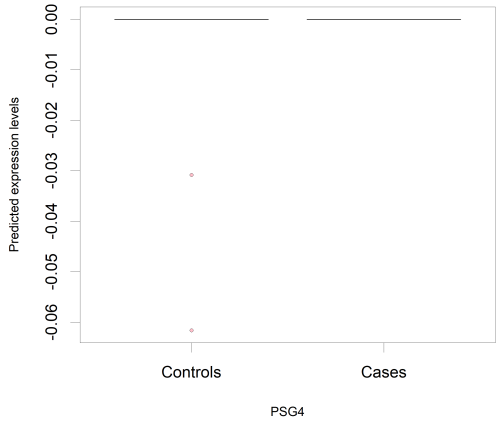
(b) CASP8



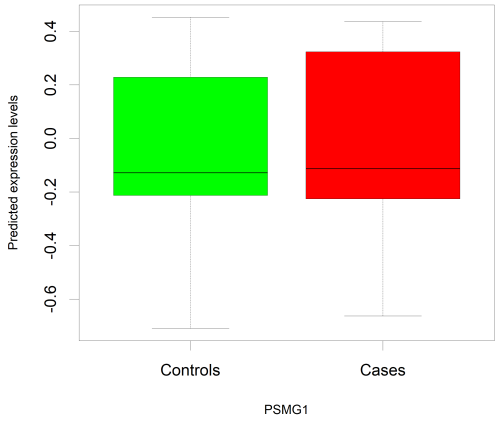
(c) EFCAB13



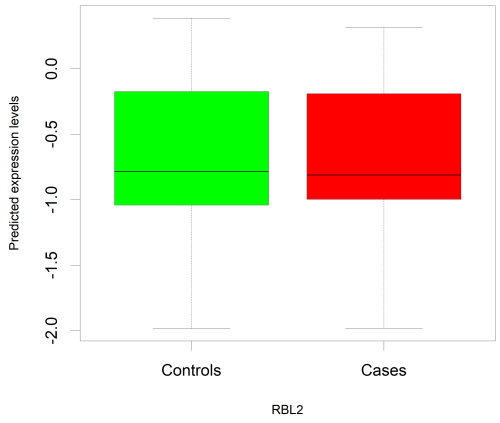
(d) OR2AE1



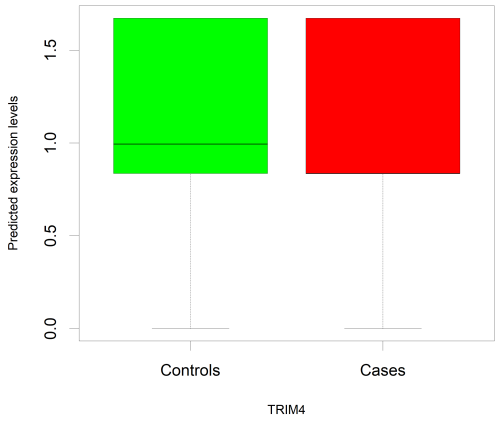
(e) PSG4



(f) PSMG1



(g) RBL2



(h) TRIM4

Figure C.1: Predicted gene expression levels in controls and cases.

Appendix D

Network analysis

Genes	Betweenness Centrality	Closeness Centrality	Clustering Coefficient	Degree	Eccentricity	Neighborhood Connectivity
ACTR10	0.010931	1	0.904762	15	1	13.66667
ACTR1A	0.010931	1	0.904762	15	1	13.66667
AHI1	0	1	0	1	1	1
AKT1	0.014229	0.741935	0.838095	15	2	14.53333
ALB	0.091992	0.766667	0.783333	16	2	14
ALG11	0	0.666667	0	1	2	2
ALK	0	0.638889	1	11	3	16
ALS2CL	0	1	0	1	1	1
ALS2CR12	0	0.442308	0	1	3	16
ANKRD34A	0	1	0	1	1	1
APOBEC3A	0	0.6	1	4	2	10.5
APRG1	0	0.666667	0	1	2	2
ARHGEF19	0	1	0	1	1	1
ATF4	0	0.638889	1	10	2	15.8
ATG10	0	1	1	3	1	3
ATG5	0	1	1	3	1	3
ATG7	0	1	1	3	1	3
ATM	0.027099	0.741935	0.780952	15	2	14.06667
ATP5A1	0.030465	0.923077	0.818182	11	2	9.545455
ATP5C1	0.077183	1	0.727273	12	1	9
ATP5G1	0.030465	0.923077	0.818182	11	2	9.545455
ATP5H	0.040314	0.923077	0.8	11	2	9.454545
ATP5I	0.008748	0.857143	0.911111	10	2	10.1
ATP5O	0.008748	0.857143	0.911111	10	2	10.1
ATP6AP1L	0	0.631579	1	5	2	10.4
ATP6V0A1	0.009091	0.75	0.892857	8	2	10
ATP6V0C	0.020996	0.857143	0.844444	10	2	9.7

