Semantic discovery and computational filtering to identify potentially novel breast cancer genes and signatures in omics data.

by

Bridget Cebisile Langa

Thesis Submitted in Fulfilment of the Requirements for the Degree:

Doctor of Philosophy (PhD)

South African National Bioinformatics Institute

Faculty of Natural Sciences

University of the Western Cape

Supervisor	**Co-Supervisor**
Dr Junaid Gamieldien	Prof. Burtram C. Fielding

December 2023

DECLARATION AND ATTRIBUTION

I, Bridget Cebisile Langa, declare that **“Semantic discovery and computational filtering to identify potentially novel breast cancer genes and signatures in omics data”** is my original work and that all the sources I have used or cited have been indicated and acknowledged by means of complete references in strict accordance with appropriate statutory rules, and the terms and conditions of the Creative Commons Attribution (https://creativecommons.org/licenses/by/4.0/).

Bridget Cebisile Langa	:	
Place	:	University of The Western Cape
Date	:	December 2023

DEDICATION

I dedicate my dissertation work to my parents, expressing a deep sense of gratitude. To my mother, NA Langa, who remains the unwavering pillar of my strength, and to my late father, BC Langa, whose words of encouragement and relentless push for tenacity continue to resonate in my ears. I am profoundly thankful for your unwavering belief in my dream.

ACKNOWLEDGEMENTS

Above all, I want to express my sincere gratitude to the Almighty God for His grace, unwavering strength, and constant guidance that have been with me throughout this journey.

- My unalloyed appreciation goes to my amiable, ever supportive, and humble supervisor, Dr Junaid Gamieldien for his extensive and invaluable contributions and instructions throughout this project. Your comments and words of encouragement have a unique place in my heart, serving as more than mere guidance, they are more than light to my path. I am most grateful.
- The contributions and support of my agile and nimble co-supervisor, Prof. Fielding, are immeasurable. Your words of hope, support and encouragement have been incredibly invaluable to me.
- I wish to extend my heartfelt gratitude to the University of the Western Cape (UWC) and the South African National Bioinformatics Institute (SANBI) for their pivotal role in making this research project a reality.
- I find it difficult to express the depth of my gratitude to my loving family, who provided me with the unwavering encouragement I needed throughout this journey. My late father, BC Langa, my mother, NA Langa, my wonderful siblings, EN Langa-Madikizela, CH Langa, CS Langa, AV Langa, and CT Langa, and finally, my nieces and nephew, MO Langa, TS Langa and A Langa, I want to convey my heartfelt thanks for your unconditional love and support.
- I would like to express my gratitude to Apostle Mohlala for his unwavering support and invaluable guidance throughout this journey. His wise words have been a source of great encouragement.
- To my colleagues and friends (Dr Esterhuysen, Prof Hughes, Dr Maboza, Dr Akurugu, Dr Pearce and Dr Cairncross) for their continued support and words of encouragement. I am grateful to you for always standing behind me.

ABSTRACT

High-throughput sequencing technologies developed rapidly in recent years. Using such platforms to sequence DNA and RNA samples has been shown to be a powerful method to analyze the genome and transcriptome of even very complex eukaryotic organisms, including humans and diseases like cancer, which results in substantial genomic and gene expression changes compared to healthy tissues. Such analyses have led to the discovery of hundreds of thousands of novel genetic and transcriptomic variations associated with disease conditions such as breast cancer. The Cancer Genome Atlas (TCGA) is one such database which maintains RNA sequencing data of all cancer related genes. However, searching such a database for aberrations that contribute to the specific disease condition can be cumbersome, especially since a relatively small set of mutations and/or expression changes are drivers of the disease, with the large majority being 'passengers'. Similarly, mutated or differentially expressed genes that are not yet known to be related to breast cancer may be incorrectly discarded as they are not 'classical' cancer genes.

In this study novel knowledge discovery and next-generation database methods that use existing knowledge of gene/protein roles in cancer-related functions, phenotypes, pathways and protein-protein interactions to predict their likely contribution to the breast cancer phenotype were developed, with an aim of computationally prioritizing breast cancer genetic, transcriptomic and structural variation.

The aim of this study was to develop a process of identifying and understanding semantics behind breast cancer data, information or content from databases and scientific literature to afford readily available information as nodes and links to end users. This was afforded by developing a biomedical knowledge graph (BORG) and exploiting bioinformatics tools to re-analyze multiple types of omics data generated for breast cancer tissue samples. This study is stratified into independent chapters that flow from Chapter 2 each with its own objectives and conclusions.

Chapter 1 introduces the expanses and complexities of breast cancer as well as the paucity of complete understanding of its interrelatedness to other diseases, pathways, and gene ontologies. It explores breast cancer statistics, related datasets and database integration, genetics and related omics, diagnosis and classification together with developments in sequencing and computational techniques. Furthermore, this chapter presents the rationale for the development of breast cancer specific knowledge graphs.

The aims of Chapter 2 were to develop a breast cancer-specific knowledge graph that integrated relevant multiple biomedical information sources into a large on-disk semantic network and to verify the validity of the conceptual data graph by preforming tertiary analyses on real breast cancer data.

To achieve this, an in-house biomedical semantic database that integrates a vast amount of curated information related to genes, disease associations, phenotypes, and pathway memberships was specialized to ‘understand’ breast cancer and cancer biology in general. Two versions of the database were developed namely, a minimal version centred around human genes and their associated functions and phenotype associations and a comprehensive database that also included similar information for rat and mouse genes. Furthermore, pathway involvement of human genes, as well as human protein-protein interactions were included in the comprehensive database. Testing the minimal database with lists of differentially expressed genes from a breast cancer RNA-seq study returned several known breast cancer genes, *IGFBP3* and *AR* with associated gene functions, phenotypes and pathways explaining the mechanism of their involvement indicated with PubMed IDs. The comprehensive database returned interesting candidates for novel genes such as *VANGL2* and *TPSAB1* with compelling evidence for roles in BCA biology.

Chapter 3 sought to test the comprehensive databases' ability to identify novel disease gene candidates, as well as evidence for their mechanisms, in the most frequently mutated subset of genes in breast cancer samples relative to other cancers in TCGA data from the Genomic Data Commons (GDC) Portal. The knowledge graph identified potentially novel BCA genes including *CSMD1, UMODL1* and *VPS13D*. In Chapter 4, the main aim was to reanalyse RNA-seq samples from TCGA using the developed semantic database in Chapter 2. The Bioconductor R package, edgeR, was used for differential expression gene analyses of read counts arising from RNA-seq between normal and tumour samples for multiple BCA subtypes. These genes were further analyzed using the graph database, which corroborated with existing databases but further elucidated several that were not previously linked to breast cancer, for example *TNXB* and *VIPR1*. Zhang et al (2023), also confirmed that TNXB could not be linked to human BCA in their research although it had mouse ortholog.

Chapter 5 hypothesizes about the ability of the data graph (BORG) to semantically discover and computationally filter genes identified in Next-Generation Sequencing (NGS) and other genomics experiments to identify potentially novel genes and pathways related to pathogenesis of breast cancer. Overall, this study contributes to the field by advancing understanding of the potential applications of graph modeling. Its findings have implications for both theoretical developments and practical applications in data science and clinical application.

KEYWORDS

Breast cancer

Cancer genomics

Omics data integration

Differential expression

Mutation

Bioinformatics analysis

ABBREVIATIONS, CONTRACTIONS, ACRONYMS AND INITIALISMS

BCA	Breast cancer
BRCA1	Breast Cancer gene 1
BRCA2	Breast Cancer gene 2
CNVs	Copy number variants
DEGs	Differentially Expressed Genes
DNA	Deoxyribonucleic acid
DNA-seq	DNA sequencing
ER-	Estrogen receptor negative
ER+	Estrogen receptor positive
FC	Fold changes
FDR	False-discovery rate
GDC	Genomic Data Commons Data Portal
GWAS	Genome-wide association studies
HBOCs	Hereditary breast-ovarian cancers
HER2	Human epidermal growth factor receptor 2
LncRNA	Long noncoding RNA
logFC	Log2 fold-change
MAF	Mutation Annotation Format
miRNA	MicroRNA’s
MRI	Magnetic resonance imaging
mRNA’s	Messenger RNA

mTOR	Mammalian target of rapamycin
NCBI	National Center for Biotechnology Information
NCI	National Cancer Institute
NCR	National Cancer Registry
ncRNA's	Non-coding RNA's
NGS	Next generation sequencing
NHGRI	National Human Genome Research Institute
PARP	Poly ADP-ribose polymerase
piRNA	Piwi-interacting RNA
pLOF	Putative loss-of function
poly-A	Polyadenylated
QNBC	Quadruple-negative breast cancer
RDBMs	Relational database management system
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
rRNA	Ribosomal RNA
SDGs	Sustainable Development Goals
SNPs	Single nucleotide polymorphisms
SNVs	Single nucleotide variations
TCGA	The Cancer Genome Atlas
TNBC	Triple-negative breast cancer
tRNA	transfer RNA
VCF	Variant Calling Format
VEGF	Vascular epithelial growth factor

WHO	World Health Organization

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

LITERATURE REVIEW

1.1 BREAST CANCER STATISTICS

Breast cancer (BCA) is one of the most common and potentially lethal diseases in women worldwide (Boyle, Leon, Maisonneuve, & Autier, 2003). Differences in incidence and mortality in different populations and different regions of the world strongly indicate that the disease is multi factorial and both environmental and genetic factors are involved (Nathanson, Wooster, & Weber, 2001). Incidence and mortality due to cancer, particularly BCA, has been increasing for the last 50 years, even though there is a decreased gap in the diagnosis of BCA at early stages. According to World Health Organization (WHO) 2012 reports, BCA is the leading cause of death in women, accounting for 23% of all cancer deaths. In Asia, one in every three women faces the risk of BCA in their lifetime as per reports of WHO 2012 (Polyak, 2014).

Breast cancer is the most frequently diagnosed cancer in the vast majority of the countries (154 of 185) and is also the leading cause of cancer death in over 100 countries; the main exceptions are Australia/New Zealand, Northern Europe, Northern America where it is preceded by lung cancer, and many countries in Sub-Saharan Africa due to elevated cervical cancer rates.

Globally, BCA incidence rates are highest in Australia/New Zealand, Northern Europe for example the United Kingdom, Sweden, Finland, and Denmark, Western Europe and

Belgium with the highest global rates, the Netherlands, and France, Southern Europe (Italy) (Figure1.1) (Bray *et al.,* 2018), and Northern America, where recent statistics demonstrated that BCA accounted for 30% of all newly diagnosed cancer cases in women (Figure 1.2) (Siegel, Miller, & Jemal, 2017).

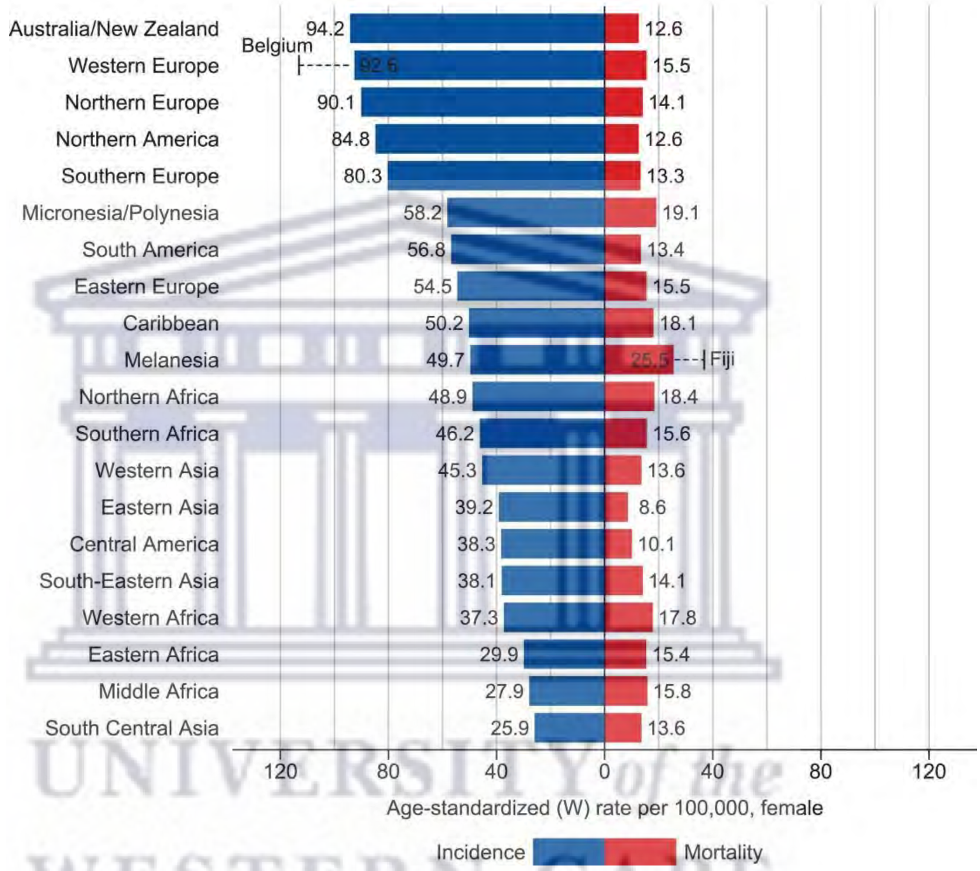
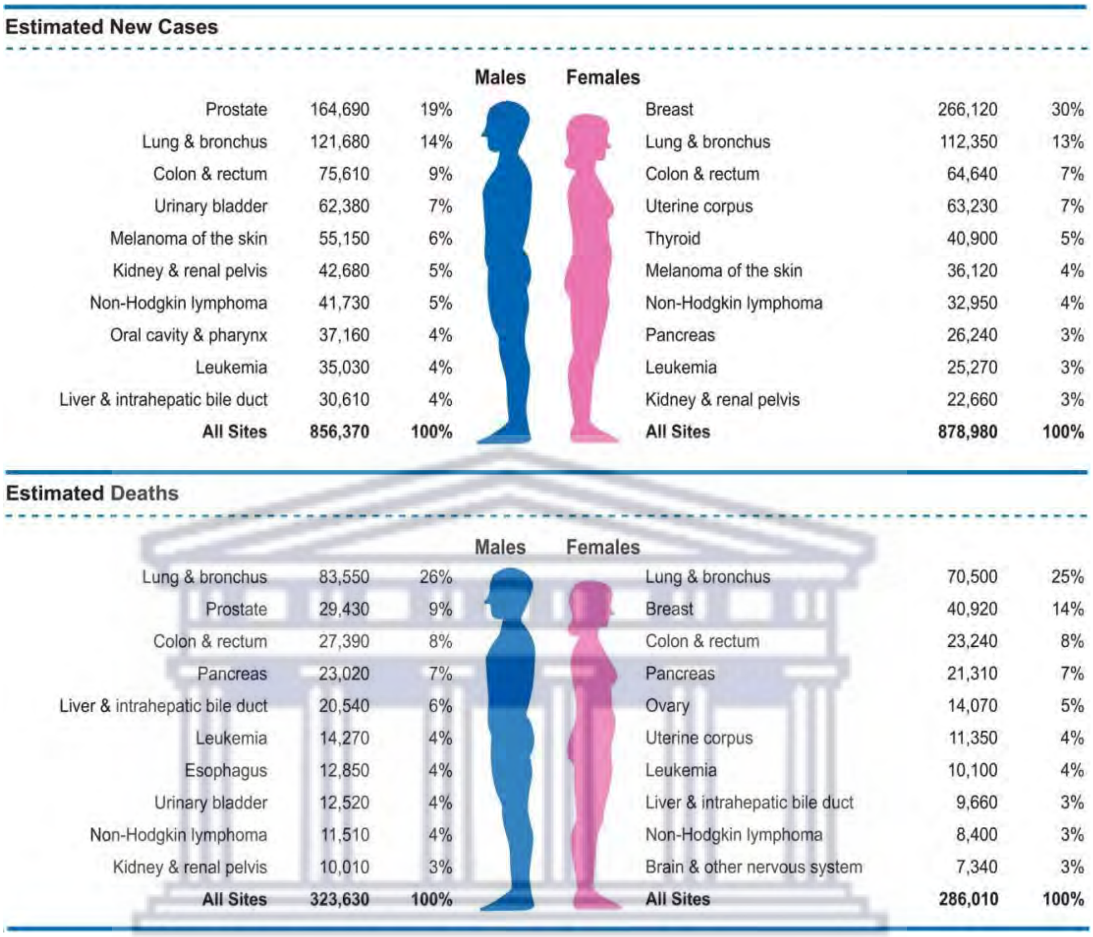


Figure 1.1: Region-specific incidence and mortality age-standardised rates for female BCA during 2018 (Bray *et al*., 2018).



Estimated New Cases

Males			Females		
Prostate	164,690	19%	Breast	266,120	30%
Lung & bronchus	121,680	14%	Lung & bronchus	112,350	13%
Colon & rectum	75,610	9%	Colon & rectum	64,640	7%
Urinary bladder	62,380	7%	Uterine corpus	63,230	7%
Melanoma of the skin	55,150	6%	Thyroid	40,900	5%
Kidney & renal pelvis	42,680	5%	Melanoma of the skin	36,120	4%
Non-Hodgkin lymphoma	41,730	5%	Non-Hodgkin lymphoma	32,950	4%
Oral cavity & pharynx	37,160	4%	Pancreas	26,240	3%
Leukemia	35,030	4%	Leukemia	25,270	3%
Liver & intrahepatic bile duct	30,610	4%	Kidney & renal pelvis	22,660	3%
All Sites	**856,370**	**100%**	**All Sites**	**878,980**	**100%**

Estimated Deaths

Males			Females		
Lung & bronchus	83,550	26%	Lung & bronchus	70,500	25%
Prostate	29,430	9%	Breast	40,920	14%
Colon & rectum	27,390	8%	Colon & rectum	23,240	8%
Pancreas	23,020	7%	Pancreas	21,310	7%
Liver & intrahepatic bile duct	20,540	6%	Ovary	14,070	5%
Leukemia	14,270	4%	Uterine corpus	11,350	4%
Esophagus	12,850	4%	Leukemia	10,100	4%
Urinary bladder	12,520	4%	Liver & intrahepatic bile duct	9,660	3%
Non-Hodgkin lymphoma	11,510	4%	Non-Hodgkin lymphoma	8,400	3%
Kidney & renal pelvis	10,010	3%	Brain & other nervous system	7,340	3%
All Sites	**323,630**	**100%**	**All Sites**	**286,010**	**100%**

Figure 1.2: Ten of the leading cancer types for new cancer cases and cancer related deaths in the U.S during 2018. Image generated by and excerpted from (Siegel, Miller, & Jemal, 2018).

In terms of mortality, BCA rates show less variability, with the highest mortality estimated in Melanesia, where Fiji has the highest mortality rates worldwide (Bray *et al.*, 2018).

Although hereditary and genetic factors, including a personal or family history of breast or ovarian cancer and inherited mutations (in *BRCA1, BRCA2,* and other BCA susceptibility genes), account for 5% to 10% of BCA cases, studies of migrants have

shown that nonhereditary factors are the major drivers of the observed international and inter-ethnic differences in incidence. Comparisons of low-risk populations migrating to high-risk populations have revealed that BCA incidence rates rise in successive generations (Johnston *et al*., 2015). Incidence has been increasing in most regions of the world, with huge inequalities between rich and poor countries. Incidence rates remain highest in more developed regions, but mortality is relatively much higher in less developed countries due to a lack of early detection and access to treatment facilities. For example, in Western Europe, BCA incidence has reached more than 90 new cases per 100 000 women annually, compared with 30 per 100 000 in eastern Africa. In contrast, BCA mortality rates in these two regions are almost identical, at about 15 per 100 000, which clearly points to a later diagnosis and much poorer survival in eastern Africa (Tao et al., 2015).

According to statistics from the National Cancer Registry (NCR) 2011, BCA was reported to have an incidence rate of 31.4 per 100 000 women in South Africa, with a reported lifetime risk of 1 in 29 women. Furthermore, it was reported in 2012 that a total of 9815 South African women were diagnosed with BCA, while a total of 3848 women died from the disease. Not surprisingly, the National Cancer Registry (NCR) 2014 included BCA among the top five cancers affecting women in South Africa, along with cervical, colorectal, uterine and lung cancer. Both breast and cervical cancer have been identified as a national priority in South Africa with increasing incidences occurring, contributing toward a governmental commitment toward the Sustainable

Development Goals (SDGs) which aims to achieve a one third reduction in premature cancer related deaths, along with other non-communicable diseases, by 2030 (Lince-Deroche et al., 2017).

The influence of BCA risk factor distribution on differences in incidence and clinical characteristics associated with ethnicity or race has received limited attention (Tariq, Latif, Zaiden, & Jasani, 2013). World Health Organization (WHO) has stated that BCA is the most frequently found cancer in women and it is affecting millions of women all over the world. However, death rates have been gradually declining after 1990 due to improvements in BCA screening, early detection, awareness and continuous improvement in treatment, Breast Cancer Deadline 2020 (Figure 1.3) (Dubey, Gupta, & Jain, 2015); (Lince-Deroche et al., 2017). Additionally, this positive trend toward early detection and awareness has also been influential in South Africa, since the 1990s. Furthermore, new technologies are being developed for the detection and diagnosis of BCA, often following 2 distinct routes. These include advances in BCA diagnosis and treatment on a global scale, along with the introduction of population-level screening (Lince-Deroche et al., 2017).

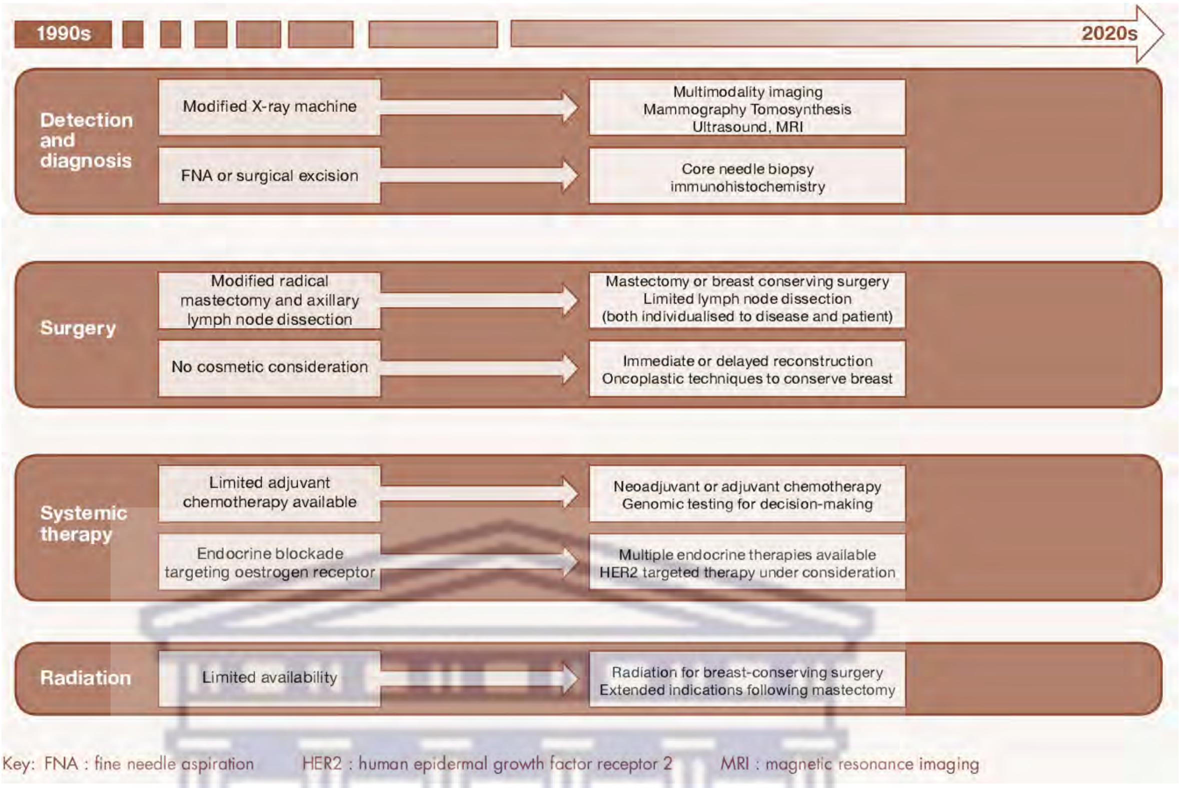

1990s
2020s
Detection and diagnosis
Modified X-ray machine
Multimodality imaging Mammography Tomosynthesis Ultrasound, MRI
FNA or surgical excision
Core needle biopsy immunohistochemistry
Surgery
Modified radical mastectomy and axillary lymph node dissection
Mastectomy or breast conserving surgery Limited lymph node dissection (both individualised to disease and patient)
No cosmetic consideration
Immediate or delayed reconstruction Oncoplastic techniques to conserve breast
Systemic therapy
Limited adjuvant chemotherapy available
Neoadjuvant or adjuvant chemotherapy Genomic testing for decision-making
Endocrine blockade targeting oestrogen receptor
Multiple endocrine therapies available HER2 targeted therapy under consideration
Radiation
Limited availability
Radiation for breast-conserving surgery Extended indications following mastectomy
Key: FNA : fine needle aspiration
HER2 : human epidermal growth factor receptor 2
MRI : magnetic resonance imaging

Figure 1.3: The chronological advances in detection and diagnosis of BCA. Image generated by and excerpted from (Lince-Deroche *et al*., 2017).

1.2 GENETICS AND BREAST CANCER

Approximately 5-10% of all cases of BCA have been found to arise from a pre-existing genetic predisposition. To date, about 25 different genes have been identified as markers of predisposition to hereditary breast and ovarian cancer, mainly involving an autosomal dominant inheritance pattern with incomplete penetrance and variable expressivity (Nielsen, van Overeem Hansen, & Sørensen, 2016). The understanding of inherited BCA susceptibility, according to Beggs *et. Al* (2008), has changed dramatically over the past 5 years, with the discovery and identification of many genes in which mutations were found to greatly influence the risk of developing BCA (Easton et.al, 2015). Furthermore, most of these genes were found to be coding for tumor suppressors, which function in genome maintenance by promoting homologous recombination repair after DNA double-strand breaks (Easton et.al, 2015).

To date, it has been found that approximately 25% of all hereditary breast-ovarian cancers (HBOCs) could be explained by the highly penetrant risk genes, BRCA*1* and *BRCA2*, and approximately 15% by other HBOC risk genes, including *RAD51C, RAD51D, ATM, CHEK2, BRIP1, PALB2, BARD1, RECQL, TP53, CDH1 and NBN1* (Easton et.al, 2015). However, this only serves to explain a combined 40% of all tested HBOCs, and does not account for the remaining 60%, for which it can be assumed that the genetic predisposition is still unknown (Beggs & Hodgson, 2008). Recently, it was also reported that pathogenic variants of the risk genes *BRCA1* and *BRCA2* conferred 40-80% lifetime risk of developing BCA, along with an additional 11-50% risk of developing ovarian cancer (Alemar *et al*., 2018). Thus, the further understanding of these genes, and their association to HBOCs, has been pivotal in BCA therapeutics, and has allowed for predictive medicines such as next-generation sequencing to arise.

1.2.1 Next-generation sequencing and the understanding of breast cancer

Next-generation sequencing (NGS) technology has steadily improved, over the past decade, and allows for rapid sequencing of millions of DNA fragments without previous sequence knowledge. Furthermore, this technology allows for high-throughput sequencing of both large and small genomic regions for many different samples, allowing it to replace the conventionally used Sanger sequencing, providing a versatile tool in BCA R&D (Kamps *et al*., 2017). More specifically, NGS provides a unique opportunity to practice predictive medicine, generally based on identification of variants that have been previously identified as being causative.

Putative loss-of function (pLOF) variants were previously reported to be a common occurrence in genomes, and therefore the improved understanding of their contribution to disease is a critical aspect with regard to the functionality of predictive medicine (Johnston *et al*., 2015). The previously identified pLOF variants commonly found occurring in genomes include nonsense, frameshift, and splice site alterations (Macarthur *et al*., 2012).

With the increasing use of next-generation sequencing technologies, specifically for the purpose predictive medicine, it is essential to be able to understand the function and predict the consequences of pLOFs, especially in individuals without pre-existing clinical diagnoses. Subsequently, although it appears that many bioinformatics approaches have been developed to classify missense alterations, very few tools exist, and are available, to assess the cellular and phenotypic impact pLOF variants (Johnston *et al*., 2015).

1.3 SEARCHING FOR THE 'MISSING HERITABILITY' OF BREAST CANCER

More than 12 % of women will be diagnosed with BCA in their lifetime. Although there have been tremendous advances in elucidating genetic risk factors underlying both familial and sporadic BCA, a vast portion of the genetic contribution to BCA aetiology remains unknown (Skol, Sasaki, & Onel, 2016). With the advent of genome-wide association studies, the next wave of discoveries was made, whereby over 80 low-penetrance and moderate-penetrance variants were identified. However, although these studies were highly successful at discovering variants associated with both familial and sporadic BCA, the variants identified to date still only serve to explain a combined 50

% of the total heritability of BCA (Skol *et al*., 2016). To identify genetic factors associated with BCA predisposition, early studies used linkage analysis and positional cloning in families with multiple affected individuals in order to discover highly penetrant susceptibility genes, such as *BRCA*1 and *BRCA2* (Harshman *et al*., 1994).

1.3.1. Genome-wide association to discover low-penetrance disease loci

In the post-genomic era, genome-wide association studies (GWAS) were conducted to provide a more powerful approach to identify common, low-penetrance disease loci without prior knowledge of location or function (Harshman *et al*., 1994). GWAS examines all or most of the genes in the genome of different individuals of a particular species to identify the extent to which the genes vary from individual to individual (Wang, Barratt, Clayton, & Todd, 2005). In short, this is achieved by studying individuals with different phenotypes and determining their genotypes at the positions of single nucleotide polymorphisms, otherwise known as SNPs. SNPs for which one variant is statistically more common in individuals belonging to a specific phenotypic group are then reported as being associated with the phenotype (Donnelly, 2008).

In humans, GWAS can identify any associations between specific genes and various diseases, including BCA (Hirschhorn & Daly, 2005). Moreover, the use of GWAS has broken the logjam, enabling genetic variants at specific loci to be associated with particular diseases. Genetic association data are now providing new routes to understanding the aetiology of disease, as well as new footholds on the long and difficult path to better treatment and disease prevention (Donnelly, 2008).

Considering that inherent power behind GWAS, and that NGS allows for simultaneous sequencing of multiple cancer susceptibility genes and, at a fraction of the cost of sequential testing, combining the two was of the utmost importance, and as such studies linking them has led to current knowledge of the genetics of BCA susceptibility (Tung, Battelli, Allen, Kaldate, & Bhatnagar, 2015).

1.4 BREAST CANCER DIAGNOSIS, CLASSIFICATION AND PROGNOSIS

For BCA diagnosis, various factors are considered critical, such as molecular classification of BCA, which is needed for both prognosis and clinical outcomes, along with those associated with worse prognosis, such as age, ethnicity, tumour grade and lack of surgery and radiation treatments (Li *et al*., 2017). Subsequently, various BCA subtypes have been defined by gene expression profiling, such as HER2-enriched, Luminal-A and Luminal-B, to name a few, which each exhibit diverse responses to various forms of treatment. While analyses of gene expression profiling was previously achieved using clustering algorithms, issues surrounding accurate identification of highly variable subtypes, such as Luminal-A, has persisted. In addition to this, the link between DNA methylation and expression levels in different BCA subtypes remains poorly understood (Yang, Shen, Yuan, Zhang, & Wei, 2017).

When prognosis is considered, the Luminal-A subtype has been shown to have more favourable outcomes when compared, across multiple databases, to most subtypes commonly found in early BCA patients. Specifically, this was apparent across 6 phase III clinical trials, namely TransATAC, GEICAM9906, CALGB9741, ABCSG08, NCIC-CTG MA.5 and NCIC-CTG MA.12 (Syrine *et al*., 2017).

Moreover, progressing the current understanding of tumour subtypes, along with molecular mechanisms, is of the utmost importance for the continual development of therapeutic strategies to the modulation of immune response. To this end, analysing neoadjuvant treatment makes it possible to assess direct response to therapies, the associated effects toward survival with the absence of disease, and overall survival rates. Moreover, it has been found that achieving a pathologically complete response following neoadjuvant chemotherapy often yields improved prognosis outcomes, particularly in patients who express HER2-postivie BCA, as well as expressing triple negative BCA (Cortazar *et al.*, 2014). In the case of HER2-overexpressing and endocrine responsive BCA, targeted therapies exhibiting only moderate levels of toxicity are continually being developed, particularly in the realm of metastatic therapy (Finn *et al.*, 2015).

BCA can be variable in their expression of the estrogen receptor, either presenting as estrogen receptor positive (ER+) or negative (ER-), separating it into two distinct categories which are widely considered to be fundamentally separate disease entities. Specifically, tumours presenting as ER- are often high grade, p53 mutated, and generally have the worse prognosis in comparison to ER+ BCA. Furthermore, unlike ER+ BCA, patients suffering from ER- tumours have very limited viable treatment option, such as targeted therapy exploiting the over-expression of the *HER2* or *ERBB2* gene in certain cases of ER- tumours (Teschendorff, Miremadi, Pinder, Ellis, & Caldas, 2007).

1.5 RECENT ADVANCES IN BREAST CANCER

Over the past few years, substantial advances have been made in the discovery of new drugs for treating BCA. Improved understanding of the biologic heterogeneity of BCA has allowed the development of more effective and individualized approach to treatment (Moulder & Hortobagyi, 2008). One of the major challenges for BCA treatment is its heterogeneous nature, which determines the therapeutic options (Polyak, 2011). The high implication of correct HER2/neu diagnostic assessment in BCA therapeutic decisions is of primary importance in clinical oncology. HER2/neu oncogene evaluation provides important prognostic information and helps clinicians to identify patients with primary or advanced metastatic cancer who are the most likely to benefit from Herceptin-targeted therapy. For this reason, this review cannot discuss HER2+ omics profiles without pointing out pathologists' efforts to correctly characterize HER2 status (Goddard *et al*., 2012).

More recently, several novel therapy strategies have emerged for the treatment of BCA. An example of this is the emergence of new agents aimed at the reversal of resistance to commonly used hormonal therapies used in the treatment of hormone receptor positive BCA. These include various novel drugs, such as Abemaciclib, Buparlisb, Everolimus and Vorinostate, to name a few, having various modes of action ranging from inhibiting cyclin dependent kinase CDK4 and CDK6 to acting as mTOR inhibitors in advanced BCA. Furthermore, advances in the realm of BCA treatment has also seen the improved understanding of resistance mechanisms, particularly in HER2+ BCA, and the emergence of immunotherapies which improve treatment outcomes. These include novel strategies such as making use of combinations of trastuzumab and PI3L, Akt and mTOR inhibitors to overcome trastuzumab resistance in HER2+ BCA, along with the use of the multi targeting TKIs Neratinib, the monoclonal antibody Patritumab, the antibody drug conjugate Trastuzumab emtansine, the farnesyl transferase inhibitor Lonafarnib, and using the peptide Nelipepimut-S in immunotherapy. Furthermore, advances have also been made with regard to triple negative BCA, whereby several novel targeted agents are emerging. These include the use of poly (ADP-ribose) polymerase inhibitors, antiangiogenic agents such as Bevacizumab, epidermal growth factor inhibitors, SRC-inhibitors and the use of the monoclonal IgG4-k antibody Pembrolizumab in immunotherapy (Tong, Wu, Cho, & To, 2018).

1.5.1 Breast cancer therapy advances

Early BCA without detectable distant metastases is considered potentially curable. Therapy has progressed substantially over the past years with a reduction in therapy intensity, both for loco-regional and systemic therapy; avoiding overtreatment but also under treatment has become a major focus. Therapy concepts follow a curative intent and need to be decided in a multidisciplinary setting, taking molecular subtype and loco-regional tumour load into account. Primary conventional surgery is not the optimal choice for all patients anymore. In triple-negative and HER2-positive early BCA, neoadjuvant therapy has become a commonly used option. Depending on clinical tumour subtype, therapeutic backbones include endocrine therapy, anti-HER2 targeting, and chemotherapy. In metastatic BCA, therapy goals are prolongation of survival and maintaining quality of life. Advances in endocrine therapies and combinations, as well as targeting of HER2, and the promise of newer targeted therapies make the prospect of long-term disease control in metastatic BCA an increasing reality (Gnant & Harbeck, 2017). To date, several novel molecular targets for the treatment of BCA are undergoing investigation. In the case of triple negative BCA, these include targets such as androgen receptor, epidermal growth factors, poly ADP-ribose polymerase (PARP) and vascular epithelial growth factor (VEGF), along with various receptors such as phosphatases, tyrosine kinases, proteases, PI3K/Akt signalling pathway, microRNA's ((miRNAs) and long noncoding RNA's (lncRNAs). In the case of kinase inhibitors, several genes have been identified as potential novel targets for drug therapy, including *CL1, CDK4, JAK2, AKT1* and *EGFR*. With respect to micro-RNA based approaches, emerging strategies involve the use of antisense oligonucleotides to inhibit onco-microRNA's, restoration of tumour suppressors by means of microRNA mimic, and finally chemically modifying microRNAs. In the case of long noncoding RNA's, oncogenes such as *HOTAIR*, *SPRY4-IT1*, *GAS5* and *PANDR* have been the focal point of ongoing research for the development of new therapeutic approaches (Mitra, 2018).

Furthermore, recent investigations surrounding biomarkers have revealed that exosomes, which have been shown to play critical roles in BCA, are stable in blood and other body fluids and thereby may be utilised as a novel biomarker (He, Zheng, Luo, & Wang, 2018). Moreover, plasma microRNA's have been identified as non-invasive, novel biomarker that may be used in the detection of BCA (Fang *et al.*, 2019). Another example of emerging biomarkers is the Autotaxin-Lysophosphatidic acid signalling axis that was shown to play an essential role in the progression and invasiveness of BCA, and that it may serve as a novel biomarker for diagnostic and prognostic purposes (Shao, Yu, He, Chen, & Liu, 2019).

Sequencing of BCA genome and transcriptome has identified BCA as a malignant disease with vast heterogeneity which is categorized into five distinct molecular subtypes including luminal A, luminal B, human epidermal growth factor receptor 2 (HER2)-enriched, basal-like, and claudin-low (Perou *et al.*, 2000). Yet, scientifically, up to ten different molecular subtypes have been identified using gene copy number and expression analyses (Curtis *et al.,* 2012). Among these, luminal-type accounts for the most part of BCA and is characterized with the typical expression of estrogen receptor (ER) and/or progesterone receptor (PR), which can be effectively targeted with hormone therapy. However, some patients have intrinsic resistance or acquired tolerance to hormone or endocrine therapy, which hampers the survival prolongation of these patients. Basal-like BCA, which is characterized with comparatively aggressive phenotype and the absent status of ER, PR and HER2, still lacks efficient treatment strategy. Thus, novel effective therapies are urgently required for BCA population (Xu *et al.,* 2017).

Approximately 15% to 20% of invasive BCA have amplification of the human epidermal growth factor receptor 2 (HER2) gene or overexpression of the HER2 protein. Before the availability of HER2-directed therapies, women with early-stage HER2-positive BCA faced a worse prognosis than those with a diagnosis of HER2-negative disease, with shorter time to disease relapse, an increased incidence of metastases, and higher mortality (Perez *et al.*, 2014). Results from large adjuvant trials showed that incorporating trastuzumab into standard adjuvant chemotherapy regimens provided substantial improvements in outcomes for women with HER2-positive BCA. Despite these impressive results, some patients will develop recurrences after trastuzumab-based adjuvant therapy, so efforts to identify more effective regimens are appropriate (Smith *et al.,* 2017). The addition of trastuzumab to paclitaxel after doxorubicin and cyclophosphamide in early-stage HER2-positive BCA results in a substantial and durable improvement in survival as a result of a sustained marked reduction in cancer recurrence (Perez et al., 2014).

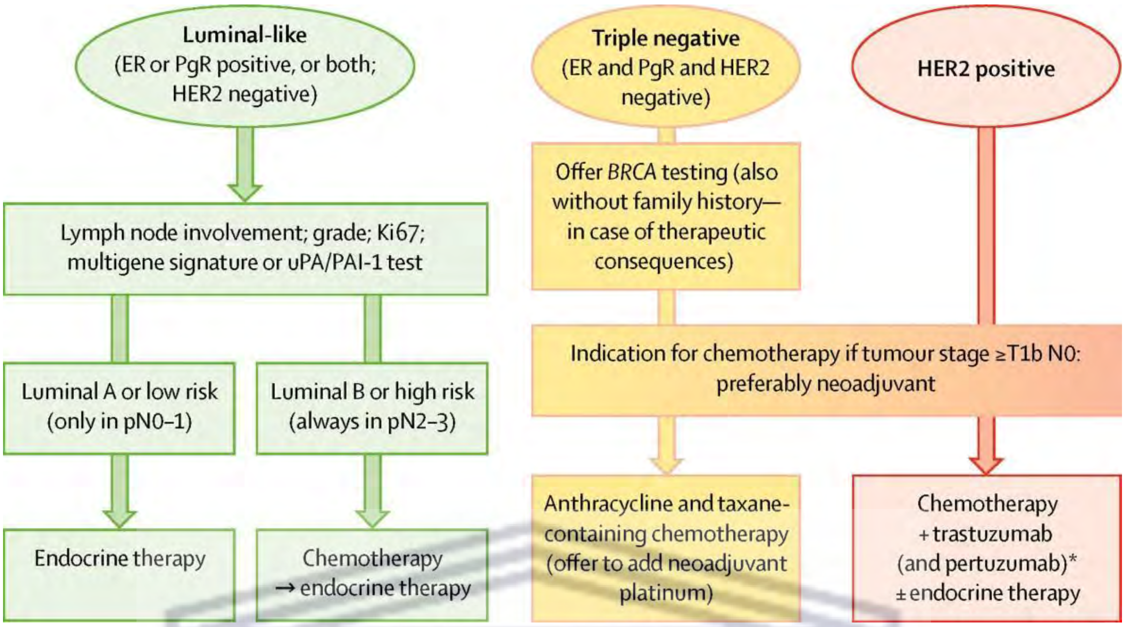

Luminal-like
(ER or PgR positive, or both; HER2 negative)
Lymph node involvement; grade; Ki67; multigene signature or uPA/PAI-1 test
Luminal A or low risk (only in pN0–1)
Luminal B or high risk (always in pN2–3)
Endocrine therapy
Chemotherapy → endocrine therapy
Triple negative
(ER and PgR and HER2 negative)
Offer BRCA testing (also without family history—in case of therapeutic consequences)
HER2 positive
Indication for chemotherapy if tumour stage ≥T1b N0: preferably neoadjuvant
Anthracycline and taxane-containing chemotherapy (offer to add neoadjuvant platinum)
Chemotherapy + trastuzumab (and pertuzumab)* ± endocrine therapy

Figure 1.4: Principles of systemic therapy in early BCA. Summary of general treatment strategies, updated after the publication of Harbeck and colleagues, 2010.

