Proteome-Transcriptome Alignment of Molecular Portraits by Self-Contained Gene Set Analysis: Breast Cancer Subtypes Case Study

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Koushik Ayaluri
Bachelor of Technology
Amity Institute of Biotechnology, Amity University

Director: Dr. Ancha Baranova, Professor
School of Systems Biology

Spring Semester 2020
George Mason University
Fairfax, VA

DEDICATION

This thesis is dedicated to my parents and all my friends for their endless love, support and encouragement.

ACKNOWLEDGEMENTS

There are many people who walked alongside me during my master's journey. They have guided me, placed opportunities in front of me and guided me to overcome challenges. I would like to thank each one of them. I would like to thank my advisor, Dr. Ancha Baranova for guiding me in this work and steering me in the right direction whenever needed. I would also like to thank Dr. Galina Glazko for all the insightful inputs which helped in this research. I would also like to thank my committee member Dr. Haw Chuan Lim and Dr. Dmitri Klimov for being supportive throughout this work.

Finally, I would like to thank my parents, whose love and guidance are with me in whatever I pursue. Very big thanks go to my friends for being there throughout this journey.

Thank you, Lord, for always being there.

This thesis is only a beginning of my journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Gene Set Analysis ..... GSA
Gene Set Network Correlation Analysis ..... GSNCA
Human epidermal growth factor receptor 2 ..... HER2
Multivariate Imputation via Chained Equations ..... MICE
Missing at Random ..... MAR
Gene Set Analysis in R ..... GSAR
Minimum Spanning Tree …… ……. MST
Principal Component Analysis……...PCA
Kolmogorov-Smirnov ……. KS
Radial Kolmogorov-Smirnov…… RKS
Rotation gene set tests……...…ROAST
Gene Set Enrichment Analysis……... GSEA
Rotation testing using Mean Ranks…… ROMER

ABSTRACT

PROTEOME-TRANSCRIPTOME ALIGNMENT OF MOLECULAR PORTRAITS BY SELF-CONTAINED GENE SET ANALYSIS: BREAST CANCER SUBTYPES CASE STUDY

Koushik Ayaluri, M.S.

George Mason University, 2020

Thesis Director: Dr. Ancha Baranova

Gene sets are formed by grouping together functionally related genes or pathways. Gene set analysis (GSA) is a method previously developed for examining transcriptome data. As the gene sets are unit of expression in transcriptome-level GSA, similarly, the unit of protein abundance may be used for proteomics GSA. *Self-contained* and *Competitive* are two GSA approaches which differ by their underlining null hypothesis. In *Self-contained approach,* each gene set is evaluated to check if it is expressed differentially between two phenotypes. In *Competitive approach,* each gene set is compared to all the genes except the genes in that set. *Competitive* approaches are rapidly becoming popular for analyzing proteomics data, as much as they were for transcriptomics data. This research applied *Self-contained GSA* test of Gene sets net correlations analysis (GSNCA) to proteomics data of 77 annotated samples of breast cancers. Regardless of significant variation in the structure of proteomics and transcriptomics data, many pathway-wide

characteristics features of breast cancer molecular subtypes were replicated at the protein level. In this work, GSA yielded a set of observations visible at proteome level, such as mitotic cell cycle process involvement in the HER2 molecular subtype. Overall, this study proves the value of Gene Sets Net Correlation Analysis (GSNCA) approach as a critical tool for analyzing proteomics data in general, and for dissecting protein-level molecular portraits of breast cancer tumors, in particular.

INTRODUCTION

In recent years, several proteomic methodologies have been developed that now make it possible to identify, characterize, and comparatively quantify the relative levels of expression of hundreds of proteins that are co-expressed in each cell type or tissue. One of the most fundamental approaches to understanding the functions of individual proteins in complex cellular processes is to correlate protein expression levels with observed biological changes [1]. Processing and analysis of this proteomics data are done through a complex multistep procedure. The molecular profiles obtained in these large-scale omics experiments are most frequently represented by gene expression levels, protein abundances or metabolite concentrations, and always require further analysis and interpretation. Most often these types of data do not offer immediate understanding of difference between phenotypes, or clearly display mechanism of the disease. To understand the fundamental biological processes underpinning phenotypic differences between normal and malignant states and to discern relevant pathophysiological mechanisms, high throughput data are usually analyzed in a context of pre-existing biological information, such as protein-protein interactions (PPIs), biological pathways, drug-protein interaction data; disease-specific databases and other relevant information sources are being commonly interrogated. Therefore, in the context of proteomics data,

we will start with the very first and highly common step of integration of proteome profiles with biological pathways [2]. The techniques for proteomics data integration that we will use are very similar to that already tested in the field of genomics data [3].

The major difficulty in applying overrepresentation analysis for proteomics data is limited sample size. Gene set analysis requires a list of entities, significantly different between two phenotypes. In proteomics datasets, underlining variances in protein expression or technical variances often preclude clear determination of differentially abundant (DA) protein lists. For example, in the case of analyzing a proteomics data set collected from patients with Parkinson's disease [4], there were 72 test patients and 72 healthy control patients, a relatively large sample size for proteomics data. When this data set was studied for overrepresentation of genes [5], after correction for multiple testing, no DA proteins between the two groups were found. Because of that, unadjusted *p*-values were used for overrepresentation analysis, which lead to high risk of false positives.