Table D.1 continued from previous page

AZGP1	0	0.442308	0	1	3	16
B3GALNT2	0	1	1	3	1	3
B3GNT1	0	1	1	3	1	3
BAZ1B	0	0.695652	1	18	2	31.61111
BBS2	0	0.6	0	1	2	3
BDH2	0	0	0	0	0	0
BRCA1	0.024925	0.741935	0.790476	15	2	14.13333
BRMS1	0	0.744186	1	21	2	31.42857
BTN3A1	0	1	1	2	1	2
BTN3A2	0	1	1	2	1	2
BTN3A3	0	1	1	2	1	2
C14orf79	0	1	1	2	1	2
C16orf13	0	0.666667	0	1	2	2
C17orf105	0	1	1	3	1	3
C1orf189	0	0.555556	0	1	2	5
C21orf58	0	1	0	1	1	1
C5orf52	0	1	1	3	1	3
C5orf56	0	1	0	1	1	1
C6orf163	0	1	0	1	1	1
C9orf3	0	0	0	0	0	0
CANX	1	1	0	2	1	1
CASP3	0.105826	0.793103	0.713235	17	2	13.35294
CASP8	0.102372	0.766667	0.725	16	2	13.5625
CBX8	6.41E-04	0.842105	0.978462	26	2	30.69231
CBY1	0	1	0	1	1	1
CCBL2	0	1	0	1	1	1
CCDC110	0	0.571429	0	1	2	4
CCDC13	0	0.833333	1	4	2	4.25
CCDC14	0	0.833333	1	4	2	4.25
CCDC18	0.4	1	0.6	5	1	3.4
CCDC66	0	0.833333	1	4	2	4.25
CCT2	1	1	0	3	1	1
CDC20	0	0.875	1	12	2	12.83333
CDC42	0	1	1	2	1	2
CDRT4	0	1	0	1	1	1
CEBPB	0	1	0	1	1	1
CEP131	0	1	0	1	1	1
CEP63	0	1	0	1	1	1
CLEC18A	0	0	0	0	0	0
CMTR2	0	0.565217	1	3	2	13
CNEP1R1	0	0.571429	0	1	2	4
CPLX1	0.666667	1	0.333333	3	1	1.666667

Table D.1 continued from previous page

CPLX4	0	0.6	0	1	2	3
CPNE1	0	1	0	1	1	1
CREBBP	0.00109	0.914286	0.970443	29	2	30.41379
CRHR1	0	1	0	1	1	1
CRIP2	0	0	0	0	0	0
CTNNB1	1	1	0	2	1	1
CUL1	0.049712	1	0.813187	14	1	11.57143
CUL3	0.049712	1	0.813187	14	1	11.57143
CYC1	0.025884	0.857143	0.844444	10	2	9.7
CYCS	0.018939	0.8	0.861111	9	2	9.777778
DCTN1	0.010931	1	0.904762	15	1	13.66667
DCTN2	0.010931	1	0.904762	15	1	13.66667
DCTN3	0.010931	1	0.904762	15	1	13.66667
DCTN4	0.010931	1	0.904762	15	1	13.66667
DCTN5	0.010931	1	0.904762	15	1	13.66667
DCTN6	0.010931	1	0.904762	15	1	13.66667
DFNA5	0	0.479167	0	1	3	20
DGKQ	0	1	0	1	1	1
DPPA4	0	0	0	0	0	0
DYNC1H1	0.002597	0.9375	0.967033	14	2	14.14286
DYNC1I2	0.002597	0.9375	0.967033	14	2	14.14286
DYNLL1	0.002597	0.9375	0.967033	14	2	14.14286
DZIP1L	0	0	0	0	0	0
ECT2L	0	0.56	1	3	2	14
EDEM2	0	1	0	1	1	1
EFCAB13	0	1	0	1	1	1
EGFR	0.002758	0.71875	0.934066	14	2	15.35714
EIF4E	0	1	0	1	1	1
EP300	0.003495	0.969697	0.926882	31	2	29.64516
ESR1	0.00109	0.914286	0.970443	29	2	30.41379
EZH2	6.65E-04	0.914286	0.980296	29	2	30.55172
FAM122A	0	0.571429	0	1	2	4
FAM149B1	0	1	0	1	1	1
FAT4	0	0.666667	0	1	2	2
FES	0	0.666667	1	2	2	4
FLOT1	0	0.666667	0	1	2	2
GALNT16	0	0.666667	0	1	2	2
GAPDH	0.005035	0.741935	0.895238	15	2	14.93333
GGH	0	1	1	3	1	3
GOSR1	0	0.789474	1	11	2	14.72727
GPR144	0	0	0	0	0	0
GPR156	0	1	0	1	1	1