1.6 PUBLIC DATASET

The amount and diversity of genomic scale data has been steadily increasing for the past several years. This increase has enabled integrative translational bioinformatics studies across these datasets (Butte & Kohane, 2006).

1.6.1 Genomics, Transcriptomics, Proteomics, and Metabolomics

There are new families of technologies that provide a comprehensive analysis of the complete, or near-complete, cellular complement of specific constituents, such as RNAs, DNAs, proteins, intermediary metabolites, etc. These have been termed "-omics" technologies, a terminology derived from the Greek suffix "-ome" which denotes a body or group—in the commonly-used sense of a complete body or group for example the "biome"—the complement of living organisms in a particular environment, or the "genome"—the complete set of genes contained in the cellular complement of chromosomes. Omics now includes genomics, transcriptomics, proteomics, and metabolomics. In the near future, we may expect extension of these technologies to include other classes of cellular molecules, such as lipids, carbohydrates, lipoproteins, etc. These technologies are extremely powerful new tools with which to study disturbances of cellular homeostasis or structural integrity at a molecular level (Aardema & Macgregor, 2002).

Fundamental biological processes can now be studied by applying the full range of OMICS technologies to the same biological sample (Morrison *et al*., 2006). Omics technologies provide the tools needed to look at the differences in DNA, RNA, proteins, and other cellular molecules between species and among individuals of a species. These types of molecular profiles can vary with cell or tissue exposure to chemicals or drugs and thus have potential use in toxicological assessments. Omics experiments can often be conducted in high-throughput assays that produce tremendous amounts of data on the functional and/or structural alterations within the cell. These new methods have already facilitated significant advances in our understanding of the molecular responses to cell and tissue damage, and of perturbations in functional cellular systems (Aardema & Macgregor, 2002).

There is every reason to expect major change during the next decade, as new technologies and knowledge become incorporated into regulatory and industrial practice. Indeed, a new sub-discipline of "toxicogenomics" has already been recognized. Toxicogenomics has been broadly defined as the study of the relationship between the structure and activity of the genome, the cellular complement of genes, and the adverse biological effects of exogenous agents. This is consistent with the broad definition of pharmacogenomics recently proposed by Lesko and Woodcock (Woodcock, 2014).

1.6.2 Multi-omics approaches to studying BCA

Recently, integrative analysis on multi-omics data to find biomarkers or pathway features highly associated with cancer has received considerable attention (Jeong, Leem, Wee, & Sohn, 2015). Considering the rich information contained in multi-omics data, many studies have investigated the interrelationships among multiple meta-dimensional data for improved biological interpretation and analysis (González-reymúndez, Campos, Gutiérrez, Lunt, & Vazquez, 2017). In a study conducted by González-reymúndez *et. al,* aimed at improving the prediction of BCA patients, they extended clinical models including prognostic and prediction factors with whole-omic data, to integrate omics profiles for gene expression and copy number variants (CNVs). Herewith, they described a modeling framework that is able to incorporate clinical risk factors, high-dimensional omics profiles, and interactions between omics and non-omics factors, such as treatment.

Omics technologies are extremely powerful new tools allowing for the study of disturbances of cellular homeostasis or structural integrity at a molecular level. Furthermore, advances in omics open new opportunities for cancer risk prediction and risk-based screening interventions (Lévesque *et al*., 2018). Recently, the power surrounding omics in BCA treatment was reported on in a study conducted by Tunali *et al.,* in 2021, where they described how omics may be utilised to achieve the goal of precision medicine and the identification of novel strategies for the treatment of cancers such as BCA, based on the underlying genetic, environmental and lifestyle factors pertaining to patients on an individual basis. Furthermore, they highlighted how these factors may be used to develop individualised treatment options with respect to particular drugs and their dosages (Tunali, Gillies, & Schabath, 2021). Similarly, a study conducted by Chakraborty *et al*., investigating the integration of multi-omics data in cancer research also reported on the emergence of system biology models which would allow for the development of tailored targeted therapies for patients, increasing onco-drug efficacy and moving towards overcoming the occurrence of resistance to conventional chemotherapeutic and immunotherapeutic strategies (Chakraborty, Hosen, Ahmed, & Shekhar, 2018). A review by Yates and Desmedt in 2017 further highlighted the necessity of an integrated approach utilizing multiple forms of omics to advance the understanding and therapeutic strategies in overcoming BCA (Yates & Desmedt, 2017).

1.7 DATA INTEGRATION AND ITS IMPORTANCE

Clearly there is a plethora of diverse data on diseases such as BCA. This data comes from diverse sources including clinical observations, biopsies, experiments and research articles, omics, and different databases *inter alia*. The onslaught of large genomic and imaging datasets is exacerbating this condition and has necessitated researchers to search for ways of coping with the acquisition, integration, storage, distribution, and analysis demands (Frey, 2018). This organization of data makes it fit for use by clinicians to expedite diagnosis and for scientists and other users to identify caveats. This process is referred to as data integration. Zipkin et al., (2021) defines data integration as a statistical modelling approach that incorporates multiple data sources within a unified analytical framework. Further, (Schaub & Abadi, 2011) and (Michener

& Jones, 2012) indicated that this methodological approach facilitates understanding of complex and interacting processes.

The benefits of data integration include successful decision making, improves collaboration and unification of systems, reduce errors and redundancy in operations, create data warehouses and data lakes, and generally improve operating intelligence. Such formatted data may be presented meaningfully in systematic reviews, white papers and expedite precision medicine (Frey, 2018). However, these data may remain in pools of legacy systems pertaining to individual institutions which still stifle efficiency. It is essential to combine legacy data with external data and in keeping up with new developments leading to database integrations.

1.8 TECHNOLOGICAL ADVANCES IN DATA AND DATABASE INTEGRATION

There are several ways to integrate data depending on the size of institutional requirements. This may begin with manual data integration, all the way through to common storage integration. In between middleware data integration, application-based integration and uniform integration are amongst the resources that can be used to develop integrated databases. Several features indicate good data integration including lots of connectors, open source, portability, and ease of use, transparent pricing, and cloud compatibility (“Reference - OGM Library,” n.d.). These features may lead to meaningful databases that can be integrated to effect efficiency.

Database integration is the process aggregating information from multiple sources that securely shares a current clean version of data which is stored and well defined according to rules across an institution. Thus, database integration provides the base to and from where all information flows within an institution. For smaller institutions on site repositories of data storage is sufficient. However, with compound institutions, such as two laboratories merger across the globe, this may be inefficient as there could be redundancies and space, hence increasing cost. This necessitates efficiency of using cloud-based databases as demonstrated by the current needs.

The benefits of database integration include universal reliability of data, holistic operations, simplified security, and ease of compliance. The technological advances date back from logbook system, punch card systems, onsite servers, and lately cloud based integrated databases. Three Apache tools namely, Apache Hadoop, Apache Spark, and Apache Cassandra, played a major role in developing open source and flexible integrated databases for theoretically unlimited scalability ("Reference - OGM Library," n.d.).

There are two main database management systems used for data integration namely, relational databases (RDBMs) and graph databases. The RDBMs are characterised by tables, rows, columns of data, constraints, and joins. These become difficult to interpret at a glance with increased information. On the other hand, graph databases are easy to interpret at a glance. Their main features include graphs characterized by nodes, relationship, and connectivity. These are the first-class entities rendering relationships more valuable than data itself. A desirable integrated database could be one that can present information in both tables and graphs, for example Neo4j.

1.9 BEYOND RAW DATA – KNOWLEDGE INTEGRATION

Scientific knowledge is created from a subjective combination of data generated and shared by various institutions emanating from clinical trials, research, experiments, information, education, decisions, intuitions, experiences *et cetera*. These data can be selected, analysed, and subsequently transformed, interpreted, and used in reasoning, decision making and also to create new knowledge (Rückemann et al., 2021). Therefore, in essence comprehensive knowledge is multi- and interdisciplinary.

There is abundant high-quality biomedical data available from numerous research efforts that is available for creating new knowledge. One of these data portals includes Genomic Data Commons Data Portal (GDC) where there is presentation of multi omics data relating to cancer (GDC data portal_ https://portal.gdc.cancer.gov/). However, this abundance simultaneously becomes the core challenge to sharing, retrieving, integration and application of data and knowledge/facts. Developments in information technology and computational sciences have eased this problem and further enabled presentation of data in concise and informative formats. One of these formats is knowledge graphs.

Design of knowledge graphs exploits information integration. However, first and foremost, data must be presented in useable formats. Generally, data may be classified into two categories i.e., structured data and unstructured data. Structured data is mostly quantitative, factual, well organised, and coded data (alphabetical, numerical, metrics or date) in a predefined tabular format. This makes it easy to search and analyse. On the other hand, unstructured data is mostly qualitative and presented in a variety of forms. This form of data is often not structured via predefined data models. Thus, it cannot be processed and analysed using conventional tools. While structured data is often fraught with loss of reasoning leading to the output, unstructured data may be thought of as residing in discourse and thus rich in explanation. However, for a comprehensive database in biomedical sciences a combination of both data types is essential.

The current developments in sharing data knowledge and information have necessitated employment of specialized information systems to facilitate the resourceful use of such items. This has resulted in substantial innovation in the area of data and knowledge integration, in the form of structured knowledgebases, the most noteworthy being those employing graph relational database systems and concepts such as semantic representation and ontology.

1.10 EXAMPLES OF KNOWLEDGE/GRAPH DATABASES

Some graph databases are able to model complex relationships that are almost impossible to depict using relational databases, due to their ability to retain relationships directly by linking 'atoms' of information (nodes) through labelled edges instead of accessing and browsing tables (Mei, Huang, Xie, & Mora, 2020). Biomedical domain has adopted graph database because of the interconnected nature of its data. This enables more informed representation models and better data integration workflows, exploration, and analysis abilities (Timón-Reina et al., 2021). Neo4j and TypeDB (https://vaticle.com/typedb) are currently amongst the most popular graph databases.

Neo4j is a popular open-source graph database that is highly scalable and schema free (NoSQL) developed in Java and it has a rich query language called Cypher. Its advantages include that it provides flexible data models that can be easily modified to diverse applications, as well as capabilities to provide real time insights. Other properties of Neo4j include representing connected and semi connected data, easy and fast retrieval and it is fully ACID (Atomicity, Consistency, Isolation and Durability) compliant ("Reference - OGM Library," n.d.).

TypeDB (https://vaticle.com/typedb) is a knowledge graph that is data oriented and ontology like, enabling complex systems making them more intelligent through logic, reasoning, and knowledge engineering (Nielsen et al., 2016). TypeDB is a database in the form of a knowledge graph that uses an intuitive ontology to model extremely complex datasets. It stores data in a way that enables machine learning by presenting meaning of information in the complete context of their relationships. The language of TypeDB is a declarative, knowledge-oriented graph query language that uses machine reasoning to retrieve explicitly stored and implicitly derived knowledge (Timón-Reina, Rincón, & Martínez-Tomás, 2021). Consequently, TypeDB allows computers to process complex information more intelligently with less human intervention (Altinok, 2020). The platform TypeDB and the language GRAQL constitutes TypeDB. TypeDB may thus be considered as a deductive database presenting data in knowledge graph format that exploits machine reasoning to simplify data processing challenges for AI applications. (Messina et al., 2018) proposed BioGrakn, a graph-based semantic database that takes advantage of the power of knowledge graphs and machine reasoning to solve problems in the domain of biomedical science. While it is claimed to model biological data in all its complexity and contextual specificity (Messina et al., 2018), it does not represent the molecular mechanisms of disease phenotypes and the biological processes and pathways that are perturbed/dysregulated in the initiation and progression of disease.

When comparing TypeDB and Neo4j, Altinok, (2020) argues that the latter is a true knowledge graph, as demonstrated by the schematic definitions in Neo4j with relations, classes, instances, and properties. In our experience, Neo4j is ‘whiteboard friendly’ and enables uncomplicated modeling of multiple biological and biomedical knowledge domains in a single knowledge graph that can be used for knowledge discovery.

1.11 OVERVIEW OF RESEARCH PLAN

In this study a bioinformatics tool is developed to enable linking data from cloud-based data lakes improving data quality while enabling healthcare practitioners, researchers and interested parties to make fast and accurate decisions by linking nodes from different sources.

BioOntological Relationship Graph database (BORG) is an example of a successful semantic integration database integrating multiple sources of genomic and biomedical knowledge (Saunders, Jalali Sefid Dashti, & Gamieldien, 2016).

In this research, Bioinformatics techniques were applied on genes implicated in cancer and BCA abstracted from multiple sources associated with genomics data, i.e., very high throughput data from TCGA on expression and from the GDC data portal the mutation frequency data and copy number variation data. An enrichment analysis was applied to the resultant data to source out relevant information. This data was then organised into a meaningful integrated data pool. In these integrated data pools were genes that were not previously linked to BCA together with myriad of cancer and non-cancer associated pathways. These genes, ontologies and pathways were organised into nodes and links to model data into graphical form enabling extrapolation and thus purporting potential functions and anomalies.

REFERENCES

- Aardema, M. J., & Macgregor, J. T. (2002). Toxicology and genetic toxicology in the new era of “toxicogenomics”: impact of “-omics” technologies, *499*, 13–25.
- Alemar, B., Gregório, C., Herzog, J., Matzenbacher Bittar, C., Oliveira Netto, C. B., Artigalas, O., Ashton-Prolla, P. (2018). Correction: BRCA1 and BRCA2 mutational profile and prevalence in hereditary breast and ovarian cancer (HBOC) probands from Southern Brazil: Are international testing criteria appropriate for this specific population (PLoS ONE 12:11(e0187630) DOI: 10.1371/j. *PLoS ONE*, *13*(5), 1–18. https://doi.org/10.1371/journal.pone.0197529
- Altinok, D. (2020). Neo4j vs Grakn | Towards Data Science.
- Beggs, A. D., & Hodgson, S. V. (2008). susceptibility, *17*(7), 855–856. https://doi.org/10.1038/ejhg.2008.235
- Boyle, P., Leon, M. E., Maisonneuve, P., & Autier, P. (2003). Cancer control in women. Update 2003. *International Journal of Gynecology and Obstetrics*, *83*(SUPPL. 1), 179–202. https://doi.org/10.1016/S0020-7292(03)90121-4
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492
- Butte, A. J., & Kohane, I. S. (2006). Creation and implications of a phenome-genome network. *Nature Biotechnology*, *24*, 55. Retrieved from https://doi.org/10.1038/nbt1150

- Chakraborty, S., Hosen, M. I., Ahmed, M., & Shekhar, H. U. (2018). Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *BioMed Research International*, *2018*(Figure 1), 1–14. https://doi.org/10.1155/2018/9836256
- Cortazar, P., Zhang, L., Untch, M., Mehta, K., Costantino, J. P., Wolmark, N., Von Minckwitz, G. (2014). Pathological complete response and long-term clinical benefit in breast cancer: The CTNeoBC pooled analysis. *The Lancet*, *384*(9938), 164–172. https://doi.org/10.1016/S0140-6736(13)62422-8
- Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, *486*, 346. Retrieved from https://doi.org/10.1038/nature10983
- Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature*, *456*, 728. Retrieved from https://doi.org/10.1038/nature07631
- Dubey, A. K., Gupta, U., & Jain, S. (2015). Breast cancer statistics and prediction methodology: A systematic review and analysis. *Asian Pacific Journal of Cancer Prevention*, *16*(10), 4237–4245. https://doi.org/10.7314/APJCP.2015.16.10.4237
- Fang, R., Zhu, Y., Hu, L., Khadka, V. S., Ai, J., Zou, H., Hu, X. (2019). Plasma MicroRNA Pair Panels as Novel Biomarkers for Detection of Early Stage Breast Cancer. *Frontiers in Physiology*, *9*(January), 1–12. https://doi.org/10.3389/fphys.2018.01879
- Finn, R. S., Crown, J. P., Lang, I., Boer, K., Bondarenko, I. M., Kulyk, S. O.,

Slamon, D. J. (2015). The cyclin-dependent kinase 4/6 inhibitor palbociclib in combination with letrozole versus letrozole alone as first-line treatment of oestrogen receptor-positive, HER2-negative, advanced breast cancer (PALOMA-1/TRIO-18): A randomised phase 2 study. *The Lancet Oncology*, *16*(1), 25–35. https://doi.org/10.1016/S1470-2045(14)71159-3

- Frey, L. J. (2018). Data integration strategies for predictive analytics in precision medicine. *Personalized Medicine*, *15*(6), 543–550. https://doi.org/10.2217/pme-2018-0035
- GDC. (n.d.). Retrieved December 3, 2021, from https://portal.gdc.cancer.gov/
- Gnant, M., Harbeck, N., & Thomssen, C. (2017). St. Gallen/Vienna 2017: a brief summary of the consensus discussion about escalation and de-escalation of primary breast cancer treatment. *Breast Care*, *12*(2), 101-106.
- Goddard, K. A. B., Bowles, E. J. A., Feigelson, H. S., Habel, L. A., Alford, S. H., McCarty, C. A., Webster, J. A. (2012). Utilization of HER2 genetic testing in a multi-institutional observational study. *American Journal of Managed Care*, *18*(11), 704–712.
- González-reymúndez, A., Campos, G. D. L., Gutiérrez, L., Lunt, S. Y., & Vazquez, A. I. (2017). Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions, (December 2016), 538–544. https://doi.org/10.1038/ejhg.2017.12
- Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., Ding, W., Skolnick, M. H. (1994). Strong Candidate for the Breast and Ovarian Cancer, (January 2016), 1–7. https://doi.org/10.1126/science.7545954
- Bossung, V., & Harbeck, N. (2010). Angiogenesis inhibitors in the management

of breast cancer. *Current Opinion in Obstetrics and Gynecology*, *22*(1), 79-86.

- He, C., Zheng, S., Luo, Y., & Wang, B. (2018). Exosome theranostics: Biology and translational medicine. *Theranostics*, *8*(1), 237–255. https://doi.org/10.7150/thno.21945
- Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*, 95. Retrieved from https://doi.org/10.1038/nrg1521
- Jeong, H., Leem, S., Wee, K., & Sohn, K.-A. (2015). Integrative network analysis for survival-associated gene-gene interactions across multiple genomic profiles in ovarian cancer. *Journal of Ovarian Research*, *8*(1), 42. https://doi.org/10.1186/s13048-015-0171-1
- Johnston, J. J., Lewis, K. L., Ng, D., Singh, L. N., Wynter, J., Brewer, C., Biesecker, L. G. (2015). Individualized Iterative Phenotyping for Genome-wide Analysis of Loss-of-Function Mutations. *The American Journal of Human Genetics*, *96*(6), 913–925. https://doi.org/10.1016/j.ajhg.2015.04.013
- Kamps, R., Brandão, R. D., van den Bosch, B. J., Paulussen, A. D. C., Xanthoulea, S., Blok, M. J., & Romano, A. (2017). Next-generation sequencing in oncology: Genetic diagnosis, risk prediction and cancer classification. *International Journal of Molecular Sciences*, *18*(2). https://doi.org/10.3390/ijms18020308
- Lévesque, E., Kirby, E., Bolt, I., Knoppers, B. M., de Beaufort, I., Pashayan, N., & Widschwendter, M. (2018). Ethical, Legal, and Regulatory Issues for the Implementation of Omics-Based Risk Prediction of Women's Cancer: Points to Consider. *Public Health Genomics*. https://doi.org/10.1159/000492663

- Li, X., Yang, J., Peng, L., Sahin, A. A., Huo, L., Ward, K. C., Meisel, J. L. (2017). Triple-negative breast cancer has worse overall survival and cause-specific survival than non-triple-negative breast cancer. *Breast Cancer Research and Treatment*, *161*(2), 279–287. https://doi.org/10.1007/s10549-016-4059-6
- Lince-Deroche, N., van Rensburg, C., Masuku, S., Rayne, S., Benn, C., & Holele, P. (2017). Breast cancer in South Africa: developing an affordable and achievable plan to improve detection and survival. *South African Health Review*, 181–188. Retrieved from url: http://www.hst.org.za/publications/south-african-health-review-2017
- Macarthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Walter, K., Jostins, L., David, N. (2012). Europe PMC Funders Group Europe PMC Funders Author Manuscripts A systematic survey of loss-of-function variants in human protein-coding genes, *335*(6070), 823–828. https://doi.org/10.1126/science.1215040.A
- Mei, S., Huang, X., Xie, C., & Mora, A. (2020). GREG - Studying transcriptional regulation using integrative graph databases. *Database*, *2020*, 1–8. https://doi.org/10.1093/database/baz162
- Messina, A., Pribadi, H., Stichbury, J., Bucci, M., Klarman, S., & Urso, A. (2018). BioGrakn: A knowledge graph-based semantic database for biomedical sciences. *Advances in Intelligent Systems and Computing*, *611*, 299–309. https://doi.org/10.1007/978-3-319-61566-0_28
- Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology & Evolution*, *27*(2), 85–93.

https://doi.org/10.1016/J.TREE.2011.11.016

- Mitra, P. (2018). Multi-drug therapy in breast cancer: are there any alternatives? *Annals of Translational Medicine*, *6*(11), 221–221. https://doi.org/10.21037/atm.2018.04.16
- Morrison, N., Cochrane, G., Faruque, N., Tatusova, T., Tateno, Y., Hancock, D., & Field, D. (2006). Concept of Sample in OMICS Technology. *OMICS: A Journal of Integrative Biology*, *10*(2), 127–137. https://doi.org/10.1089/omi.2006.10.127
- Moulder, S., & Hortobagyi, G. N. (2008). Advances in the treatment of breast cancer. *Clinical Pharmacology and Therapeutics*, *83*(1), 26–36. https://doi.org/10.1038/sj.clpt.6100449
- Nathanson, K. N., Wooster, R., & Weber, B. L. (2001). Breast cancer genetics: What we know and what we need. *Nature Medicine*, *7*, 552. Retrieved from https://doi.org/10.1038/87876
- Nielsen, F. C., van Overeem Hansen, T., & Sørensen, C. S. (2016). Hereditary breast and ovarian cancer: new genes in confined pathways. *Nature Reviews Cancer*, *16*, 599. Retrieved from https://doi.org/10.1038/nrc.2016.72
- Perez, E. A., Romond, E. H., Suman, V. J., Jeong, J. H., Sledge, G., Geyer, C. E., Wolmark, N. (2014). Trastuzumab plus adjuvant chemotherapy for human epidermal growth factor receptor 2 - Positive breast cancer: Planned joint analysis of overall survival from NSABP B-31 and NCCTG N9831. *Journal of Clinical Oncology*, *32*(33), 3744–3752. https://doi.org/10.1200/JCO.2014.55.5730
- Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C.

A., Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, *406*, 747. Retrieved from https://doi.org/10.1038/35021093

- Polyak, K. (2011). Review series introduction Heterogeneity in breast cancer. *J.Clin.Invest.*, *121*(10), 2011–2013. https://doi.org/10.1172/JCI60534.3786
- Polyak, K. (2014). Breast cancer statistics and markers. *Journal of Cancer Research and Therapeutics*, *10*(3), 506–511. https://doi.org/10.4103/0973-1482.137927
- Reference - OGM Library. (n.d.). Retrieved December 8, 2021, from https://neo4j.com/docs/ogm-manual/current/reference/
- Rückemann, C. P., Pavani, R., Kovacheva, Z., Gersbeck-Schierholz, B., Hülsmann, F., & Naydenova, I. (2021). Delegates' Summit: – Best Practice and Definitions – Concepts of Cognostic Addressing Structured and Non-structured Data The Eleventh Symposium on Advanced Computation and Information in Natural and Applied Sciences (SACINAS) The International Conference.
- Saunders, C. J., Jalali Sefid Dashti, M., & Gamieldien, J. (2016). Semantic interrogation of a multi knowledge domain ontological model of tendinopathy identifies four strong candidate risk genes. *Scientific Reports*, *6*(November 2015), 1–10. https://doi.org/10.1038/srep19820
- Schaub, M., & Abadi, F. (2011). Integrated population models: A novel analysis framework for deeper insights into population dynamics. *Journal of Ornithology*, *152*(1), S227–S237. https://doi.org/10.1007/S10336-010-0632-7/TABLES/1
- Easton DF, Pharoah PD, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, Devilee P, Meindl A, Couch FJ, Southey M, Goldgar DE, Evans

DG, Chenevix-Trench G, Rahman N, Robson M, Domchek SM, Foulkes WD. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med. 2015 Jun 4;372(23):2243-57. doi: 10.1056/NEJMsr1501341. Epub 2015 May 27. PMID: 26014596; PMCID: PMC4610139.

- Shao, Y., Yu, Y., He, Y., Chen, Q., & Liu, H. (2019). Serum ATX as a novel biomarker for breast cancer. *Medicine (United States)*, *98*(13). https://doi.org/10.1097/MD.0000000000014973
- Siegel, R. L., Miller, K. D., & Jemal, A. (2017). Cancer Statistics, 2017. *CA: A Cancer Journal for Clinicians*, *67*(1), 7–30. https://doi.org/10.3322/caac.21387
- Siegel, R. L., Miller, K. D., & Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, *68*(1), 7–30. https://doi.org/10.3322/caac.21442
- Skol, A. D., Sasaki, M. M., & Onel, K. (2016). The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past BRCA and towards clinical relevance. *Breast Cancer Research*, 1–8. https://doi.org/10.1186/s13058-016-0759-4
- Smith, J. W., Buyse, M. E., Rastogi, P., Geyer, C. E., Jacobs, S. A., Patocskai, E. J., Wolmark, N. (2017). Epirubicin With Cyclophosphamide Followed by Docetaxel with Trastuzumab and Bevacizumab as Neoadjuvant Therapy for HER2-Positive Locally Advanced Breast Cancer or as Adjuvant Therapy for HER2-Positive Pathologic Stage III Breast Cancer: A Phase II Trial O. *Clinical Breast Cancer*, *17*(1), 48-54. e3. https://doi.org/10.1016/j.clbc.2016.07.008
- Syrine, A., Ihem, B., Meher, N., Olfa, A., Aida, G., Hatem, B., Amor, G. (2017). Prognostic implications of the intrinsic molecular subtypes in male breast cancer. *Journal of B.U.ON.*, *22*(2), 377–382.

https://doi.org/10.1016/j.breast.2015.07.008

- Tao, Z., Shi, A., Lu, C., Song, T., Zhang, Z., & Zhao, J. (2015). Breast cancer: epidemiology and etiology. *Cell biochemistry and biophysics*, *72*, 333-338.
- Tariq, K., Latif, N., Zaiden, R., & Jasani, N. (2013). Breast Cancer and Racial Disparity Between Caucasian and African American Women, *11*(8), 505–509.
- Teschendorff, A. E., Miremadi, A., Pinder, S. E., Ellis, I. O., & Caldas, C. (2007). An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. *Genome Biology*, *8*(8). https://doi.org/10.1186/gb-2007-8-8-r157
- Timón-Reina, S., Rincón, M., & Martínez-Tomás, R. (2021). An overview of graph databases and their applications in the biomedical domain. *Database*, *2021*(5), 1–22. https://doi.org/10.1093/database/baab026
- Tong, C. W. S., Wu, M., Cho, W. C. S., & To, K. K. W. (2018). Recent Advances in the Treatment of Breast Cancer. *Frontiers in Oncology*, *8*, 227. https://doi.org/10.3389/fonc.2018.00227
- Tunali, I., Gillies, R. J., & Schabath, M. B. (2021). Application of Radiomics and Artificial Intelligence for Lung Cancer Precision Medicine. *Cold Spring Harbor Perspectives in Medicine*, *11*(8). https://doi.org/10.1101/CSHPERSPECT.A039537
- Tung, N., Battelli, C., Allen, B., Kaldate, R., & Bhatnagar, S. (2015). Frequency of Mutations in Individuals with Breast Cancer Referred for BRCA1 and BRCA2 Testing Using Next-Generation Sequencing With a 25-Gene Panel, 25–33. https://doi.org/10.1002/cncr.29010
- Wang, W. Y. S., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-

wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, *6*, 109. Retrieved from https://doi.org/10.1038/nrg1522

- Woodcock, J. (2014). Pharmacogenomic-guided drug development: regulatory perspective, (February 2002). https://doi.org/10.1038/sj/tpj/
- Xu, H., Yu, S., Liu, Q., Yuan, X., Mani, S., Pestell, R. G., & Wu, K. (2017). Recent advances of highly selective CDK4/6 inhibitors in breast cancer. *Journal of Hematology & Oncology*, *10*(1), 97. https://doi.org/10.1186/s13045-017-0467-2
- Yang, L., Shen, Y., Yuan, X., Zhang, J., & Wei, J. (2017). Analysis of breast cancer subtypes by AP-ISA biclustering. *BMC Bioinformatics*, *18*(1), 1–13. https://doi.org/10.1186/s12859-017-1926-z
- Yates, L. R., & Desmedt, C. (2017). Translational genomics: Practical applications of the genomic revolution in breast cancer. *Clinical Cancer Research*, *23*(11), 2630–2639. https://doi.org/10.1158/1078-0432.CCR-16-2548
- Zipkin, E. F., Zylstra, E. R., Wright, A. D., Saunders, S. P., Finley, A. O., Dietze, M. C., Tingley, M. W. (2021). Addressing data integration challenges to link ecological processes across scales. *Frontiers in Ecology and the Environment*, *19*(1), 30–38. https://doi.org/10.1002/fee.2290

CHAPTER 2

THE DEVELOPMENT OF BREAST CANCER KNOWLEDGE GRAPH / DATABASE AND THE PROOF OF CONCEPT

2.1 INTRODUCTION

Bioinformatics and computational resources provide powerful tools for research and the diagnosis of complex diseases, with sophisticated software and pipelines for secondary and tertiary data analyses having been developed to assist in the study of such diseases (Sadedin et al., 2015). While the software and analysis pipelines required for extracting genetic variations from a genomics experiment have reached a relatively advanced stage of development (Meehan et al., 2019), the tools for identifying a subset of relevant candidates for a particular study are still in their infancy. Even after extensive filtering based on potential disease-causing effects and prior knowledge, omics studies generate more potential candidates than can be experimentally confirmed. This challenge is equally applicable to the investigation of inherited diseases and somatic disorders such as cancer and is particularly difficult in complex diseases, where multifactorial variations collectively contribute to the disease phenotype (Marian, 2012; Tam et al., 2019). Novel approaches are essential to prioritize candidates for further investigation, and developing specialized computational and bioinformatics resources and software tools is therefore necessary to extract meaningful biological and clinical insights from next-generation sequencing datasets. To address this need, methods and concepts for effectively mining functional gene annotations continue to evolve, as they play a crucial role in deciphering the relationships between genes/genotypes and phenotypes (Beebe & Kennedy, 2016).

Biomedical knowledge graphs for enhancing genomics discovery

In the pre-genomic era, researchers primarily used a hypothesis-driven approach, which involved in-depth characterization of a disease's phenotype, leading to the identification of a single or a small set of disease gene candidates, which was then subjected to sequencing, differential expression analyses, or functional studies (Kann, 2009). While this method was widely practiced and contributed significantly to our understanding of genetic diseases in the pre-genomic era, it has largely been replaced in the next-generation sequencing (NGS) era. However, since most diseases are complex and many involve multiple phenotypic presentations and diverse genetic abnormalities. This process has shifted from hypothesis-driven approaches to more data-driven and automated methods, with the goal of efficiently prioritizing candidates for further investigation in the NGS era.

We posit that the assessment of genomic candidate genes should involve a comprehensive evaluation of their connections to known biomolecular functions and phenotypes. We therefore argue that there is a need for the development of carefully constructed knowledge graphs linking diverse formal biomedical ontologies, gene and protein functions, clinical phenotypes and known genotypes from diverse sources to enable more seamless and thorough interrogation of the multitude of information sources needed to make sense of candidates from a post-genomic scale study (Zhang et al., 2018). The argument is that each potentially functional candidate variant or differentially expressed gene generated through an NGS study should undergo a rigorous assessment, akin to what a biomedical scientist would perform - if it were feasible to manually consult the vast amount of relevant prior knowledge in the scientific literature in an integrated manner. In this study, we proposed to adopt automation and a robust data mining approach, facilitated by a comprehensive database that formally represents the relationships between genes and phenotypes, as well as the existing knowledge about the disease of interest and its phenotypic characteristics. We hypothesize that this approach would enable a knowledge-driven method for prioritizing omics candidates in general and promises to assist in the prioritization of genes and pathways, preventing the premature discarding of potentially promising candidates, whether they be from genetic, expression, epigenetic, or other genome-scale studies. This strategy becomes particularly valuable when traditional statistical methods are not applicable (Singhal et al., 2016).

Previously, Saunders et al. (2016) identified candidate genes associated with tendinopathy by using our lab's biomedical knowledge graph, the BioOntological Relationship Graph (BORG) database to re-screen differentially expressed human genes for potential links to tendinopathy. After prioritization, they identified four strong candidate genes that are not only differentially expressed in tendinopathy but also functionally related to clinical phenotypes and to genes previously implicated in other connective tissue diseases.

High-value biomedical knowledge sources

The Gene Ontology (GO) and its associated annotation project is currently the most comprehensive and reliable source of functional annotations. This resource involves both automated text mining tools and curation, identifying experimental findings in scientific literature to associate specific GO terms, formally describing molecular functions, biological processes, or cellular locations, with gene products. While manually curated annotations are considered of the highest quality, annotations automatically extracted from texts have also proven to be valuable for functional annotations (Camon et al., 2004). In disease research, gene knockout experiments in model organisms serve as an extremely important source of functional annotations for implicating genes in the development of particular phenotypes (Albert & Kruglyak, 2015). Databases for mouse and rat genomes are good examples in this regard, recording gene annotations that reflect observable morphological, physiological, and behavioural characteristics arising in gene knockout models over the lifespan of the animal (Kaldunski et al., 2022). Crucially, both databases employ the Mammalian Phenotype Ontology, a resource rich in community-accepted annotation terms that are also relevant to human disease research (Twigger et al., 2007).

Chapter synopsis

In this study, we hypothesized that existing knowledge about breast cancer and its typical phenotypic presentation should be considered comprehensively alongside gene functional annotations known to be involved in somatic oncology. This holistic knowledge-driven approach would help identify candidates that satisfy multiple criteria and are, therefore, more likely to be involved in the disease, while potentially uncovering previously unknown biological mechanisms underlying the disease. However, the multi-relational nature of such a strategy posed a significant challenge to relational database management systems (RDBMS), which are ill-suited for handling the semantic complexity and high interconnectedness of modern biological information. To address these limitations, emerging technologies like graph databases, which enable the modeling of highly complex data and knowledge, are becoming increasingly important (Storey & Song, 2017).

This chapter presents an implementation of a semantic model of breast cancer within our in-house biomedical knowledgebase, built on the flexible and robust Neo4J graph database management system (http://www.neo4j.org). It introduces a novel approach to integrate and mine extensive biomedical knowledge for the purpose of prioritizing candidates and provides a proof-of-concept for this approach in BCA. We demonstrate the utility of our breast cancer knowledge graph database in recommending the candidacy, potential disease relevance, and mechanism of action of differentially expressed genes from a previously published dataset.

2.2 MATERIALS AND METHODS

2.2.1 BioOntological Relationship Graph (BORG) database

The in-house biomedical semantic database seamlessly integrates a vast amount of curated information related to genes, disease associations, phenotypes, and pathway memberships. This wealth of data is structured into a comprehensive on-disk semantic network, drawing inspiration from the principles in Sowa (2011). The Neo4J graph database (http://www.neo4j.org) serves as the foundation for this database, facilitating the storage of individual "knowledge atoms" and the relationships between them in the most intuitive and natural manner. In essence, this database resembles a large "concept map" that organizes genes and their roles in pathogenesis, e.g., breast cancer. It captures the intricate web of connections and associations in a way that resembles how a biomedical scientist would conceptualize and reason about these elements (see Figure 2.1).

The foundational concept of this study's semantic database centers around human genes and was further developed to establish connections to their known orthologs in the mouse and rat genomes, together with protein-protein interactions through meticulously annotated links that accurately describe the semantic relationships between them. Within this newly developed database, various bio-ontologies were incorporated to serve as crucial reference points for integration and furnish domain-specific terms essential for constructing queries. A significant focus of this study was the development of a specialized version of the BORG database tailored to support genomic research on breast cancer.

We developed two versions of the database. The first was centred only around human genes and their functions, associated phenotypes, and known disease involvement (Figure 2.1, Table 2.1).

1. Human genes are linked to Gene Ontology (GO) terms based on annotations provided by the GO consortium (Berardini et al., 2010).
2. Disease Ontology terms based on curated associations mined from the NCBI's Gene Reference into Function database (Schriml et al., 2012)
3. Human Phenotype Ontology (Robinson and Mundlos, 2010) terms based on the phenotypes that are documented to be associated with human genes in the OMIM database.

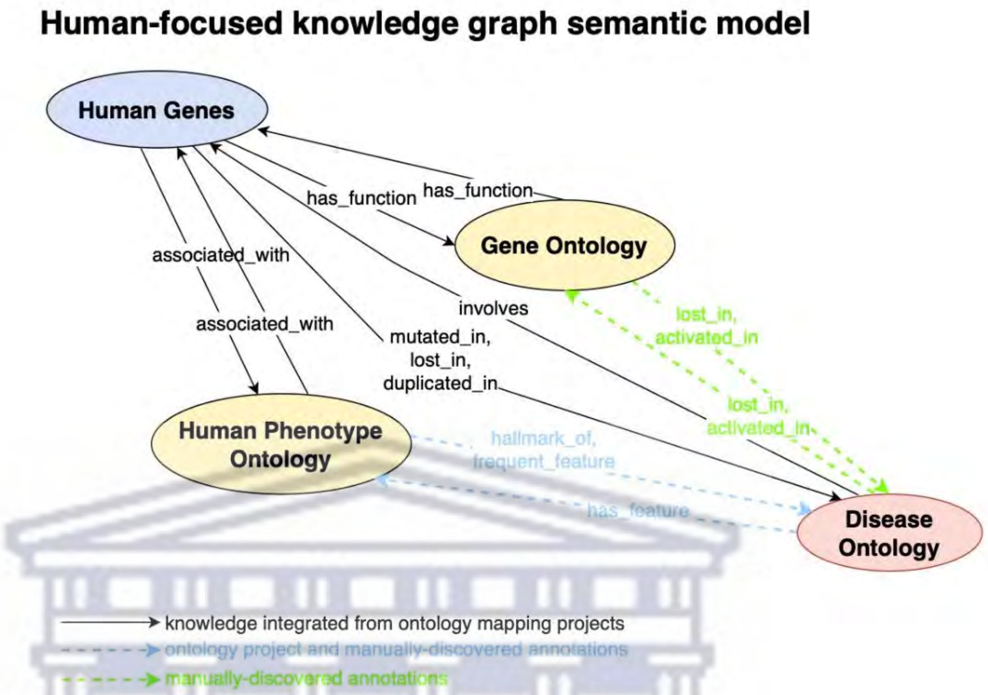


Figure 2.1: The minimal iteration of the knowledge graph centered around human genes and their functions with associated phenotypes and known disease involvement related to human breast cancer. The edges linking the nodes show knowledge integrated from ontology mapping projects (solid black lines), ontology projects and manually-discovered annotations (dotted blue lines), and manually-discovered annotations (dotted green lines).

The second, more comprehensive iteration of the database included mouse and rat genes and their functions and knockout phenotypes, pathway involvement of human genes, as well as human protein-protein interactions (Figure 2.2, Table 2.2):

1. Mammalian Phenotype Ontology (Smith and Eppig, 2012) terms, which describe the phenotypes that arise when the gene is knocked out in mice (Bult et al., 2013) or/and rat (Nigam et al., 2013).

2. Pathway Ontology (Green and Karp, 2006) terms, which models gene product involvement in pathways at a conceptual rather than structural level.