The overrepresentation analysis of gene sets has been designed to improve our understanding of tediously long differentially expressed (DE) gene lists, which is a typical output of transcriptomics research. This type of analysis compares sets of genes annotated to pathways in a Gene Ontology (GO) categories, or Kyoto Encyclopedia of genes and genomes (KEGG), or Molecular Signature Database (MSigDB), or any other pathway database to a list of those genes that are significantly Differentially expressed (DE) between two phenotypes, using standard statistical tests for enrichment [6]. This

overrepresentation analysis is also commonly used for examining proteomics data. Here we will use an approach collectively known as Gene Set Analysis (GSA). GSA approaches are characterized into two groups: *self-contained* or *competitive,* on the basis of the null hypotheses they test [7]. In proteomics GSA, significantly differentially abundant (DA) proteins are extracted from proteomics data, in the place of significantly DE genes. Competitive GSA approaches are progressively becoming popular and are more comprehensively analyzed in the context of proteomics data as much as in context of transcriptomics data [2].

The various techniques similar to Gene Set Enrichment Analysis (GSEA) applied for proteomics data are known as protein set enrichment analysis (PSEA)) [8] and PSEA-Quant (an example of protein set enrichment analysis allowing to compare proteomic profiles from one or more conditions) [9]. According to our knowledge, *self-contained approaches* are not frequently applied to proteomics data, even though they have a low Type I error rate and high power compared to the *competitive* approach [10].

In a previous recently published work, self-contained GSA tests were successfully applied in an analysis of the proteomes of the four consensus molecular subtypes (CMSs) previously established by transcriptome dissection of colon carcinoma specimens [GLAZKO G et al., 2019]. Self-contained approaches compare whether a gene set is differentially expressed between two phenotypes, while competitive approaches compare a gene set against its complement that contains all genes except genes in the set. Inspired by this work, were undertook an effort to understand the power and the applicability of

self-contained GSA tests on proteomics data by applying these tests on breast cancer proteomics data.

Breast cancer is a heterogenous disease with distinct molecular portraits, with subtype-dependent clinical outcomes [11]. When breast cancer is detected in the early stage, and is in the localized stage, with suitable treatment the 5-year relative survival rate is 100%. The size of a breast cancer and how far it has spread are two of the most important factors in predicting the prognosis. Initial clinical characterization of the breast cancer is defined by TNM (Tumor, Nodes, Metastasis) stage. T describes the size of the tumor (area of cancer) and its stage. N describes whether the cancer has spread to the lymph nodes that are involved. M describes whether the cancer has spread to a different part of the body. A significant limitation of the breast cancer TNM staging system is that it does not account for biologic factors known to have predictive and prognostic value such as tumor grade, estrogen receptor (ER) and progesterone receptor (PR) status, and HER2 status [12]. So, the treatments and the therapies that follow are decided by the different molecular phenotype of the breast cancer.

In the last two decades whole transcriptome analysis became routinely used to dissect cancer molecular subtypes correlating with clinical outcomes. Starting with the seminal paper of Golub et al [13], defining finer subclasses of the leukemias, there has been a steady growth in similarly designed research [14,15]. The Breast cancer is differentiated into four different molecular subtypes based on mRNA expression into Luminal A, Luminal B, Basal-like (triple-negative) and HER2 positive. The basal-like

subtype is characterized by the expansion of keratins typically found in basal epithelial cells, such as keratins CK5 and CK17, low expression of the ER gene cluster and HER2 gene cluster and high expression of the proliferative gene cluster [16]. HER2 (*ERBB2*) gene amplification and its corresponding overexpression are present in 15–30% of invasive breast cancers and is associated with poor prognosis [17]. Signaling pathways activated by HER2 include mitogen-activated protein kinase (MAPK), phosphoinositide 3-kinase (PI3K/Akt), phospholipase C γ, protein kinase C (PKC), signal transducer and activator of transcription (STAT) [18]. At the RNA and protein level, Luminal A and B subtypes are largely distinguished by the expression of two main biological processes: proliferation/cell cycle-related and luminal/hormone-regulated pathways [19]. When compared to the Luminal A tumors, the Luminal B tumors have higher expression of proliferation/cell cycle-related genes or proteins (e.g. MKI67 and AURKA) and lower expression of several luminal-related genes or proteins such as the progesterone receptor (PR) [19] and FOXA1, but not the estrogen receptor [19], which is found similarly expressed between the two luminal breast carcinoma subtypes, and, therefore, may help to distinguish luminal from non-luminal disease only.

As protein expression links genotype to phenotype, for more detailed characterization of breast cancer molecular subtypes, respective proteomes were also analyzed. Proteomic data are commonly collected using high-resolution accurate mass tandem mass spectrometry (MS/MS) that includes extensive peptide fractionation and phosphopeptide enrichment [20]. Although mass spectrometry-based proteomics has the advantage of detecting thousands of proteins from a single experiment, it faces certain

challenges such as the presence of missing values, indicating lack of quantification for certain protein in some but not all samples. This is a common issue in proteomic experiments, which arises due to sample complexity and variation in sampling from one run to another [21]. To overcome the missing value problem, proteins that are too sparsely quantified must be removed from consideration. In my preliminary run, I have removed the proteins which have more than 5% of missing data. After filtering the data, some more missing values remained in the dataset. These values may be imputed using certain statistical approaches [22]. Here I implemented the imputation method by MICE package and imputed the missing values by median imputation.

Here I reanalyze previously published proteomes of breast cancer molecular subtypes to explain to what extent transcriptionally identified breast cancer molecular subtypes are detected at the proteome level with gene sets network correlations analysis (GSNCA) a self-contained GSA test and find if any new pathways are detected with the GSNCA.