Table D.1 continued from previous page

GPRC5D	0	1	0	1	1	1
GSTM3	0	1	1	3	1	3
GSTM4	0	1	1	3	1	3
GSTM5	0	1	1	3	1	3
GYLTL1B	0	1	1	3	1	3
H2AFX	0.002919	0.969697	0.937634	31	2	29.80645
HAPLN4	0	1	0	1	1	1
HDAC1	0.003274	0.969697	0.931183	31	2	29.70968
HDAC2	0.001773	0.941176	0.956322	30	2	30.13333
HDAC3	0.001544	0.941176	0.96092	30	2	30.2
HIST1H2AC	0.003019	0.969697	0.935484	31	2	29.77419
HIST1H2AD	0.003019	0.969697	0.935484	31	2	29.77419
HIST1H2AJ	0.003019	0.969697	0.935484	31	2	29.77419
HIST1H2BA	0.004043	1	0.919355	32	1	29.5
HIST1H2BB	0.004043	1	0.919355	32	1	29.5
HIST1H2BD	0.004043	1	0.919355	32	1	29.5
HIST1H2BH	0.004043	1	0.919355	32	1	29.5
HIST1H2BJ	0.004043	1	0.919355	32	1	29.5
HIST1H2BK	0.004043	1	0.919355	32	1	29.5
HIST1H2BL	0.004043	1	0.919355	32	1	29.5
HIST1H2BM	0.004043	1	0.919355	32	1	29.5
HIST1H2BN	0.004043	1	0.919355	32	1	29.5
HIST1H2BO	0.004043	1	0.919355	32	1	29.5
HIST2H2AC	0.003019	0.969697	0.935484	31	2	29.77419
HIST2H2BE	0.004043	1	0.919355	32	1	29.5
HLA-DQA2	0	0.882353	1	13	2	14.46154
HOMER3	0	1	0	1	1	1
HRAS	0.041065	0.766667	0.716667	16	2	13.4375
HSF2	0	0.666667	1	2	2	4
HSP90AA1	0.25	1	0.5	4	1	2.5
HSPA4	0.25	1	0.5	4	1	2.5
HSPA5	3.95E-04	0.657143	0.981818	11	2	15.54545
HSPA8	0.001425	0.8125	0.977778	10	2	11.6
INS	0	1	0	1	1	1
IST1	0	0.6	0	1	2	3
JUN	0.001418	0.941176	0.963218	30	2	30.23333
KDM6B	2.33E-04	0.888889	0.992063	28	2	30.82143
KEAP1	0	0.875	1	12	2	12.83333
KIAA1377	0	0.833333	1	4	2	4.25
KIF3A	0	0.882353	1	13	2	14.46154
KLC1	0	0.882353	1	13	2	14.46154
KLF5	0	0.511111	1	4	3	17.25