3. Human protein-protein interactions from BIOGRID (http://www.thebiogrid.org).

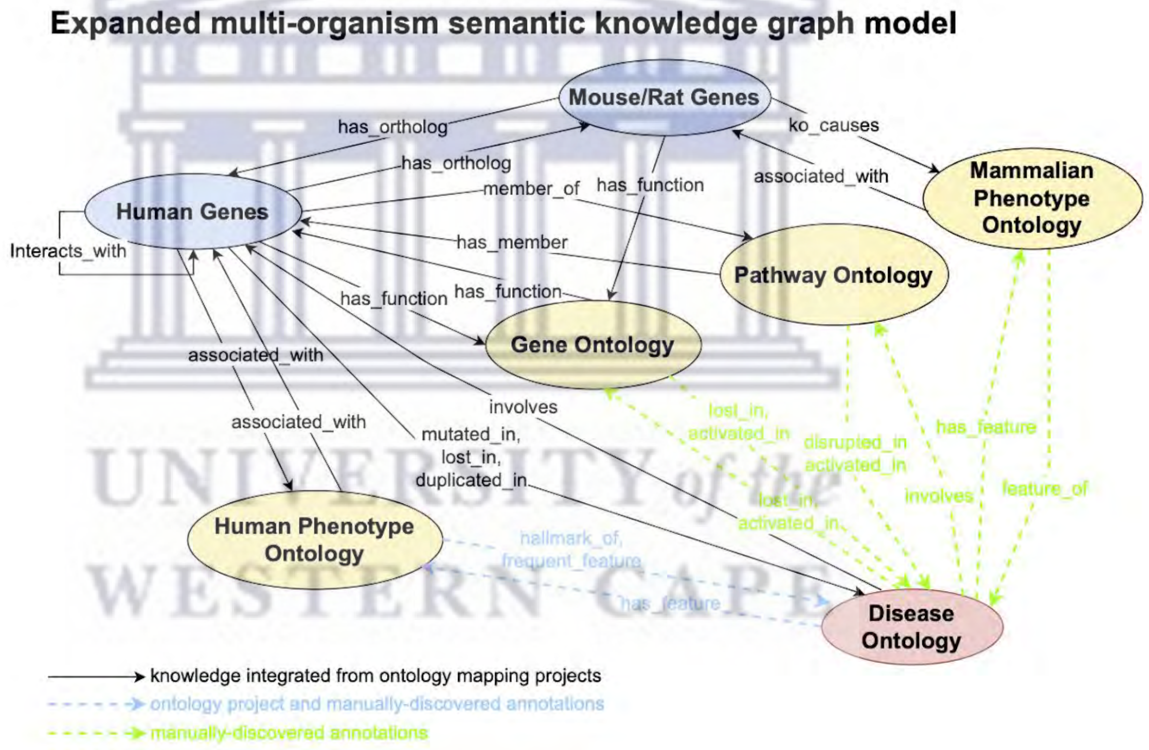


Figure 2.2 The comprehensive iteration of the knowledge graph expanded from figure 2.1, including mouse and rat genes and their functions and knockout phenotypes, the pathway involvement of human genes, as well as human protein-protein interactions. The edges linking the nodes show knowledge integrated from ontology mapping

projects (solid black lines), ontology projects and manually-discovered annotations (dotted blue lines), and manually-discovered annotations (dotted green lines).

Structuring knowledge using ontologies offers a significant advantage: when a gene is associated with a very specific term through hierarchical relationships. This hierarchical structure enhances the ability to extract relevant information from the database when conducting searches to implicate a preliminary list of genes in a disease of interest. Additionally, it's worth noting that storing information as a directed network enables human genes to "inherit" knockout phenotype annotations from model organisms through transitive associations.

2.2.2 Development of a semantic network representation of BCA gene-phenotype-disease relationships

It is now widely acknowledged that epistasis, the intricate interplay between genes, plays a crucial role in shaping the phenotypic expression of complex diseases. Consequently, the hypothesis in this study was that any gene with prior evidence of involvement in a phenotype related to a specific disease could potentially be a novel disease gene. Similarly, genes with functions similar to those of previously associated genes in the context of a disease logically become potential candidates for that disease. However, we were acutely aware that specific phenotypes and functions are conceptually related as 'parent' and 'child' terms within relevant ontologies. This awareness meant that even with extensive manual filtering and the use of the BORG database, we might overlook potential candidates due to gaps in our knowledge of the hierarchical relationships between these phenotypes and functions. Additionally, biases stemming from preconceived ideas about relevant phenotypes and functions could lead to the incorrect rejection of potentially significant candidates.

Table 2.1 (a): Ontology terms used to transitively link genes to the breast cancer terms in the minimal BORG semantic database.

TERM ID	ONTOLOGY TERM
Gene Ontology	
GO:0005925	focal adhesion
GO:0048545	response to steroid hormone
GO:0097305	response to alcohol
GO:0050678	regulation of epithelial cell proliferation
GO:0048729	tissue morphogenesis
GO:0042772	DNA damage response, signal transduction resulting in transcription
GO:0000079	regulation of cyclin-dependent protein serine/threonine kinase activity
GO:1903555	regulation of tumor necrosis factor superfamily cytokine production
GO:0072395	signal transduction involved in cell cycle checkpoint
GO:0006977	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest
GO:0006282	regulation of DNA repair
GO:0071158	positive regulation of cell cycle arrest
GO:0043407	negative regulation of MAP kinase activity
GO:0032200	telomere organization
GO:0000187	activation of MAPK activity
GO:0097191	extrinsic apoptotic signaling pathway
GO:0072331	signal transduction by p53 class mediator
GO:0043409	negative regulation of MAPK cascade
GO:0038127	ERBB signaling pathway
GO:0043410	positive regulation of MAPK cascade
GO:0043405	regulation of MAP kinase activity
GO:2001233	regulation of apoptotic signaling pathway
GO:0000165	MAPK cascade
GO:0007173	epidermal growth factor receptor signaling pathway

GO:0008285	negative regulation of cell proliferation
GO:0043066	negative regulation of apoptotic process
GO:0008284	positive regulation of cell proliferation
GO:0043065	positive regulation of apoptotic process
GO:0008083	growth factor activity
GO:0042981	regulation of apoptotic process
GO:0006974	cellular response to DNA damage stimulus
GO:0001664	G-protein coupled receptor binding
GO:0048020	CCR chemokine receptor binding
GO:0005159	insulin-like growth factor receptor binding
GO:0052742	phosphatidylinositol kinase activity
GO:0004693	cyclin-dependent protein serine/threonine kinase activity
GO:0097472	cyclin-dependent protein kinase activity
GO:0004709	MAP kinase kinase kinase activity
GO:0043560	insulin receptor substrate binding
GO:0004713	protein tyrosine kinase activity
GO:0030330	DNA damage response, signal transduction by p53 class mediator
GO:0043516	regulation of DNA damage response, signal transduction by p53 class mediator
Human Phenotype Ontology	
HP:0003002	Breast carcinoma
HP:0001428	Somatic mutation
HP:0003220	Abnormality of chromosome stability
HP:0100013	Neoplasm of the breast

Table 2.1 (b): Ontology terms used to transitively link genes to the breast cancer terms in the comprehensive database.

TERM ID	**ONTOLOGY TERM**
Pathways (PW)	

PW:0000624	breast cancer pathway
PW:0000232	PI3K-Akt signaling
PW:0001317	cell cycle pathway
PW:0000718	p53 signaling
PW:0001252	prolactin signaling
PW:0000525	ras signaling
PW:0000501	thyroid hormone signaling
PW:0000829	chemokine signaling
PW:0000386	Rap1 signaling
PW:0000008	wnt signaling
PW:0001515	hippo signaling
PW:0000007	MAPK signaling
PW:0000180	mtor signaling
PW:0000814	Toll-like receptor signaling
PW:0000542	AMPK signaling
PW:0000303	p53-dependent G1/S DNA damage checkpoint pathway
PW:0000304	p53-independent G1/S DNA damage checkpoint pathway
Mammalian Phenotype Ontology (MP)	
MP:0002166	altered tumor susceptibility
MP:0010639	altered tumor pathology
MP:0002019	abnormal tumor incidence
MP:0003077	abnormal cell cycle
MP:0003448	altered tumor morphology
MP:0010307	abnormal tumor latency
MP:0006035	abnormal mitochondrial morphology
MP:0000858	altered metastatic potential
MP:0003566	abnormal cell adhesion
MP:0008058	abnormal DNA repair
MP:0002006	tumorigenesis
MP:0005076	abnormal cell differentiation

MP:0002166	altered tumor susceptibility
MP:0003447	decreased tumor growth/size

2.2.3 Semantic representation of BCA disease biology through cross-ontology mapping

To mitigate potential human errors and make optimal use of the rich semantic relationships within the BORG database for candidate prioritization, we leveraged another important and innovative feature of storing the semantic network within the Neo4J graph database. This feature enabled us to establish links between terms from different ontologies in a semantically correct manner. Specifically, we could link terms relevant to breast cancer from the two phenotype ontologies, the Gene Ontology (GO) and pathway ontologies to the "breast cancer" term in the Disease Ontology (as illustrated in Figures 2.1 and 2.2). This fulfilled our objective of developing a semantically sound search method to discover transitive associations between genes and diseases based on their known involvement in relevant phenotypes, functions, or pathways related to human breast cancer.

In our case, a mutated or differentially expressed gene not previously implicated in breast cancer but known to play a significant functional role related to the disease or its associated phenotypes could be considered a potential novel BCA gene, biomarker, or drug target. Automation was essential for this process, necessitating the selection of a core set of concepts that could be used to establish transitive links between genes and diseases via an intermediate set of functions, phenotypes, and pathways. Simultaneously, we had to carefully choose the highest-level term on the ontology hierarchy, enabling a gene linked to a highly specialized term to be automatically implicated in disease involvement through its transitive association with the parent term, as depicted in Figure 2.2.

2.2.4 Mining the 'BCA-BORG' database using path-based transitive association queries

The BORG database is designed to structure individual facts in a manner that aligns with how humans naturally think about them. Researchers can ask complex questions of the system based on the underlying meaning of the data. This capability enables *in-silico* experimentation through intricate querying, retrieval of annotations, and the semantic discovery of genotype-to-phenotype associations. The hierarchical structure of biological ontologies, as discussed earlier, further assists in identifying transitive associations that may not always be obvious but are biologically plausible and correct.

One of the most powerful features of the BORG database is that it allows researchers to discover transitive links between genes and diseases. This feature returns the semantic relationships between all concepts (genes or terms) in the discovered path, such as "associated with" or "feature of," providing a comprehensible human-readable report that explains the biological relevance of the link. As previously mentioned, a significant use case for this feature is the ability to uncover potential associations that may not be immediately apparent but have biological validity.

In the context of breast cancer, this querying facility operates by performing a directed "walk" on the semantic network to find all paths between a candidate gene of interest and the disease term within the database (Figure 2.3). It can find the shortest path, all paths, or all paths shorter than a pre-specified length. Reports are generated on a per-gene basis and prove particularly useful when filtering a large list of candidates, as only genes with at least one path leading to the disease will be returned. These reports incorporate information from various knowledge domains, which can be used to further prioritize candidates manually based on the automatically discovered evidence. The most appealing aspect of this query facility is its capacity to uncover transitive associations that might have been overlooked when consulting the literature or individual databases directly.

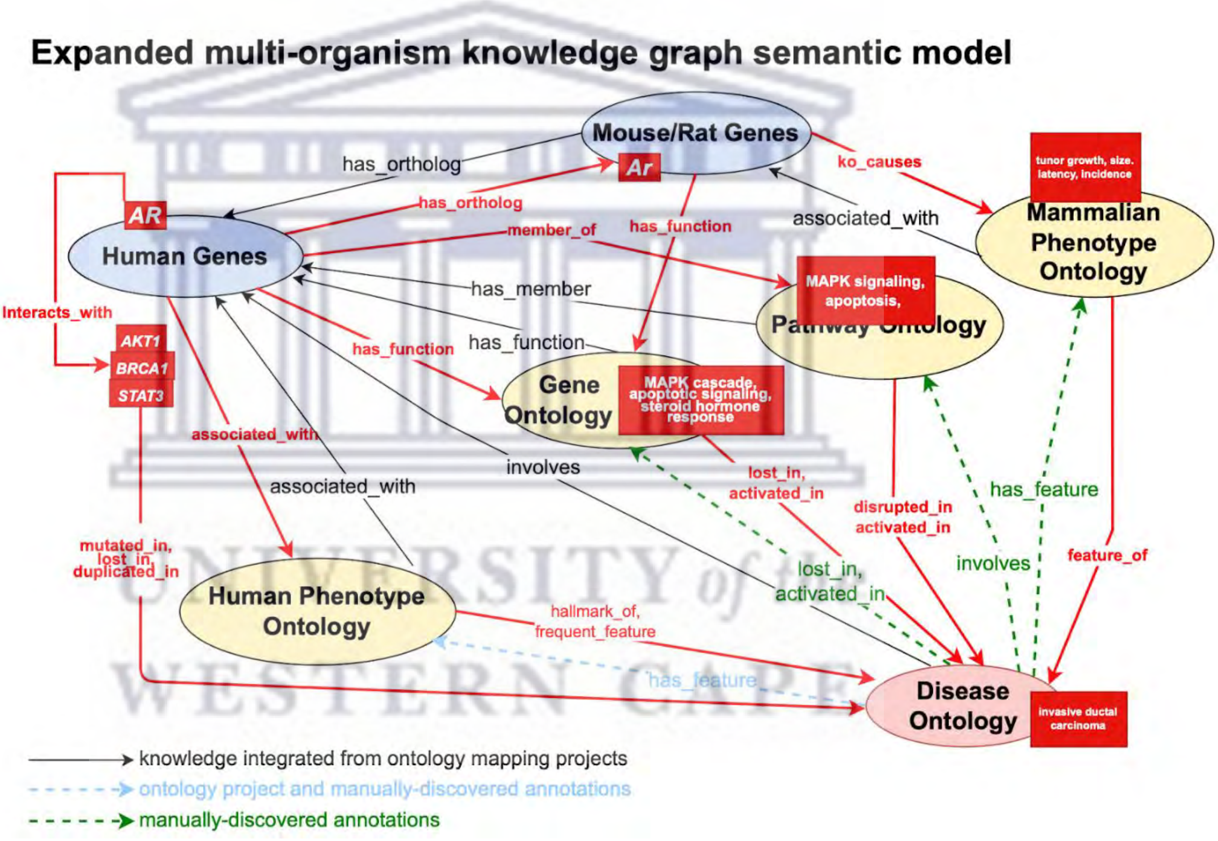


Figure 2.3. Example of a path-based walk identifying transitive evidence for a gene's potential role in breast cancer, as illustrated by the red arrows and blocks.

2.2.5 Testing the knowledge discovery potential of 'BCA-BORG'

Eswaran et al. (2012) successfully mapped a breast cancer transcriptomic landscape through mRNA sequencing, yielding comprehensive digital transcriptomes of TNBC, non-TNBC, and HER2-positive breast cancers. The heterogeneity of BCA and the need for a better understanding of its genetic landscape, cellular hierarchy, and molecular basis to improve diagnosis, prognosis, and treatment were highlighted in this study. Furthermore, novel transcripts and differentially expressed transcripts across the three breast cancer subtypes were elucidated. This study generated 1.2 billion reads, with 2617 transcripts differentially expressed between TNBC and Non-TNBC groups and 3087 transcripts differentially expressed between TNBC and HER2-positive groups from 17 individual human breast cancer tissues.

With the ground-work done in this study, there is a need for the development of more robust yet concise methodologies linking and presenting inter-relationships between genetics, cellular hierarchy, and molecular mechanisms to improve diagnosis, prognosis, and treatment. In the current study, the development of data graphs was deemed appropriate to enhance this information.

The objective of the current study was to evaluate the ability of the minimal 'BCA-BORG' system to uncover new genes that can potentially be associated with BCA, yet previously not definitively linked to BCA and reported in large studies and also accessible in databases after enrichment analysis (https://maayanlab.cloud/Enrichr/). BORG was expected to return these genes with either breast cancer associated pathways, ontologies, disease and or "guilt by association." Furthermore, to test if the system can return evidence supporting the published assertions.

The putative genes implicated in BCA were extracted from Eswaran et al. (2012). These genes included those that were upregulated and those that were downregulated. These genes were stratified using Log2FC (≥1 and ≤ -1, for upregulated and downregulated, respectively) and FDR (<0.05). The clusters of genes were filtered using pathways, ontologies, and disease on Enrichr. The adjusted P value of < 0.05 was used to include relevant genes. The genes meeting the above criteria were selected to train BORG. The 'all paths' option was selected.

This analysis identified functions that were overrepresented among these genes. From this set of overrepresented functions, we selected an appropriate parent term that serves as a representative and transversely links all the identified terms and their related functions to the disease.

2.3 RESULTS AND DISCUSSION

2.3.1 RESULTS

To identify gene functions relevant to the development of or predisposition to breast cancer and to implicate novel genes sharing similar impactful functional mutations, an analysis was conducted in BORG, genes strongly associated with BCA from the Eswaran study were analyzed. Several statistically overrepresented gene functions were identified through this analysis and subsequently yielded the compilation of Gene Ontology (GO), Human Phenotype (HP), Pathways (PW), and Mammalian Phenotype (MP) terms, as detailed in Table 2.1 (a) and (b).

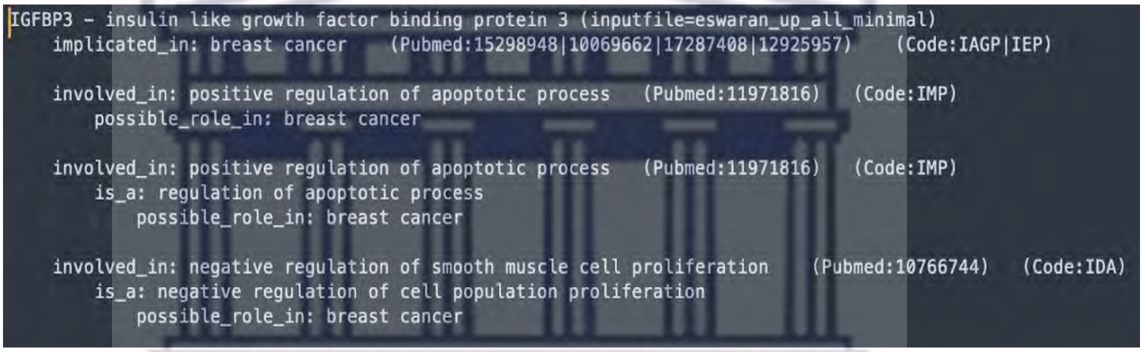
Table 2.2: Genes returned by the BORG minimal and comprehensive knowledge graphs.

Genes	Minimal database	Comprehensive database
Upregulated	*IGFBP3, MET**	*ABI1, ACTN1, ANP32E, CDH3, **CHST11**, DDIT4, E2F3, EEF2, FLNA, **FZD8**, GAPDH, **HNRNPA3**, IGFBP3, KIF1B, KIF5B, KRT16, MET*, MFGE8, NRP2, **PLEC, PRKX**, PTP4A3, RORA, S100A8, S100A9, SOX11, **VANGL2**, YES1*
Downregulated	*AGR2, AR*, ESR1, FOXA1, RHOB*	*AGR2, AR*, BCL2, **ELP2**, ERBB4, ESR1**, FBP1**, FOXA1, GATA3, **KDM4B**, **RHOB**, SPDEF, **TPSAB1**, ZNF703*

Potential novel genes are presented in bold print. The asterisk (*) indicates genes referred to in the section discussing the strengths of integrating the two databases.

The gene data from Eswaran was run on BORG as two databases namely the minimal and comprehensive databases. Both knowledge graphs iterated genes together with relevant attributes including their association with human breast cancer and PubMed IDs. Selected results from the system will be presented and discussed in the subsequent sections. A comprehensive list of the BORG results is presented in an "Additional Files" folder.(https://drive.google.com/drive/u/0/folders/1yLjvBIaqMNYHNun8uw3nBlckQN_0UmtZ)

Amongst the differentially expressed genes returned by the minimal database were *IGFBP3* and *MET* (upregulated); and *AGR2 AR*, *ESR1*, *FOXA1*, and *RHOB* (downregulated). These genes are known to be implicated in breast cancer and the system returned additional information explaining their molecular roles in oncology. For example, the minimal model annotated the *IGFBP3* (insulin like growth factor binding protein 3) transcript as being involved in positive regulation of apoptotic process and negative regulation of smooth muscle cell proliferation, which indicate possible roles in the development of BCA (Figure 2.4).

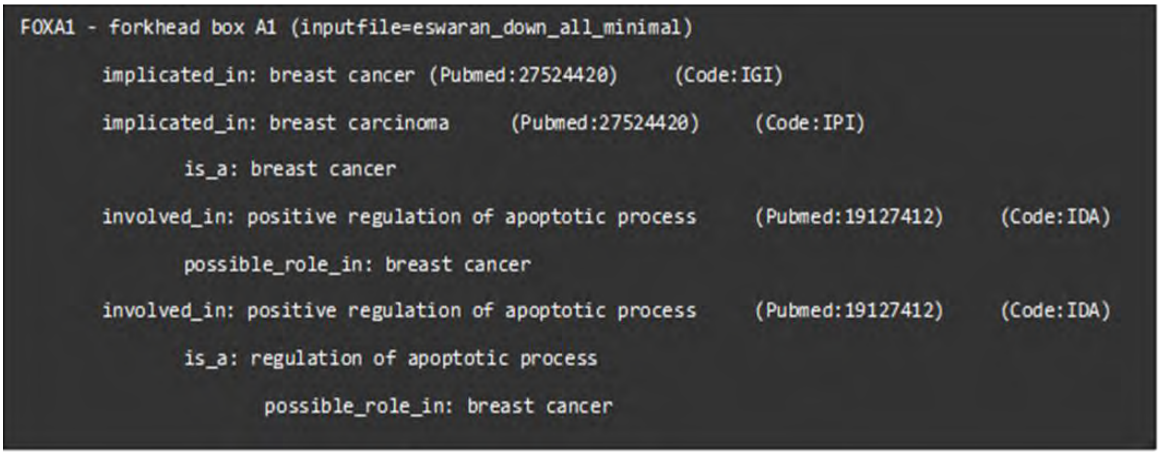


```
IGFBP3 - insulin like growth factor binding protein 3 (inputfile=eswaran_up_all_minimal)
    implicated_in: breast cancer    (Pubmed:15298948|10069662|17287408|12925957)    (Code:IAGP|IEP)

    involved_in: positive regulation of apoptotic process   (Pubmed:11971816)   (Code:IMP)
        possible_role_in: breast cancer

    involved_in: positive regulation of apoptotic process   (Pubmed:11971816)   (Code:IMP)
        is_a: regulation of apoptotic process
            possible_role_in: breast cancer

    involved_in: negative regulation of smooth muscle cell proliferation    (Pubmed:10766744)   (Code:IDA)
        is_a: negative regulation of cell population proliferation
            possible_role_in: breast cancer
```

Figure 2.4. The result from the minimal database for the upregulated insulin-like growth factor binding protein 3 gene.

With downregulated genes, Forkhead box A1 (*FOXA1*) is a gene that encodes a protein belonging to the FOX family of transcription factors. The minimal BORG annotations further indicated that this gene is implicated in BCA via positive regulation of apoptotic process (Figure 2.5).

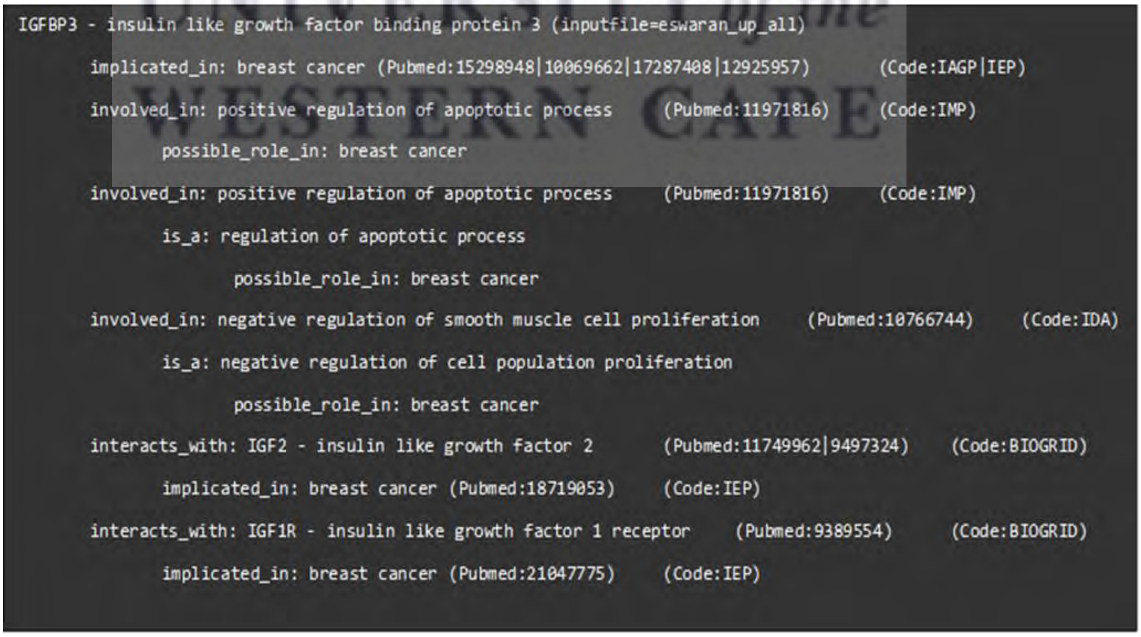
These results indicated that the minimal model works effectively in identifying information of genes in relation to breast cancer, but we hypothesized that the complete model would perform better.



```
FOXA1 - forkhead box A1 (inputfile=eswaran_down_all_minimal)
        implicated_in: breast cancer (Pubmed:27524420)     (Code:IGI)
        implicated_in: breast carcinoma     (Pubmed:27524420)     (Code:IPI)
                is_a: breast cancer
        involved_in: positive regulation of apoptotic process     (Pubmed:19127412)     (Code:IDA)
                possible_role_in: breast cancer
        involved_in: positive regulation of apoptotic process     (Pubmed:19127412)     (Code:IDA)
                is_a: regulation of apoptotic process
                        possible_role_in: breast cancer
```

Figure 2.5. The result from the minimal database for the upregulated Forkhead box A1 gene.

The comprehensive database returned additional information on the above known genes. For an example, *IGFBP3* (Figure 2.6) was further reported by the system to interact with two other genes namely, *IGF2*, the insulin like growth factor 2 and *IGF1R,* the insulin like growth factor 1 receptor, which have both been previously reported to be involved in breast cancer.



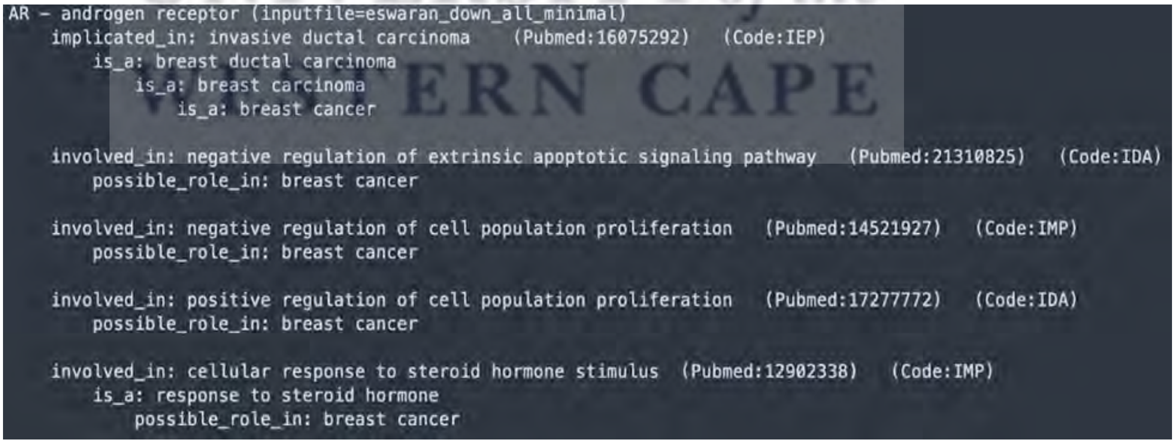
```
IGFBP3 - insulin like growth factor binding protein 3 (inputfile=eswaran_up_all)
        implicated_in: breast cancer (Pubmed:15298948|10069662|17287408|12925957)      (Code:IAGP|IEP)
        involved_in: positive regulation of apoptotic process     (Pubmed:11971816)     (Code:IMP)
                possible_role_in: breast cancer
        involved_in: positive regulation of apoptotic process     (Pubmed:11971816)     (Code:IMP)
                is_a: regulation of apoptotic process
                        possible_role_in: breast cancer
        involved_in: negative regulation of smooth muscle cell proliferation     (Pubmed:10766744)     (Code:IDA)
                is_a: negative regulation of cell population proliferation
                        possible_role_in: breast cancer
        interacts_with: IGF2 - insulin like growth factor 2       (Pubmed:11749962|9497324)     (Code:BIOGRID)
                implicated_in: breast cancer (Pubmed:18719053)     (Code:IEP)
        interacts_with: IGF1R - insulin like growth factor 1 receptor     (Pubmed:9389554)      (Code:BIOGRID)
                implicated_in: breast cancer (Pubmed:21047775)     (Code:IEP)
```

Figure 2.6. The result from the comprehensive database for the downregulated insulin like growth factor binding protein 3 gene.

Demonstration of the strengths of integrating the two databases

The comprehensive database prioritised genes not returned by the minimal version or provided much more compelling evidence for a gene's candidacy for involvement in BCA. Examples of genes returned by these both systems include *AR* as upregulated and *MET* downregulated as shown in Figures 2.7, 2.8, 2.9, and 2.10 below.

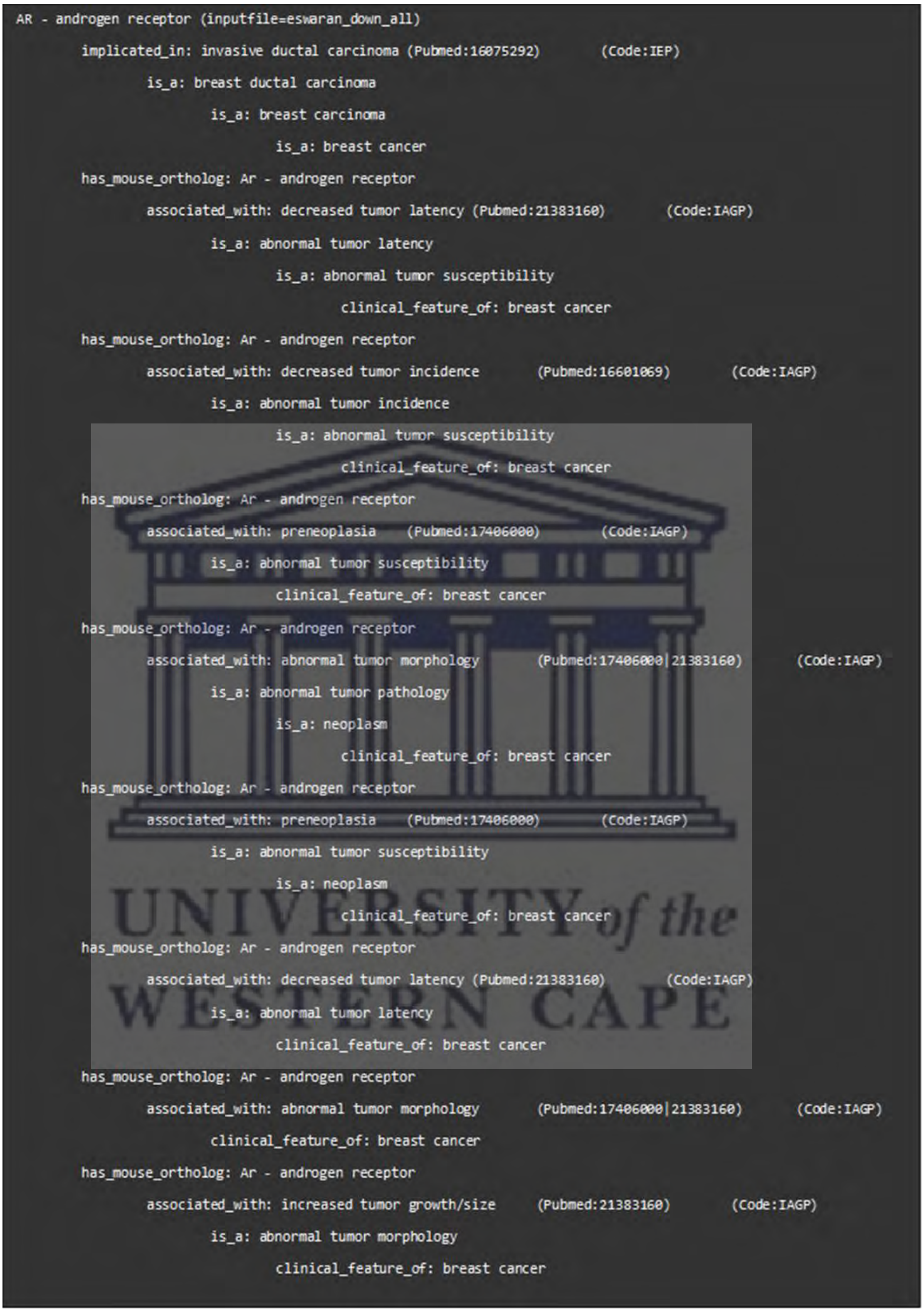
The minimal database annotated the *AR* gene with attributes related to human breast cancer including implication to the disease, pathways, and processes. This gene is known to be implicated in invasive ductal carcinoma, and is involved in processes including negative regulation of cell population proliferation, positive regulation of cell population proliferation, and cellular response to steroid hormone stimulus. It is further implicated in negative regulation of extrinsic apoptotic signaling pathway.



```
AR - androgen receptor (inputfile=eswaran_down_all_minimal)
    implicated_in: invasive ductal carcinoma    (Pubmed:16075292)   (Code:IEP)
        is_a: breast ductal carcinoma
            is_a: breast carcinoma
                is_a: breast cancer

    involved_in: negative regulation of extrinsic apoptotic signaling pathway   (Pubmed:21310825)   (Code:IDA)
        possible_role_in: breast cancer

    involved_in: negative regulation of cell population proliferation   (Pubmed:14521927)   (Code:IMP)
        possible_role_in: breast cancer

    involved_in: positive regulation of cell population proliferation   (Pubmed:17277772)   (Code:IDA)
        possible_role_in: breast cancer

    involved_in: cellular response to steroid hormone stimulus  (Pubmed:12902338)   (Code:IMP)
        is_a: response to steroid hormone
            possible_role_in: breast cancer
```

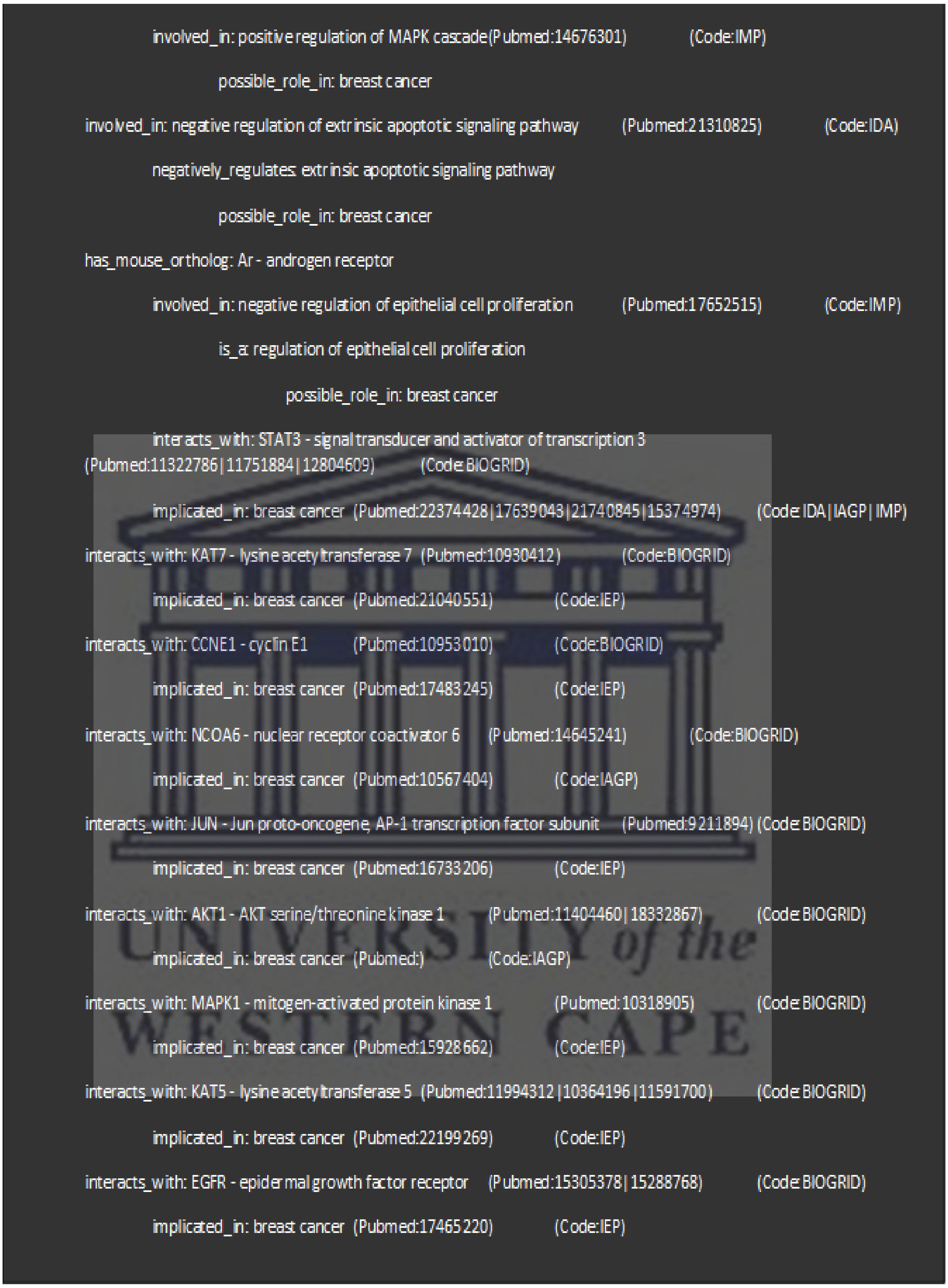
Figure 2.7. The result from the minimal database for the downregulated androgen receptor gene.