MATERIALS AND METHODS

1. MICE package

MICE (Multivariate Imputation via Chained Equations) is a multiple imputation technique. MICE operate under the assumption that given the variables used in the imputation procedure, the missing data are Missing at Random (MAR) [23]. It imputes data on a variable by variable basis by specifying an imputation model per variable.

2. Self-contained GSA tests

2.1 KS and RKS

There are two variants of the Kolmogorov-Smirnov test, one that tests the null hypothesis of mean vector equality (KS) while the other that tests the variance vector equality (RKS) between the two phenotypes. The Radial Kolmogorov-Smirnov is sensitive to the alternatives which show similarity in mean vectors and difference in the scale. These two tests are available in GSAR Bioconductor in R [24].

2.2 ROAST

Rotation gene set tests (ROAST) checks if the coefficient is non-zero for all the genes using a linear modeling framework [25]. It gives information about the correlations between the genes and also has the ability to use different alternative hypotheses to test the direction of change for a gene [25]. This test uses a parametric resampling method

called rotation and hence has better p-values even for the small sample sizes. It is available in the limma Bioconductor of R.

2.3 GSNCA

Given two conditions, the Gene Sets Net Correlation Analysis is used to account for the net correlation structure for a gene between them [26]. It is available in R as a function named GSNCA test in a Bioconductor package GSAR. Here GSNCA test was applied to the gene sets to find the differential expression (DE), differential variability (DV) and differential co-expression (DC) between the cancer subtypes.

For a pathway to be included as a DE or a DV pathway, it should have the Benjamini - Hochberg value adjusted to <0.01 after multiple testing corrections. A pathway needs to include more than 60% each of upregulated and downregulated genes and the original pathways should account for at least 50%. Its addition to this, a pathway should have the Benjamini - Hochberg value <0.1 after correction, to be included in DC pathway [2].

3. Competitive GSA tests

3.1 GSEA

The Gene Set Enrichment Analysis is said to be the first competitive GSA test method [27]. It tests the null hypothesis that the phenotypic association between the genes in a gene set is random. As local and global test statistic, it uses signal to noise ratio and weighted Kolmogorov-Smirnov tests respectively. It may be accessed on the website (http://software.broadinstitute.org/gsea/index.jsp).

3.2 ROMER

As the name suggests, the Rotation testing using Mean Ranks (ROMER) tests if the genes in a gene set are randomly associated with the phenotype as the null hypothesis. This is similar to that of GSEA except for the fact that it uses rotations to get the p-values instead of permutations as seen in ROAST. This is available in the limma package of Bioconductor [28].

4. Minimum spanning tree

The co-expression networks are produced using the minimum spanning tree (MST) method. In MST2 the vertices correspond to gene in the gene set and set of edges connecting pairs of vertices with weights estimated by some correlation distance measure. The connection network MST2 provides the minimal set of critical interactions between genes, which we interpret as a functional interaction network. A gene that is strongly associated with most of the other genes in the series of genes appears to occupy a central role and has a relatively high degree in MST2, as the shortest paths connecting the vertices of the first and second MSTs continue to pass through this gene. A gene with low inter-gene associations, by comparison, most likely occupies a non-centric role in the MST2 and has a degree of 2. This property of the MST2 makes it a powerful graphical visualization method, by highlighting the most influential genes, to analyze the full correlation network obtained from gene expression data [24].

5. Data Set

The breast cancer proteomes consist of 77 tumor samples and 3 biological replicates and 6807 proteins were downloaded from [20]. The data is already normalized and TCGA identifiers as well as clinical information were available for each sample [20]. This cohort includes a balanced representation of PAM-50 defined intrinsic subtypes of breast cancers including 19 basal-like, 13 HER2 (ERBB2)-enriched tumors, 23 luminal A and 25 luminal B tumors.

RESULTS

When the samples for all four-breast cancer molecular subtypes were analyzed by PCA based on their proteome features, the separation of subtypes was relatively poor (Fig.1). Here we embark on finding out if there are any protein-level pathways from MSigDB C2 curated gene sets that were differentially expressed between breast cancer molecular subtype.