Table D.1 continued from previous page

KLHDC10	0	1	0	1	1	1
KLHDC7A	0	1	0	1	1	1
KLHL5	0	0.875	1	12	2	12.83333
L3MBTL3	0	1	0	1	1	1
LMO4	0	1	0	1	1	1
LRRC37A	0	1	0	1	1	1
LRRC37A2	0	1	0	1	1	1
LRRC3B	1	1	0	2	1	1
MAN2C1	0	1	0	1	1	1
MAP1LC3A	0	1	1	3	1	3
MAPK8	0.005035	0.741935	0.895238	15	2	14.93333
METTL10	0	0.666667	0	1	2	2
MMP24	0	1	0	1	1	1
MTHFD1L	0	1	0	1	1	1
MUTYH	0	0.511111	1	4	3	16.5
MYC	0.034153	0.793103	0.742647	17	2	13.58824
MYRF	0	0	0	0	0	0
NCBP1	0.004452	0.866667	0.945455	11	2	11.36364
NCBP2	0.004452	0.866667	0.945455	11	2	11.36364
NDUFS7	1	1	0	2	1	1
NR1H3	0	1	0	1	1	1
NUDT17	0	1	0	1	1	1
NUP107	0.001425	0.8125	0.977778	10	2	11.7
OGFOD3	0	1	0	1	1	1
OR1E2	0	1	1	2	1	2
OR2AE1	0	1	0	1	1	1
OXLD1	0	1	0	1	1	1
PACS1	0	1	0	1	1	1
PAIP1	0	1	0	1	1	1
PDLIM4	0	0.45098	0	1	3	17
PIDD1	0	0.534884	1	4	3	17
PILRA	0	1	0	1	1	1
PILRB	0	1	0	1	1	1
PLEKHD1	1	1	0	2	1	1
POLR2J	0	0.8	1	24	2	31.125
PPFIA1	0	0.75	1	2	2	2.5
PPFIA2	0	0.75	1	2	2	2.5
PPP2R1A	1	1	0	4	1	1
PRADC1	0	0.555556	0	1	2	5
PRSS46	0	1	0	1	1	1
PSG1	0.05	0.714286	0.666667	3	2	3
PSG11	0	0.555556	0	1	2	5

Table D.1 continued from previous page

PSG4	0.75	1	0.2	5	1	1.8
PSG6	0	0.625	1	2	2	4
PSG8	0	0.625	1	2	2	4
PSMA1	1	1	0	2	1	1
PSMG1	0	0.666667	0	1	2	2
PSORS1C1	0	1	0	1	1	1
PSORS1C2	0	1	0	1	1	1
PTDSS2	0	0	0	0	0	0
PTPN11	0	1	0	1	1	1
RAB7A	0	1	0	1	1	1
RAN	0	1	0	1	1	1
RBBP4	0.002423	0.969697	0.944086	31	2	29.90323
RBBP7	0.002423	0.969697	0.944086	31	2	29.90323
RBL2	0	0.522727	1	5	3	16.6
RBX1	0.009419	0.933333	0.923077	13	2	12.30769
RCCD1	0	1	0	1	1	1
RHOA	0	1	1	2	1	2
RHOD	0	1	1	2	1	2
RIC8A	0	0.681818	1	8	2	15
RMND1	0	1	0	1	1	1
RPLP2	0	0.764706	1	9	2	11.88889
RPS27A	0.009419	0.933333	0.923077	13	2	12.30769
SAMD13	0	1	0	1	1	1
SERTAD4	0	0.571429	0	1	2	4
SH3TC2	0	1	0	1	1	1
SIKE1	0	0.571429	0	1	2	4
SIRT1	8.74E-04	0.914286	0.975369	29	2	30.48276
SKP1	0.049712	1	0.813187	14	1	11.57143
SLC22A5	0	1	0	1	1	1
SLC39A9	0	0.666667	0	1	2	2
SMIM7	0	0.666667	1	2	2	3
SMIM8	0	1	0	1	1	1
SMN2	0	0.684211	1	7	2	12.14286
SNRPD1	0.01422	0.928571	0.878788	12	2	10.91667
SNRPD2	0.01422	0.928571	0.878788	12	2	10.91667
SNRPD3	0.01422	0.928571	0.878788	12	2	10.91667
SNRPE	0.056955	1	0.769231	13	1	10.23077
SNRPF	0.056955	1	0.769231	13	1	10.23077
SNRPG	0.056955	1	0.769231	13	1	10.23077
SNUPN	0.005495	0.8125	0.933333	10	2	11.4
SNX32	0	0	0	0	0	0
SPANXN1	0	0.571429	0	1	2	4