These annotations were further expanded by the comprehensive database to include mouse associations to several phenotypes, pathways and processes broadly associated with BCA. These included annotations associating *AR* with decreased tumor latency, decreased tumor incidence, preneoplasia, abnormal tumor morphology, negative regulation of extrinsic apoptotic signaling pathway, cellular response to steroid hormone stimulus, *inter alia*. In addition to the above, the comprehensive database identified numerous interactions with *AR* including: *STAT3*, *KAT7*, *CCNE1*, *KMT2D*, *inter alia*. It is of interest that the comprehensive model also returned interactions with rare genes such as *MLH3* which is not commonly associated with BCA.



```
AR - androgen receptor (inputfile=eswaran_down_all)
        implicated_in: invasive ductal carcinoma (Pubmed:16075292)       (Code:IEP)
                is_a: breast ductal carcinoma
                        is_a: breast carcinoma
                                is_a: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: decreased tumor latency (Pubmed:21383160)       (Code:IAGP)
                        is_a: abnormal tumor latency
                                is_a: abnormal tumor susceptibility
                                        clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: decreased tumor incidence       (Pubmed:16601069)       (Code:IAGP)
                        is_a: abnormal tumor incidence
                                is_a: abnormal tumor susceptibility
                                        clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: preneoplasia    (Pubmed:17406000)       (Code:IAGP)
                        is_a: abnormal tumor susceptibility
                                clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: abnormal tumor morphology       (Pubmed:17406000|21383160)       (Code:IAGP)
                        is_a: abnormal tumor pathology
                                is_a: neoplasm
                                        clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: preneoplasia    (Pubmed:17406000)       (Code:IAGP)
                        is_a: abnormal tumor susceptibility
                                is_a: neoplasm
                                        clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: decreased tumor latency (Pubmed:21383160)       (Code:IAGP)
                        is_a: abnormal tumor latency
                                clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: abnormal tumor morphology       (Pubmed:17406000|21383160)       (Code:IAGP)
                        clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: increased tumor growth/size     (Pubmed:21383160)       (Code:IAGP)
                        is_a: abnormal tumor morphology
                                clinical_feature_of: breast cancer
```

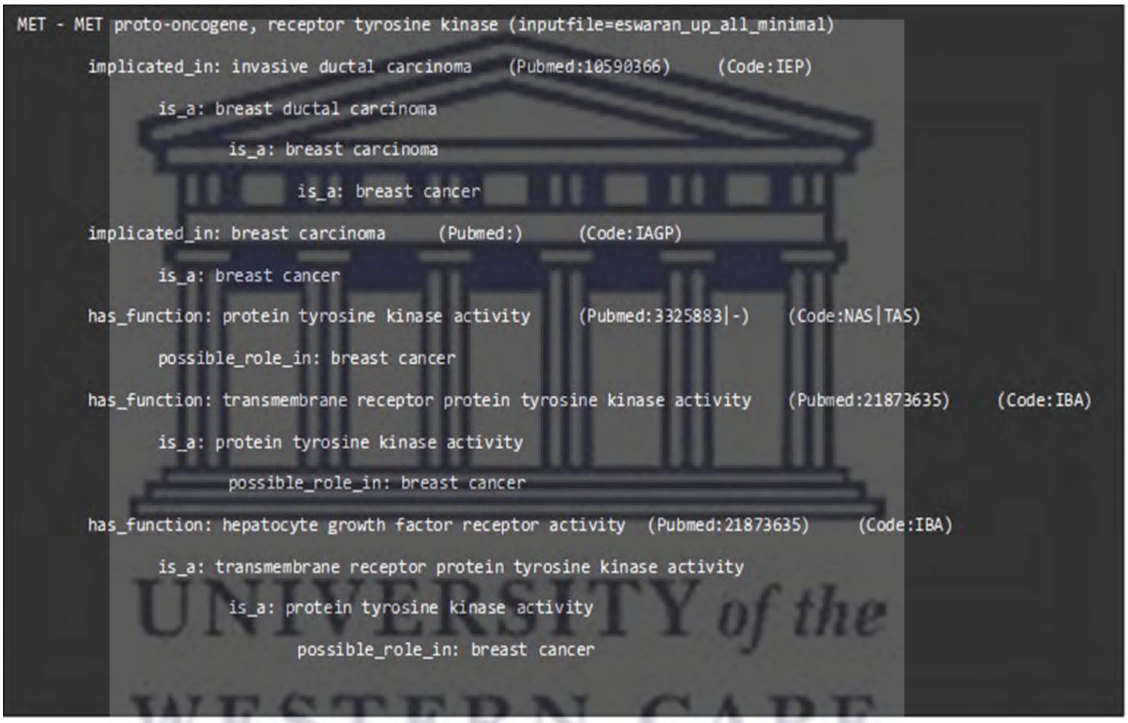
Figure 2.8. The result from the comprehensive database for the downregulated androgen receptor gene.



```
          involved_in: positive regulation of MAPK cascade(Pubmed:14676301)          (Code:IMP)
                    possible_role_in: breast cancer
involved_in: negative regulation of extrinsic apoptotic signaling pathway       (Pubmed:21310825)          (Code:IDA)
          negatively_regulates: extrinsic apoptotic signaling pathway
                    possible_role_in: breast cancer
has_mouse_ortholog: Ar - androgen receptor
          involved_in: negative regulation of epithelial cell proliferation       (Pubmed:17652515)          (Code:IMP)
                    is_a: regulation of epithelial cell proliferation
                              possible_role_in: breast cancer
          interacts_with: STAT3 - signal transducer and activator of transcription 3
(Pubmed:11322786|11751884|12804609)        (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:22374428|17639043|21740845|15374974)       (Code:IDA|IAGP|IMP)
interacts_with: KAT7 - lysine acetyltransferase 7  (Pubmed:10930412)          (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:21040551)          (Code:IEP)
interacts_with: CCNE1 - cyclin E1        (Pubmed:10953010)          (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:17483245)          (Code:IEP)
interacts_with: NCOA6 - nuclear receptor coactivator 6      (Pubmed:14645241)          (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:10567404)          (Code:IAGP)
interacts_with: JUN - Jun proto-oncogene, AP-1 transcription factor subunit     (Pubmed:9211894) (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:16733206)          (Code:IEP)
interacts_with: AKT1 - AKT serine/threonine kinase 1       (Pubmed:11404460|18332867)        (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:)          (Code:IAGP)
interacts_with: MAPK1 - mitogen-activated protein kinase 1          (Pubmed:10318905)        (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:15928662)          (Code:IEP)
interacts_with: KAT5 - lysine acetyltransferase 5  (Pubmed:11994312|10364196|11591700)       (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:22199269)          (Code:IEP)
interacts_with: EGFR - epidermal growth factor receptor    (Pubmed:15305378|15288768)        (Code:BIOGRID)
          implicated_in: breast cancer  (Pubmed:17465220)          (Code:IEP)
```

Figure 2.8 (continued**).** The result from the comprehensive database for the downregulated androgen receptor gene.

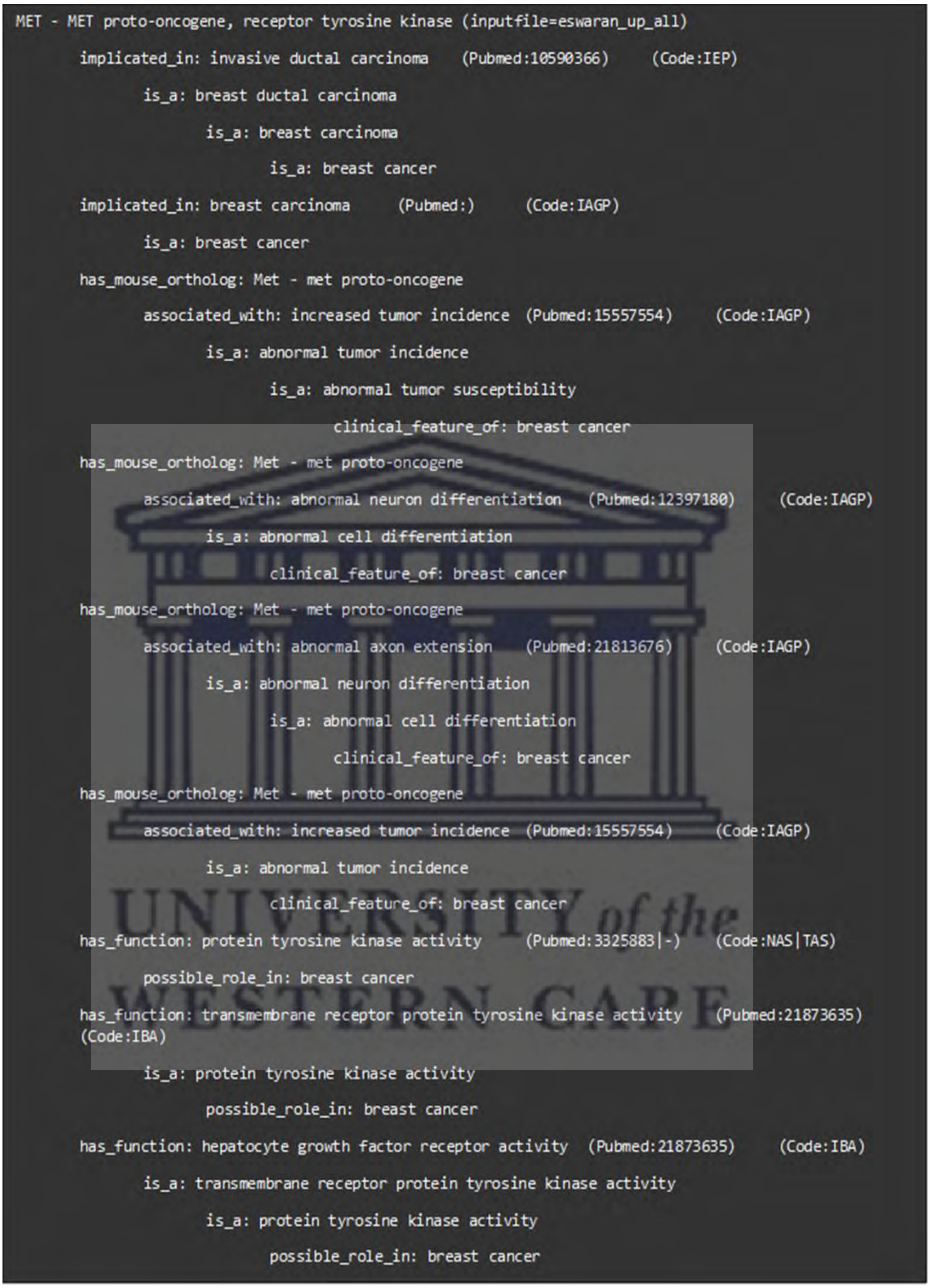
The minimal database annotated *MET* (MET proto-oncogene, receptor tyrosine kinase) with the following attributes: it is implicated in invasive ductal carcinoma and broadly in breast cancer. Furthermore, *MET* has functions in protein tyrosine kinase activity, in transmembrane receptor protein tyrosine kinase activity, and in hepatocyte growth factor receptor activity. The latter two functions are both protein tyrosine kinase activity with a possible role in BCA (Figure 2.9).



```
MET - MET proto-oncogene, receptor tyrosine kinase (inputfile=eswaran_up_all_minimal)
        implicated_in: invasive ductal carcinoma    (Pubmed:10590366)     (Code:IEP)
                is_a: breast ductal carcinoma
                        is_a: breast carcinoma
                                is_a: breast cancer
        implicated_in: breast carcinoma     (Pubmed:)      (Code:IAGP)
                is_a: breast cancer
        has_function: protein tyrosine kinase activity     (Pubmed:3325883|-)    (Code:NAS|TAS)
                possible_role_in: breast cancer
        has_function: transmembrane receptor protein tyrosine kinase activity    (Pubmed:21873635)     (Code:IBA)
                is_a: protein tyrosine kinase activity
                        possible_role_in: breast cancer
        has_function: hepatocyte growth factor receptor activity  (Pubmed:21873635)     (Code:IBA)
                is_a: transmembrane receptor protein tyrosine kinase activity
                        is_a: protein tyrosine kinase activity
                                possible_role_in: breast cancer
```

Figure 2.9. The result from the minimal database for the upregulated *MET* proto-oncogene, receptor tyrosine kinase gene.

The annotations on this gene was further enhanced in the comprehensive database with information from mouse and rat orthologs, and interactions with several genes (Figure 2.10). The mouse and rat orthologs are associated with increased and abnormal tumor incidence; which is a clinical feature of BCA. It further showed that *MET* is associated with abnormal neuron differentiation and abnormal axon extension. The functions are similar in both minimal and comprehensive databases albeit with additional information on mouse and rat orthologs in the latter database. The comprehensive database further identified protein to protein interactions including with *STAT3*, *CDH1*, *GRB2*, *CASP3*, and *EGFR* all implicated in BCA.



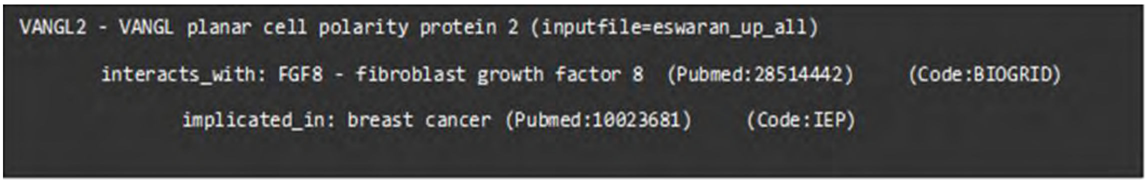
```
MET - MET proto-oncogene, receptor tyrosine kinase (inputfile=eswaran_up_all)
    implicated_in: invasive ductal carcinoma    (Pubmed:10590366)    (Code:IEP)
        is_a: breast ductal carcinoma
            is_a: breast carcinoma
                is_a: breast cancer
    implicated_in: breast carcinoma    (Pubmed:)    (Code:IAGP)
        is_a: breast cancer
    has_mouse_ortholog: Met - met proto-oncogene
        associated_with: increased tumor incidence  (Pubmed:15557554)    (Code:IAGP)
            is_a: abnormal tumor incidence
                is_a: abnormal tumor susceptibility
                    clinical_feature_of: breast cancer
    has_mouse_ortholog: Met - met proto-oncogene
        associated_with: abnormal neuron differentiation   (Pubmed:12397180)    (Code:IAGP)
            is_a: abnormal cell differentiation
                clinical_feature_of: breast cancer
    has_mouse_ortholog: Met - met proto-oncogene
        associated_with: abnormal axon extension   (Pubmed:21813676)    (Code:IAGP)
            is_a: abnormal neuron differentiation
                is_a: abnormal cell differentiation
                    clinical_feature_of: breast cancer
    has_mouse_ortholog: Met - met proto-oncogene
        associated_with: increased tumor incidence  (Pubmed:15557554)    (Code:IAGP)
            is_a: abnormal tumor incidence
                clinical_feature_of: breast cancer
    has_function: protein tyrosine kinase activity    (Pubmed:3325883|-)    (Code:NAS|TAS)
        possible_role_in: breast cancer
    has_function: transmembrane receptor protein tyrosine kinase activity    (Pubmed:21873635)
    (Code:IBA)
        is_a: protein tyrosine kinase activity
            possible_role_in: breast cancer
    has_function: hepatocyte growth factor receptor activity  (Pubmed:21873635)    (Code:IBA)
        is_a: transmembrane receptor protein tyrosine kinase activity
            is_a: protein tyrosine kinase activity
                possible_role_in: breast cancer
```

Figure 2.10. The result from the comprehensive database for the upregulated *MET* proto-oncogene, receptor tyrosine kinase gene.

Potentially novel genes implicated in BCA

The comprehensive Breast cancer semantic system further annotated novel genes with relatively scant information. This could mean these genes are novel to breast cancer pathogenesis. Protein-protein interactions are crucial in cancer for elucidating disease mechanisms and, hence, implicating novel genes in disease pathogenesis.

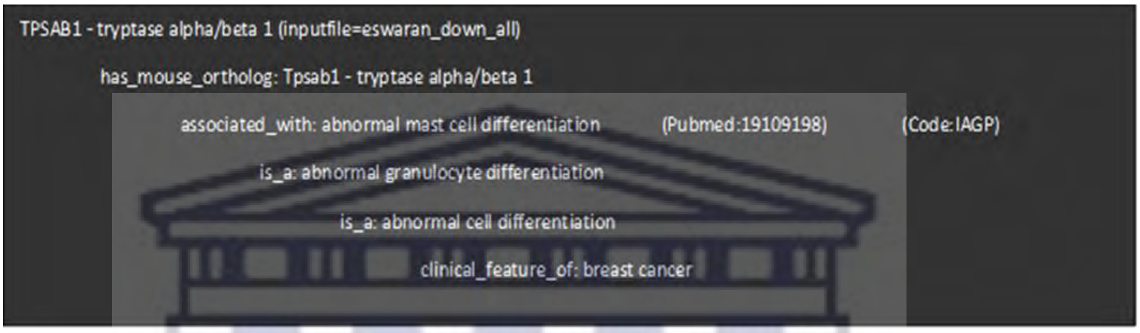
Examples of novel genes picked up by our Breast cancer semantic system include: *PLEC*, *VANGL2, CHST11*, *FZD8*, *HNRNPA3* among the upregulated genes, and *ELP2*, *FBP1*, *KDM4B*, *RHOB*, and *TPSAB1* as downregulated genes. However, some of these genes had paucity of information, e.g., *VANGL2* and *TPSAB1*. The *VANGL2* (VANGL planar cell polarity protein 2) gene product interacts with *FGF8*, a fibroblast growth factor 8, which is known to be involved in breast cancer, thereby implicating the candidate gene. *FGF8* is implicated in cell growth, proliferation, and differentiation, which are crucial processes in both normal development and cancer progression. Liu et al. (2014) reported that the expression levels of *FGF8* correlate with those of *HBXIP* in clinical breast cancer tissues, where *HBXIP* activates the LXRs/SREBP-1c/FAS signaling cascade. This enhances the abnormal lipid metabolism and growth of breast cancer cells (Zhao, 2016).



```
VANGL2 - VANGL planar cell polarity protein 2 (inputfile=eswaran_up_all)
      interacts_with: FGF8 - fibroblast growth factor 8  (Pubmed:28514442)     (Code:BIOGRID)
            implicated_in: breast cancer (Pubmed:10023681)     (Code:IEP)
```

Figure 2.11 The result from the comprehensive database for the upregulated *VANGL* planar cell polarity protein 2 gene.

The *TPSAB1* (tryptase alpha/beta 1) gene was annotated by the comprehensive model as having a mouse ortholog that is associated with abnormal mast cell differentiation. This gene was further annotated as associated with abnormal granulocyte differentiation and abnormal cell differentiation, which are clinical features of breast cancer. This indicates that this gene may have roles in the breast cancer phenotype.



```
TPSAB1 - tryptase alpha/beta 1 (inputfile=eswaran_down_all)
        has_mouse_ortholog: Tpsab1 - tryptase alpha/beta 1
              associated_with: abnormal mast cell differentiation        (Pubmed:19109198)        (Code:IAGP)
                    is_a: abnormal granulocyte differentiation
                          is_a: abnormal cell differentiation
                                clinical_feature_of: breast cancer
```

Figure 2.12 The result from the comprehensive database for the downregulated tryptase alpha/beta 1 gene.

There are also genes prioritised by the system that have not yet been associated with breast cancer. For example, *HNRNPA3* (heterogeneous nuclear ribonucleoprotein A3) has a mouse ortholog that is associated with abnormal neuron differentiation, abnormal DNA repair, mitotic nondisjunction, and increased mitotic index, which are attributes of breast cancer. The system further indicated that this gene interacts with *ESR1* - estrogen receptor 1, *ICAM1* - intercellular adhesion molecule 1, and *SNW*1 - SNW domain containing 1. These interactions, processes, and associations are commonly at the core of human breast cancer development. The detailing in the iterations mentioned above indicate that the comprehensive Breast cancer semantic system has merit in genomic studies, especially in elucidating potentially novel disease genes.

2.3.1.1 Semantic discovery strongly implicates previously hypothesized BCA genes in the disease

When we applied the novel path-based guilt-by-indirect-association approach to analyze genes previously implicated in the development of breast cancer, we were able to extract a significant amount of information from the BCA-BORG semantic network. This information would have been exceedingly challenging to uncover without the automation and reasoning approach provided by our system. In some instances, the system provided both functional and phenotypic evidence, as illustrated in Figure 2.4.

One particularly notable aspect of the system's performance was its ability to traverse multiple levels within an ontology, identifying transitive links that were not immediately obvious but were biologically relevant and semantically correct. Moreover, in addition to uncovering evidence implicating a gene of interest as a potential disease gene, the system had the potential to suggest possible mechanisms related to the gene. This comprehensive approach not only supported the gene-disease association but also provided insights into the potential biological mechanisms underlying the gene's involvement in breast cancer.

2.4 CONCLUSION

The evidence presented in this chapter provides strong support for the idea that ontology-driven semantic discovery has the potential to implicate novel genes in breast cancer, particularly when those genes are differentially expressed or carry high-impact functional mutations. Further details and insights on this topic will be presented in subsequent chapters, where the system's effectiveness in identifying and understanding the functional relationships between genes and the disease will be explored in more depth.

REFERENCES

- Albert, F. W., & Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212. https://doi.org/10.1038/nrg3891
- Beebe, K., & Kennedy, A. D. (2016). Sharpening Precision Medicine by a Thorough Interrogation of Metabolic Individuality. *Computational and Structural Biotechnology Journal*, *14*, 97–105. https://doi.org/10.1016/j.csbj.2016.01.001
- Berardini, T. Z., Khodiyar, V. K., Lovering, R. C., and Talmud, P. (2010). The gene ontology in 2010: extensions and refinements. Nucleic acids research, 38(Databa): D331–D335.
- Bult, C. J., Eppig, J. T., Blake, J. A., Kadin, J. A., Richardson, J. E., and, M. G. D. G. (2013). The mouse genome database: genotypes, phenotypes, and models of human disease. Nucleic Acids Res, 41(Database issue): D885–D891.
- Brunt, L., Greicius, G., Rogers, S., Evans, B. D., Virshup, D. M., Wedgwood, K. C., & Scholpp, S. (2021). Vangl2 promotes the formation of long cytonemes to enable distant Wnt/β-catenin signaling. Nature Communications, 12(1), 2058.
- Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., & Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database: Sharing knowledge in Uniprot with Gene Oncology. *Nucleic Acids Research*, *32*(DATABASE ISS.), 262–266. https://doi.org/10.1093/nar/gkh021

- Eswaran, J., Cyanam, D., Mudvari, P., Reddy, S. D. N., Pakala, S. B., Nair, S. S., & Kumar, R. (2012). Transcriptomic landscape of breast cancers through mRNA sequencing. Scientific reports, 2(1), 264.
- Green, M. and Karp, P. (2006). The outcomes of pathway database computations depend on pathway ontology. Nucleic acids research, 34(13):3687–3697.
- Kaldunski, M. L., Smith, J. R., Hayman, G. T., Brodie, K., De Pons, J. L., Demos, W. M., Gibson, A. C., Hill, M. L., Hoffman, M. J., Lamers, L., Laulederkind, S. J. F., Nalabolu, H. S., Thorat, K., Thota, J., Tutaj, M., Tutaj, M. A., Vedi, M., Wang, S. J., Zacher, S., … Kwitek, A. E. (2022). The Rat Genome Database (RGD) facilitates genomic and phenotypic data integration across multiple species for biomedical research. Mammalian Genome, 33(1), 66–80. https://doi.org/10.1007/s00335-021-09932-x
- Kann, M. G. (2009). Advances in translational bioinformatics: Computational approaches for the hunting of disease genes. *Briefings in Bioinformatics*, *11*(1), 96–110. https://doi.org/10.1093/bib/bbp048
- Khamis, A., Raoult, D., & La Scola, B. (2004). rpoB gene sequencing for identification of Corynebacterium species. *Journal of Clinical Microbiology*, *42*(9), 3925–3931. https://doi.org/10.1128/JCM.42.9.3925-3931.2004
- Kiehn, O., & Car. (2017). 乳鼠心肌提取 HHS Public Access. *Physiology & Behavior*, *176*(3), 139–148. https://doi.org/10.1097/MED.0000000000000150.Next-generation

- Liu, F., You, X., Wang, Y., Liu, Q., Liu, Y., Zhang, S., ... & Ye, L. (2014). The oncoprotein HBXIP enhances angiogenesis and growth of breast cancer through modulating FGF8 and VEGF. Carcinogenesis, 35(5), 1144-1153.
- Marian, A. J. (2012). Challenges in Medical Applications of Whole Exome/Genome Sequencing Discoveries. *Trends in Cardiovascular Medicine*, *22*(8), 219–223. https://doi.org/10.1016/j.tcm.2012.08.001
- Meehan, C. J., Goig, G. A., Kohl, T. A., Verboven, L., Dippenaar, A., Ezewudo, M., Farhat, M. R., Guthrie, J. L., Laukens, K., Miotto, P., Ofori-Anyinam, B., Dreyer, V., Supply, P., Suresh, A., Utpatel, C., van Soolingen, D., Zhou, Y., Ashton, P. M., Brites, D., Van Rie, A. (2019). Whole genome sequencing of Mycobacterium tuberculosis: current standards and open issues. *Nature Reviews Microbiology*, *17*(9), 533–545. https://doi.org/10.1038/s41579-019-0214-5
- Nigam, R., Laulederkind, S. J. F., Hayman, G. T., Smith, J. R., Wang, S.-J., Lowry, T. F., Petri, V., De Pons, J., Tutaj, M., Liu, W., Jayaraman, P., Munzenmaier, D. H., Worthey, E. A., Dwinell, M. R., Shimoyama, M., and Jacob, H. J. (2013). Rat genome database: a unique resource for rat, human, and mouse quantitative trait locus data. Physiol Genomics, 45(18):809–816.
- Robinson, P. N., & Mundlos, S. (2010). The human phenotype ontology. Clinical genetics, 77(6), 525-534.
- Sadedin, S. P., Dashnow, H., James, P. A., Bahlo, M., Bauer, D. C., Lonie, A., Lunke, S., Macciocca, I., Ross, J. P., Siemering, K. R., Stark, Z., White, S. M., Taylor, G., Gaff, C., Oshlack, A., & Thorne, N. P. (2015). Cpipe: A shared

variant detection pipeline designed for diagnostic settings. *Genome Medicine*, *7*(1), 1–10. https://doi.org/10.1186/S13073-015-0191-X/FIGURES/4

- Saunders, C. J., Jalali Sefid Dashti, M., & Gamieldien, J. (2016). Semantic interrogation of a multi knowledge domain ontological model of tendinopathy identifies four strong candidate risk genes. Scientific reports, 6(1), 19820.
- Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W. W., Mazaitis, M., Felix, V., & Kibbe, W. A. (2012). Disease Ontology: a backbone for disease semantic integration. Nucleic acids research, 40(D1), D940-D946.
- Singhal, A., Simmons, M., & Lu, Z. (2016). Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. *PLoS Computational Biology*, *12*(11), 1–19. https://doi.org/10.1371/journal.pcbi.1005017
- Smith, C. L., & Eppig, J. T. (2012). The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. Mammalian genome, 23, 653-668.
- Sowa, J. F. (2011). Future directions for semantic systems. Intelligence-based systems engineering, 23-47.
- Storey, V. C., & Song, I. Y. (2017). Big data technologies and Management: What conceptual modeling can do. *Data and Knowledge Engineering*, *108*(February), 50–67. https://doi.org/10.1016/j.datak.2017.01.001
- Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics*, *20*(8), 467–484. https://doi.org/10.1038/s41576-019-0127-1

- Thomas, P. D., Mi, H., & Lewis, S. (2007). Ontology annotation: mapping genomic regions to biological function. *Current Opinion in Chemical Biology*, *11*(1), 4–11. https://doi.org/10.1016/j.cbpa.2006.11.039
- Twigger, S. N., Shimoyama, M., Bromberg, S., Kwitek, A. E., & Jacob, H. J. (2007). The Rat Genome Database, update 2007 - Easing the path from disease to data and back again. *Nucleic Acids Research*, *35*(SUPPL. 1), 658–662. https://doi.org/10.1093/nar/gkl988
- Zhang, H., Guo, Y., Li, Q., George, T. J., Shenkman, E., Modave, F., & Bian, J. (2018). An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. *BMC Medical Informatics and Decision Making*, *18*(2), 129–147. https://doi.org/10.1186/S12911-018-0636-4/FIGURES/11
- Zhao, Y., Li, H., Zhang, Y., Li, L., Fang, R., Li, Y., & Ye, L. (2016).
- Oncoprotein HBXIP modulates abnormal lipid metabolism and growth of breast cancer cells by activating the LXRs/SREBP-1c/FAS signaling cascade. Cancer research, 76(16), 4696-4707.

CHAPTER 3

ANALYSIS OF GENOMIC MUTATIONS

3.1 INTRODUCTION

Cancer is generally a heterogeneous disease having diverse genetic variations (Shao et al., 2019). It has been known for decades that this diversity may be demonstrated by single-nucleotide polymorphisms (SNPs) which are amongst the most frequently encountered forms of human genetic variation (Redon et al., 2009). These are good resources for mapping complex genetic traits (Marth et al., 1999). Ahmad & Shah 2021 demonstrated that SNPs can be used for diagnosis of breast cancer (Ahmad & Shah, 2021).

Single nucleotide variations (SNVs) and copy number variations (CNVs) are currently the two major types of genomic alterations associated with tumorigenesis (Xu, 2021).

Research has further shown evidence that copy number variations (CNVs) of certain genes are involved in development and progression of many cancers (Shao et al., 2019). They indicated a caveat whether the CNVs and multiple cancers correlation can be considered a general phenomenon. However genomic mutation is the key element influencing gene expression and function, and hence greatly contributes to the phenotype (Hollander et al., 2018).

Genetic structural variation in the genome may also occur due to large chromosome aberrations besides CNVs and SNVs. (Redon et al., 2009).

While SNPs were previously regarded as the predominant form of structural variation and accounted for much phenotypic variation however, CNVs have taken a centre stage in research (Gökçümen & Lee, 2009).

Copy number variation (CNV) is defined as polymorphism in the human genome involving DNA fragments larger than 1kb. The CNV sites provide hotspots of somatic alterations in various cancers accounting for a crucial part of genetic structural variation (Sebat et al., 2004); (Shao et al., 2019).

Patterns of somatic mutations in cancers has provided valuable information about their role in tumorigenesis, and thus can be explored for intervention in these diseases (Murakami et al., 2021). TP53 is an example of a gene associated with tumour suppression and promoting activities (Mair et al., 2016).

SNVs have been widely used in evolutionary history studies of cancer because they are accumulated gradually without reverse mutations. This gradual accumulation is demonstrated in the phylogenetic trees of cancers such as prostate cancer, renal cancer and other cancer types (Gao et al., 2020). A number of SNVs are linked with altered human traits and genetic diseases through alteration of the normal activity of existing regulatory elements (Bozhilov et al., 2021). SNVs and CNVs are both structural variants and can play a role in modulation of the upregulation and downregulation of genes in some cases.

In this study frequencies, upregulation and downregulation of the SNVs and CNVs genes implicated in breast cancer are compared to other cancers. The genes of interest are those that are not well reported either broadly across all cancer types or specifically in BCA.

3.2 RESEARCH AIMS AND OBJECTIVES

The aim of the study was to compare mutation frequency of all genes in breast cancer samples in the Genomic Data Commons (GDC) Portal (GDC data portal_ https://portal.gdc.cancer.gov/) using the TCGA data to all other cancers to identify those substantially more frequently mutated in the former using the developed semantic database (BORG).

The candidate genes in the different mutation categories were subjected to enrichment analysis to discover non-classical cancer genes that are associated with breast cancer. Those genes not assigned to any cancer-relevant enriched functions and pathways were then analysed using the here developed semantic database to evaluate potential novel roles in cancer biology.

3.3 MATERIALS AND METHODS

The Genomic Data Commons Portal (GDC data portal_ https://portal.gdc.cancer.gov/) repository was used to manually curate and perform exploratory analysis by identifying variations in cancer cells that may play a vital role in breast tissue carcinogenesis development. The data gleaned from the GDC portal helped in identifying both high- and low-frequency cancer drivers such as mutations. The portal provides access to valuable DNA sequence data and generates associated Variant Calling Format (VCF) and Mutation Annotation Format (MAF) files that identify somatic mutations such as point mutations, missense mutations, nonsense mutations, and insertions and deletions (indels) of nucleotides in the DNA. The portal also provides access to Copy Number Variation (CNV) data to identify amplified and attenuated gene expression due to chromosomal duplications, losses, insertions and deletions (GDC data portal_ https://portal.gdc.cancer.gov/).

3.3.1 Selection criteria for subsequent analysis

The top 500 frequently mutated genes were selected from the DGC portal (TCGA) out of 11519 of all the cancer patients (Table 3.1 in Appendix 1). The focus was in (i) the number of the mutations in the Breast Cancer cohort (*TCGA-ductal and lobular neoplasm-primary tumor-genes)* i.e. the number of cases where the gene is mutated/cases tested for simple Somatic mutations (SNVs), (ii) the number of mutations overall in the same gene i.e. number of cases where genes contain simple Somatic mutations/number of cases tested for simple Somatic mutations, (iii) the number of duplications i.e. number of cases where CNV gain events are observed in gene/number of cases tested for Copy Number Alteration in gene and (iv) the number of gene loses i.e. number of cases where CNV gain events are observed in gene/ number of cases tested for Copy Number Alteration in gene. We then selected the top 500 frequently mutated genes (GDC portal).

Firstly, the data was from the samples of affected cases in the breast cancer cohort matched against the sample of affected cases in all cancer cohorts. Furthermore, to streamline data SNVs and CNVs were selected as the other main parameters. Thereafter for each of the SNVs and CNVs gains and losses were used as further defining parameters.

The frequencies of breast cancer mutation were calculated from the number of affected breast cancer cohort to the total of cancer cohorts in the data. Thereafter the frequencies of each SNVs and CNVs gains in genes were calculated from patients with positive gains and the total number of patients where these gains were calculated. Similarly, percentage losses were calculated for each of the 500 selected genes. These constituted the frequencies of the mutations in these genes.

3.4 RESULTS AND DISCUSSION

Selection criteria for subsequent analysis

The top 500 genes from the DGC portal comparing mutations in breast cancer patients and all cancers are reflected in Table 3.1 in the link. Out of all cancer patients (11 519), 954 patients had single nucleotide mutations. The percentage mutation for each gene SNV was calculated by dividing the number of specific SNV by the total number of SNV mutation in all BCA patients. There were 1029 affected patients with CNV amongst all cancer cohorts. In order to calculate % CNV per gene the number of specific CNV mutation was divided from the total number of CNVs. After cleaning the results according to the selected parameters and applying the cut off value (5%) genes of interest were obtained. This resulted in 17 SNVs with higher frequencies in BCA than other cancers and 135 CNVs that were frequently upregulated and 124 frequently down regulated CNVs, respectively. The fact that these genes are frequently mutated already but there are differences in mutation frequencies between breast cancer and other cancers is what makes it interesting.

Hereafter, these genes were then further analysed in Enrichr to see which classes of these genes are frequently duplicated or lost and which pathways are frequently duplicated and lost. On Enrichr, $p \leq 0.05$ was considered significant. To achieve this, the associated pathways and ontologies were analysed, yielding statistically favourable results. The analyses yielded 23 duplicated and 11 lost genes without apparent association with breast cancer Table 3.2 and Table 3.4 (Appendix I). For example, *CACNA1E* is duplicated in almost 10% of breast cancer patients while *RB1* has loss in 12% of the same cohort.

Enrichr and scientific literature elucidated some genes not previously describes as classical cancer genes where others had no link with breast cancer. These genes include *GOLGA6L6*, *XKR4*, *REXO1L1P*, *NBPF12*. Selected mutated genes not yet linked to breast cancer will be discussed.

GOLGA6L6 has not been associated with either cancer or any disorder (GeneCards, accessed on the 11/19/2021). While Enrichr did not link this gene with any pathway, GeneCards showed protein interactions associated with this gene (Figure 3.1). In this study SNVs in *GOLGA6L6* has been linked to 5% of BC patients compared to 2% of all other cancers. Furthermore, *GOLGA6L6* CNV was also lost in 9% and duplicated in 5% of breast cancer samples.

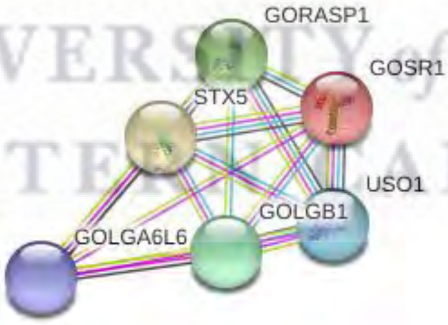


Figure 3.1: Selected Interacting proteins linked to *GOLGA6L6* Gene expression (GeneCards).

We found downregulation of CNVs in *XKR4* gene with the frequency of 6.32% in BCA compared to other cancers. In as much as this gene is not associated with any disorders in Gene cards, literature has linked this gene to papillary thyroid carcinoma (Zhan et al., 2014). SNVs in this gene was below the set threshold of 5% while both upregulation and downregulation of CNVs was above threshold at 9% and 6%, respectively.

In this study only SNV for the *REXO1L1P* gene was of significance with a frequency of 1.5 in breast cancer compared to all other cancers. However, looking at breast cancer mutations separately, these SNVs were 3.04%. The CNVs were insignificant. Chen et al., 2017 indicated that the SNV in this gene is associated with stop-gains in BC.

According to GeneCards *NBPF12* is not yet associated with any disorder or pathway. However, recently the Neuroblastoma Breakpoint Family (NBPF) has been linked to increased brain size and neuropsychiatric diseases, including autism and schizophrenia (Benton et al., 2021). In this study the SNV in this gene and its CNV down regulation were insignificant while the CNV upregulation was 18% in breast cancer patients.

To date the *ERICH3* gene has not yet been conclusively associated with any disorder and its associated pathways are still to be elucidated. In this study, SNV for *ERICH3* and its CNV upregulation were below the set thresholds, while the CNV downregulation 5.54% in breast cancer cohort.

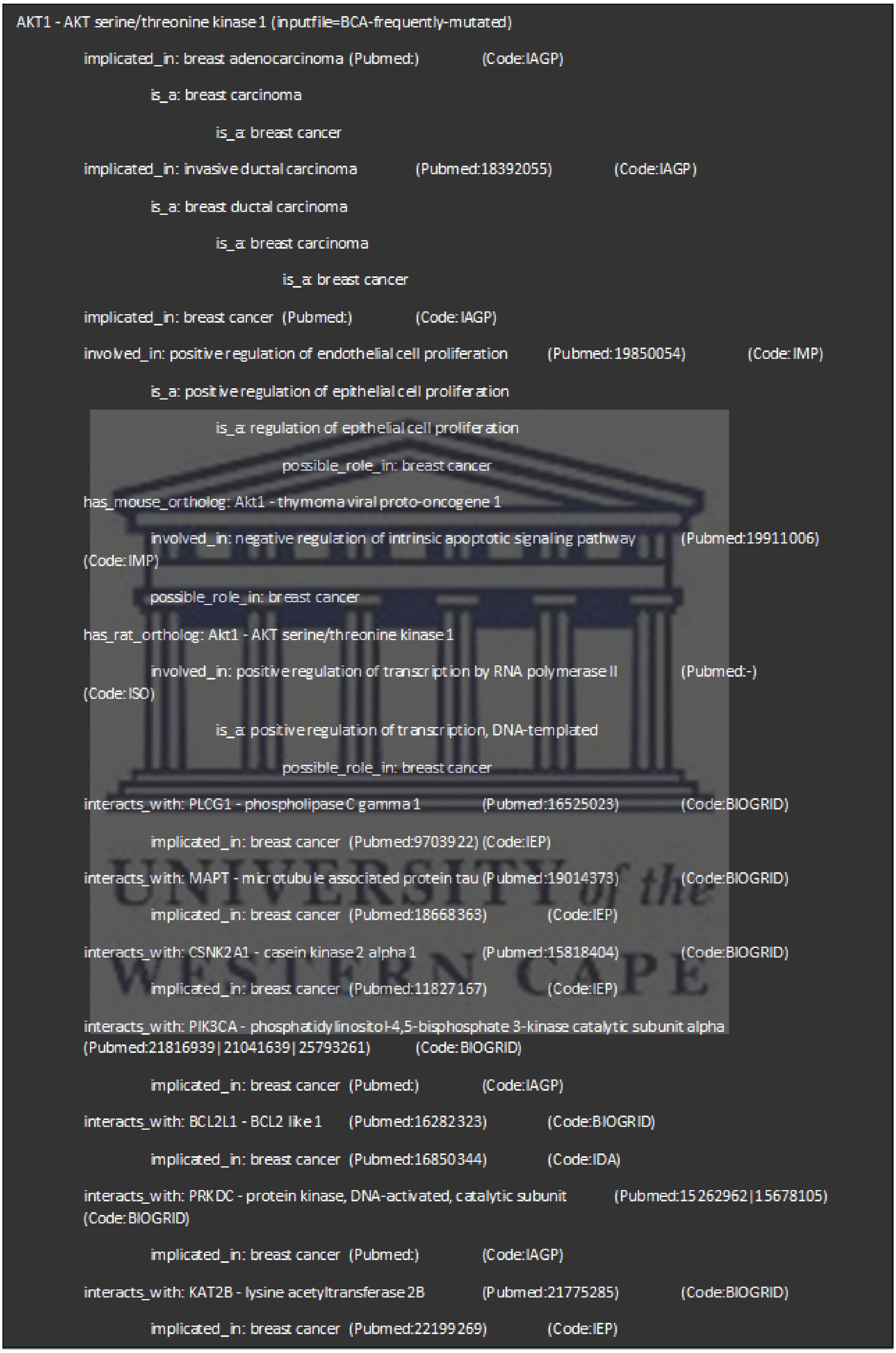
A selection in both highly relevant genes and insignificant genes were then run through the developed semantic BORG database to see if any of the latter returned with better annotations.

Semantic analysis results for frequently mutated genes

The knowledge graph returned several genes including the frequently mutated, the upregulated and the down-regulated ones. Selected genes are presented below. A comprehensive list of the BORG results is presented in "Additional Files" (https://drive.google.com/drive/u/0/folders/1vkBCRNBApPGzzM1LRIqxj19vz3KE6B97). Amongst the genes returned are known BCA genes and genes that were previously not associated with BCA. Some of these genes are implicated in BCA via protein–protein interactions while others may be associated via extrinsic and intrinsic apoptotic signaling pathways, cytochrome release, cell proliferation, cyclin-dependent protein kinase activity, *inter alia.* Some of the novel genes returned are implicated by prior knowledge from rat and mouse orthologs. Amongst the known BCA genes are *AKT1*, *CBFB*, *CDH1*, *GATA3*, *MAP2K4*, *MAP3K1*, *PIK3CA* and *RUNX1*.

BORG annotated *AKT1* with diverse biological processes associated with human breast cancer including effects in breast adenocarcinoma, invasive ductal carcinoma, negative regulation of extrinsic apoptotic signaling pathway in absence of ligand, negative regulation of oxidative stress-induced intrinsic apoptotic signaling pathway, positive regulation of smooth muscle cell proliferation, positive regulation of transcription, DNA-templated, negative regulation of apoptotic process, amongst others. These processes are involved in carcinogenesis as well as the progression of the disease. For an example the negative regulation of extrinsic/intrinsic apoptotic signaling pathway inhibits apoptosis of breast cancer cells immortalising them (Yang et al., 2021). On the other hand invasive ductal carcinoma correlates with angiogenesis-related factors in metastatic breast cancer (Li et al., 2021).