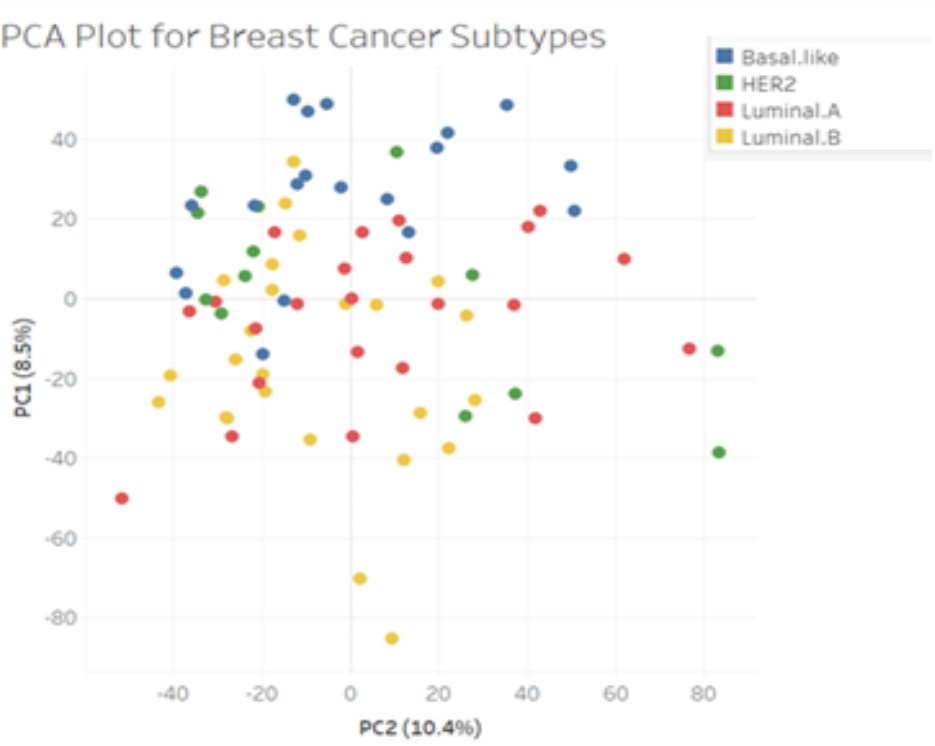

Figure 1. PCA plot for Breast cancer subtypes. The number in parenthesis indicates the percent of variance explained by each Principal Component.

1. Breast cancer subtype specific insights extracted using GSNCA

Using GSNCA analysis, five pathways were differentially co-expressed between basal subtype and other three subtypes, seven pathways were differentially co-expressed between HER2 subtype and other three subtypes, three pathways differentially co-expressed between Luminal A and other three subtypes and seven pathways differentially co-expressed between Luminal B and other three subtypes. Significant pathways were those with Benjamini-Hochberg adjusted p-value <0.01.

1.1 Pathways differentially expressed between Basal subtype and other breast cancer subtypes

Previous transcriptome analysis showed that the genes associated with signal transduction, angiogenesis, cell cycle and proliferation, cell survival, DNA replication and recombination, motility and invasion, and NFkB signaling are overexpressed in basal tumors [29]. The 5 differentially co-expressed pathways between basal subtype and the other three breast cancer subtypes are ‘Doane breast cancer classes downregulated’, ’Ginestier breast cancer 20Q13 amplification upregulated’, ’Nikolsky breast cancer 7Q21 Q22 amplicon’, ’Roylance breast cancer 16Q copy number upregulated’ and ‘Yang breast cancer ESR1 upregulated’ (**Table 1**) discovered from the MSigDb C2 curated gene sets with Benjamini-Hochberg adjusted p-value <0.01. As we can observe from the (**Table 2**), all the hub genes for both groups are upregulated. The hub genes are defined as genes

with the largest weights and show that these genes correspond frequently to major and specific pathway regulators, as well as to genes that are most affected by the biological difference between two conditions.

Table 1: GSNCA highlighted pathways differentially co-expressed between basal subtype and other breast cancer subtypes and their GSNCA P-values

Pathway Names	**GSNCA P-value**
DOANE_BREAST_CANCER_CLASSES_DN	0.008
GINESTIER_BREAST_CANCER_20Q13_AMPLIFICATION_UP	0.004
NIKOLSKY_BREAST_CANCER_7Q21_Q22_AMPLICON	0.004
ROYLANCE_BREAST_CANCER_16Q_COPY_NUMBER_UP	0.001
YANG_BREAST_CANCER_ESR1_UP	0.002

Table 2: Hub genes differ between Basal subtype and other breast cancer subtype

	Hub Genes			
Pathway Names	**Group-1 (Basal Subtype)**	**Up or Down**	**Group-2 (Other three subtypes)**	**Up or Down**
DOANE_BREAST_CANCER_CLASSES_DN	**PARVB**	**Up**	**FERMT1**	**Up**
GINESTIER_BREAST_CANCER_20Q13_AMPLIFICATION_UP	**CFD**	**Up**	**SEC62**	**Up**
NIKOLSKY_BREAST_CANCER_7Q21_Q22_AMPLICON	**MVD**	**Up**	**ZC3H18**	**Up**
ROYLANCE_BREAST_CANCER_16Q_COPY_NUMBER_UP	**PLCG2**	**Up**	**CPNE2**	**Up**
YANG_BREAST_CANCER_ESR1_UP	**INPP4B**	**Up**	**CA12**	**Up**