Table D.1 continued from previous page

SPATA18	0	1	0	1	1	1
STAT3	0.020516	0.766667	0.808333	16	2	14.1875
STXBP4	0	1	0	1	1	1
SUMO1	2.33E-04	0.761905	0.987013	22	2	30.90909
SYTL3	0	0.666667	1	2	2	3
TBC1D32	0	1	0	1	1	1
TBX5	0	0.666667	0	1	2	2
TM6SF1	0	1	0	1	1	1
TM6SF2	0	1	0	1	1	1
TMC4	0	1	0	1	1	1
TMCO1	0	1	1	2	1	2
TMEM136	0	1	1	3	1	3
TMEM42	0	1	1	3	1	3
TMEM5	0	1	1	3	1	3
TMEM80	0	1	0	1	1	1
TP53	0.168869	0.884615	0.563158	20	2	11.85
TRIM4	0	0.875	1	12	2	12.83333
TRIOBP	0	1	1	2	1	2
TSPAN5	0	0	0	0	0	0
UBA52	0.009419	0.933333	0.923077	13	2	12.30769
UBB	0	0.875	1	12	2	12.83333
UBC	0	0.875	1	12	2	12.83333
UBD	0	0.666667	1	7	2	13.42857
UBE2C	0.009419	0.933333	0.923077	13	2	12.30769
UBLCP1	0	0.666667	0	1	2	2
UQCRH	0.001894	0.8	0.972222	9	2	10.44444
USP19	0	0.666667	1	2	2	4
WDR90	0	0.6	0	1	2	3
YBEY	0	1	0	1	1	1
ZFYVE21	0	1	1	2	1	2
ZNF165	0.833333	1	0.166667	4	1	1.5
ZNF334	0	0	0	0	0	0
ZNF404	0	1	0	1	1	1
ZNF735P	0	1	0	1	1	1
ZNF839	0	1	1	2	1	2
ZSWIM5	0	0	0	0	0	0

Table D.1: Detailed Network statistics of each node in the STRING network.