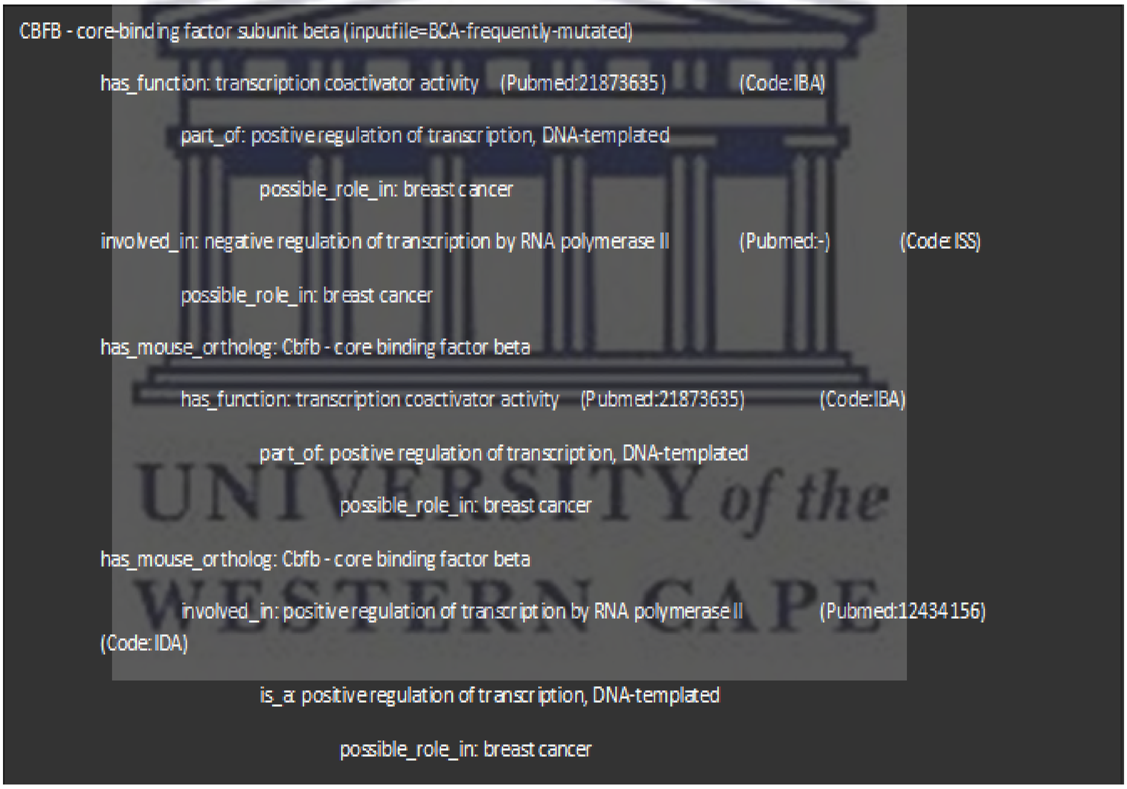
The system also annotated *AKT1* (Figure 3.2) interactions with numerous proteins including *IL13RA2*, *EGFR*, *PLCG1*, *MAPT*, *CSNK2A1*, *PIK3CA*, *BCL2L1*, *PRKDC*, *KAT2B*, *NOS3*, *PTEN*, *ESR1*, *SKP2*, *TSC1*, *IKBKB*, *XIAP*, *NCOR2*, *NF2 - NF2*, *CDKN1B*, *TERT*, *BAD*, *TSC2*, *SMAD4*, *PAK1*, *FANCA*, *BRCA1*. Furthermore, *AKT1* returned with rat and mouse orthologs for an example where it is associated with thymoma viral proto-oncogene 1 which increases in increased mammary adenocarcinoma incidence and increased mammary gland tumor incidence.



```
AKT1 - AKT serine/threonine kinase 1 (inputfile=BCA-frequently-mutated)
          implicated_in: breast adenocarcinoma (Pubmed:)          (Code:IAGP)
                    is_a: breast carcinoma
                              is_a: breast cancer
          implicated_in: invasive ductal carcinoma          (Pubmed:18392055)          (Code:IAGP)
                    is_a: breast ductal carcinoma
                              is_a: breast carcinoma
                                        is_a: breast cancer
          implicated_in: breast cancer  (Pubmed:)          (Code: IAGP)
          involved_in: positive regulation of endothelial cell proliferation       (Pubmed:19850054)          (Code: IMP)
                    is_a: positive regulation of epithelial cell proliferation
                              is_a: regulation of epithelial cell proliferation
                                        possible_role_in: breast cancer
          has_mouse_ortholog: Akt1 - thymoma viral proto-oncogene 1
                    involved_in: negative regulation of intrinsic apoptotic signaling pathway          (Pubmed:19911006)
          (Code: IMP)
                    possible_role_in: breast cancer
          has_rat_ortholog: Akt1 - AKT serine/threonine kinase 1
                    involved_in: positive regulation of transcription by RNA polymerase II          (Pubmed:-)
          (Code: ISO)
                              is_a: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
          interacts_with: PLCG1 - phospholipase C gamma 1          (Pubmed:16525023)          (Code:BIOGRID)
                    implicated_in: breast cancer  (Pubmed:9703922) (Code:IEP)
          interacts_with: MAPT - microtubule associated protein tau (Pubmed:19014373)          (Code:BIOGRID)
                    implicated_in: breast cancer  (Pubmed:18668363)          (Code:IEP)
          interacts_with: CSNK2A1 - casein kinase 2 alpha 1          (Pubmed:15818404)          (Code:BIOGRID)
                    implicated_in: breast cancer  (Pubmed:11827167)          (Code:IEP)
          interacts_with: PIK3CA - phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
          (Pubmed:21816939|21041639|25793261)          (Code: BIOGRID)
                    implicated_in: breast cancer  (Pubmed:)          (Code:IAGP)
          interacts_with: BCL2L1 - BCL2 like 1     (Pubmed:16282323)          (Code:BIOGRID)
                    implicated_in: breast cancer  (Pubmed:16850344)          (Code:IDA)
          interacts_with: PRKDC - protein kinase, DNA-activated, catalytic subunit          (Pubmed:15262962|15678105)
          (Code:BIOGRID)
                    implicated_in: breast cancer  (Pubmed:)          (Code:IAGP)
          interacts_with: KAT2B - lysine acetyltransferase 2B          (Pubmed:21775285)          (Code:BIOGRID)
                    implicated_in: breast cancer  (Pubmed:22199269)          (Code:IEP)
```

Figure 3.2. Semantic analysis results of *AKT1* gene (frequently mutated).

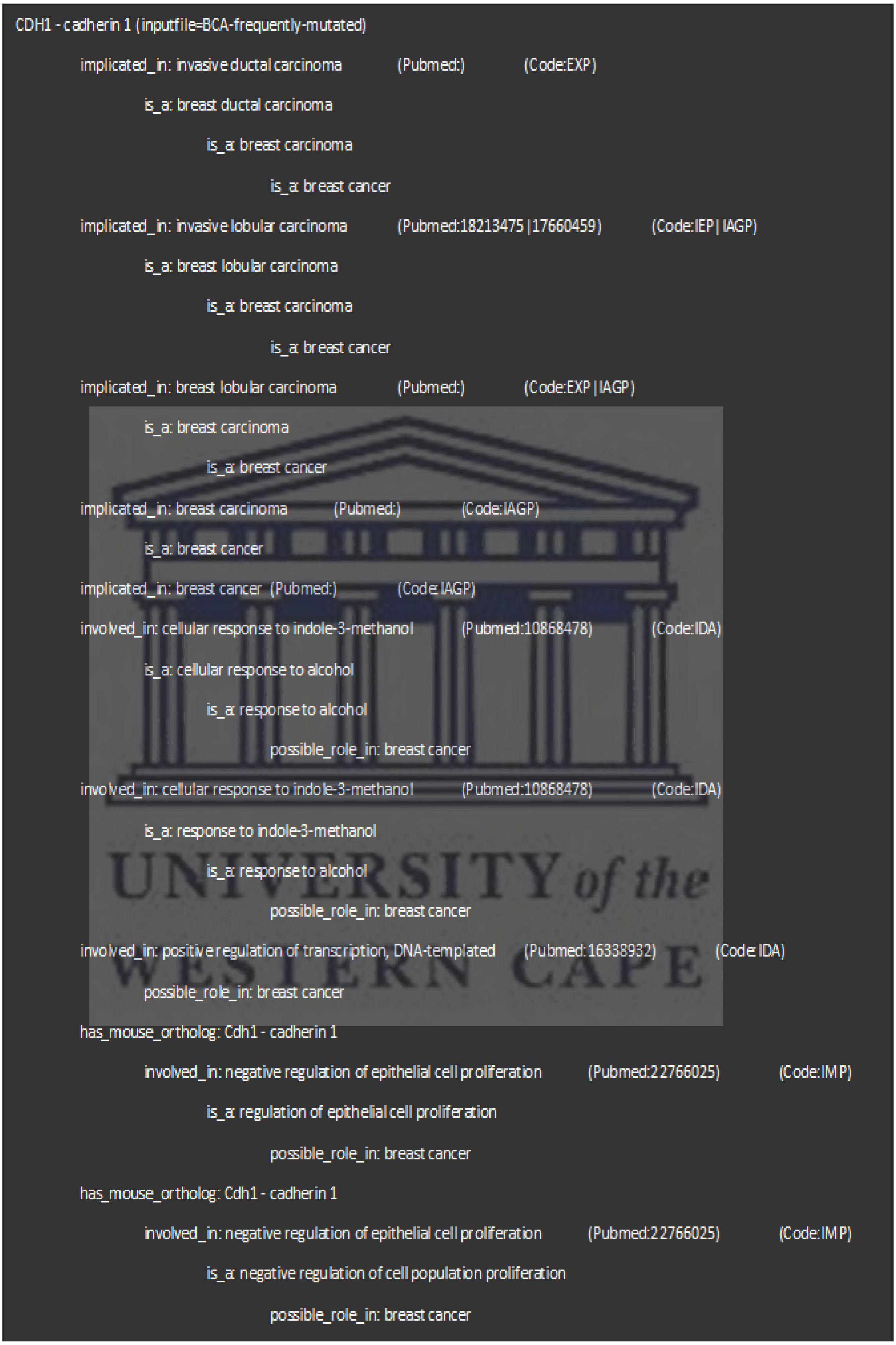
CBFB (Figure 3.3) returned with only one ortholog that has transcription coactivator activity (Pubmed:21873635). It is involved in negative regulation of transcription by RNA polymerase II, thus implicated in BCA. The protein–protein interactions associated with human breast cancer genes that were returned for *CBFB* were *MYC* and *RUNX3*.This gene had several rat and mouse orthologs associated with breast cancer functions including involvement in negative regulation of transcription by RNA polymerase II and abnormal osteoblast differentiation in mouse. In rats this gene returned with association transcription coactivator activity.



```
CBFB - core-binding factor subunit beta (inputfile=BCA-frequently-mutated)
          has_function: transcription coactivator activity    (Pubmed:21873635)          (Code:IBA)
                    part_of: positive regulation of transcription, DNA-templated
                              possible_role_in: breast cancer
          involved_in: negative regulation of transcription by RNA polymerase II          (Pubmed:-)          (Code: ISS)
                    possible_role_in: breast cancer
          has_mouse_ortholog: Cbfb - core binding factor beta
                    has_function: transcription coactivator activity    (Pubmed:21873635)          (Code:IBA)
                              part_of: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
          has_mouse_ortholog: Cbfb - core binding factor beta
                    involved_in: positive regulation of transcription by RNA polymerase II          (Pubmed:12434156)
          (Code: IDA)
                              is_a: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
```

Figure 3.3. Semantic analysis results of *CBFB* gene (frequently mutated).

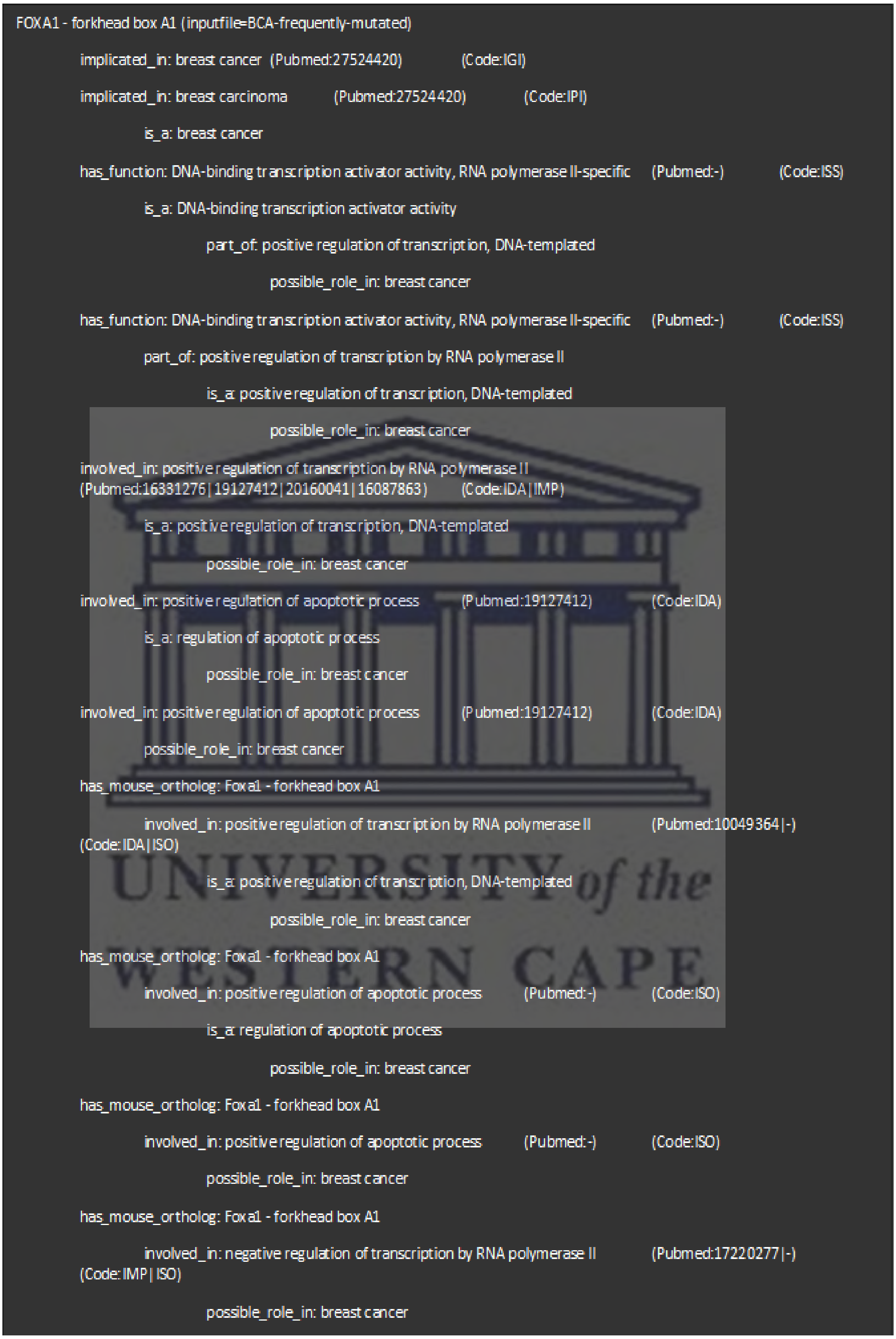
CDH1 (Figure 3.4) was shown to be previously associated with invasive ductal carcinoma, invasive lobular carcinoma and breast lobular carcinoma, and cellular response to indole-3-methanol, positive regulation of transcription, DNA-templated. A clinical case study previously reported the importance of *CDH1* germline variants in patients with lobar breast cancer and gastrointestinal cancers (Adib et al., 2022). The protein – protein interactions returned by the system included *F11R*, *RB1CC1*, *STAT1*, *NF2*, *DDX3X*, *CTSB*, *CASP3*, *DDR1*, *SKP2*, *PTEN*, *EGFR*. CDH1 returned with several biological functions associated with cancer in rat and mouse, e.g. abnormal tumor vascularization, regulation of morphogenesis, increased metastatic potential in rats and mouse amongst others.



```
CDH1 - cadherin 1 (inputfile=BCA-frequently-mutated)
    implicated_in: invasive ductal carcinoma      (Pubmed:)      (Code:EXP)
        is_a: breast ductal carcinoma
            is_a: breast carcinoma
                is_a: breast cancer
    implicated_in: invasive lobular carcinoma      (Pubmed:18213475 |17660459)      (Code:IEP| IAGP)
        is_a: breast lobular carcinoma
            is_a: breast carcinoma
                is_a: breast cancer
    implicated_in: breast lobular carcinoma      (Pubmed:)      (Code:EXP | IAGP)
        is_a: breast carcinoma
            is_a: breast cancer
    implicated_in: breast carcinoma      (Pubmed:)      (Code:IAGP)
        is_a: breast cancer
    implicated_in: breast cancer  (Pubmed:)      (Code:IAGP)
    involved_in: cellular response to indole-3-methanol      (Pubmed:10868478)      (Code:IDA)
        is_a: cellular response to alcohol
            is_a: response to alcohol
                possible_role_in: breast cancer
    involved_in: cellular response to indole-3-methanol      (Pubmed:10868478)      (Code:IDA)
        is_a: response to indole-3-methanol
            is_a: response to alcohol
                possible_role_in: breast cancer
    involved_in: positive regulation of transcription, DNA-templated      (Pubmed:16338932)      (Code:IDA)
        possible_role_in: breast cancer
    has_mouse_ortholog: Cdh1 - cadherin 1
        involved_in: negative regulation of epithelial cell proliferation      (Pubmed:22766025)      (Code:IMP)
            is_a: regulation of epithelial cell proliferation
                possible_role_in: breast cancer
    has_mouse_ortholog: Cdh1 - cadherin 1
        involved_in: negative regulation of epithelial cell proliferation      (Pubmed:22766025)      (Code:IMP)
            is_a: negative regulation of cell population proliferation
                possible_role_in: breast cancer
```

Figure 3.4. Semantic analysis results of *CDH1*gene (frequently mutated).

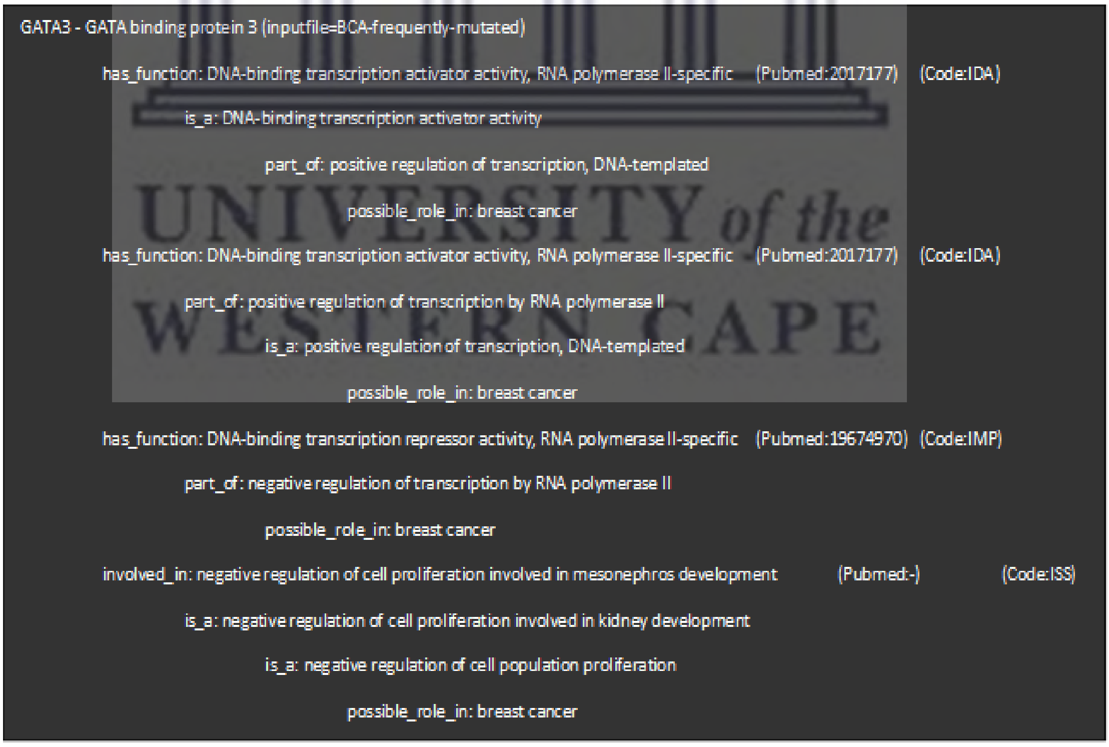
FOXA1(Figure 3.5) returned without protein–protein interactions however with numerous biological processes associated with cancer in human, and rat and mice. The system yielded functions that are associated with transcription activation, apoptotic processes, and uncontrolled growth. These functions were represented by DNA-binding transcription activator activity, RNA polymerase II-specific, positive regulation of transcription by RNA polymerase II, positive regulation of apoptotic process. There were similarities between human, rat and mouse gene functions that were returned by our system in as far as DNA-binding transcription activator activity, RNA polymerase II-specific, negative regulation of transcription by RNA polymerase II and positive regulation of apoptotic process. However, there were additional biological processes associated with cancer returned by the system via ortholog knockouts in rats and mice. These include positive regulation of intracellular estrogen receptor signaling pathway, decreased incidence of tumors by chemical induction, increased incidence of tumors by chemical induction, increased hepatocellular carcinoma incidence *inter alia*.



```
FOXA1 - forkhead box A1 (inputfile=BCA-frequently-mutated)
        implicated_in: breast cancer  (Pubmed:27524420)        (Code:IGI)
        implicated_in: breast carcinoma       (Pubmed:27524420)        (Code:IPI)
                is_a: breast cancer
        has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:-)        (Code:ISS)
                is_a: DNA-binding transcription activator activity
                        part_of: positive regulation of transcription, DNA-templated
                                possible_role_in: breast cancer
        has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:-)        (Code:ISS)
                part_of: positive regulation of transcription by RNA polymerase II
                        is_a: positive regulation of transcription, DNA-templated
                                possible_role_in: breast cancer
        involved_in: positive regulation of transcription by RNA polymerase II
        (Pubmed:16331276|19127412|20160041|16087863)      (Code:IDA|IMP)
                is_a: positive regulation of transcription, DNA-templated
                        possible_role_in: breast cancer
        involved_in: positive regulation of apoptotic process      (Pubmed:19127412)        (Code:IDA)
                is_a: regulation of apoptotic process
                        possible_role_in: breast cancer
        involved_in: positive regulation of apoptotic process      (Pubmed:19127412)        (Code:IDA)
                possible_role_in: breast cancer
        has_mouse_ortholog: Foxa1 - forkhead box A1
                involved_in: positive regulation of transcription by RNA polymerase II        (Pubmed:10049364|-)
        (Code: IDA|ISO)
                        is_a: positive regulation of transcription, DNA-templated
                                possible_role_in: breast cancer
        has_mouse_ortholog: Foxa1 - forkhead box A1
                involved_in: positive regulation of apoptotic process       (Pubmed:-)        (Code:ISO)
                        is_a: regulation of apoptotic process
                                possible_role_in: breast cancer
        has_mouse_ortholog: Foxa1 - forkhead box A1
                involved_in: positive regulation of apoptotic process       (Pubmed:-)        (Code:ISO)
                        possible_role_in: breast cancer
        has_mouse_ortholog: Foxa1 - forkhead box A1
                involved_in: negative regulation of transcription by RNA polymerase II        (Pubmed:17220277|-)
        (Code: IMP| ISO)
                        possible_role_in: breast cancer
```

Figure 3.5. Semantic analysis results of *FOXA1*gene (frequently mutated)

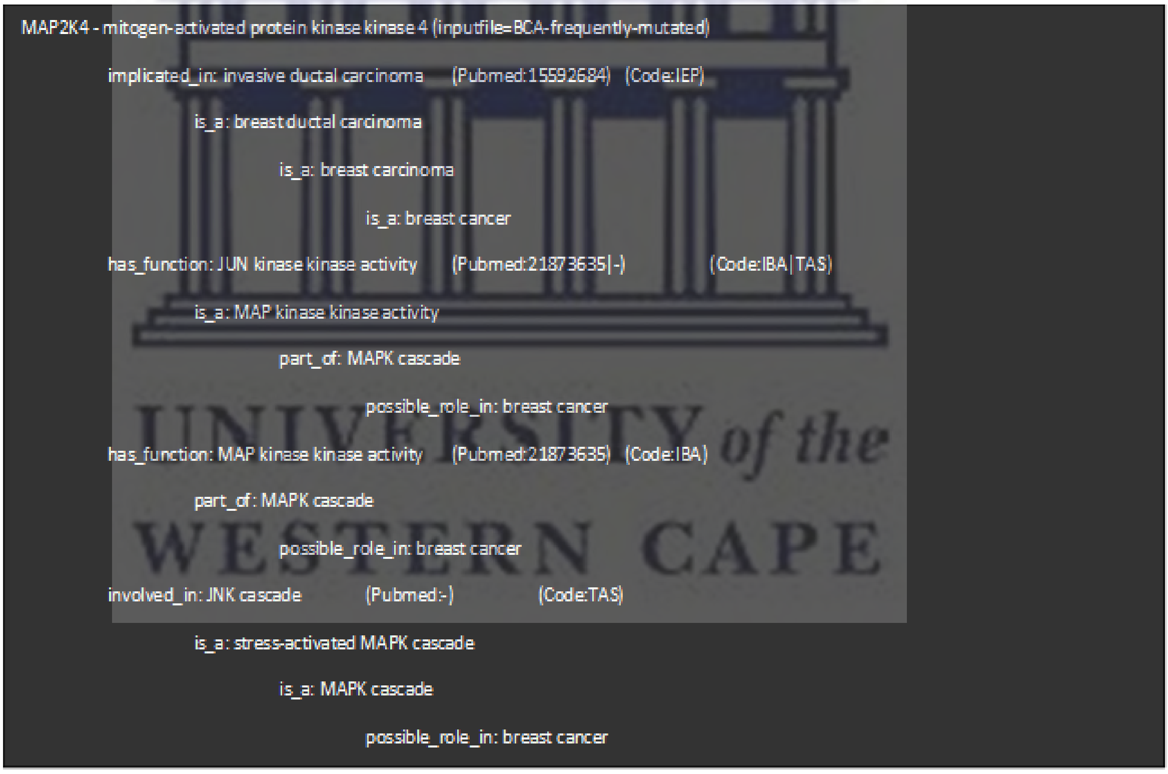
GATA3 (Figure 3.6) returned linking to numerous biological processes, protein-to-protein interactions and with mouse and rat orthologs that differ to human in function. The functions associated cancer included DNA-binding transcription activator activity, transcription repressor activity, negative regulation of cell proliferation, negative regulation of apoptotic process, regulation of transcription and negative regulation of mammary gland epithelial cell proliferation. Protein-protein interactions with known BCA genes were with *CDK2* and *BRCA1*. There were also numerous relevant functions associated with rat and mouse orthologs, including regulation of neuron apoptotic process, regulation of endothelial cell apoptosis, regulation of interleukin-2 production and interferon-gamma production, and regulation of cell proliferation.



```
GATA3 - GATA binding protein 3 (inputfile=BCA-frequently-mutated)
        has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:2017177)    (Code:IDA)
                is_a: DNA-binding transcription activator activity
                        part_of: positive regulation of transcription, DNA-templated
                                possible_role_in: breast cancer
        has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:2017177)    (Code:IDA)
                part_of: positive regulation of transcription by RNA polymerase II
                        is_a: positive regulation of transcription, DNA-templated
                                possible_role_in: breast cancer
        has_function: DNA-binding transcription repressor activity, RNA polymerase II-specific    (Pubmed:19674970)   (Code:IMP)
                part_of: negative regulation of transcription by RNA polymerase II
                        possible_role_in: breast cancer
        involved_in: negative regulation of cell proliferation involved in mesonephros development        (Pubmed:-)          (Code:ISS)
                is_a: negative regulation of cell proliferation involved in kidney development
                        is_a: negative regulation of cell population proliferation
                                possible_role_in: breast cancer
```

Figure 3.6. Semantic analysis results of *GATA3*gene (frequently mutated).

MAP2K4 (Figure 3.7) was shown to be previously associated with invasive ductal carcinoma. Cancer-relevant functions returned included JUN kinase kinase activity, MAP kinase kinase activity and MAPK cascade. Previously, inhibition of the *MAP2K4* signal axis has been reported to have an effect in regulation of the proliferation and apoptosis of cancer cells (B. Wang et al., 2022). The Breast cancer semantic system returned MAP2K4 protein-protein interactions with known BCA gene products, including *EGFR*, *MAPK1*, JUN, *MAP3K1*, *AKT1*, *MAP3K8*. The rat and mouse evidence yielded by the system included: regulation of multiple apoptotic processes and positive regulation of JUN kinase activity.

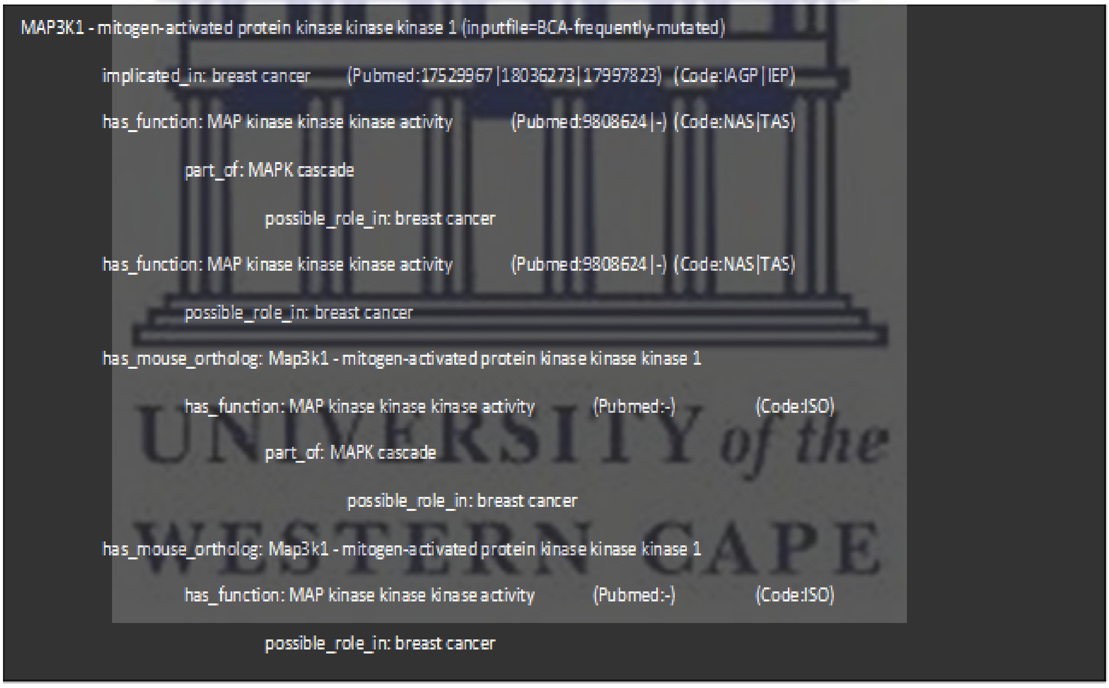


```
MAP2K4 - mitogen-activated protein kinase kinase 4 (inputfile=BCA-frequently-mutated)
        implicated_in: invasive ductal carcinoma    (Pubmed:15592684)  (Code:IEP)
                is_a: breast ductal carcinoma
                        is_a: breast carcinoma
                                is_a: breast cancer
        has_function: JUN kinase kinase activity    (Pubmed:21873635|-)          (Code:IBA|TAS)
                is_a: MAP kinase kinase activity
                        part_of: MAPK cascade
                                possible_role_in: breast cancer
        has_function: MAP kinase kinase activity    (Pubmed:21873635)  (Code:IBA)
                part_of: MAPK cascade
                        possible_role_in: breast cancer
        involved_in: JNK cascade          (Pubmed:-)           (Code:TAS)
                is_a: stress-activated MAPK cascade
                        is_a: MAPK cascade
                                possible_role_in: breast cancer
```

Figure 3.7. Semantic analysis results of *MAP2K4* gene (frequently mutated).

MAP3K1 (Figure 3.8) biological processes that were associated with human breast cancer included the following functions: MAP kinase kinase kinase activity and MAP kinase kinase kinase activity. Literature has reported that *MAP3K1* as one of the MAPK family serine-threonine kinase that is often mutated in human cancer with prognosis in breast (Cheukfai et al., 2022). The *MAP3K* interactions with known BCA genes included: *BRCA1*; *KAT5*; *PAK1*; *GSTP1*; *JUN*; *TP53*; *RHOA*; *IKBKB*; *GRB2*; *MAPK1* and *MAP2K1*.

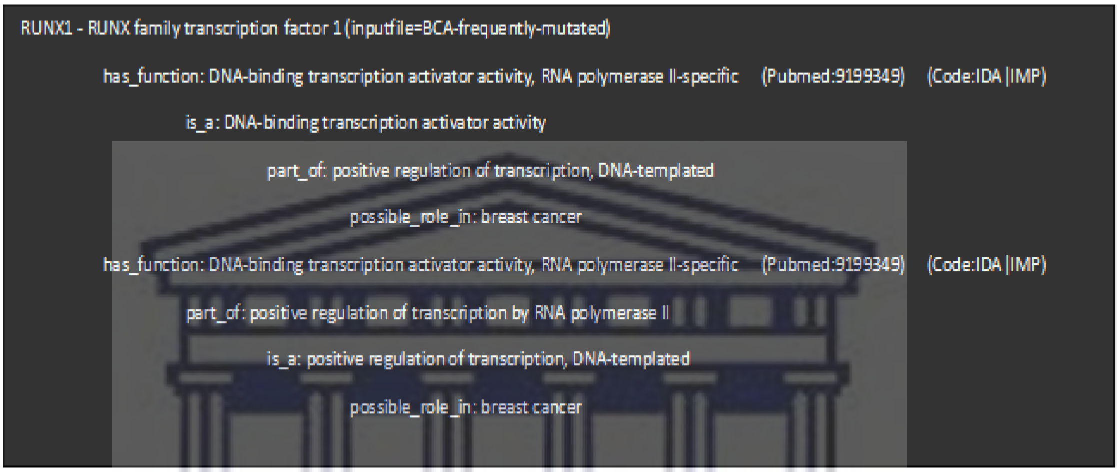
The rat and mouse evidence yielded by the system included the following functions: JUN kinase kinase kinase activity; MAP kinase kinase kinase activity; positive regulation of JUN kinase activity and JNK cascade.



```
MAP3K1 - mitogen-activated protein kinase kinase kinase 1 (inputfile=BCA-frequently-mutated)
          implicated_in: breast cancer      (Pubmed:17529967|18036273|17997823)  (Code:IAGP|IEP)
          has_function: MAP kinase kinase kinase activity        (Pubmed:9808624|-)  (Code:NAS|TAS)
                    part_of: MAPK cascade
                              possible_role_in: breast cancer
          has_function: MAP kinase kinase kinase activity        (Pubmed:9808624|-)  (Code:NAS|TAS)
                    possible_role_in: breast cancer
          has_mouse_ortholog: Map3k1 - mitogen-activated protein kinase kinase kinase 1
                    has_function: MAP kinase kinase kinase activity        (Pubmed:-)          (Code:ISO)
                              part_of: MAPK cascade
                                        possible_role_in: breast cancer
          has_mouse_ortholog: Map3k1 - mitogen-activated protein kinase kinase kinase 1
                    has_function: MAP kinase kinase kinase activity        (Pubmed:-)          (Code:ISO)
                              possible_role_in: breast cancer
```

Figure 3.8. Semantic analysis results of *MAP3K1* gene (frequently mutated).

The system returned *RUNX1* (Figure 3.9) with human, rat and mouse orthologs together with protein-to-protein interactions with known BCA genes. *RUNX1* was shown to be functionally relevant based on functions in regulation of transcription and interactions with the previously BCA-associated proteins: *FANCD2*; SOX2; *NCOR2*; *MYC*; *CDK1*; *SMARCA4*; *VDR*; *KAT6B*. Furthermore, rat and mouse knockout evidence included abnormal tumor morphology, increased lymphoma incidence, increased chronic myelocytic and promyelocytic leukemia and lymphoma incidence, increased incidence of tumors by chemical induction, and increased metastatic potential.



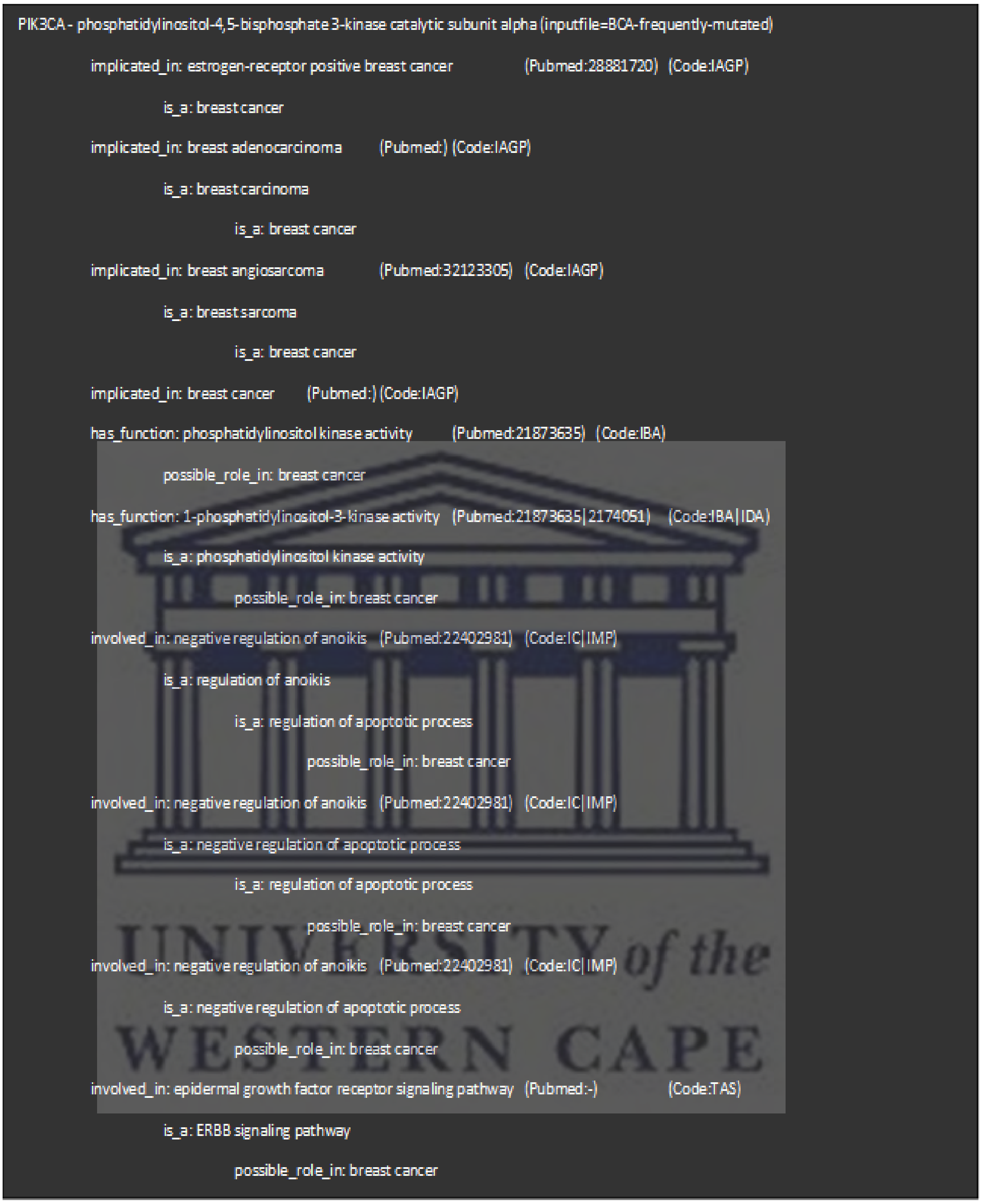
```
RUNX1 - RUNX family transcription factor 1 (inputfile=BCA-frequently-mutated)
          has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:9199349)    (Code:IDA |IMP)
                    is_a: DNA-binding transcription activator activity
                              part_of: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
          has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:9199349)    (Code:IDA |IMP)
                    part_of: positive regulation of transcription by RNA polymerase II
                              is_a: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
```

Figure 3.9. Semantic analysis results of *RUNX1* gene (frequently mutated).

The system annotated *PIK3CA* (Figure 3.10) human, rat and mouse orthologs together with protein-to-protein interactions. In humans *PIK3CA* is implicated in estrogen-receptor positive breast cancer; breast adenocarcinoma, breast angiosarcoma with functions in phosphatidylinositol kinase activity, negative regulation of anoikis. The regulation of anoikis is involved in the regulation of the apoptotic process while others may impact cancer via ERBB signaling pathway.

PIK3CA returned with the following protein-protein interactions: *IL13RA2*; *AKT1*, *ESR1*, *STAT1*, *GRB2*, *KRAS*, *ATR* - *ATR*. These interactions are in breast cancer via different mechanisms.

This system returned similar functions to human ortholog from rat and mouse orthologs. In addition to these functions, rat and mouse orthologs were also involved in negative regulation of fibroblast apoptotic process, negative regulation of neuron apoptotic process, regulation of genetic imprinting, response to dexamethasone, decreased incidence of induced tumors, increased fibroadenoma incidence, increased skin hamartoma incidence, increased mammary gland tumor incidence, increased sarcoma incidence; malignancy amongst others. Some of these features play a possible role in breast cancer development while others are clinical features of breast cancer.