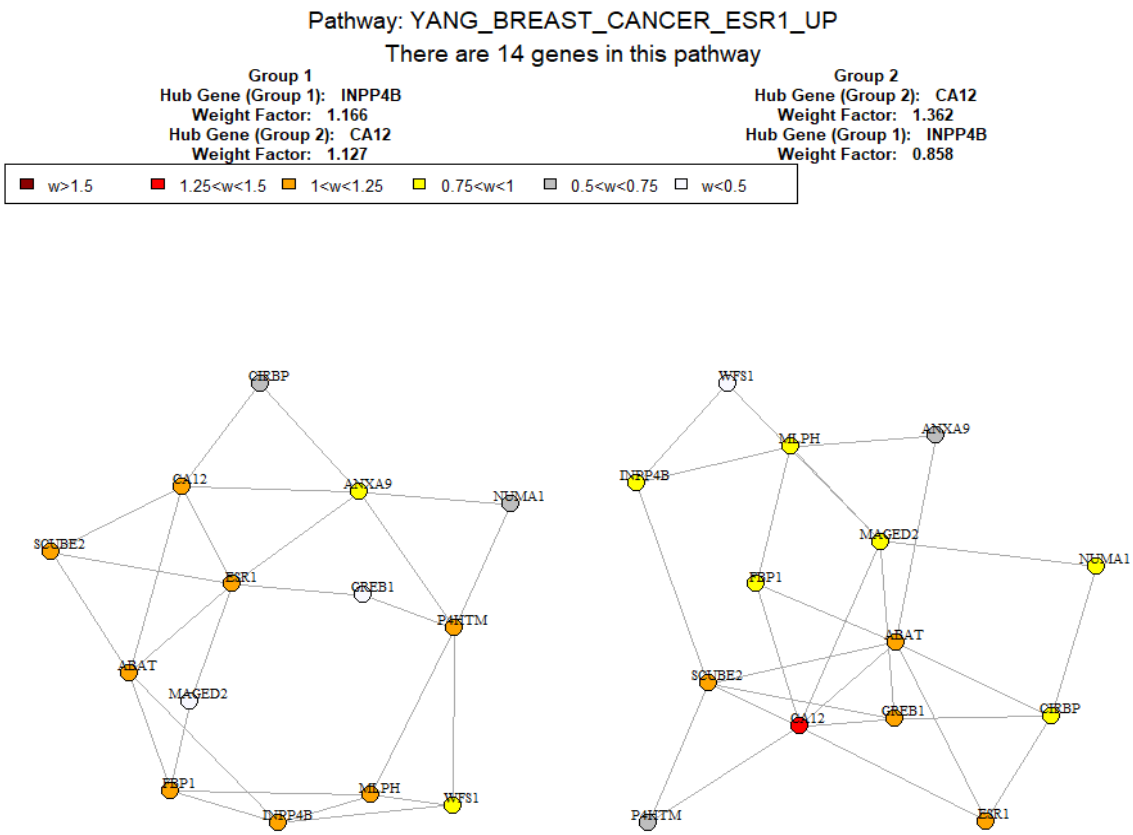

Figure 2. Different co-expression network configuration of Yang Breast Cancer ESR1 Up in Basal subtype vs other subtypes.

Below the differences in 'Yang breast cancer ESR1 upregulated' gene set representation in two types of breast cancer are explained in detail. For basal subtype, the hub protein of this pathway is INPP4B, which encodes for inositol polyphosphate 4-phosphate type II. This enzyme is involved in phosphatidylinositol signaling pathway. It plays an important role in the late stages of macropinocytosis by dephosphorylating phosphatidylinositol 3,4-bisphosphate in membrane ruffles. For the other breast carcinoma subtypes, the hub protein is CA12, which participates in a variety of biological

processes, such as respiration, calcification, acid-base balance, bone resorption, and the formation of aqueous humor, cerebrospinal fluid, saliva, and gastric acid. The carbonic anhydrase XII (CA12) gene encodes a zinc metalloenzyme, which is responsible for acidification of the microenvironment of cancer cells, commonly taking place in estrogen receptor alpha positive (ER alpha +ve) breast tumors [30].

INPP4B, which is a hub gene of group-1 breast cancers, interacts with *MLPH*, *WFS1, FBP1* and *ABAT* genes, among which both *MLPH* and *FBP1* are upregulated. *INPP4B* acts as a tumor suppressor by negatively regulating normal and malignant mammary epithelial cell proliferation through regulation of the PI3K/Akt signaling pathway. The loss of INPP4B protein is a marker of aggressive basal-like breast cancer [31]. It interacts with *MLPH* gene which is upregulated in Group 1 breast cancers (basal subtype), and encodes a protein called melanophilin. *FBP1* gene, which is also upregulated only in group 1 breast cancers, acts as a rate-limiting enzyme in gluconeogenesis by catalyzing the hydrolysis of fructose 1,6-bisphosphate to fructose 6-phosphate in the presence of divalent cations. In breast cancer, overexpression of FBP1 protein may repress tumor growth, migration and glycolysis by targeting P4HTM [32].

Overall, a set of proteome-derived pathways, that I have detected in basal breast cancer subtype samples was somewhat aligned with the previously known transcriptional characteristics of basal subtype of breast cancer, with an addition of overexpression of *P4HTM* gene, which is not common in the breast cancer as the involvement of the *P4HTM* gene in the breast cancer is currently poorly explained.

1.2 Pathways differentially expressed between HER2 subtype and other breast cancer subtypes

Human epidermal growth factor receptor 2 (HER2) shows high levels of protein expression in HER2-positive cancers, which represent an aggressive type of breast tumors. Previous transcriptomic analysis showed the major features of HER2-positive subtype of the breast cancer are ubiquitin mediated proteolysis, RHO-family GTPase signaling, M-phase signaling, integrin signaling and TGF-beta signaling [33], with activation of a number of signaling pathways such as mitogen-activated protein kinase (MAPK), phosphoinositide 3-kinase (PI3K/Akt), phospholipase C γ, protein kinase C (PKC), signal transducer and activator of transcription [43]. The seven differentially co-expressed pathways between HER2-positive breast cancer and other three subtypes were 'Framer breast cancer cluster 2','Lien breast carcinoma metaplastic vs ductal downregulated', 'Pujana breast cancer with BRCA1 mutated upregulated', 'Smid breast cancer luminal B upregulated', 'Smid breast cancer relapse in brain downregulated', 'Sortiriou breast cancer grade 1 vs 3 upregulated', and 'Vantveer breast cancer ESR1 upregulated' (**Table 3**). All of these pathways were discovered among the MSigDb C2 curated gene sets with Benjamini-Hochberg adjusted p-value <0.01. As we can observe from the (**Table 4)**, all the hub genes for both comparison groups were upregulated.