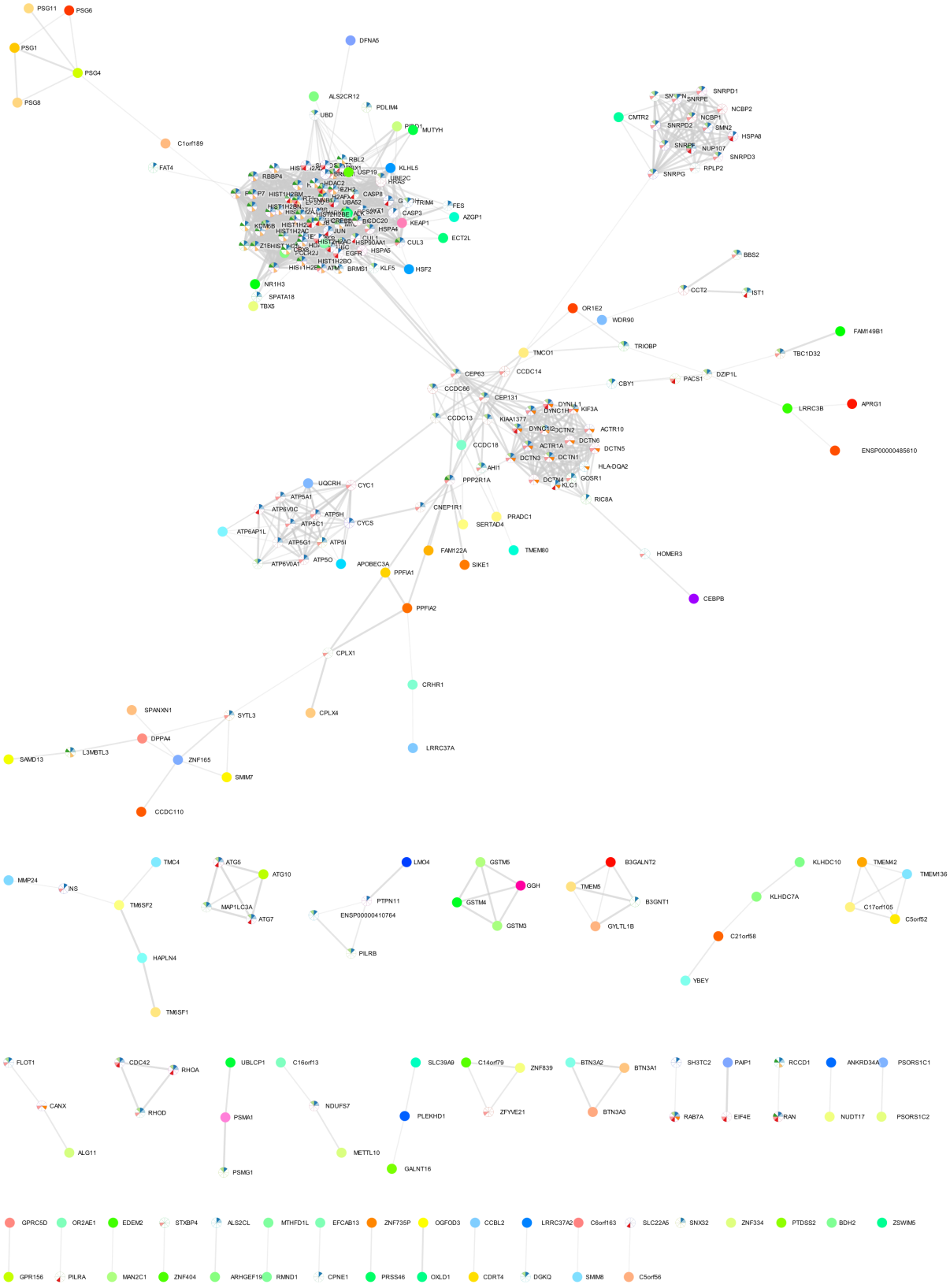


Figure D.1: String interaction network of 275 genes, after expanding using a cutoff of 0.7.

Appendix E

Enrichment analysis

description	**genes**
organelle organization	SIRT1, RBX1, GOSR1, NUP107, GAPDH, SKP1, NDUFS7, BBS2, PDLIM4, ACTR10, KDM6B, DCTN3, RBL2, CREBBP, EP300, KIAA1377, CUL3, ATP6V0A1, STAT3, RAB7A, TP53, CBX8, RPS27A, HIST1H2BA, ATM, HIST1H2BD, ATP5O, WDR90, SPATA18, ATP5H, HSPA4, HDAC3, UBB, ATP5I, CYCS, RHOD, CCDC13, ALS2CL, EZH2, PPP2R1A, HIST1H2AJ, ATP6V0C, FES, DZIP1L, HIST2H2AC, HSP90AA1, CEP63, HIST1H2AD, BAZ1B, CTNNB1, ATG7, UBE2C, DYNC1H1, ATP5C1, HIST1H2BK, CASP8, DCTN1, AHI1, ATG5, HIST2H2BE, ACTR1A, JUN, CDC20, RBBP4, HDAC1, MAP1LC3A, HIST1H2BL, HIST1H2AC, RBBP7, SUMO1, DYNLL1, ATP5G1, RCCD1, CCDC66, MAPK8, DYNC1I2, TMEM80, TBC1D32, ATP5A1, CDC42, TRIOBP, KIF3A, UBA52, CEP131, BRMS1, RHOA, ESR1, CNEP1R1, DCTN2, KLC1, BRCA1, HDAC2, L3MBTL3, H2AFX, IST1, UBC, RAN, AKT1, HIST1H2BJ, HIST1H2BO, HIST1H2BM, HIST1H2BH, MYC, HIST1H2BB, HIST1H2BN
cellular component assembly	SIRT1, SNRPD3, RBX1, NUP107, SKP1, NDUFS7, BBS2, DCTN3, CREBBP, EP300, KIAA1377, CUL3, RAB7A, SNRPF, TP53, RPS27A, SNRPG, DGKQ, HIST1H2BA, HIST1H2BD, WDR90, CCT2, SNRPD1, HSPA4, HDAC3, UBB, RHOD, CCDC13, PPP2R1A, CUL1, PSMG1, DZIP1L, HSP90AA1, CEP63, SNRPD2, CTNNB1, ATG7, UBE2C, DYNC1H1, HIST1H2BK, CASP8, DCTN1, AHI1, ATG5, HIST2H2BE, ACTR1A, CDC20, RBBP4, HDAC1, MAP1LC3A, NCBP1, FLOT1, UBD, HIST1H2BL, KLF5, RBBP7, SMN2, DYNLL1, CCDC66, DYNC1I2, TMEM80, TBC1D32, CDC42, KIF3A, UBA52, PILRB, CEP131, RHOA, SNRPE, DCTN2, ENSP00000410764, H2AFX, IST1, UBC, SNUPN, HIST1H2BJ, HIST1H2BO, HIST1H2BM, CBY1, HIST1H2BH, MYC, HIST1H2BB, HIST1H2BN