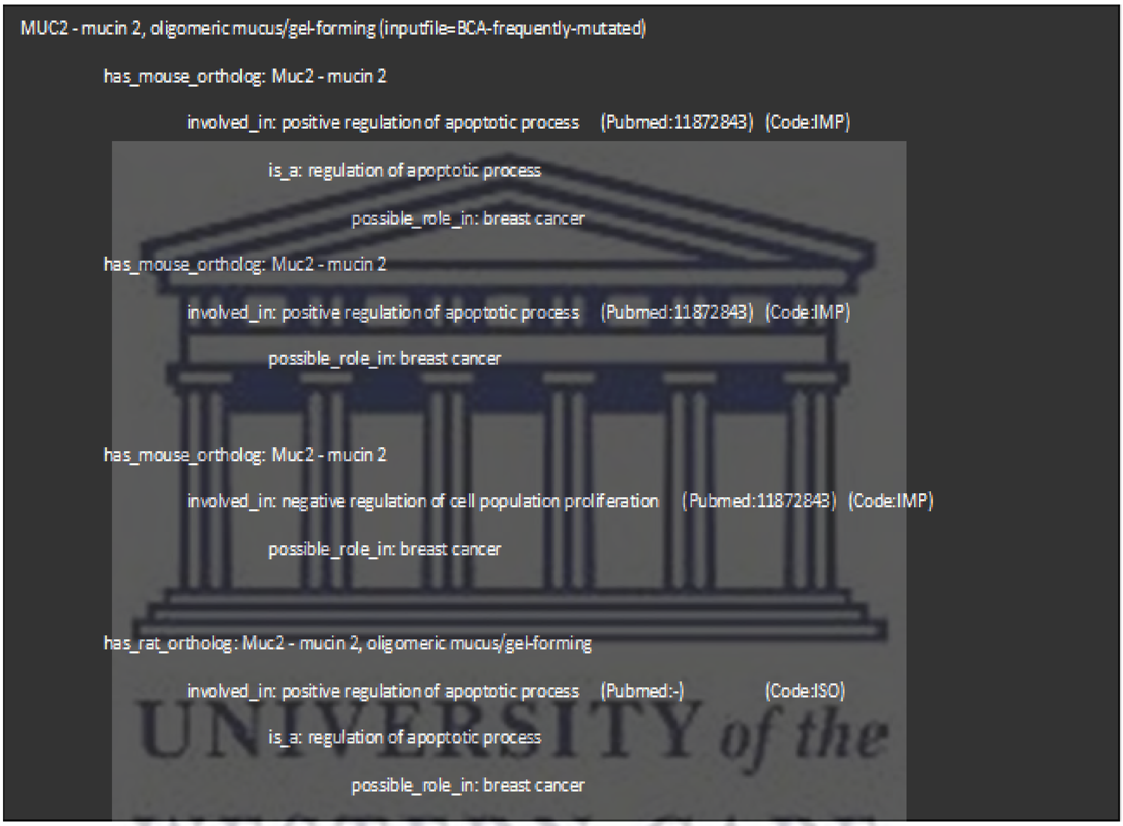


```
PIK3CA - phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (inputfile=BCA-frequently-mutated)
        implicated_in: estrogen-receptor positive breast cancer        (Pubmed:28881720)  (Code:IAGP)
                is_a: breast cancer
        implicated_in: breast adenocarcinoma      (Pubmed:) (Code:IAGP)
                is_a: breast carcinoma
                        is_a: breast cancer
        implicated_in: breast angiosarcoma        (Pubmed:32123305)  (Code:IAGP)
                is_a: breast sarcoma
                        is_a: breast cancer
        implicated_in: breast cancer      (Pubmed:) (Code:IAGP)
        has_function: phosphatidylinositol kinase activity      (Pubmed:21873635)  (Code:IBA)
                possible_role_in: breast cancer
        has_function: 1-phosphatidylinositol-3-kinase activity  (Pubmed:21873635|2174051)   (Code:IBA|IDA)
                is_a: phosphatidylinositol kinase activity
                        possible_role_in: breast cancer
        involved_in: negative regulation of anoikis   (Pubmed:22402981)  (Code:IC|IMP)
                is_a: regulation of anoikis
                        is_a: regulation of apoptotic process
                                possible_role_in: breast cancer
        involved_in: negative regulation of anoikis   (Pubmed:22402981)  (Code:IC|IMP)
                is_a: negative regulation of apoptotic process
                        is_a: regulation of apoptotic process
                                possible_role_in: breast cancer
        involved_in: negative regulation of anoikis   (Pubmed:22402981)  (Code:IC|IMP)
                is_a: negative regulation of apoptotic process
                        possible_role_in: breast cancer
        involved_in: epidermal growth factor receptor signaling pathway  (Pubmed:-)          (Code:TAS)
                is_a: ERBB signaling pathway
                        possible_role_in: breast cancer
```

Figure 3.10. Semantic analysis results of *PIK3CA* gene (frequently mutated).

MUC2 (Figure 3.11) returned with protein – protein interaction with MLH1 which is implicated in breast cancer.

MUC2 had rat and mouse orthologs where it is involved in positive regulation of apoptotic process, negative regulation of cell population proliferation, abnormal mitochondrial crista morphology, dilated mitochondria and increased intestinal adenocarcinoma incidence. Others have a putative role in breast carcinogenesis. Some of these features play a possible role in breast cancer development while others are clinical features of breast cancer.

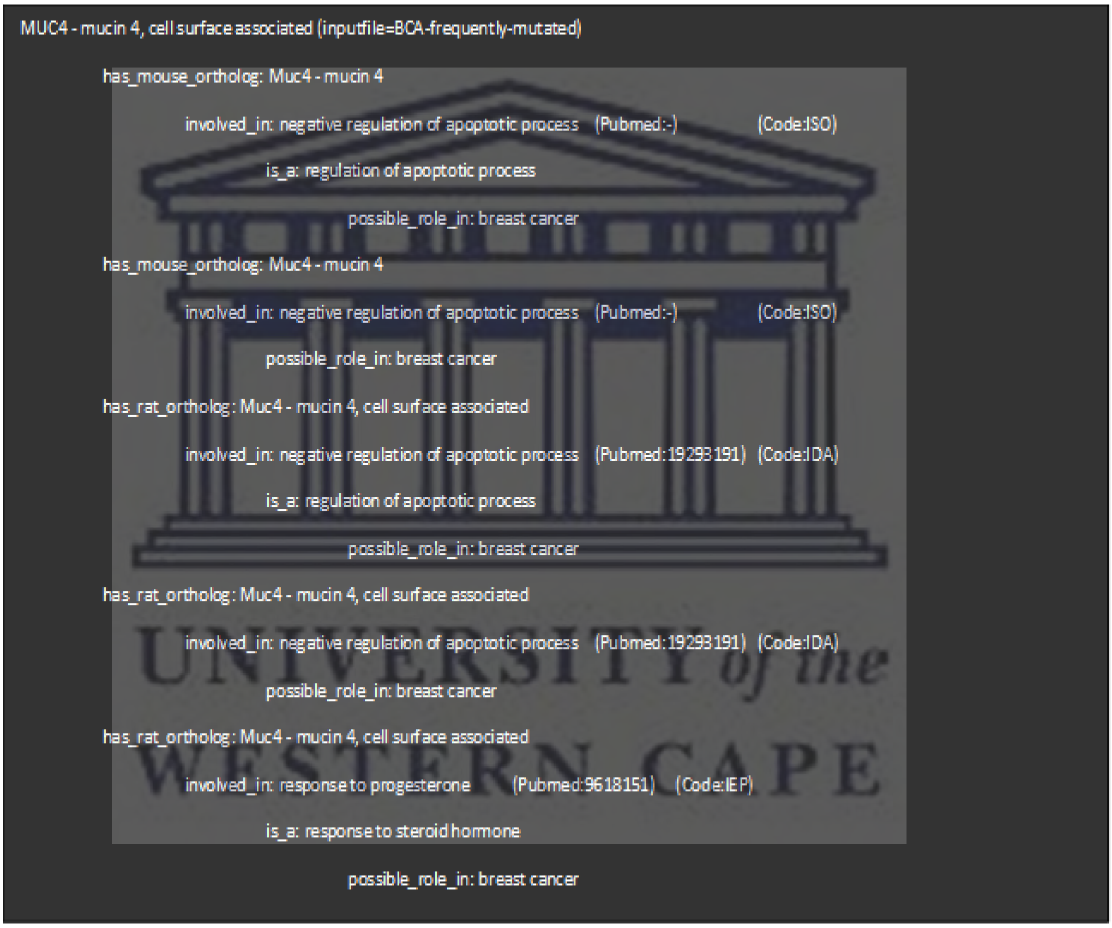


```
MUC2 - mucin 2, oligomeric mucus/gel-forming (inputfile=BCA-frequently-mutated)
        has_mouse_ortholog: Muc2 - mucin 2
                involved_in: positive regulation of apoptotic process    (Pubmed:11872843)  (Code:IMP)
                        is_a: regulation of apoptotic process
                                possible_role_in: breast cancer
        has_mouse_ortholog: Muc2 - mucin 2
                involved_in: positive regulation of apoptotic process    (Pubmed:11872843)  (Code:IMP)
                        possible_role_in: breast cancer

        has_mouse_ortholog: Muc2 - mucin 2
                involved_in: negative regulation of cell population proliferation    (Pubmed:11872843)  (Code:IMP)
                        possible_role_in: breast cancer

        has_rat_ortholog: Muc2 - mucin 2, oligomeric mucus/gel-forming
                involved_in: positive regulation of apoptotic process    (Pubmed:-)          (Code:ISO)
                        is_a: regulation of apoptotic process
                                possible_role_in: breast cancer
```

Figure 3.11. Semantic analysis results of *MUC2* gene (frequently mutated)

MUC4 (Figure 3.12) returned with several functions but had no protein-protein interactions. *MUC4* functions include negative regulation of apoptotic process, response to progesterone which may have a possible role in BCA. Furthermore, *MUC4* returned with known clinical features of breast cancer in rat and mouse including decreased incidence of tumors by chemical induction.

While *MUC4* is overexpressed in metastatic breast cancer patients (Dreyer et al., 2022). It is the *MUC5B* gene that is commonly abnormally expressed in breast cancer tissue. *MUC2*, *MUC5A* and *MUC5B* are commonly indicated in colorectal cancers (Iranmanesh et al., 2021).



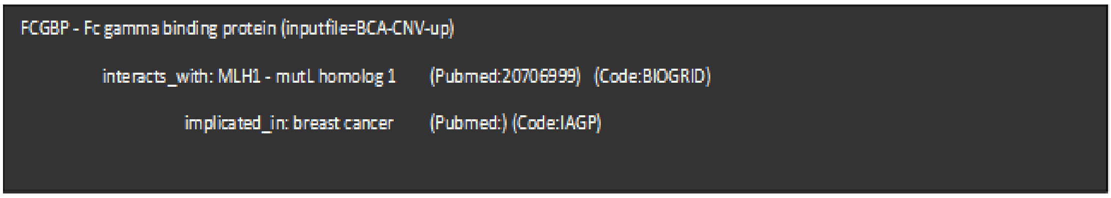
```
MUC4 - mucin 4, cell surface associated (inputfile=BCA-frequently-mutated)
        has_mouse_ortholog: Muc4 - mucin 4
                involved_in: negative regulation of apoptotic process    (Pubmed:-)            (Code:ISO)
                        is_a: regulation of apoptotic process
                                possible_role_in: breast cancer
        has_mouse_ortholog: Muc4 - mucin 4
                involved_in: negative regulation of apoptotic process    (Pubmed:-)            (Code:ISO)
                        possible_role_in: breast cancer
        has_rat_ortholog: Muc4 - mucin 4, cell surface associated
                involved_in: negative regulation of apoptotic process    (Pubmed:19293191)  (Code:IDA)
                        is_a: regulation of apoptotic process
                                possible_role_in: breast cancer
        has_rat_ortholog: Muc4 - mucin 4, cell surface associated
                involved_in: negative regulation of apoptotic process    (Pubmed:19293191)  (Code:IDA)
                        possible_role_in: breast cancer
        has_rat_ortholog: Muc4 - mucin 4, cell surface associated
                involved_in: response to progesterone        (Pubmed:9618151)    (Code:IEP)
                        is_a: response to steroid hormone
                                possible_role_in: breast cancer
```

Figure 3.12. Semantic analysis results of *MUC4* gene (frequently mutated).

Semantic analysis results for the upregulated CNVs

The breast cancer semantic system retuned the following genes: *FCGBP*, *KIAA1549L*, *MAGEA12*, *MALAT1*, *PDZD2*, *QSER1*, *RLF*, *TANC2*, and *UMODL1*.

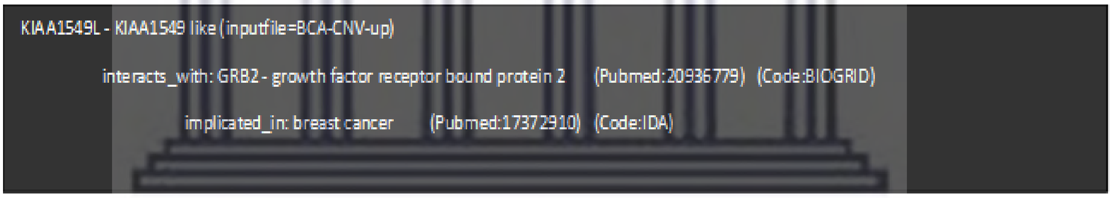
The *FCGBP* **(**Figure 3.13) gene returned with protein-protein interaction with *MLHI*. This interaction is implicated in breast cancer development. However, *FCGBP* is mostly implicated in colorectal and prostate cancer progression (Wang et al., 2021).



```
FCGBP - Fc gamma binding protein (inputfile=BCA-CNV-up)
        interacts_with: MLH1 - mutL homolog 1     (Pubmed:20706999)   (Code:BIOGRID)
                implicated_in: breast cancer      (Pubmed:) (Code:IAGP)
```

Figure 3.13. Semantic analysis results of *FCGBP* gene (Up regulated CNV)

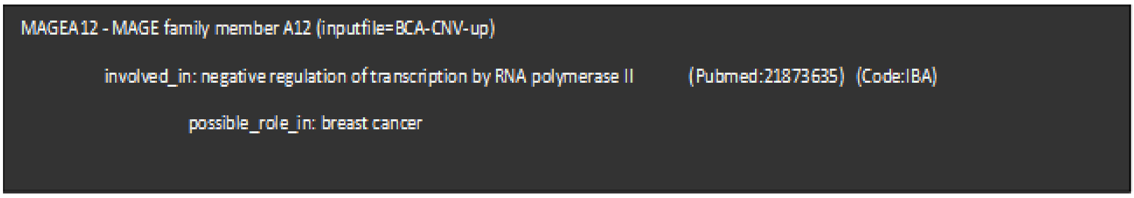
KIAA1549L (Figure 3.14) gene returned with protein-protein interaction with GRB2 I, an interaction implicated in breast cancer development.



```
KIAA1549L - KIAA1549 like (inputfile=BCA-CNV-up)
        interacts_with: GRB2 - growth factor receptor bound protein 2     (Pubmed:20936779)   (Code:BIOGRID)
                implicated_in: breast cancer      (Pubmed:17372910)   (Code:IDA)
```

Figure 3.14. Semantic analysis results of *KIAA1549L* gene (Up regulated CNV)

The system returned with the information that *MAGEA12* (Figure 3.15) is involved in negative regulation of transcription by RNA polymerase II which is implicated in human breast cancer.

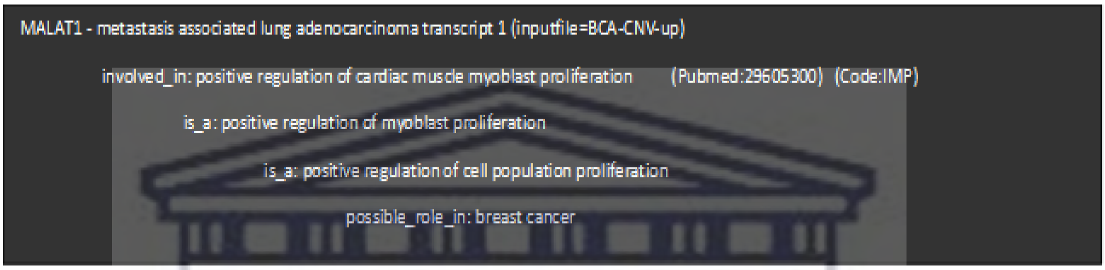


```
MAGEA12 - MAGE family member A12 (inputfile=BCA-CNV-up)
        involved_in: negative regulation of transcription by RNA polymerase II       (Pubmed:21873635)   (Code:IBA)
                possible_role_in: breast cancer
```

Figure 3.15. Semantic analysis results of *MAGEA12* gene (Up regulated CNV)

The system returned *MALAT1* (Figure 3.16) with that this gene is involved in positive regulation of cardiac muscle myoblast proliferation. Furthermore, it returned that this gene is involved in positive regulation of cell population proliferation with a conclusion that it may have a possible role in human breast cancer.

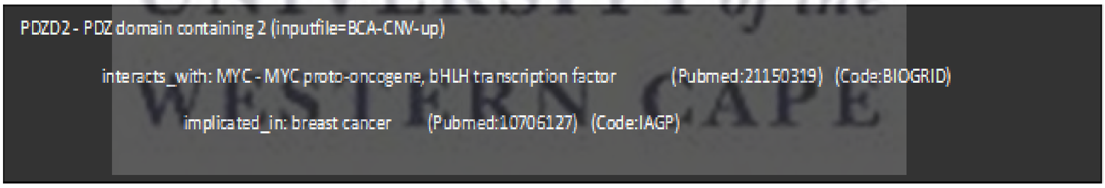
MALAT1 is a gene that was originally identified in pulmonary adenocarcinoma (Huang et al., 2021).



```
MALAT1 - metastasis associated lung adenocarcinoma transcript 1 (inputfile=BCA-CNV-up)
        involved_in: positive regulation of cardiac muscle myoblast proliferation      (Pubmed:29605300)  (Code:IMP)
                is_a: positive regulation of myoblast proliferation
                        is_a: positive regulation of cell population proliferation
                                possible_role_in: breast cancer
```

Figure 3.16. Semantic analysis results of *MALAT1* gene (Up regulated CNV)

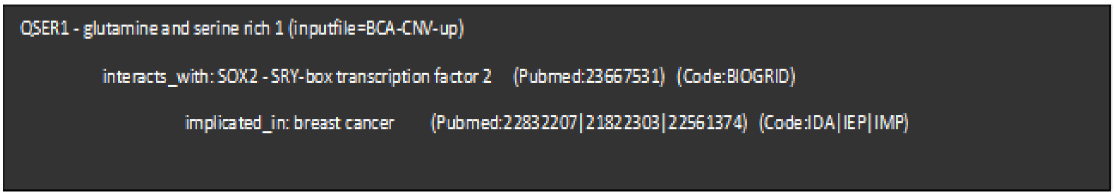
The system returned *PDZD2* (Figure 3.17) with the information that its protein interacts with *MYC*-MYC proto-oncogene and therefore may be implicated in human breast cancer.



```
PDZD2 - PDZ domain containing 2 (inputfile=BCA-CNV-up)
        interacts_with: MYC - MYC proto-oncogene, bHLH transcription factor      (Pubmed:21150319)  (Code:BIOGRID)
                implicated_in: breast cancer      (Pubmed:10706127)  (Code:IAGP)
```

Figure 3.17. Semantic analysis results of *PDZD2* gene (Up regulated CNV)

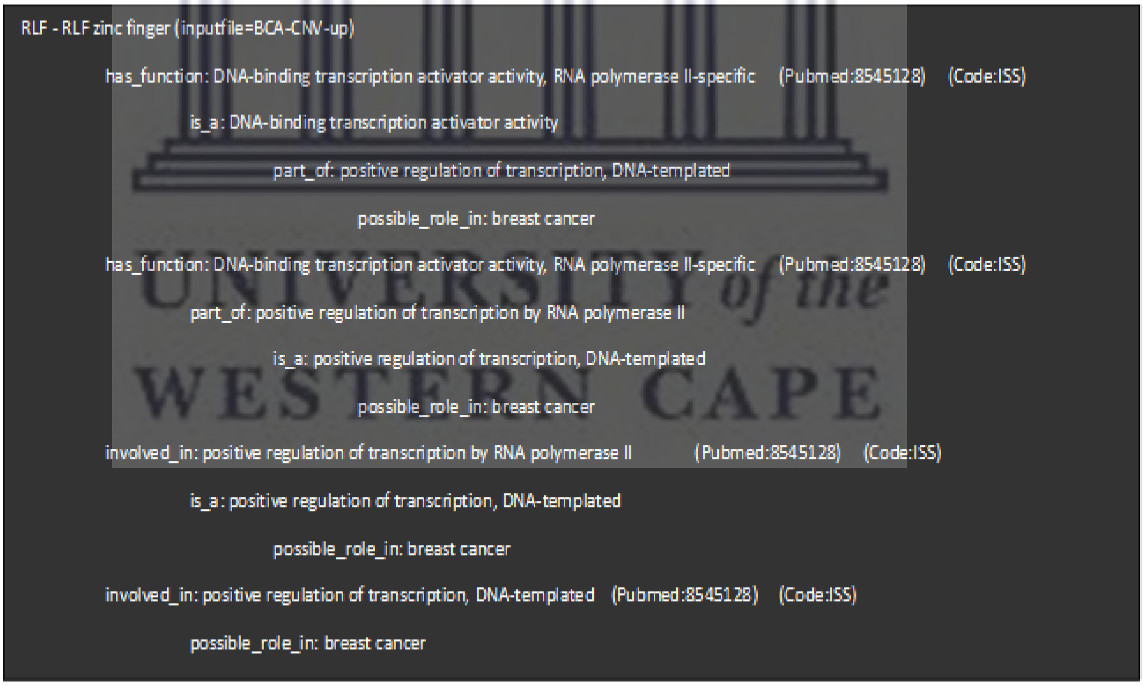
The information presented by the system included that *QSER1* (Figure 3.18) transcripts reacts with SOX2 - SRY-box transcription factor 2. The system concluded that due to this interaction this gene may be implicated in human breast cancer development.



```
QSER1 - glutamine and serine rich 1 (inputfile=BCA-CNV-up)
            interacts_with: SOX2 - SRY-box transcription factor 2    (Pubmed:23667531)  (Code:BIOGRID)
                        implicated_in: breast cancer      (Pubmed:22832207|21822303|22561374)  (Code:IDA|IEP|IMP)
```

Figure 3.18. Semantic analysis results of *QSER1* gene (Up regulated CNV)

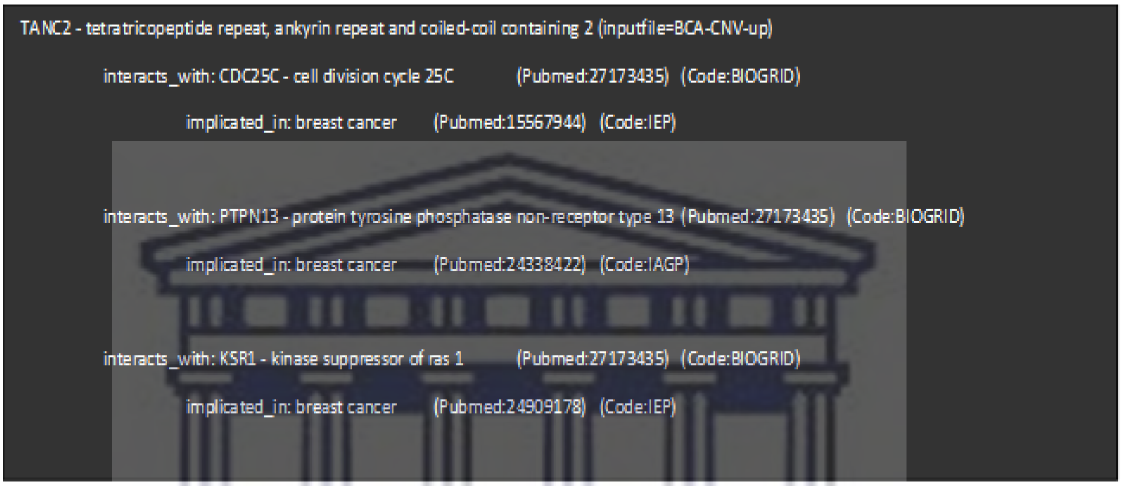
The system returned with that the *RLF* (Figure 3.19) gene has DNA-binding transcription activator activity that is RNA polymerase II-specific. It is involved in positive regulation of transcription by RNA polymerase II, and positive regulation of transcription of DNA-template. It further concluded that as such, the *RLF* gene is implicated in breast cancer development.



```
RLF - RLF zinc finger (inputfile=BCA-CNV-up)
            has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:8545128)    (Code:ISS)
                        is_a: DNA-binding transcription activator activity
                                    part_of: positive regulation of transcription, DNA-templated
                                                possible_role_in: breast cancer
            has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:8545128)    (Code:ISS)
                        part_of: positive regulation of transcription by RNA polymerase II
                                    is_a: positive regulation of transcription, DNA-templated
                                                possible_role_in: breast cancer
            involved_in: positive regulation of transcription by RNA polymerase II        (Pubmed:8545128)    (Code:ISS)
                        is_a: positive regulation of transcription, DNA-templated
                                    possible_role_in: breast cancer
            involved_in: positive regulation of transcription, DNA-templated    (Pubmed:8545128)    (Code:ISS)
                        possible_role_in: breast cancer
```

Figure 3.19. Semantic analysis results of *RLF* gene (Up regulated CNV)

The *TANC2 (*Figure 3.20) gene was returned with the following annotations. It interacts with *CDC25C* - cell division cycle 25C, *PTPN13* - protein tyrosine phosphatase non-receptor type 13, and kinase suppressor of ras 1. It further concluded that these interactions may be implicated in human breast cancer development.

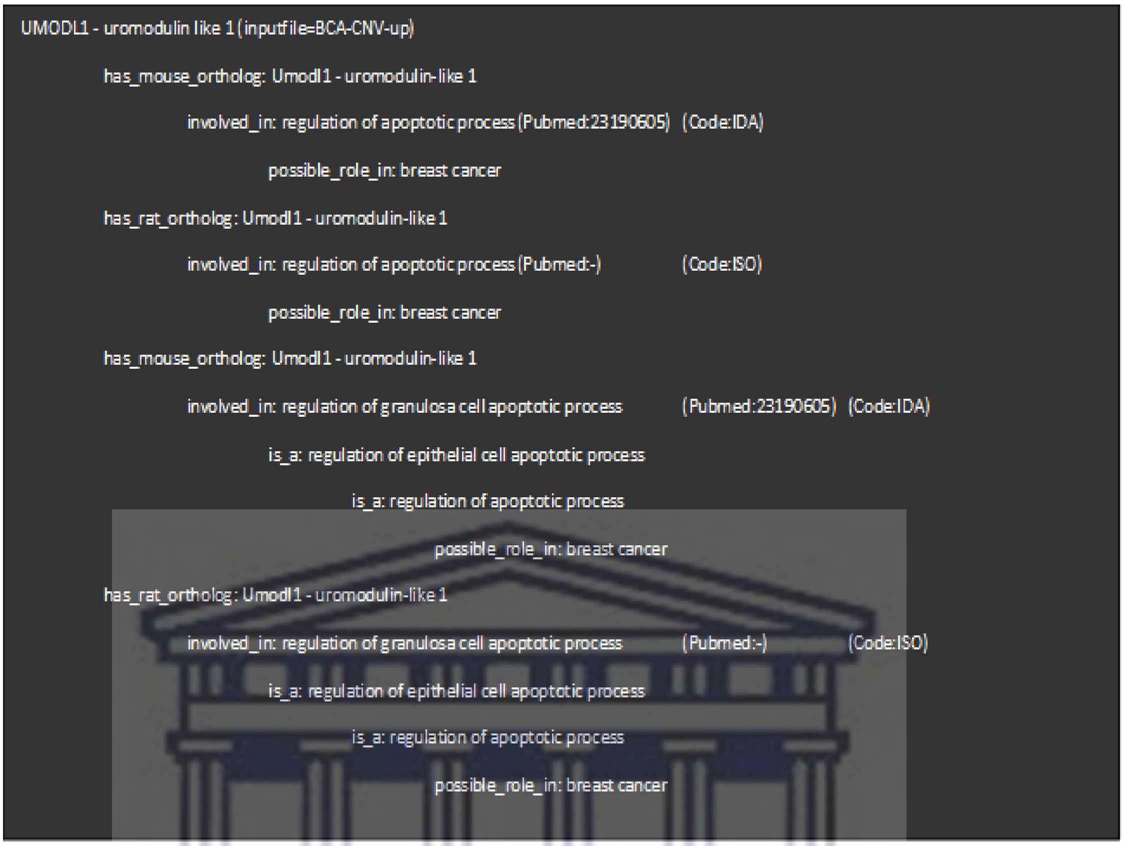


```
TANC2 - tetratricopeptide repeat, ankyrin repeat and coiled-coil containing 2 (inputfile=BCA-CNV-up)
        interacts_with: CDC25C - cell division cycle 25C        (Pubmed:27173435)  (Code:BIOGRID)
                implicated_in: breast cancer     (Pubmed:15567944)  (Code:IEP)

        interacts_with: PTPN13 - protein tyrosine phosphatase non-receptor type 13 (Pubmed:27173435)  (Code:BIOGRID)
                implicated_in: breast cancer     (Pubmed:24338422)  (Code:IAGP)

        interacts_with: KSR1 - kinase suppressor of ras 1       (Pubmed:27173435)  (Code:BIOGRID)
                implicated_in: breast cancer     (Pubmed:24909178)  (Code:IEP)
```

Figure 3.20. Semantic analysis results of *TANC2* gene (Up regulated CNV)

UMODL1 (Figure 3.21) was returned with rat and mouse orthologs. This gene was reported to be involved in regulation of apoptotic process and regulation of granulosa cell apoptotic process. Both these biological processes implicate the gene in human breast cancer development.

This gene is not commonly associated with breast cancer development, however when its upregulated it is associated with lung cancer metastasis and ovarian degradation (Davenport et al., 2021).



```
UMODL1 - uromodulin like 1 (inputfile=BCA-CNV-up)
        has_mouse_ortholog: Umodl1 - uromodulin-like 1
                involved_in: regulation of apoptotic process (Pubmed:23190605)   (Code:IDA)
                        possible_role_in: breast cancer
        has_rat_ortholog: Umodl1 - uromodulin-like 1
                involved_in: regulation of apoptotic process (Pubmed:-)          (Code:ISO)
                        possible_role_in: breast cancer
        has_mouse_ortholog: Umodl1 - uromodulin-like 1
                involved_in: regulation of granulosa cell apoptotic process      (Pubmed:23190605)   (Code:IDA)
                        is_a: regulation of epithelial cell apoptotic process
                                is_a: regulation of apoptotic process
                                        possible_role_in: breast cancer
        has_rat_ortholog: Umodl1 - uromodulin-like 1
                involved_in: regulation of granulosa cell apoptotic process      (Pubmed:-)          (Code:ISO)
                        is_a: regulation of epithelial cell apoptotic process
                                is_a: regulation of apoptotic process
                                        possible_role_in: breast cancer
```

Figure 3.21. Semantic analysis results of *UMODL1* gene (Up regulated CNV)

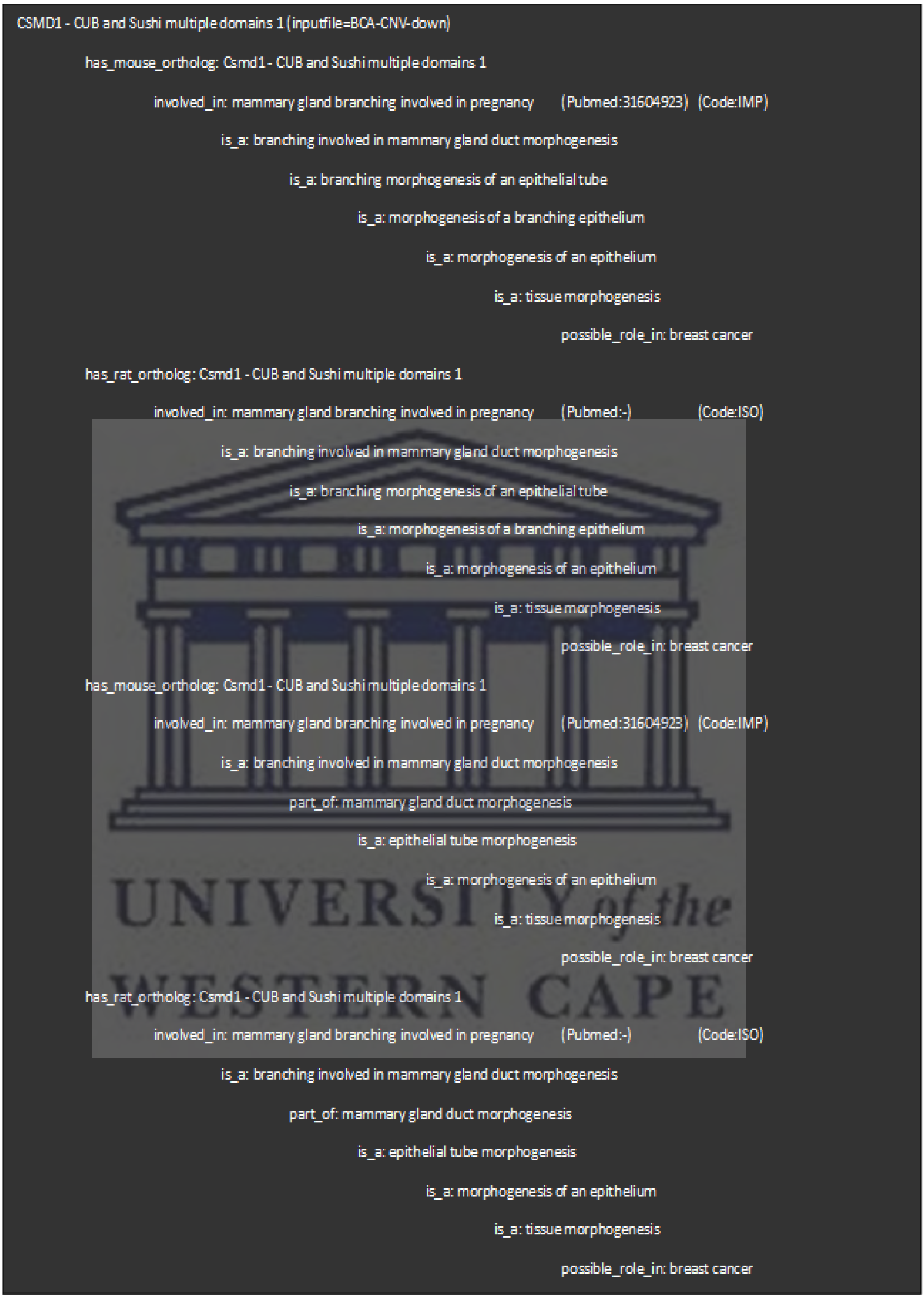
Semantic analysis results for the downregulated CNVs

The breast cancer semantic system returned the following downregulated CNV genes: *CSMD1*, *MALAT1*, *VPS13D* and *ZNF292*.

The system did not return the human orthologs for the *CSMD1* gene (Figure 3.22). Therefore, this gene may be considered a novel gene in human breast cancer.

CSMD1 returned with rat and mouse orthologs with no protein-to-protein annotations. The system reported that this gene is implicated in mammary gland branching involved in pregnancy where mammary gland duct, epithelial tube, epithelial branching and general breast tissues morphogenesis occurs. The Breast cancer semantic system reported that all these biological processes are implicated in breast cancer development in these mammals.

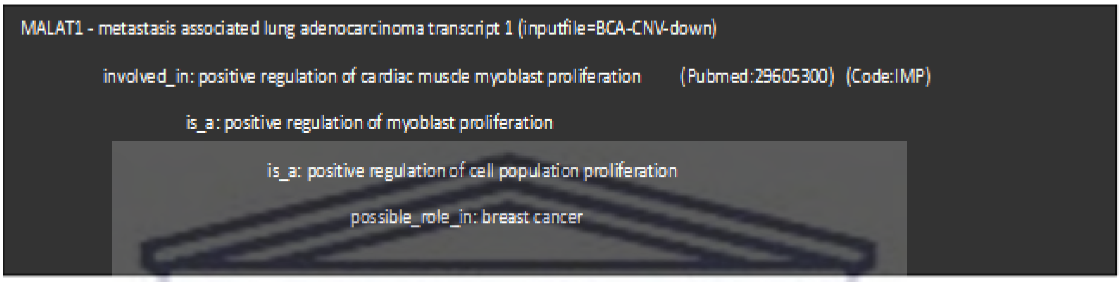
Overexpression of *CSMD1* is known to have tumor suppressor effect in breast cancer therefore it’s under expression may be associated with breast cancer proliferation (Gialeli et al., 2021).



```
CSMD1 - CUB and Sushi multiple domains 1 (inputfile=BCA-CNV-down)
    has_mouse_ortholog: Csmd1 - CUB and Sushi multiple domains 1
        involved_in: mammary gland branching involved in pregnancy     (Pubmed:31604923)  (Code:IMP)
            is_a: branching involved in mammary gland duct morphogenesis
                is_a: branching morphogenesis of an epithelial tube
                    is_a: morphogenesis of a branching epithelium
                        is_a: morphogenesis of an epithelium
                            is_a: tissue morphogenesis
                                possible_role_in: breast cancer
    has_rat_ortholog: Csmd1 - CUB and Sushi multiple domains 1
        involved_in: mammary gland branching involved in pregnancy     (Pubmed:-)         (Code:ISO)
            is_a: branching involved in mammary gland duct morphogenesis
                is_a: branching morphogenesis of an epithelial tube
                    is_a: morphogenesis of a branching epithelium
                        is_a: morphogenesis of an epithelium
                            is_a: tissue morphogenesis
                                possible_role_in: breast cancer
    has_mouse_ortholog: Csmd1 - CUB and Sushi multiple domains 1
        involved_in: mammary gland branching involved in pregnancy     (Pubmed:31604923)  (Code:IMP)
            is_a: branching involved in mammary gland duct morphogenesis
                part_of: mammary gland duct morphogenesis
                    is_a: epithelial tube morphogenesis
                        is_a: morphogenesis of an epithelium
                            is_a: tissue morphogenesis
                                possible_role_in: breast cancer
    has_rat_ortholog: Csmd1 - CUB and Sushi multiple domains 1
        involved_in: mammary gland branching involved in pregnancy     (Pubmed:-)         (Code:ISO)
            is_a: branching involved in mammary gland duct morphogenesis
                part_of: mammary gland duct morphogenesis
                    is_a: epithelial tube morphogenesis
                        is_a: morphogenesis of an epithelium
                            is_a: tissue morphogenesis
                                possible_role_in: breast cancer
```

Figure 3.22. Semantic analysis results of *CSMD1* gene (downregulated CNV)

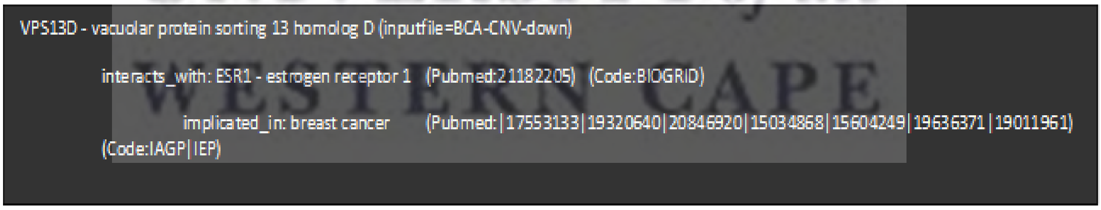
The system annotated *MALAT1* gene (Figure 3.23) as potentially downregulated in some cancers while in others it is regulated as observed above. Similarly, to upregulated conditions of the gene it is involved in positive regulation of cardiac muscle myoblast proliferation.



```
MALAT1 - metastasis associated lung adenocarcinoma transcript 1 (inputfile=BCA-CNV-down)
        involved_in: positive regulation of cardiac muscle myoblast proliferation     (Pubmed:29605300)  (Code:IMP)
                is_a: positive regulation of myoblast proliferation
                        is_a: positive regulation of cell population proliferation
                                possible_role_in: breast cancer
```

Figure 3.23. Semantic analysis results of *MALAT1*gene (downregulated CNV).

The *VPS13D* gene (Figure3.24) returned the information that it interacts with: ESR1. Furthermore, the system surmised that this interaction implicated the gene in human breast cancer.