Table 3. GSNCA highlighted pathways differentially co-expressed between HER2 subtype and other breast cancer subtypes and their GSNCA P-values

Pathway Names	GSNCA P-values
FARMER_BREAST_CANCER_CLUSTER_2	0.001
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN	0.003
PUJANA_BREAST_CANCER_WITH_BRCA1_MUTATED_UP	0.001
SMID_BREAST_CANCER_LUMINAL_B_UP	0.001
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_DN	0.003
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	0.002
VANTVEER_BREAST_CANCER_ESR1_UP	0.004

Table 4. Hub genes differ between HER2 subtype and other breast cancer subtypes

	Hub Genes			
Pathway Names	**Group-1 (Her2 Subtype)**	**Up or down**	**Group-2 (Other three subtypes)**	**Up or down**
FARMER_BREAST_CANCER_CLUSTER_2	**CENPE**	**Up**	**KIF11**	**Up**
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN	**SPINT2**	**Up**	**FOXA1**	**Up**
PUJANA_BREAST_CANCER_WITH_BRCA1_MUTATED_UP	**DNMT1**	**Up**	**NCAPD2**	**Up**
SMID_BREAST_CANCER_LUMINAL_B_UP	**ERBB4**	**Up**	**FOXA1**	**Up**
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_DN	**KIAA1324**	**Up**	**FOXA1**	**Up**
SOTIRIOU_BREAST_CANCER_GRADE_1_VS_3_UP	**CEP55**	**Up**	**NCAPH**	**Up**
VANTVEER_BREAST_CANCER_ESR1_UP	**SCAMP1**	**Up**	**TBC1D9**	**Up**

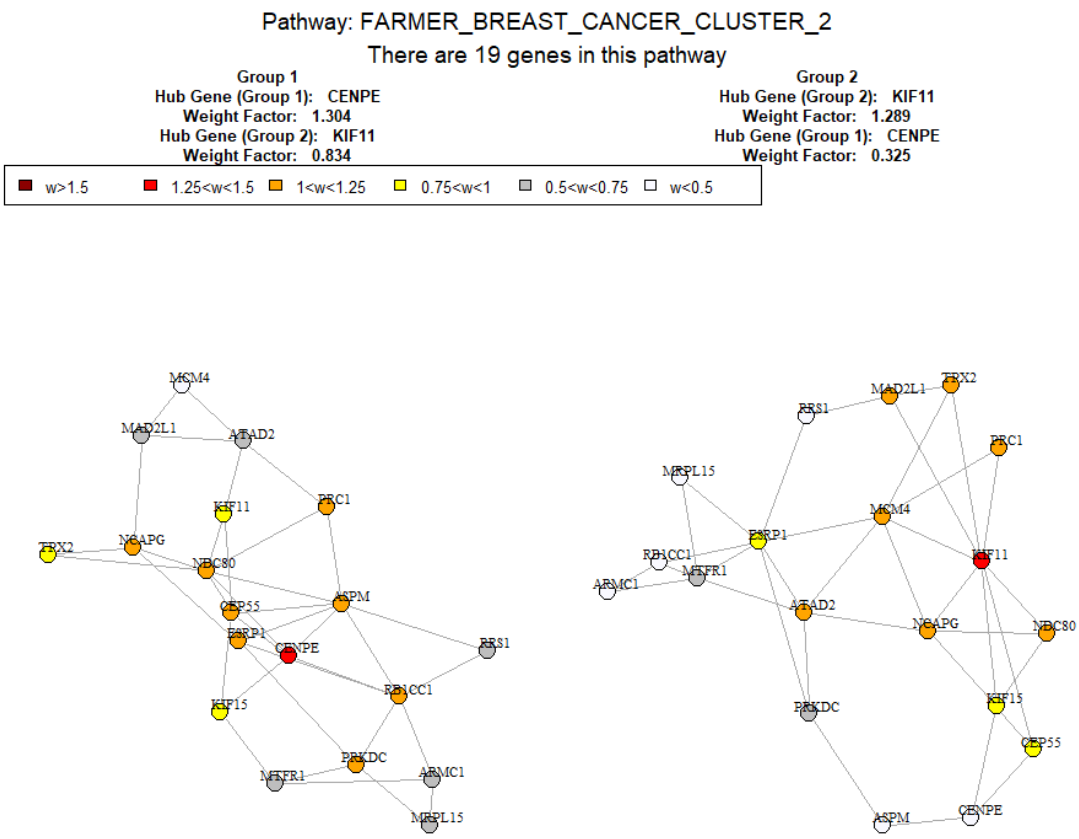

Figure 3. Different co-expression network configuration of Farmer Breast Cancer Cluster 2 in HER2 subtype vs other subtypes

Below the differences in the 'Farmer Breast Cancer Cluster 2' pathways revealed by the group comparison depicted above are explained in detail. For HER2 subtype of the breast tumors, the hub protein is CENPE. This protein plays a crucial role in the cell-cycle promotion from metaphase to anaphase [34]. In breast cancer, *CENPE* upregulation is strongly correlated with disease-specific survival [35]. The progression of breast cancer cells to G2-M phase is assumed to be heavily dependent on CENPE. Notably, this gene also plays similar function in the prostate cancer cells. For the other breast carcinoma

subtypes, the hub gene of the same pathway is *KIF11*, which encodes a member of kinesin-like protein family involved in several spindle dynamics. The major function of this gene is in the processes of centrosome separation, chromosome positioning and in establishing a polar spindle during cell mitosis [30]. The overexpression of *KIF11* is common in the advanced stage of the malignancy [36].