Table E.1 continued from previous page

description	**genes**
chromosome organization	SIRT1, RBX1, NUP107, SKP1, KDM6B, RBL2, CREBBP, EP300, CUL3, TP53, CBX8, RPS27A, HIST1H2BA, ATM, HIST1H2BD, HDAC3, UBB, EZH2, PPP2R1A, HIST1H2AJ, HIST2H2AC, HSP90AA1, HIST1H2AD, BAZ1B, CTNNB1, HIST1H2BK, HIST2H2BE, CDC20, RBBP4, HDAC1, HIST1H2BL, HIST1H2AC, RBBP7, RCCD1, UBA52, BRMS1, ESR1, DCTN2, HDAC2, L3MBTL3, H2AFX, UBC, RAN, HIST1H2BJ, HIST1H2BO, HIST1H2BM, HIST1H2BH, MYC, HIST1H2BB, HIST1H2BN
cellular localization	SNRPD3, DCTN6, GOSR1, NUP107, SKP1, BBS2, CANX, PPFIA1, ACTR10, DCTN3, GGH, CUL3, ATP6V0A1, STAT3, RAB7A, SNRPF, TP53, RPS27A, SNRPG, EGFR, ATM, ATP5O, ALB, CCT2, CPLX4, DCTN5, SNRPD1, ATP5H, HSPA4, UBB, CPLX1, ATP5I, TSPAN5, RHOD, SNX32, PACS1, CYC1, CPNE1, EZH2, RPLP2, HSPA5, PPP2R1A, NCBP2, ATP6V0C, DZIP1L, HSP90AA1, CEP63, SNRPD2, CTNNB1, ATG7, DYNC1H1, ATP5C1, DCTN1, AHI1, ACTR1A, NCBP1, STXBP4, FLOT1, KLF5, SMN2, SUMO1, DYNLL1, ATP5G1, DYNC1I2, INS, TBC1D32, ATP5A1, CDC42, KIF3A, UBA52, CEP131, RHOA, SNRPE, ESR1, CNEP1R1, DCTN2, DCTN4, CCDC14, EIF4E, HSPA8, IST1, HOMER3, UBC, RAN, C14orf79, PPFIA2, AKT1, SNUPN, SYTL3
antigen processing and presentation of exogenous peptide antigen via MHC class II	DCTN6, CANX, ACTR10, DCTN3, RAB7A, DCTN5, DYNC1H1, DCTN1, ACTR1A, HLA-DQA2, DYNLL1, DYNC1I2, KIF3A, DCTN2, DCTN4, KLC1
intracellular transport	SNRPD3, DCTN6, GOSR1, NUP107, ACTR10, DCTN3, CUL3, STAT3, RAB7A, SNRPF, TP53, RPS27A, SNRPG, ATP5O, DCTN5, SNRPD1, ATP5H, HSPA4, UBB, ATP5I, RHOD, SNX32, CYC1, RPLP2, HSPA5, NCBP2, HSP90AA1, SNRPD2, DYNC1H1, ATP5C1, DCTN1, ACTR1A, NCBP1, STXBP4, SMN2, DYNLL1, ATP5G1, DYNC1I2, INS, ATP5A1, CDC42, KIF3A, UBA52, CEP131, SNRPE, DCTN2, DCTN4, EIF4E, HSPA8, HOMER3, UBC, RAN, C14orf79, PPFIA2, AKT1, SNUPN, SYTL3
cellular response to stress	KEAP1, SIRT1, RBX1, SKP1, CANX, KDM6B, GSTM3, RBL2, CREBBP, EP300, CUL3, TP53, CBX8, RPS27A, EGFR, ATM, ATG10, POLR2J, SPATA18, ALB, HDAC3, UBB, CEBPB, CYCS, CCDC13, CASP3, EZH2, HSPA5, CUL1, HSP90AA1, CEP63, ATF4, PIDD1, PTPN11, BAZ1B, ATG7, DCTN1, HSF2, ATG5, JUN, MUTYH, EDEM2, MAP1LC3A, STXBP4, FLOT1, RBBP7, SUMO1, MAPK8, UBA52, RHOA, USP19, HRAS, PSMA1, BRCA1, HDAC2, HSPA8, H2AFX, UBC, AKT1, MYC, TMCO1