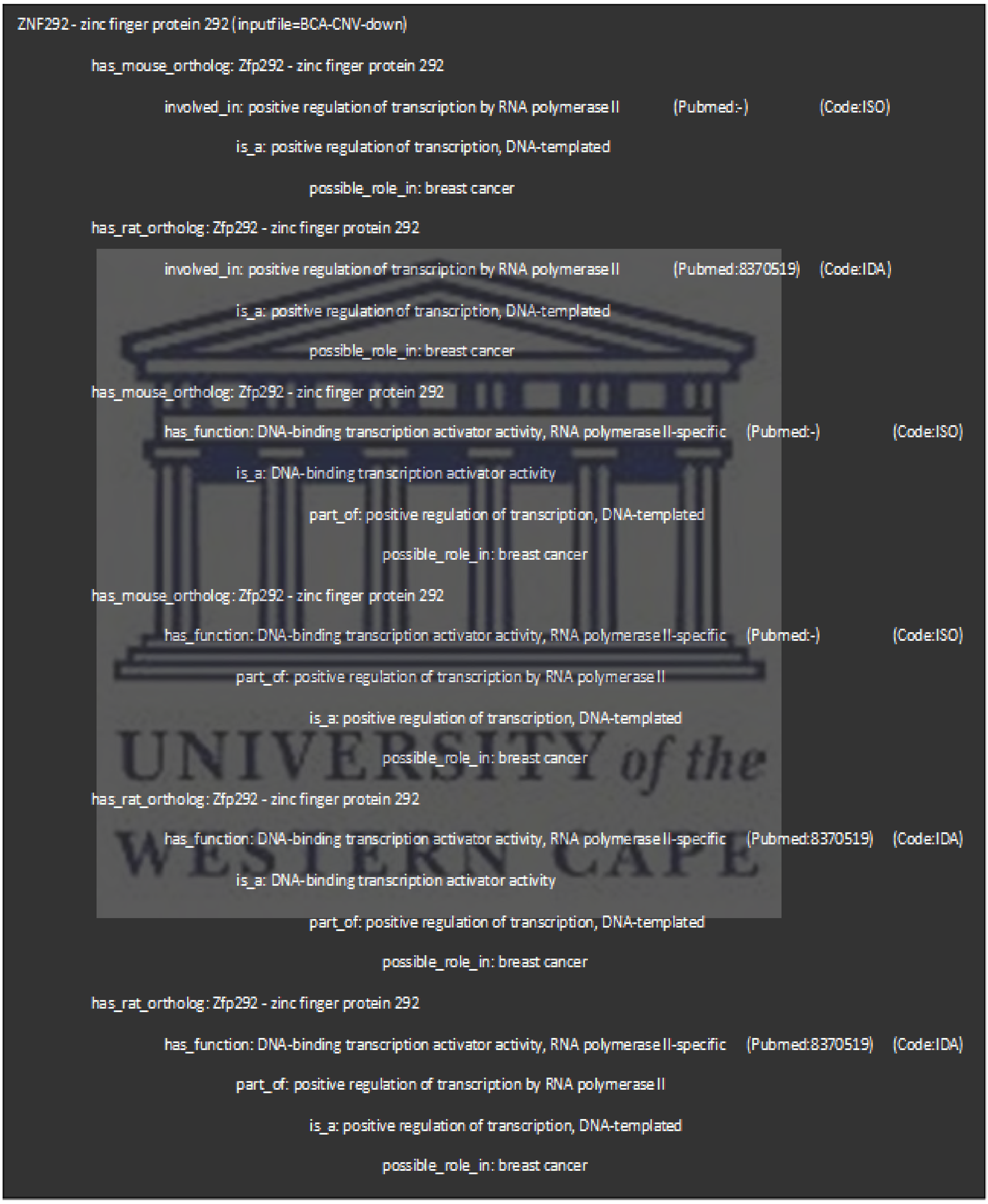


```
VPS13D - vacuolar protein sorting 13 homolog D (inputfile=BCA-CNV-down)
        interacts_with: ESR1 - estrogen receptor 1   (Pubmed:21182205)  (Code:BIOGRID)
                implicated_in: breast cancer     (Pubmed:|17553133|19320640|20846920|15034868|15604249|19636371|19011961)
        (Code:IAGP|IEP)
```

Figure 3.24. Semantic analysis results of *VPS13D* gene (downregulated CNV).

Th*e ZNF292* gene (Figure3.25) returned with only with rat and mouse orthologs implicated in positive regulation of transcription by RNA polymerase II. This DNA-binding transcription activator activity may have a possible role in breast cancer of these mammals.

Nirgude et al, reported this gene as a tumor suppressor gene and therefore it may be considered as enhancing carcinogenesis when under expressed (Nirgude et al., 2022).



```
ZNF292 - zinc finger protein 292 (inputfile=BCA-CNV-down)
        has_mouse_ortholog: Zfp292 - zinc finger protein 292
                involved_in: positive regulation of transcription by RNA polymerase II        (Pubmed:-)           (Code:ISO)
                        is_a: positive regulation of transcription, DNA-templated
                                possible_role_in: breast cancer
        has_rat_ortholog: Zfp292 - zinc finger protein 292
                involved_in: positive regulation of transcription by RNA polymerase II        (Pubmed:8370519)    (Code:IDA)
                        is_a: positive regulation of transcription, DNA-templated
                                possible_role_in: breast cancer
        has_mouse_ortholog: Zfp292 - zinc finger protein 292
                has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:-)           (Code:ISO)
                        is_a: DNA-binding transcription activator activity
                                part_of: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
        has_mouse_ortholog: Zfp292 - zinc finger protein 292
                has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:-)           (Code:ISO)
                        part_of: positive regulation of transcription by RNA polymerase II
                                is_a: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
        has_rat_ortholog: Zfp292 - zinc finger protein 292
                has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:8370519)    (Code:IDA)
                        is_a: DNA-binding transcription activator activity
                                part_of: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
        has_rat_ortholog: Zfp292 - zinc finger protein 292
                has_function: DNA-binding transcription activator activity, RNA polymerase II-specific    (Pubmed:8370519)    (Code:IDA)
                        part_of: positive regulation of transcription by RNA polymerase II
                                is_a: positive regulation of transcription, DNA-templated
                                        possible_role_in: breast cancer
```

Figure 3.25. Semantic analysis results of *ZNF292* gene (downregulated CNV)

3.5 CONCLUSION

Given the number of mutated genes linked to breast cancer that are not classical cancer genes prioritized by the BCA knowledge graph indicates the utility of this tool. It of course remains to determine pathways associated with these non-classical cancer genes after having implicated them in BCA and to verify their role in immortalisation and proliferation of breast cancer cells. That said, the system and concept have potential for use in other diseases that are associated with gene mutation and their expression.

The here generated Breast cancer semantic system proved effective because it returned known genes implicated in human cancer development substantiating it with sound biological information published in scientific journals. Furthermore, it pointed out novel genes that are not previously implicated in human breast cancer while linked to other mammalian breast cancers as well as other cancers. Some of the information that came through may be used to identify the stages in carcinogenesis for an example, metastasis. This means that the developed system has a potential of indicating novel breast cancer genes, which can inform BCA diagnosis and potentially drug development.

REFERENCES

- Adib, E., El Zarif, T., Nassar, A. H., Akl, E. W., Abou Alaiwi, S., Mouhieddine, T. H., Esplin, E. D., Hatchell, K., Nielsen, S. M., Rana, H. Q., Choueiri, T. K., Kwiatkowski, D. J., & Sonpavde, G. (2022). CDH1 germline variants are enriched in patients with colorectal cancer, gastric cancer, and breast cancer. *British Journal of Cancer*, *126*(5), 797–803. https://doi.org/10.1038/s41416-021-01673-7
- Ahmad, M., & Shah, A. A. (2021). *Predictive role of single nucleotide polymorphism (rs11614913) in the development of breast cancer in Pakistani population*. *May*, 212–227. https://doi.org/10.2217/pme-2019-0086
- Bozhilov, Y. K., Downes, D. J., Telenius, J., Marieke Oudelaar, A., Olivier, E. N., Mountford, J. C., Hughes, J. R., Gibbons, R. J., & Higgs, D. R. (2021). A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-23980-6
- Cheukfai, L. I., Zhang, G., Wang, Y., Chen, B., Li, K., Cao, L., Ren, C., Wen, L., Jia, M., Mok, H., Lai, J., Xiao, W., Li, X., & Liao, N. (2022). Spectrum of MAP3K1 mutations in breast cancer is luminal subtype-predominant and related to prognosis. *Oncology Letters*, *23*(2), 1–12. https://doi.org/10.3892/ol.2022.13187
- Davenport, V., Horstmann, C., Patel, R., Wu, Q., & Kim, K. (2021). An Assessment of InP/ZnS as Potential Anti-Cancer Therapy: Quantum Dot Treatment Increases Apoptosis in HeLa Cells. *Journal of Nanotheranostics*, *2*(1), 16–32. https://doi.org/10.3390/jnt2010002

- Dreyer, C. A., Vorst, K. Vander, Free, S., Rowson-Hodel, A., & Carraway, K. L. (2022). The role of membrane mucin MUC4 in breast cancer metastasis. *Endocrine-Related Cancer*, *29*(1), R17–R32. https://doi.org/10.1530/ERC-21-0083
- Gao, G., Wang, Z., Qu, X., & Zhang, Z. (2020). Prognostic value of tumor-infiltrating lymphocytes in patients with triple-negative breast cancer: A systematic review and meta-analysis. *BMC Cancer*, *20*(1). https://doi.org/10.1186/s12885-020-6668-z
- Gialeli, C., Tuysuz, E. C., Staaf, J., Guleed, S., Paciorek, V., Mörgelin, M., Papadakos, K. S., & Blom, A. M. (2021). Complement inhibitor CSMD1 modulates epidermal growth factor receptor oncogenic signaling and sensitizes breast cancer cells to chemotherapy. *Journal of Experimental and Clinical Cancer Research*, *40*(1), 1–21. https://doi.org/10.1186/s13046-021-02042-1
- Gökçümen, Ö., & Lee, C. (2009). Copy number variants (CNVs) in primate species using array-based comparative genomic hybridization. *Methods*, *49*(1), 18–25. https://doi.org/10.1016/j.ymeth.2009.06.001
- Hollander, M., Hamed, M., Helms, V., & Neininger, K. (2018). MutaNET: A tool for automated analysis of genomic mutations in gene regulatory networks. *Bioinformatics*, *34*(5), 864–866. https://doi.org/10.1093/bioinformatics/btx687
- Huang, Y., Zhou, Z., Zhang, J., Hao, Z., He, Y., Wu, Z., Song, Y., Yuan, K., Zheng, S., Zhao, Q., Li, T., & Wang, B. (2021). lncRNA MALAT1 participates in metformin inhibiting the proliferation of breast cancer cell. *Journal of Cellular and Molecular Medicine*, *25*(15), 7135–7145. https://doi.org/10.1111/jcmm.16742

- Iranmanesh, H., Majd, A., Mojarad, E. N., Zali, M. R., & Hashemi, M. (2021). *Investigating the Relationship Between the Expression Level of Mucin Gene Cluster (MUC2, MUC5A, and MUC5B) and Clinicopathological Characterization of Colorectal Cancer*. https://doi.org/10.31661/gmj.v10i0.2030
- Li, J., Du, J., Wang, Y., & Jia, H. (2021). A Coagulation-Related Gene-Based Prognostic Model for Invasive Ductal Carcinoma. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/FGENE.2021.722992/FULL
- Mair, B., Konopka, T., Kerzendorfer, C., Sleiman, K., Salic, S., Serra, V., Muellner, M. K., Theodorou, V., & Nijman, S. M. B. (2016). Gain- and Loss-of-Function Mutations in the Breast Cancer Gene GATA3 Result in Differential Drug Sensitivity. *PLoS Genetics*, *12*(9), 1–26. https://doi.org/10.1371/journal.pgen.1006279
- Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitziel, N. O., Hillier, L., Kwok, P.-Y., & Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery. *Nature Genetics*, *23*(4), 452–456. https://doi.org/10.1038/70570
- Murakami, F., Tsuboi, Y., Takahashi, Y., Horimoto, Y., Mogushi, K., Ito, T., Emi, M., Matsubara, D., Shibata, T., Saito, M., & Murakami, Y. (2021). Short somatic alterations at the site of copy number variation in breast cancer. *Cancer Science*, *112*(1), 444–453. https://doi.org/10.1111/cas.14630
- Nirgude, S., Desai, S., & Choudhary, B. (2022). Curcumin alters distinct molecular pathways in breast cancer subtypes revealed by integrated

miRNA/mRNA expression analysis. *Cancer Reports*, *June 2021*, 1–18. https://doi.org/10.1002/cnr2.1596

- Ramaiah, M. J., Tangutur, A. D., & Manyam, R. R. (2021). Epigenetic modulation and understanding of HDAC inhibitors in cancer therapy. *Life Sciences*, *277*(January), 119504. https://doi.org/10.1016/j.lfs.2021.119504
- Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., George, H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., Juan, R., Gratacos, M., Huang, J., Kalaitzopoulos, D., Komura, D., Okamura, K., Tchinda, J. (2009). *Europe PMC Funders Group Global variation in copy number in the human genome*. *444*(7118), 444–454. https://doi.org/10.1038/nature05329.Global
- Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Månér, S., Massa, H., Walker, M., Chi, M., Navin, N., Lucito, R., Healy, J., Hicks, J., Ye, K., Reiner, A., Gilliam, T. C., Trask, B., Patterson, N., … Wigler, M. (2004). Large-scale copy number polymorphism in the human genome. *Science*, *305*(5683), 525–528. https://doi.org/10.1126/science.1098918
- Shao, X., Lv, N., Liao, J., Long, J., Xue, R., Ai, N., Xu, D., & Fan, X. (2019). Copy number variation is highly correlated with differential gene expression: A pan-cancer study. *BMC Medical Genetics*, *20*(1), 1–14. https://doi.org/10.1186/s12881-019-0909-5
- Shao, Y., Yu, Y., He, Y., Chen, Q., & Liu, H. (2019). Serum ATX as a novel biomarker for breast cancer. *Medicine (United States)*, *98*(13). https://doi.org/10.1097/MD.0000000000014973

- Tetsu, O., & McCormick, F. (2003). Proliferation of cancer cells despite CDK2 inhibition. *Cancer Cell*, *3*(3), 233–245. https://doi.org/10.1016/S1535-6108(03)00053-9
- Wang, B., Yang, S., Jia, Y., Yang, J., Du, K., Luo, Y., Li, Y., Wang, Z., Liu, Y., & Zhu, B. (2022). PCAT19 Regulates the Proliferation and Apoptosis of Lung Cancer Cells by Inhibiting miR-25-3p via Targeting the MAP2K4 Signal Axis. *Disease Markers*, *2022*, 1–18. https://doi.org/10.1155/2022/2442094
- Wang, K., Guan, C., Shang, X., Ying, X., Mei, S., Zhu, H., Xia, L., & Chai, Z. (2021). A bioinformatic analysis: the overexpression and clinical significance of FCGBP in ovarian cancer. *Aging*, *13*(5), 7416–7429. https://doi.org/10.18632/aging.202601
- Xu, B. (2021). *Integrative Clonal Evolution Analysis on SNV and CNV Levels in Multifocal Breast Cancer*. 1–17.
- Yang, L., Zhao, S., Zhu, T., & Zhang, J. (2021). GPRC5A Is a Negative Regulator of the Pro-Survival PI3K/Akt Signaling Pathway in Triple-Negative Breast Cancer. *Frontiers in Oncology*, *10*(February), 1–10. https://doi.org/10.3389/fonc.2020.624493
- Zhan, M., Chen, G., Pan, C. M., Gu, Z. H., Zhao, S. X., Liu, W., Wang, H. N., Ye, X. P., Xie, H. J., Yu, S. S., Liang, J., Gao, G. Q., Yuan, G. Y., Zhang, X. M., Zuo, C. L., Su, B., Huang, W., Ning, G., Chen, S. J., … Song, H. D. (2014). Genome-wide association study identifies a novel susceptibility gene for serum TSH levels in Chinese populations. *Human Molecular Genetics*, *23*(20), 5505–5517. https://doi.org/10.1093/hmg/ddu250

- Zhang, J., Gan, Y., Li, H., Yin, J., He, X., Lin, L., Xu, S., Fang, Z., Kim, B., Gao, L., Ding, L., Zhang, E., Ma, X., Li, J., Li, L., Xu, Y., Horne, D., Xu, R., Yu, H., … Huang, W. (2022). Inhibition of the CDK2 and Cyclin A complex leads to autophagic degradation of CDK2 in cancer cells. *Nature Communications*, *13*(1), 1–16. https://doi.org/10.1038/s41467-022-30264-

CHAPTER 4

DIFFERENTIAL EXPRESSION

4.1 INTRODUCTION

4.1.1 RNA-Sequencing and Cancer

Next generation sequencing (NGS) technologies have been utilised for genome, DNA sequencing (DNA-seq), chromatin immunoprecipitation (Chip-seq), methylation (Methyl-seq), and more importantly RNA sequencing (RNA-seq) (Jazayeri, Saadat, Ramezani, & Kaviani, 2015). RNA sequencing (RNA-seq) is a technique which utilises various methods of either high throughput sequencing or next generation sequencing to examine the quantity and sequences of RNA in a sample. Moreover, this technique analyses the transcriptome of gene expression patterns encoded within our RNA (Jazayeri et al., 2015). Particularly, the transcriptome contains coding messenger RNA (mRNA's), along with non-coding RNA's (ncRNA's) such as microRNA's (miRNA) long-coding RNA (LncRNA), Ribosomal RNA (rRNA), and transfer RNA (tRNA).

Nevertheless, RNA-seq allows for the investigation and discovery of total cellular RNA's, including mRNA, rRNA and transfer RNA (tRNA), providing key insights into the transcriptome, and thus functional genomic protein expression. More specifically, the technique can provide insights into specific cellular gene activity, such as which gene are expressed, the nature of their expression, along with when they are switched on or off (Ozsolak & Milos, 2011). Hence, a number of techniques utilize the functionality of RNA-seq, including SNP identification, RNA editing, differential gene expression profiling, and transcriptional profiling (Han et al., 2015) and (Jazayeri et al., 2015). With particular regard to the transcriptome, RNA-Seq provides an in-depth view, and can be divided into several steps. These include RNA extraction, library construction, sequencing, and data analysis (Kukurba & Montgomery, 2015). Initially, high quality RNA must be extracted from cancerous tissues, and will be composed of rRNA, mRNA and various types of ncRNA (Kukurba & Montgomery, 2015) and (L. Wang, Feng, Wang, Wang, & Zhang, 2009). To achieve this, a number of separation techniques have been developed. For example, the 3′ polyadenylated (poly-A) tail of mRNA may be extracted using oligo-dT primer beads by selecting for poly-A RNAs (termed a poly-A library). Hereafter, a proportion of the lncRNA are excluded from the poly-A library due to the absence of a poly-A tail (C. Wang, Lu, Emanuel, Babcock, & Zhuang, 2019). Another method involves removal of rRNA using commercially available kits, such as Ribo-Zero and RiboMinus (Peano et al., 2013). Specifically, these kits selectively isolate small RNAs, including miRNA and piwi-interacting RNA (piRNA), which are short (15–30 nt), sparse, and lack a poly-A tail (Kukurba & Montgomery, 2015). Once RNA is isolated from the total RNA content, it may be converted to a library of complementary DNA (cDNA) fragments with adaptors ligated to one or both ends (C. Wang et al., 2019). Subsequently, these adaptors may be ligated to small RNAs and finally undergo reverse transcription. Long RNA or cDNA molecules must first be fragmented into smaller pieces, however, in order to be compatible with Next Generation Sequencing (NGS) technologies.

4.1.2 Gene expression in breast cancer

Breast cancer has been reported to be the most commonly occurring cancer on a global scale, accounting for a reported 11.6% of new cancer cases, and 6.6% of all cancer related deaths (Vishnubalaji, Sasidharan Nair, Ouararhni, Elkord, & Alajez, 2019). Global statistics reported in GLOBOCAN 2022 demonstrated that breast cancer was the most frequently diagnosed, and leading cause of cancer-related deaths in women. For the past decades, researchers have tried to stratify BC in order to find improved means for early diagnosis and ultimately better therapeutic approach. Research within the pathology department worldwide has attempted to find clear pathological stratification on the basis of the difference in gene expression profile Five major groups were characterised as follows: the first two groups are called Luminal A and B and were ER positive. The remaining three were ER negative and grouped as follows: A "Basal-like breast cancer characterised by the lack of expression of ER, PgR and HER-2 as well as increased expression of basal cytokeratins CKs 5/6 and 17. The second type is erbB2 like/HER-2 like with high expression of erb2 and thirdly the normal like BC showing molecular characteristics of normal tissue (Sørlie et al., 2001).

Gene expression analyses has previously classified breast cancer into five main molecular subtypes, namely the luminal A, luminal B, HER2-enriched, triple negative, and non-cancerous breast tissue subtypes (Vishnubalaji et al., 2019) and (Desmedt et al., 2012). Additionally, genome-wide association studies have led to the identification of a number of novel breast cancer variants. These new variants include hereditary risk factors, such as BRCA1, BCRA2, TP53 and PALB2, to name a few (Vishnubalaji et al., 2019) (Desmedt et al., 2012).

Mutations in the BRCA1 and BRCA2 genes in particular have been linked to an estimated 5-10% of breast cancer cases (Pharoah *et al.* 2008). The identification and testing for BRCA mutations have led to the creation and implementation of potentially life-saving strategies, including the use of magnetic resonance imaging (MRI) scanning in breast cancer surveillance and more timely use of chemotherapeutics, along with surgical options (Vishnubalaji et al., 2019) (Desmedt et al., 2012). Thus, the continued understanding of the transcriptome has led to improved understanding of the driving factors involved with breast cancer, identifying new targets for improved prognostic outcomes.

4.1.3 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA), collaboration between the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI), aims to generate comprehensive, multi-dimensional maps of the key genomic changes in major types and subtypes of cancer. A three-year pilot project initiated in 2006 confirmed that an atlas of changes could be created for specific cancer types. It also showed that a national network of research and technology teams working on distinct but related projects could pool the results of their efforts, create an economy of scale and develop an infrastructure for making the data publicly accessible. Importantly, it proved that making the data freely available would enable researchers around the world to make and validate important discoveries. The success of the pilot led the National Institutes of Health to commit major resources to TCGA to collect and characterize additional tumour types. TCGA finalized tissue collection with matched tumour and normal tissues from 11,000 patients, allowing for the comprehensive characterization of 33 cancer types and subtypes, including 10 rare cancers. To achieve this, TGCA aimed to investigate catalogue upwards of 30 human cancers on the bases of cancer-causing genomic alterations, by means of large-scale genome sequencing and integrated multi-dimensional analyses and making these data available to all researchers (Chandran et al., 2016) and (Tomczak, Czerwińska, & Wiznerowicz, 2015).

Presently, TCGA data accounts for over 1.2 Petabytes of information, including multiple forms of datasets, such as whole genome sequencing, whole exome sequence, methylation, RNA expression, proteomic, and clinical data. Access to these datasets are intended for multiple independent research groups to analyse data with the ultimate goal of accelerating the discovery of various biomarkers associated with cancer initiation, progression and response to therapy (Chandran et al., 2016) and (Tomczak et al., 2015).

Datasets provided by TCGA consist of both publicly available and protected datasets. Multiple portals are used to release publicly available TCGA datasets, including the TCGA data portal, the cBIO cancer genomic portal, and University of California, Santa Cruz cancers genome browser. Additionally, these datasets may also be directly downloaded from FIREHOUSE, hosted by the Broad Institute, and via the Sage Bionetworks Synapse repository. Moreover, a number of tools may be used to analyse datasets, including the portals' GUI interfaces, or R packages such as TCGABioLinks (Chandran et al., 2016) and (Tomczak et al., 2015).

4.2 RESEARCH AIMS AND OBJECTIVES

The aims of this chapter were:

To determine which differentially expressed genes are potentially involved for the onset and progression of breast cancer in a developed Breast cancer semantic system. Particular focus was to identify genes that have not yet been described to be involved in breast cancer, or other cancers. Therefore, this was achieved by reanalysing publicly available RNA-seq data from multiple subtypes of breast cancer and normal breast tissues from The Cancer Genome Atlas and performing semantic analysis on these using the complete breast cancer knowledge graph, as in Chapter 2.

4.3 MATERIALS AND METHODS

A manually curated set of already consented and anonymized patient RNA-seq samples from TCGA was re-analysed using an R-based differential expression analysis pipeline. EdgeR was used to filter data based on the p values.

Differentially expressed genes were subjected to classical enrichment analysis using the Enrichr web service and genes that were not assigned to any enriched functions or pathways were then analysed using the semantic database to evaluate potential novel roles in cancer biology.

4.3.1 Data Curation

Selection of Publicly Available Datasets

The Cancer Genome Atlas (TCGA) repository was used to manually curate raw gene counts of RNA-Seq data from breast tissue samples. Subsequently, this database was used to evaluate tumour tissues associated with breast cancer and normal breast tissues against one another.

Table 3.1 shows the dataset generated by the TCGA Research Network ("The Cancer Genome Atlas Program - National Cancer Institute,"). The samples were curated using their allocated TCGA sample IDs from the Genomic Data Commons Data Portal ("GDC Data Portal," n.d.). Normal-adjacent-tumour breast and primary breast tumour samples were curated and the TCGA research paper (The Cancer Genome Atlas Network, 2012) was used to obtain the correct molecular subtypes of tumour samples according to PAM50 classification.

Table 4.1: Summary of breast tissue samples curated from The Cancer Genome Atlas Data Repository.

***TCGA–The Cancer Genome Atlas Dataset**	**Tissue Type**	**Data type**	**Data Repository**
Paired/Unpaired samples	*Normal-Adjacent-Tumour Breast Tissue *Estrogen-Positive Primary Tumour *Her2-PositivePrimary Tumour *Triple Negative Primary Tumour *Triple Positive Primary Tumour	**Raw RNA-Seq counts**	**TCGA**

4.3.2 Discovery of Differentially Expressed Genes (DEGs)

Paired TCGA Samples

The Bioconductor R package, edgeR, was used for differential expression gene analyses of read counts arising from RNA-Seq between normal and tumour samples.

To achieve this, 30 paired samples (normal-adjacent-tumour and primary tumour sample from a single patient, from different patients) were used for a differential gene expression signal to determine if the signal was present, and to ensure correct implementation and application of the edgeR package. (Workflow link in Appendix II)

4.3.3 Processing of Gene Expression Data and Statistical Analysis

Differential expression analysis

Differential expression analysis was performed by comparing all groups' subtypes to healthy subjects (Normal versus tumor), for P-values and fold changes (FC). Next, a filter was created to remove genes with low FC or insignificant P-values (P >0.05). Additionally, the filter accounted for log-fold change, otherwise known as false discovery rate (FDR), whereby 2-fold was considered 4X more or 4X less in differential expression, shown by Log FC ≥2 or ≤—2. Additionally, data was filtered using FDR, significant P-values (P ≤0.05). Following this, InteractiVenn (www.interactivenn.net) tool was used to compare multiple data for each subtype to identify genes shared between subtypes, specifically normal versus HER2, normal versus triple positive, normal versus triple negative, normal versus ER positive PR negative, and normal versus ER positive PR positive.

An R script was used to filter for significant genes based on the following criteria: both p-value and FDR ≤ 0.05 and LogFC between ≤ -2 or ≥ 2. Bar charts showing the distribution of the original raw data and the filtered data for each of the five breast cancer subtypes (Normal_vs_ ERPositivePRNegative, Normal_vs_ ERPositivePRPositive, Normal_vs_ HER2, Normal_vs_ TripleNegative and Normal_vs_ TriplePositive) were plotted in excel. Another R script was used to check for and extract all the gene IDs that are common to the five filtered subtypes.

4.3.4 Functional Enrichment Analysis

Shared genes were analysed further using Enrichr for enrichment analysis to determine which pathways the different genes are involved in. Common and unique DEGs were identified, genes that were DE in a subtype and linked to cancer were of interest. On Enrichr, p≤0.05 was considered significant. To achieve this, Pathways and Ontologies, specifically, were analysed. Genes in each group that were not involved in any enriched pathway or function relevant breast cancer were selected for semantic annotation using the BCA knowledge graph.

4.4 RESULTS AND DISCUSION

4.4.1 Overlapping genes between breast cancer subtypes

Venn diagrams generated indicated 1069 overlapping genes, "Additional Files" (https://drive.google.com/drive/u/0/folders/1T6O8tUKF_PjuvCN8RU-meN40s-oPh7__)

that are shared between breast cancer subtypes, specifically normal versus HER2, normal versus triple positive, normal versus triple negative, normal versus ER positive PR negative, and normal versus ER positive PR positive (Figure 1). There were 271 unique genes for Normal versus ER positive PR negative, 267 unique genes for normal versus Triple positive, 271 unique genes for normal versus ER positive PR positive, 642 unique genes for normal versus HER2 and 1193 unique genes for normal versus Triple negative.

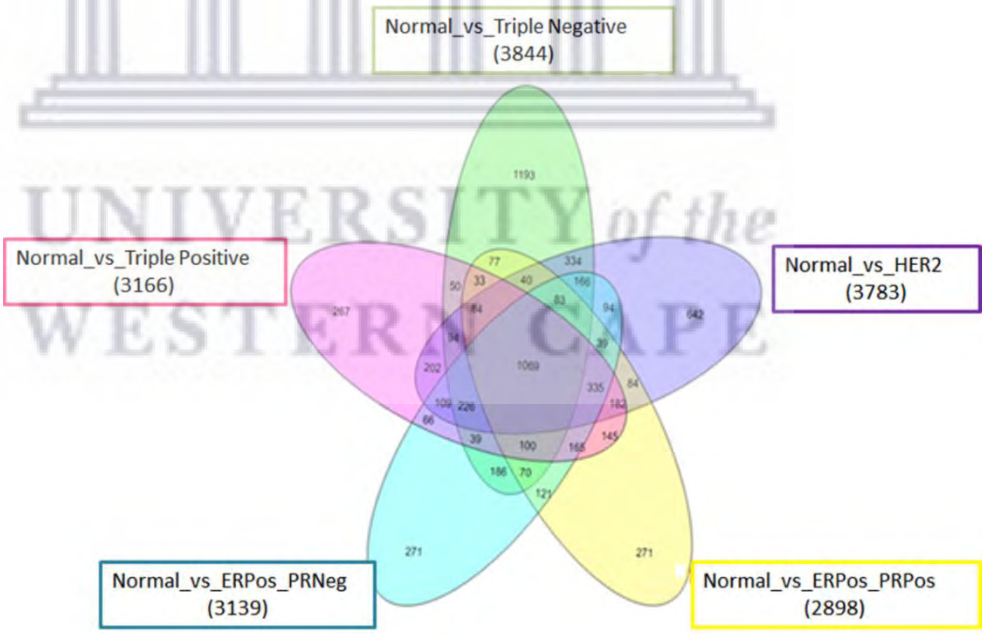

Normal_vs_Triple Negative (3844)
Normal_vs_Triple Positive (3166)
Normal_vs_HER2 (3783)
Normal_vs_ERPos_PRNeg (3139)
Normal_vs_ERPos_PRPos (2898)
1193
77
334
50
33
40
166
83
94
267
64
642
34
39
202
1069
335
84
109
226
182
66
39
100
168
145
186
70
121
271
271

Figure 4.1: Venn diagram showing common and unique DEGs in Breast cancer (LogFC between ≤ -2 or ≥ 2 and significant P-values (P ≤0.05) between all five breast cancer subtypes. There are 1069 common DEGs across five breast cancer subtypes.

Amongst the 1069 genes were those that were downregulated and those that were upregulated. Figures 4.2 and 4.3 present the interactions of the downregulated and upregulated across the different breast cancer subtypes. The interaction revealed 413 and 638 upregulated and down regulated genes, respectively.

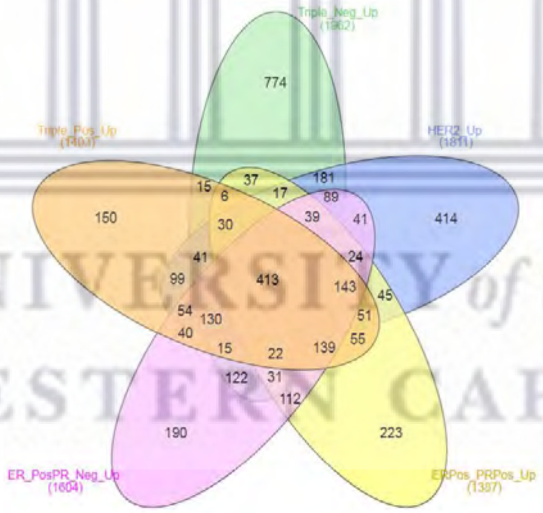


Figure 4.2: Venn diagram showing common and unique upregulated DEGs in breast cancer (LogFC between ≤ -2 or ≥ 2 and significant P-values (P ≤0.05) between all five breast cancer subtypes with 413 genes in common across the five breast cancer subtypes.

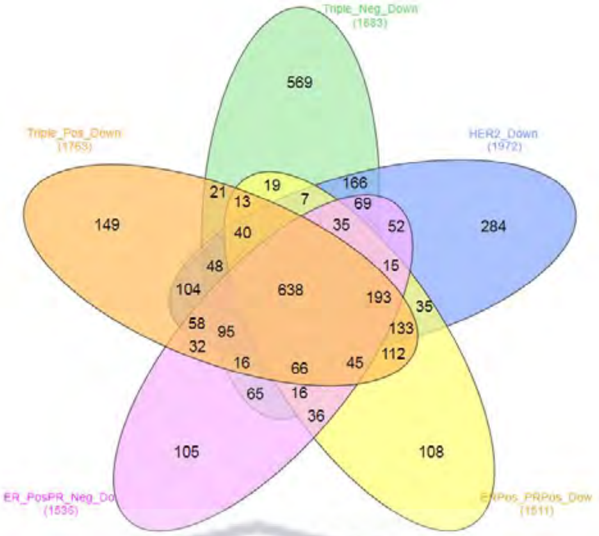
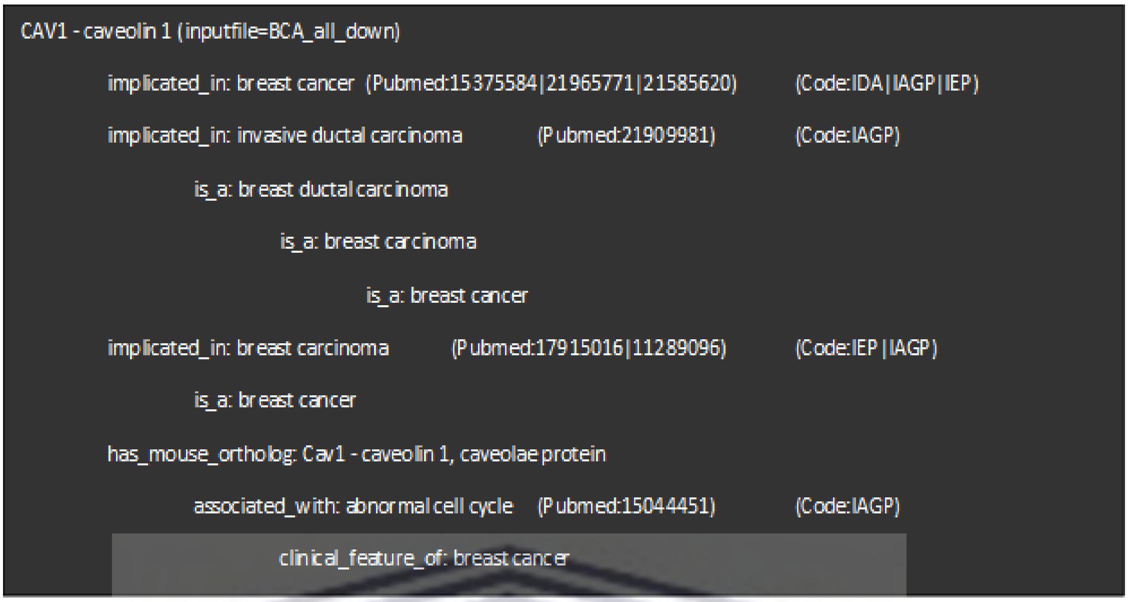


Figure 4.3: Venn diagram showing common and unique downregulated DEGs in breast cancer (LogFC between ≤ -2 or ≥ 2 and significant P-values (P ≤0.05) between all five breast cancer subtypes with 638 genes in common across the five breast cancer subtypes.

4.4.2 Semantic analysis of DEGs

Selected DEGs were run through the comprehensive BCA semantic database as described in Chapter 2. Selected results from the system will be presented and discussed in the subsequent sections. A comprehensive list of the BORG results is presented in an "Additional Files" folder (https://drive.google.com/drive/u/0/folders/1xTBbzr6g2-_JvmuLGhBnUS2h6aj1a0V8).

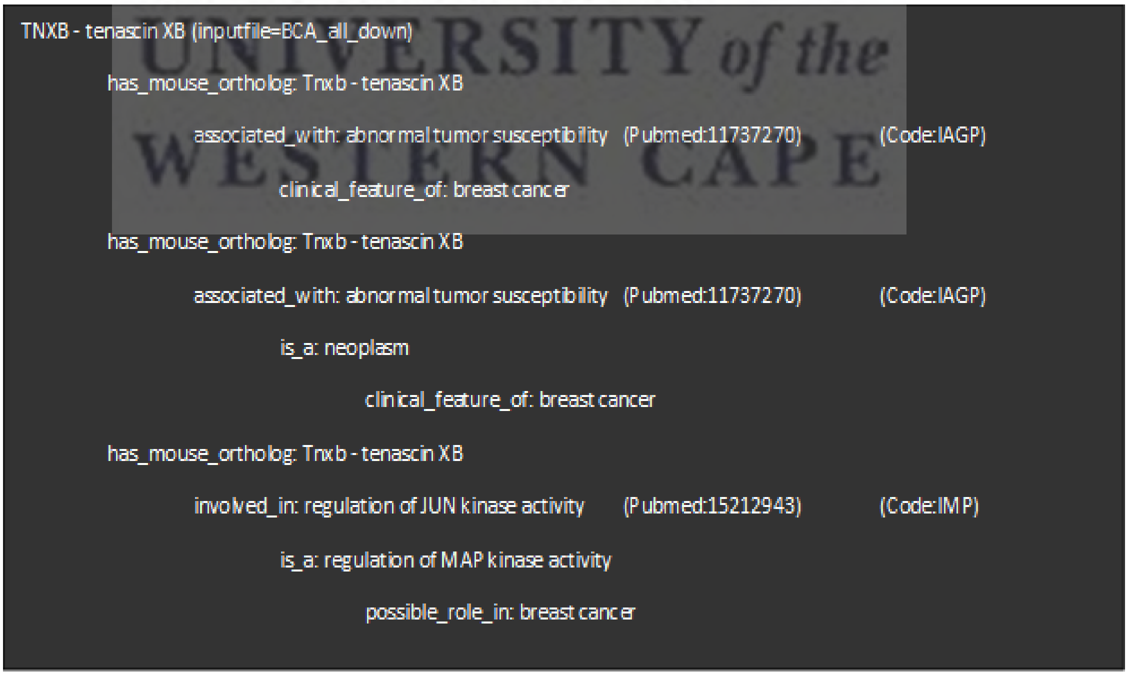
Among the selected genes returned from BORG common across all BCA subtypes, three genes had numerous protein-to-protein interactions, namely *CAV1*, *TP63* and *NR4A1*. The transcripts of the *CAV1* gene (figure 4.5) returned with 19 protein-to-protein interactions. For an example CAV1 interacts with *TLR4* - toll like receptor 4~~,~~ *MAPK1* - mitogen-activated protein kinase 1, *PTEN* - phosphatase and tensin homolog.



```
CAV1 - caveolin 1 (inputfile=BCA_all_down)
        implicated_in: breast cancer  (Pubmed:15375584|21965771|21585620)        (Code:IDA|IAGP|IEP)
        implicated_in: invasive ductal carcinoma          (Pubmed:21909981)          (Code:IAGP)
                is_a: breast ductal carcinoma
                        is_a: breast carcinoma
                                is_a: breast cancer
        implicated_in: breast carcinoma          (Pubmed:17915016|11289096)          (Code:IEP|IAGP)
                is_a: breast cancer
        has_mouse_ortholog: Cav1 - caveolin 1, caveolae protein
                associated_with: abnormal cell cycle     (Pubmed:15044451)          (Code:IAGP)
                        clinical_feature_of: breast cancer
```

Figure 4.4. Semantic analysis results of *CAV1*gene across all BCA subtypes.

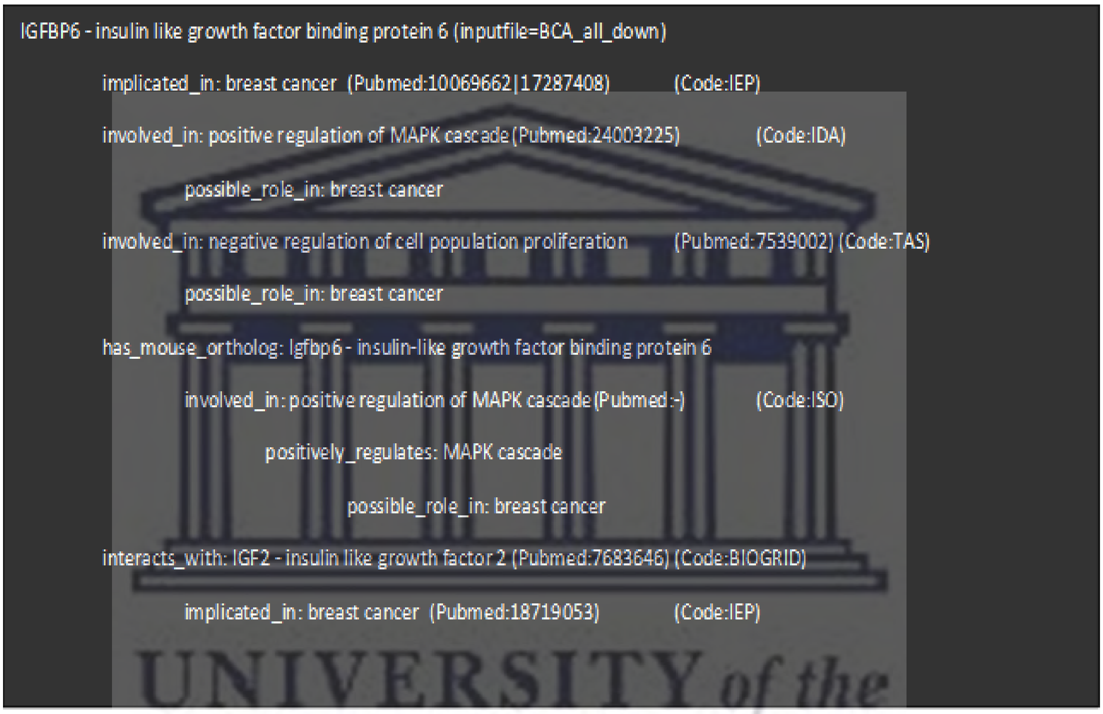
TNXB (Figure 4.6) returned with no reported protein-to-protein interactions yet was associated with abnormal tumor susceptibility, which is relevant to cancer. To date literature has not linked this gene to BCA indicating that it could potentially be a novel effector.



```
TNXB - tenascin XB (inputfile=BCA_all_down)
        has_mouse_ortholog: Tnxb - tenascin XB
                associated_with: abnormal tumor susceptibility   (Pubmed:11737270)          (Code:IAGP)
                        clinical_feature_of: breast cancer
        has_mouse_ortholog: Tnxb - tenascin XB
                associated_with: abnormal tumor susceptibility   (Pubmed:11737270)          (Code:IAGP)
                        is_a: neoplasm
                                clinical_feature_of: breast cancer
        has_mouse_ortholog: Tnxb - tenascin XB
                involved_in: regulation of JUN kinase activity      (Pubmed:15212943)          (Code:IMP)
                        is_a: regulation of MAP kinase activity
                                possible_role_in: breast cancer
```

Figure 4.5. Semantic analysis results of *TNXB* gene across all BCA subtypes.