The *CENPE* hub gene interacts with *ESRP1, CEP55, NDC80, KIF15, ASPM* and *RB1CC1* genes. In breast cancer, overexpression of the *NDC80* gene is associated with poor clinical outcomes. NDC80 plays a crucial role in tumorigenesis, as it is a potential mitotic target for breast cancer [37]. However, only *CEP55, ASPM, ESRP1, CENPE, RB1CC1* and *PRKDC* genes are upregulated in the group 1 tumors, which is HER2 subtype phenotype. The overexpression of PRKDC gene plays a critical role in regulating cell cycle and chromosomal segregation which promotes tumorigenesis and results in poor survival of HER2 subtype breast cancer [38]. RB1CC1 gene which is upregulated inhibits PTK2/FAK1 and PTK2B/PYK2 kinase activity, affecting their downstream signaling pathways [39,40]. The protein encoded by RB1CC1 interacts with signaling pathways to regulate cell growth, cell proliferation, apoptosis, autophagy, and cell migration. ASPM plays a crucial role in controlling the function of mitotic spindle [41]. ASPM overexpression similar to the CENPE is involved in the regulation of G2/M cell cycle progression [42]. So, I am assuming ASPM and CENPE could play crucial role in the progression of HER2 subtype breast carcinoma.

Overall, a set of proteome-derived pathways deregulated in HER2 breast cancer subtype was aligned with known transcriptomic characteristics of the HER2 breast cancer

subtype phenotype such as upregulated genes involved in G2/M cell cycle progression, inhibition of PTK2B/PYK2 kinase activity [39,40].

1.3 Pathways differentially expressed between Luminal A subtype and other breast cancer subtypes

Luminal A breast cancer is hormone receptor positive, HER2 negative and exhibit low levels of protein Ki-67 which assists in how rapid the cancer cells multiply. They represent 50% - 60% of all the breast cancers. The previous transcriptomic analysis of the luminal A breast cancer subtype show expression of luminal epithelial cytokeratins (CK) 8 and 18 and show expression of the genes encoding the estrogen-receptor and related proteins such as LIV1, FOXA1, XBP1, GATA3, BCL2, erbB3 and erbB4 [44]. The three differentially co-expressed pathways of the Luminal A subtype and other three breast cancer subtypes are 'Poola invasive breast cancer', 'Smid breast cancer relapse in brain downregulated', 'Vantveer breast cancer ESR1 upregulated' **(Table 5)** discovered from the MSigDb C2 curated gene sets with Benjamini-Hochberg adjusted p-value <0.01. As we can observe from the **(Table 6),** all the hub genes for both the groups are upregulated.

Table 5. GSNCA highlighted pathways differentially co-expressed between Luminal A subtype and other breast cancer subtypes and their GSNCA P-values

Pathway Names	GSNCA P-values
VANTVEER_BREAST_CANCER_ESR1_UP	0.004
POOLA_INVASIVE_BREAST_CANCER_DN	0.005
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_DN	0.008

Table 6. Hub genes differ between Luminal A subtype and other breast cancer subtypes

	Hub Genes			
Pathway Names	**Group-1 (Luminal-A)**	**Up or down**	**Group-2 (Other three subtypes)**	**Up or down**
POOLA_INVASIVE_BREAST_CANCER_DN	**TF**	**Up**	**RABEP1**	**Up**
SMID_BREAST_CANCER_RELAPSE_IN_BRAIN_DN	**GATA3**	**Up**	**TBC1D9**	**Up**
VANTVEER_BREAST_CANCER_ESR1_UP	**FOXA1**	**Up**	**TBC1D9**	**Up**

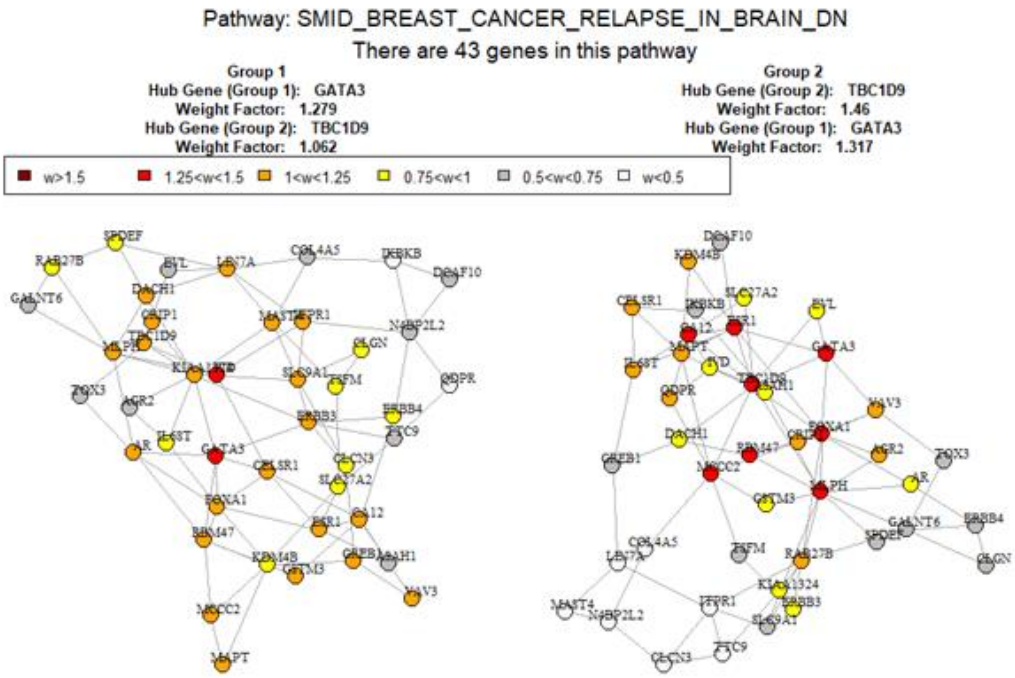

Figure 4. Different co-expression network configuration of Smid breast cancer relapse in brain downregulated in Luminal A subtype vs other subtypes