Table E.1 continued from previous page

description	**genes**
regulation of mitotic cell cycle phase transition	RBX1, SKP1, DCTN3, RBL2, EP300, CUL3, TP53, EGFR, ATM, EZH2, PPP2R1A, CUL1, HSP90AA1, CEP63, PIDD1, UBE2C, DYNC1H1, DCTN1, ACTR1A, CDC20, UBD, DYNLL1, DYNC1I2, CEP131, DCTN2, PSMA1, BRCA1, AKT1
symbiotic process	KEAP1, PILRA, SIRT1, RBX1, NUP107, GAPDH, SKP1, CANX, CREBBP, EP300, STAT3, RAB7A, TP53, RPS27A, EGFR, UBB, PACS1, RPLP2, CUL1, ATP6V0C, HSP90AA1, CTNNB1, ATG7, CASP8, ATG5, JUN, HDAC1, SUMO1, DYNLL1, DYNC1I2, CDC42, UBA52, RHOA, KLC1, EIF4E, HSPA8, H2AFX, IST1, UBC, RAN
protein-containing complex assembly	SNRPD3, RBX1, NUP107, SKP1, NDUFS7, CREBBP, EP300, CUL3, SNRPF, TP53, RPS27A, SNRPG, DGKQ, HIST1H2BA, HIST1H2BD, CCT2, SNRPD1, HSPA4, UBB, PPP2R1A, CUL1, PSMG1, HSP90AA1, SNRPD2, CTNNB1, UBE2C, HIST1H2BK, CASP8, HIST2H2BE, RBBP4, HDAC1, NCBP1, HIST1H2BL, RBBP7, SMN2, UBA52, PILRB, SNRPE, ENSP00000410764, H2AFX, UBC, SNUPN, HIST1H2BJ, HIST1H2BO, HIST1H2BM, CBY1, HIST1H2BH, MYC, HIST1H2BB, HIST1H2BN
protein-DNA complex assembly	RBX1, RPS27A, HIST1H2BA, HIST1H2BD, UBB, HIST1H2BK, HIST2H2BE, RBBP4, HIST1H2BL, RBBP7, UBA52, H2AFX, UBC, HIST1H2BJ, HIST1H2BO, HIST1H2BM, HIST1H2BH, HIST1H2BB, HIST1H2BN
chromatin assembly or disassembly	SIRT1, TP53, HIST1H2BA, HIST1H2BD, BAZ1B, HIST1H2BK, HIST2H2BE, RBBP4, HDAC1, HIST1H2BL, RBBP7, H2AFX, HIST1H2BJ, HIST1H2BO, HIST1H2BM, HIST1H2BH, HIST1H2BB, HIST1H2BN
regulation of cell cycle	SIRT1, RBX1, SKP1, DCTN3, RBL2, EP300, CUL3, STAT3, TP53, EGFR, ATM, CASP3, EZH2, PPP2R1A, CUL1, HSP90AA1, CEP63, PIDD1, PTPN11, CTNNB1, UBE2C, DYNC1H1, DCTN1, ACTR1A, JUN, CDC20, RBBP4, STXBP4, UBD, RBBP7, DYNLL1, DYNC1I2, INS, CDC42, CEP131, RHOA, USP19, HRAS, DCTN2, PSMA1, BRCA1, EIF4E, HSPA8, H2AFX, AKT1, MYC

Table E.1: List of genes involved in the enriched GO processess.