The rat and mouse orthologs of *IGFBP6* (Figure 4.7), insulin like growth factor binding protein 6 (IGFBP6) are involved in positive regulation of MAPK cascade, negative regulation of cell population proliferation and positive regulation of stress-activated MAPK cascade, all of which have potential roles in tumour biology.

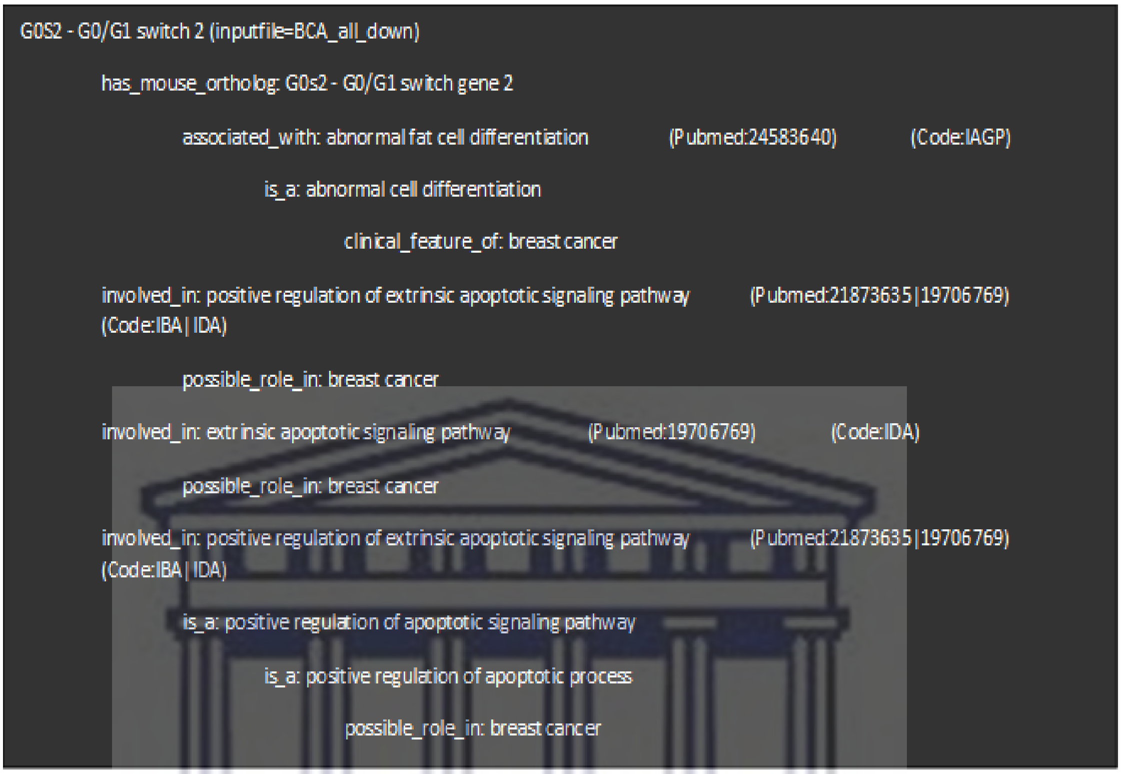


```
IGFBP6 - insulin like growth factor binding protein 6 (inputfile=BCA_all_down)
        implicated_in: breast cancer  (Pubmed:10069662|17287408)        (Code:IEP)
        involved_in: positive regulation of MAPK cascade (Pubmed:24003225)        (Code:IDA)
                possible_role_in: breast cancer
        involved_in: negative regulation of cell population proliferation      (Pubmed:7539002) (Code:TAS)
                possible_role_in: breast cancer
        has_mouse_ortholog: Igfbp6 - insulin-like growth factor binding protein 6
                involved_in: positive regulation of MAPK cascade (Pubmed:-)        (Code:ISO)
                        positively_regulates: MAPK cascade
                                possible_role_in: breast cancer
        interacts_with: IGF2 - insulin like growth factor 2 (Pubmed:7683646) (Code:BIOGRID)
                implicated_in: breast cancer  (Pubmed:18719053)        (Code:IEP)
```

Figure 4.6. Semantic analysis results of *IGFBP6* gene across all BCA subtypes.

G0S2 (Figure 4.8), the G0/G1 switch 2 (*G0S2*) gene, has rat and mouse orthologs. *G0S2 had no reported protein-to-protein interactions* but is implicated in breast cancer by being associated with cellular differentiation and is involved in regulation of the extrinsic apoptotic signaling pathway.

ALDH2, aldehyde dehydrogenase 2 family member, has rat and mouse orthologs and has protein-to-protein interaction with the *SOD2* superoxide dismutase, is associated with decreased tumor latency, negative regulation of apoptosis process and responses to ethanol and progesterone.



```
G0S2 - G0/G1 switch 2 (inputfile=BCA_all_down)
        has_mouse_ortholog: G0s2 - G0/G1 switch gene 2
                associated_with: abnormal fat cell differentiation        (Pubmed:24583640)        (Code:IAGP)
                        is_a: abnormal cell differentiation
                                clinical_feature_of: breast cancer
        involved_in: positive regulation of extrinsic apoptotic signaling pathway        (Pubmed:21873635|19706769)
        (Code:IBA|IDA)
                possible_role_in: breast cancer
        involved_in: extrinsic apoptotic signaling pathway        (Pubmed:19706769)        (Code:IDA)
                possible_role_in: breast cancer
        involved_in: positive regulation of extrinsic apoptotic signaling pathway        (Pubmed:21873635|19706769)
        (Code:IBA|IDA)
                is_a: positive regulation of apoptotic signaling pathway
                        is_a: positive regulation of apoptotic process
                                possible_role_in: breast cancer
```

Figure 4.7. Semantic analysis results of *G0S2* across all BCA subtypes.

We found that the *AR* gene is the androgen receptor gene to be differentially expressed in the triple negative breast cancer (TNBC) subtype. *AR* expression was previously implicated in breast cancer, including TNBC where it is implicated in cell growth, epithelial-to-mesenchymal transition, angiogenesis and immunity, migration, and apoptosis (Lehmann et al., 2020), affecting disease prognosis (Maqbool, Bekele, & Fekadu, 2022). The BCA semantic system annotated AR as implicated in invasive ductal carcinoma with rat and mouse ortholog associated with preneoplasia, decreased tumor latency, abnormal tumor morphology amongst other attributes. Furthermore, AR was annotated as being involved in regulation of apoptotic signaling, positive regulation of transcription, *inter alia*. Finally, the system identified several AR interactions with known BCA genes including *STAT3*, *KAT7*, *CCNE1*, *NCOA6*, *JUN*, and *FOXA1*.

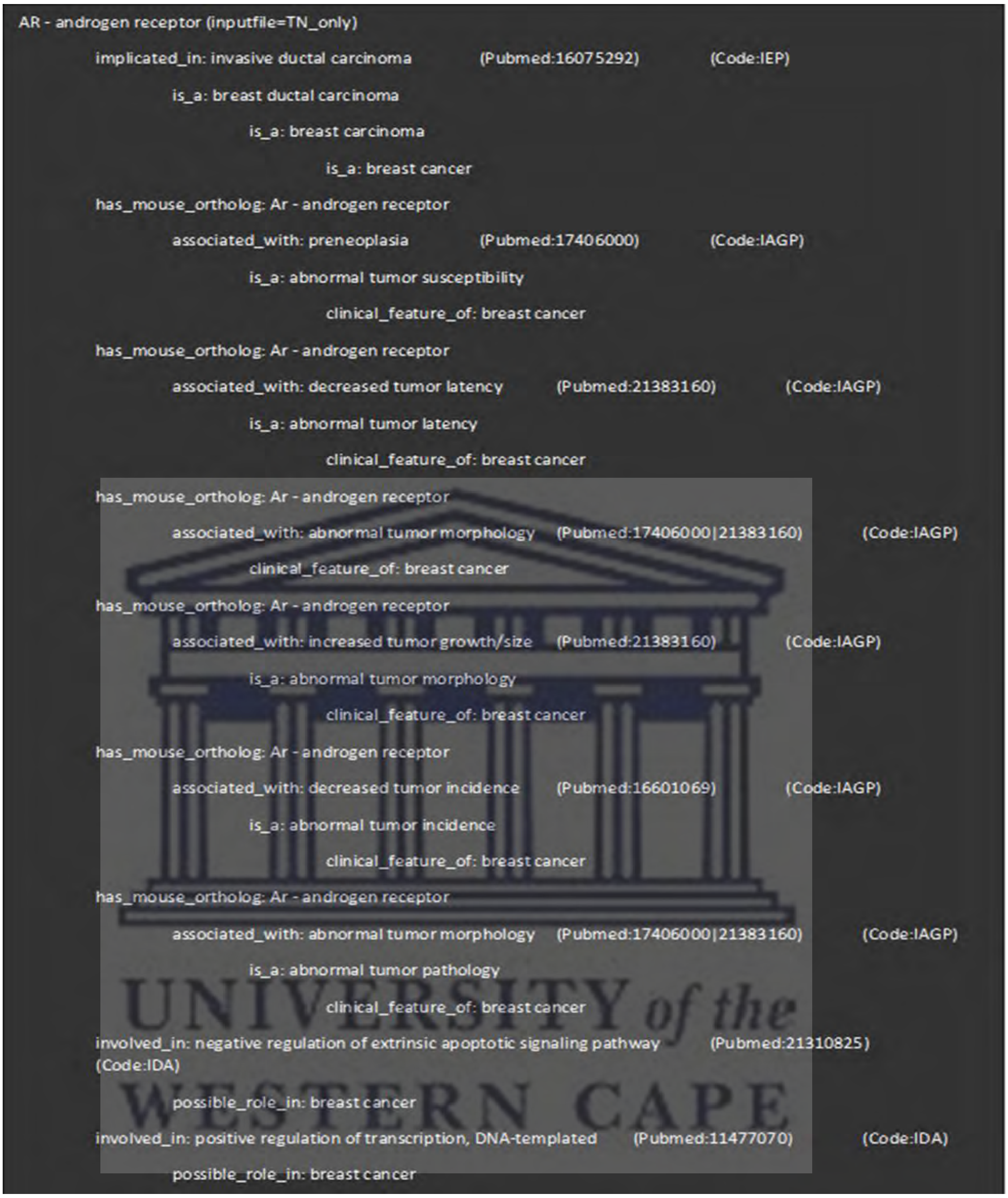
The prevalence of AR expression in all invasive BCA is ± 80% and in about 30% of TNBC patients (Rampurwala, Wisinski, & O'Regan, 2016), (Qattan, Al-Tweigeri, & Suleman, 2022), (Anestis, Zoi, Papavassiliou, & Karamouzis, 2020).

In AR positive BCA, AR may be exploited as a therapeutic target with drugs, for example PROTAC as in prostate cancer (Han et al., 2019) and AR antagonists, such as enzalutamide and bicalutamide in breast cancer (Bhattarai, Saini, Gogineni, & Aneja, 2020). Also, Lehmann et al. demonstrated that preclinical patients with androgen receptor triple-negative breast cancer (TNBC) cells are sensitive to AR antagonists (Lehmann et al., 2020). Thus, AR antagonist may be exploited to improve prognostic outcomes. Furthermore, some studies have reported AR expression association with positive effects in BCA such as the decrease proliferation of TNBC cells and expression of cell-cycle regulator Cyclin D1 (Shen et al., 2017). These are positive effects, indicating the positive potential of the *AR gene* in BCA. However, other studies indicated a negative effect of AR expression, for an example, AR-mediated downregulation of G-protein coupled estrogen receptor expression is associated witth promoting the proliferation in TNBC cells (Zhu et al., 2016), (Anestis, Zoi, Papavassiliou, & Karamouzis, 2020).

On both accounts, AR provides a potential target for regulating these BCA where it is expressed. Furthermore, the accumulating evidence suggests that androgen signaling plays an important role in BCA and androgen receptor (AR) is emerging as a practical marker and therapeutic target as well as a prognostic indicator (Yuan et al., 2017), (Liman et al., 2022), (Yi, Hong, Ohrr, & Yi, 2014), (Park et al., 2011). Bhatarrai et al. indicated that in as much as 10–43% of TNBCs with Androgen receptor (AR) benefit from AR expression the remaining 67%–90% of TNBCs not expressing AR, do not benefit from AR antagonists. Some studies have reported worse prognosis for AR negative-TNBC patients compared to those with AR-positive TNBC (Bhattarai et al., 2020), (Rakha et al., 2007).

There is a higher prevalence of younger patients diagnosed with AR-negative triple negative breast cancers compared to AR-positive patients, with an average age ranging between 35 and 49 (Park et al., 2011), (Davis et al., 2018). Furthermore, Davies et al. indicated higher prevalence of downregulation of AR express in African American patients compared to whites (Davis et al., 2018). AR negative TNBC patients have decreased survival rate compared to the *AR* positive TNBC patients (Anestis et al., 2020), (Davis et al., 2018).

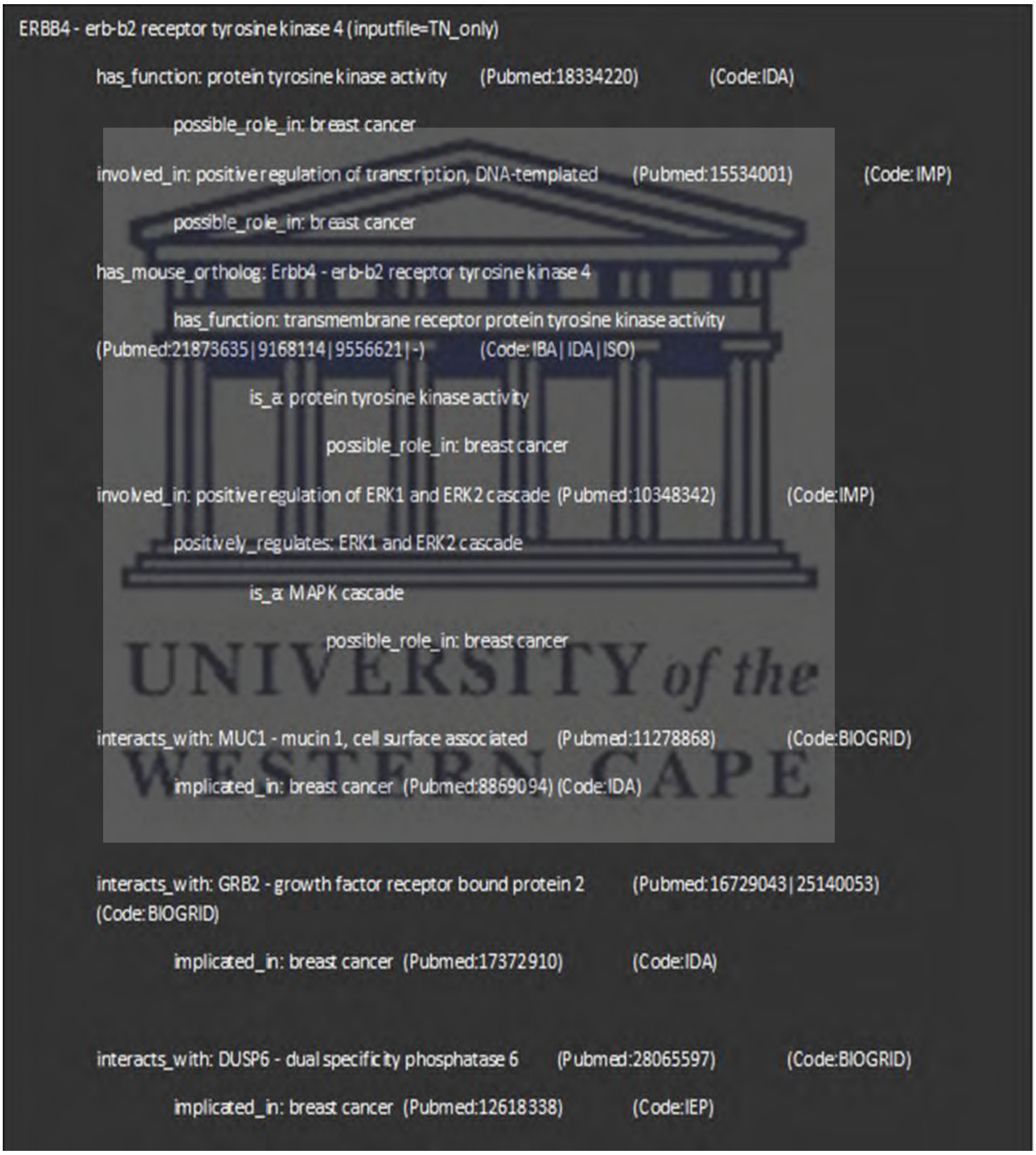
The presence and absence of *AR* indicates that it can be considered as an independent and essential biomarker for the prevalence of and prognostic factor for triple negative breast cancer. Triple-negative breast cancer (TNBC) is a heterogeneous collection of biologically diverse cancers (Lehmann et al., 2011; Vtorushin, Dulesova, & Krakhmal, 2001). Upregulation and downregulation of AR suggests of classification of TNBC into two clades, AR+TNBC and AR-TNBC. Furthermore, the prevalence of, and the therapeutic difficulties in treating AR-TNBC may suggest this clade as a potentially independent BC subtype, the quadruple negative breast cancer (QNBC) (Date et al., 2016).



```
AR - androgen receptor (inputfile=TN_only)
        implicated_in: invasive ductal carcinoma        (Pubmed:16075292)        (Code:IEP)
                is_a: breast ductal carcinoma
                        is_a: breast carcinoma
                                is_a: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: preneoplasia           (Pubmed:17406000)        (Code:IAGP)
                        is_a: abnormal tumor susceptibility
                                clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: decreased tumor latency        (Pubmed:21383160)        (Code:IAGP)
                        is_a: abnormal tumor latency
                                clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: abnormal tumor morphology    (Pubmed:17406000|21383160)        (Code:IAGP)
                        clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: increased tumor growth/size    (Pubmed:21383160)        (Code:IAGP)
                        is_a: abnormal tumor morphology
                                clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: decreased tumor incidence      (Pubmed:16601069)        (Code:IAGP)
                        is_a: abnormal tumor incidence
                                clinical_feature_of: breast cancer
        has_mouse_ortholog: Ar - androgen receptor
                associated_with: abnormal tumor morphology    (Pubmed:17406000|21383160)        (Code:IAGP)
                        is_a: abnormal tumor pathology
                                clinical_feature_of: breast cancer
        involved_in: negative regulation of extrinsic apoptotic signaling pathway        (Pubmed:21310825)
        (Code:IDA)
                possible_role_in: breast cancer
        involved_in: positive regulation of transcription, DNA-templated        (Pubmed:11477070)        (Code:IDA)
                possible_role_in: breast cancer
```

Figure 4.8. Semantic analysis results of *AR* gene, which was differentially expressed in triple negative breast cancer.

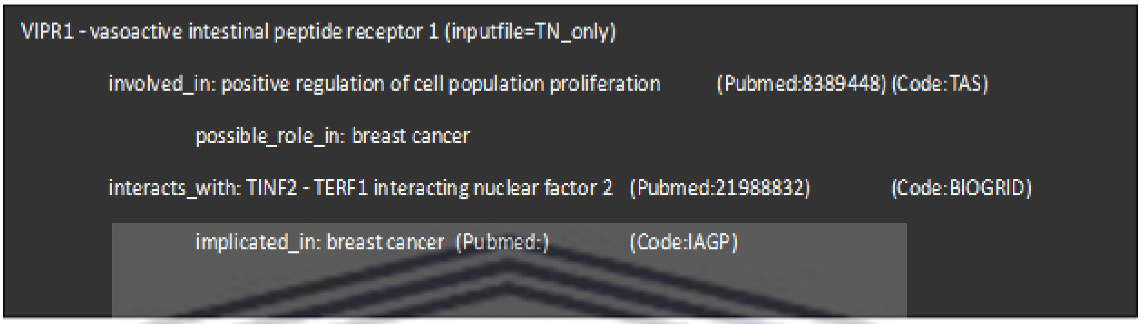
Another gene of interest prioritized by our system was *ERBB4* (Figure 4.10). This gene encodes the erb-b2 receptor tyrosine kinase 4 with a function in protein tyrosine kinase activity that is involved in positive and negative regulation of cell population proliferation. The system annotated that this gene has mouse and rat orthologs that have roles in breast cancer, positive regulation of transcription, *inter alia*. Furthermore, annotations included interactions with several BCA genes including *MUC1*, *GRB2*, and *DUSP*.



```
ERBB4 - erb-b2 receptor tyrosine kinase 4 (inputfile=TN_only)
        has_function: protein tyrosine kinase activity     (Pubmed:18334220)          (Code:IDA)
                possible_role_in: breast cancer
        involved_in: positive regulation of transcription, DNA-templated      (Pubmed:15534001)          (Code: IMP)
                possible_role_in: breast cancer
        has_mouse_ortholog: Erbb4 - erb-b2 receptor tyrosine kinase 4
                has_function: transmembrane receptor protein tyrosine kinase activity
        (Pubmed:21873635|9168114|9556621|-)          (Code: IBA|IDA|ISO)
                        is_a: protein tyrosine kinase activity
                                possible_role_in: breast cancer
        involved_in: positive regulation of ERK1 and ERK2 cascade (Pubmed:10348342)          (Code:IMP)
                positively_regulates: ERK1 and ERK2 cascade
                        is_a: MAPK cascade
                                possible_role_in: breast cancer

        interacts_with: MUC1 - mucin 1, cell surface associated     (Pubmed:11278868)          (Code:BIOGRID)
                implicated_in: breast cancer  (Pubmed:8869094) (Code:IDA)

        interacts_with: GRB2 - growth factor receptor bound protein 2       (Pubmed:16729043|25140053)
        (Code: BIOGRID)
                implicated_in: breast cancer  (Pubmed:17372910)          (Code:IDA)


        interacts_with: DUSP6 - dual specificity phosphatase 6       (Pubmed:28065597)          (Code:BIOGRID)
                implicated_in: breast cancer  (Pubmed:12618338)          (Code:IEP)
```

Figure 4.9. Semantic analysis results of *ERBB4* gene in triple negative breast cancer.

BORG prioritization included genes with paucity of information, which that could indicate novelty implicating them to BCA, for an example *VIPR1 (*Figure 4.11). This gene is implicated through interaction with TINF2, which has known roles in BCA. To date literature has not linked *VIPR1*gene to BCA.



```
VIPR1 - vasoactive intestinal peptide receptor 1 (inputfile=TN_only)
        involved_in: positive regulation of cell population proliferation      (Pubmed:8389448) (Code:TAS)
                possible_role_in: breast cancer
        interacts_with: TINF2 - TERF1 interacting nuclear factor 2   (Pubmed:21988832)          (Code:BIOGRID)
                implicated_in: breast cancer  (Pubmed:)          (Code:IAGP)
```

Figure 4.10. Semantic analysis results of *VIPR1*gene in triple negative breast cancer.

4.5 CONCLUSION

The currently designed semantic database is sensitive enough to prioritize known genes associated with BCA together with novel genes that have not been associated with the disease. The latter genes are scantily reported in literature. Therefore, further analysis needs to be done to elucidate their association with breast cancer. This includes linking *AR* gene expression with triple negative breast cancer leading to further classification of this subtype into QNBC.

This tool maybe used across different diseases for discovering novel genes that may be associated with the discovery of novel biomarkers and potential drug targets.

REFERENCES

- Anestis, A., Zoi, I., Papavassiliou, A. G., & Karamouzis, M. V. (2020). Androgen Receptor in Breast Cancer-Clinical and Preclinical Research Insights. *Aclinical and Preclinical Research Insights." Molecules 25.2 (2020): 358.*, *25*(2), 358. https://doi.org/10.3390/molecules25020358
- Bhattarai, S., Saini, G., Gogineni, K., & Aneja, R. (2020). Quadruple-negative breast cancer: novel implications for a new disease. *Breast Cancer Research*. https://doi.org/10.1186/s13058-020-01369-5
- Botta, A., Visconti, V. V., Fontana, L., Bisceglia, P., Bengala, M., Massa, R., Novelli, G. (2021). A 14-Year Italian Experience in DM2 Genetic Testing: Frequency and Distribution of Normal and Premutated CNBP Alleles. *Frontiers in Genetics*, *12*(June), 1–10. https://doi.org/10.3389/fgene.2021.668094
- Chandran, U. R., Medvedeva, O. P., Barmada, M. M., Blood, P. D., Chakka, A., Luthra, S., Jacobson, R. S. (2016). TCGA expedition: A data acquisition and management system for TCGA data. *PLoS ONE*, *11*(10), 1–14. https://doi.org/10.1371/journal.pone.0165395
- Date, J., Hon, C., Singh, B., Sahin, A., Du, G., Wang, J., Lee, P. (2016). Breast cancer molecular subtypes: from TNBC to QNBC. *Am J Cancer Res*, *6*(9), 1864–1872. Retrieved from www.ajcr.us/
- Davis, M., Tripathi, S., Hughley, R., He, Q., Bae, S., Karanam, B., Yates, C. (2018). AR negative triple negative or “quadruple negative” breast cancers in African American women have an enriched basal and immune signature. *PLOS ONE*, *13*(6), e0196909. https://doi.org/10.1371/JOURNAL.PONE.0196909

- Desmedt, C., Majjaj, S., Kheddoumi, N., Singhal, S. K., Haibe-Kains, B., El Ouriaghli, F., Sotiriou, C. (2012). Characterization and clinical evaluation of CD10 + stroma cells in the breast cancer microenvironment. *Clinical Cancer Research*, *18*(4), 1004–1014. https://doi.org/10.1158/1078-0432.CCR-11-0383
- Gu, J., Xu, T., Huang, Q.-H., Zhang, C.-M., & Chen, H.-Y. (2019). <p>HMGB3 silence inhibits breast cancer cell proliferation and tumor growth by interacting with hypoxia-inducible factor 1α</p>. *Cancer Management and Research*, *Volume 11*, 5075–5089. https://doi.org/10.2147/cmar.s204357
- Han, L., Diao, L., Yu, S., Xu, X., Li, J., Zhang, R., Liang, H. (2015). The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer Cell*, *28*(4), 515–528. https://doi.org/10.1016/j.ccell.2015.08.013
- Jazayeri, S. B., Saadat, S., Ramezani, R., & Kaviani, A. (2015). Incidence of primary breast cancer in Iran: Ten-year national cancer registry data report. *Cancer Epidemiology*, *39*(4), 519–527. https://doi.org/10.1016/j.canep.2015.04.016
- Jeon, M., You, D., Bae, S. Y., Kim, S. W., Nam, S. J., Kim, H. H., Lee, J. E. (2017). Dimerization of EGFR and HER2 induces breast cancer cell motility through STAT1-dependent ACTA2 induction. *Oncotarget*, *8*(31), 50570–50581. https://doi.org/10.18632/oncotarget.10843
- Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, *2015*(11), 951–969. https://doi.org/10.1101/pdb.top084970
- Lehmann, B. D., Abramson, V. G., Sanders, M. E., Mayer, E. L., Haddad, T. C., Nanda, R., Pietenpol, J. A. (2020). TBCRC 032 IB/II Multicenter Study:

Molecular Insights to AR Antagonist and PI3K Inhibitor Efficacy in Patients with AR þ Metastatic Triple-Negative Breast Cancer. *CLINICAL CANCER RESEARCH*, *26*(9), 2111–2123. https://doi.org/10.1158/1078-0432.CCR-19-2170

- Lehmann, B. D., Bauer, J. A., Chen, X., Sanders, M. E., Chakravarthy, A. B., Shyr, Y., & Pietenpol, J. A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation*, *121*(7), 2750–2767. https://doi.org/10.1172/JCI45014
- Liman, A. A., Kabir, B., Abubakar, M., Abdullahi, S., Ahmed, S. A., & Shehu, S. M. (2022). Triple-negative Breast Cancer (TNBC) and Its Luminal Androgen Receptor (LAR) Subtype: A Clinicopathologic Review of Cases in a University Hospital in Northwestern Nigeria. https://doi.org/10.4103/njcp.njcp_437_20
- Ma, R. M., Yang, F., Huang, D. P., Zheng, M., & Wang, Y. L. (2019). The Prognostic Value of the Expression of SMC4 mRNA in Breast Cancer. *Disease Markers*, *2019*. https://doi.org/10.1155/2019/2183057
- Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, *12*(2), 87–98. https://doi.org/10.1038/nrg2934
- Park, S., Koo, J. S., Kim, M. S., Park, H. S., Lee, J. S., Lee, J. S., Lee, K. S. (2011). Androgen receptor expression is significantly associated with better outcomes in estrogen receptor-positive breast cancers. *Annals of Oncology*, *22*(8), 1755–1762. https://doi.org/10.1093/ANNONC/MDQ678
- Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Petiti, L., Tagliabue, L., Landini, P. (2013). An efficient rRNA removal method for RNA sequencing in

GC-rich bacteria. *Microbial Informatics and Experimentation*, *3*(1), 1–11. https://doi.org/10.1186/2042-5783-3-1

- Qattan, A., Al-Tweigeri, T., & Suleman, K. (2022). Translational implications of dysregulated pathways and microRNA regulation in quadruple-negative breast cancer. Biomedicines, 10(2), 366.
- Rakha, E. A., El-Sayed, M. E., Green, A. R., Lee, A. H. S., Robertson, J. F., & Ellis, I. O. (2007). Prognostic markers in triple-negative breast cancer. *Cancer*, *109*(1), 25–32. https://doi.org/10.1002/CNCR.22381
- Rampurwala M, Wisinski KB, O'Regan R. Role of the androgen receptor in triple-negative breast cancer. Clin Adv Hematol Oncol. 2016 Mar;14(3):186-93. PMID: 27058032; PMCID: PMC5221599.
- Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10869–10874. https://doi.org/10.1073/pnas.191367098
- Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkologia*, *1A*, A68–A77. https://doi.org/10.5114/wo.2014.47136
- Vishnubalaji, R., Sasidharan Nair, V., Ouararhni, K., Elkord, E., & Alajez, N. M. (2019). Integrated Transcriptome and Pathway Analyses Revealed Multiple Activated Pathways in Breast Cancer. *Frontiers in Oncology*, *9*(September), 1–13. https://doi.org/10.3389/fonc.2019.00910
- Vtorushin, S., Dulesova, A., & Krakhmal, N. (2001). Luminal androgen receptor (LAR) subtype of triple-negative breast cancer: molecular,

morphological, and clinical features. *Journal of Zhejiang University-SCIENCE B*, *23*(8), 617–624. https://doi.org/10.1631/jzus.B2200113

- Wang, C., Lu, T., Emanuel, G., Babcock, H. P., & Zhuang, X. (2019). Imaging-based pooled CRISPR screening reveals regulators of lncRNA localization. *Proceedings of the National Academy of Sciences of the United States of America*, *166*(22), 10842–10851. https://doi.org/10.1073/pnas.1903808116
- Wang, L., Feng, Z., Wang, X., Wang, X., & Zhang, X. (2009). DEGseq: An R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, *26*(1), 136–138. https://doi.org/10.1093/bioinformatics/btp612
- Wang, Y., Wang, L., Li, X., Qu, X., Han, N., Ruan, M., & Zhang, C. (2021). Decreased CSTA expression promotes lymphatic metastasis and predicts poor survival in oral squamous cell carcinoma. *Archives of Oral Biology*, *126*(March), 105116. https://doi.org/10.1016/j.archoralbio.2021.105116
- Yi, S.-W., Hong, J.-S., Ohrr, H., & Yi, J.-J. (2014). Agent Orange exposure and disease prevalence in Korean Vietnam veterans: The Korean veterans' health study. *Environmental Research*, *133*, 56–65. Retrieved from http://10.0.3.248/j.envres.2014.04.027
- Yuan, C., Horng, C., Lee, C., Chiang, N., Tsai, F., Lu, C., Chen, F. (2017). Epigallocatechin gallate sensitizes cisplatin-resistant oral cancer CAR cell apoptosis and autophagy through stimulating AKT/STAT3 pathway and suppressing multidrug resistance 1 signaling. *Environmental Toxicology*, *32*(3), 845–855. Retrieved from http://10.0.3.234/tox.22284
- Zhang, Kai, et al. "Circular RNA PDK1 targets miR-4731-5p to enhance TNXB expression in ligamentum flavum hypertrophy." The FASEB Journal 37.5 (2023): e22877.

- Zhu, C., Hu, H., Li, J., Wang, J., Wang, K., & Sun, J. (2020). Identification of key differentially expressed genes and gene mutations in breast ductal carcinoma in situ using RNA-seq analysis. *World Journal of Surgical Oncology*, *18*(1), 1–10. https://doi.org/10.1186/s12957-020-01820-z

CHAPTER 5

CONCLUDING REMARKS

BCA is a heterogeneous complex of disease having a spectrum of many subtypes with distinct biological features that lead to differences in response patterns to various treatment modalities and clinical outcomes. For the past two decades, researchers have tried to stratify BCA in order to find improved means for early diagnosis and ultimately better therapeutic approach (Yersal & Barutca, 2014). Due to the complexity of this disease there is still disjointed and paucity of information. Therefore, a development of semantic discovery databases with artificial intelligence is crucial. Such databases will enable end users accessing in-depth information throughout the stages of disease development.

This PhD study developed a comprehensive semantic data graph (BORG) addressing pathogenesis of BCA exploiting computationally filtered genes from various databases including literature to identify novel genes and pathways relating to this disease. This may expedite early detection and intervention curtailing the scourge due to this disease.

It has the potential to assist in improving the understanding of the disease and possibly early detection by respective associations. Interestingly, this tool revealed several genes not previously linked to BCA connecting them through guilt by association through protein-protein interactions with genes with known roles in BCA.

Furthermore, this tool may not be limited only to understanding BCA but may also be extended to other diseases with minor variations, using the semantic modeling methods presented in Chapter 2. In a dynamic graph format, this system may enable an end user to pick up a node or link and detailed information with BCA subtypes, ontologies, pathways etc., may come up.

To date, some of the genes picked out by this system as novel still come with inconclusive information associating them with BCA. For an example, Lu et.al (2022) could not associate *HNRNPA3* gene with the disease. The discovery of genes implicated to BCA through "guilt by association" also opens avenue for future research elucidating their role in the disease. Moreover, the overall description of genes from the here developed semantic discovery and computational filtration may pave way for drug repurposing.

The limitation is the speed of generating information in this day and age which may make data appear outdated. However, the idea behind this study was to tap on artificial intelligence. Therefore, the use of the here-developed semantic system will automatically be updated.

REFERENCES

Lu, Y., Wang, X., Gu, Q. *et al*. Heterogeneous nuclear ribonucleoprotein A/B: an emerging group of cancer biomarkers and therapeutic targets. *Cell Death Discov.* **8**, 337 (2022). https://doi.org/10.1038/s41420-022-01129-8

Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol. 2014 Aug 10;5(3):412-24. doi: 10.5306/wjco.v5.i3.412. PMID: 25114856; PMCID: PMC4127612.

APPENDIX I

Table 3.1: Comparison of frequencies of SNVs mutation between breast cancer and other cancers. "Additional Files" (https://drive.google.com/drive/u/0/folders/1HxuY4NhCHKQTAHCzle9omVtUhCjyE6e-)

Symbol	# SSM Affected Cases in Cohort	Total BCA SNV	% BCA mutated	# SSM Affected	Total All CA SN	% SNV All Cancers	BCA vs All (percent difference)
CDH1	149	954	15.62%	397	11,519	3.45%	453.2%
GATA3	134	954	14.05%	370	11,519	3.21%	437.4%
MAP3K1	89	954	9.33%	347	11,519	3.01%	309.7%
CBFB	28	954	2.94%	114	11,519	0.99%	297.1%
PIK3CA	331	954	34.70%	1,416	11,519	12.29%	282.3%
MAP2K4	43	954	4.51%	208	11,519	1.81%	249.8%
AKT1	32	954	3.35%	160	11,519	1.39%	241.2%
GOLGA6L6	49	954	5.14%	254	11,519	2.21%	233.1%
RUNX1	48	954	5.03%	259	11,519	2.25%	223.7%
MUC2	109	954	11.43%	716	11,519	6.22%	183.9%
FOXA1	32	954	3.35%	218	11,519	1.89%	177.0%
FAM230B	24	954	2.52%	164	11,519	1.42%	177.0%
MUC4	193	954	20.23%	1,340	11,519	11.63%	173.9%
ST6GALNAC3	38	954	3.98%	286	11,519	2.48%	160.3%
REXO1L1P	29	954	3.04%	229	11,519	1.99%	152.9%
CCDC168	51	954	5.35%	405	11,519	3.52%	152.2%
NBPF12	38	954	3.98%	308	11,519	2.67%	148.8%

Table 3.2: Upregulated CNVs_Enrichment Analysis

Genes not on Enrichr	Gene Prev. described in Breast Cancer (Y/N)	Gene Prev. described in any other Cancer (Y/N)	Enrchr pathways KEGG
TANC2	No	No	No data
FER1L6	No	No	No data
PKHD1L1	No	No	No data

PHF20L1	No disorders were found for PHF20L1 Gene	No disorders were found for PHF20L1 Gene	No data
NBPF12	No	No	No data
PCNXL2	No	No	No data
MALAT1	No	Yes	No data
CCDC168	No	Yes	No data
XKR4	No disorders were found for XKR4 Gene	No disorders were found for XKR4 Gene	No data
POTEM	No disorders were found for POTEM Gene	No disorders were found for POTEM Gene	No data
KIAA1549L	No disorders were found for KIAA1549L Gene	No disorders were found for KIAA1549L Gene	No data
QSER1	No disorders were found for QSER1 Gene	No disorders were found for QSER1 Gene	No data
ADAMTSL3	Yes	Yes	No data
GOLGA6L10	No	Yes	No data
FAT3	No	No	No data
UMODL1	No	No	No data
MAGEA12	No	Yes	No data
NBEA	No	No	No data
FCGBP	No	Yes	No data
PDZD2	No	Yes	No data
RLF	No	No	No data
KIAA2018	No	No	No data
NBEAL1	No	Yes	No data

Table 3.4: Downregulated CNVs_Enrichment Analysis

Genes not on Enrichr	**Gene Prev. described in Breast Cancer (Y/N)**	**Gene Prev. described in any other Cancer (Y/N)**	**Enrchr pathways KEGG**
VPS13D	No	No	No data
CSMD1	No	Yes	No data
GOLGA6L6	No	No	No data
NBEA	No	No	No data
ZNF292	No	Yes	No data
CSMD2	No	Yes	No data
FAT3	No	No	No data
XKR4	No disorders were found for XKR4 Gene	No disorders were found for XKR4 Gene	No data

CMYA5	No	No	No data
ERICH3	No disorders were found for ERICH3 Gene	No disorders were found for ERICH3 Gene	No data
MALAT1	No	Yes	No data

APPENDIX II

R (EdgeR) WORKFLOW:

EdgeR workflow for differential expression gene analyses of read counts arising from RNA-Seq between normal and tumour samples.

```
#WORKFLOW on R (EdgeR):
files <- dir()  # tell R your files are in the working directory (only
htseq.count files)
x <- readDGE(files,header=FALSE,columns=c(1,2),comment.char="_")    # read
files into edgeR format
dim(x)    # check that all lines (genes) and samples read in correctly,
dimensions (dim) of datagroup
<as.factor(c("N","N","N","N","N","N","N","N","N","N","TN","TN","TN","TN","TN
","TN","TN","TN","TN","TN"))
data <- DGEList(counts = x, group = group)   # assigning tissue type (group
label) to each sample read into R
data$samples    # check each sample correctly labelled
count_table<- data $ counts # gives you a table of all raw counts
keep    <-    filterByExpr(data,design=NULL,group=group,min.count    =
1,min.total.count = 10)   # 30 in my case filtering out rows (rowSums) that
equal zero, or don't have at least a count of 1 transcript per sample
table(keep)   #checking how many genes were TRUE or FALSE based on conditions
provided in filtering step
y <- data[keep,]  #removing the non-informative genes/transcripts from dataset
for DE analysis
z <- calcNormFactors(y,method = "upperquartile")   #normalizing/scaling data
to the top 25% of the raw counts (3rd or upper quartile in data distribution
bell curve)
a <- estimateCommonDisp(z)   #estimate common dispersion of normalized data
(z)
```

```
b <- estimateTagwiseDisp(a) #estimate data dispersion based on assigned labels of samples, but use previous variable (a)
de <- exactTest(b, pair = c("N","TN"))   #find differentially expressed genes between two groups N and TN (10 samples each)
tt<- topTags(de, n=nrow(de))    #ranking of DE genes with logFC and adjusted p-value (FDR); n can equal any number (desired top number of genes e.g. Top 100 DE genes, then n = 100)
write.csv(tt,"DE_genes.csv")    #write DE genes (ranked) to csv file to view in Excel, with logFC, p-values, etc.
```