Below the differences in 'Smid breast cancer relapse in brain downregulated' pathways will be discussed in detail. For luminal A subtype breast cancer group 1 the hub gene is GATA3 which is an important expression gene from the transcriptional analysis. GATA3 is a transcriptional activator which binds with the enhancer of the T-cell receptor alpha and delta genes [45]. In luminal subtype breast cancer, the GATA3 upregulates protooncogenes with increased expression of the ER target genes and may promote tumorigenesis [46]. While for the other subtypes pathway the hub gene is TBC1D9 which is also known as multidrug resistance gene (MDR1). 41.2% of the breast cancer tumors

express TBC1D9 [47]. TBC1D9 gene expression plays a significant role in supporting resistance to adjuvant chemotherapy in women with breast cancer.

The *GATA3* is a hub gene of group 1 co-expression pathway (**Fig 4**). It interacts with *FOXA1, RBM47, AR, CELSR1, ERBB3* & *KIAA1324* genes which are upregulated genes in Luminal A subtype of the breast cancer. If a breast cancer tumor exhibits the expression of these four transcription factors (GATA3, ER-alpha, FOXA1 and XBP1), it is actually classified as Luminal A subtype as its very definition. GATA3 acts as an important marker of the all luminal breast cancers [48]. In the Luminal A subtype breast cancer, a combination of FOXA1 and GATA3 transcription factors controls the gene expression pattern. ESR1, which encodes an estrogen receptor 1, also exhibits overexpression in the Group 1 (Luminal A) cancer subtype. There is also a small network in the Group 1, where upregulated *LIN7A, DACH1, SLC9A1, MAST4* and *ITPR1* genes are interlinked. This small subnetwork is not prominent in the Group 2 gene network. Notably, this subnetwork was previously noted as associated with brain cancer relapse. This subnetwork includes *MAST4*, a mast cell biomarker implicated in the metastasis of breast cancer cells, and *DACH1*, a cell fate determination factor [49].

Overall, a set of proteome derived pathways deregulated in the Luminal A subtype breast cancer aligns well with the previously described transcriptional characteristics of the Luminal A subtype breast cancer.

1.4 Pathways differentially expressed between Luminal B subtype and other breast cancer subtypes

Luminal B subtype breast cancers are characterized by higher expression of Estrogen receptor, Progesterone receptors and Ki-67 protein. A total of 15-20% of the breast cancers belong to Luminal-B subtype; they have generally more aggressive phenotype [44]. The major difference between the Luminal A and B subtypes is that Luminal B subgroup has increased expression of proliferation-related genes such as avian myeloblastosis viral oncogene homolog (*MYB*), gamma glutamyl hydrolase (*GGH*), lysosome-associated transmembrane protein 4-beta (*LAPTMB4*), nuclease sensitive element binding protein 1 (*NSEP1*) and cyclin E1 (*CCNE1*) [44]. Previous transcriptomic analysis of the Luminal-B subtype breast tumors pointed at increased expression of growth receptor signaling genes [50]. The seven differentially co-expressed pathways of Luminal B subtype when compared with other three subtypes of breast cancer are 'Bertucci medullary vs Ductal breast cancer downregulated', 'Doane breast cancer classes upregulated', 'Lien breast carcinoma metaplastic vs Ductal downregulated', 'Smid breast cancer Luminal B upregulated', 'Smid breast cancer relapse in bone upregulated', 'Vantveer breast cancer poor prognosis' and 'Yang breast cancer ESR1 downregulated' **(Table 7).** All these pathways were discovered in the MSigDb C2 curated gene sets with Benjamini-Hochberg adjusted p-value <0.01. As we can observe from the **(Table 8)**, all the hub genes for both comparison groups were upregulated.

Table 7. GSNCA highlighted pathways differentially co-expressed between Luminal B subtype and other breast cancer subtypes and their GSNCA P-values

Pathway Names	GSNCA P-Values
BERTUCCI_MEDULLARY_VS_DUCTAL_BREAST_CANCER_DN	0.01
DOANE_BREAST_CANCER_CLASSES_UP	0.008
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN	0.01
SMID_BREAST_CANCER_LUMINAL_B_UP	0.001
SMID_BREAST_CANCER_RELAPSE_IN_BONE_UP	0.003
VANTVEER_BREAST_CANCER_POOR_PROGNOSIS	0.002
YANG_BREAST_CANCER_ESR1_DN	0.008

Table 8. Hub genes differ between Luminal B subtype and other breast cancer subtypes

	Hub Genes			
Pathway Names	**Group-1 (Luminal-B)**	**Up or down**	**Group-2 (other three subtypes)**	**Up or down**
BERTUCCI_MEDULLARY_VS_DUCTAL_BREAST_CANCER_DN	TPM2	Up	TPM2	Up
DOANE_BREAST_CANCER_CLASSES_UP	CDK12	Up	FOXA1	Up
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL_DN	TTC39A	Up	FOXA1	Up
SMID_BREAST_CANCER_LUMINAL_B_UP	ESR1	Up	GATA3	Up
SMID_BREAST_CANCER_RELAPSE_IN_BONE_UP	TOM1L1	Up	GATA3	Up
VANTVEER_BREAST_CANCER_POOR_PROGNOSIS	PRC1	Up	MCM6	Up
YANG_BREAST_CANCER_ESR1_DN	CDH3	Up	TRIM2	Up

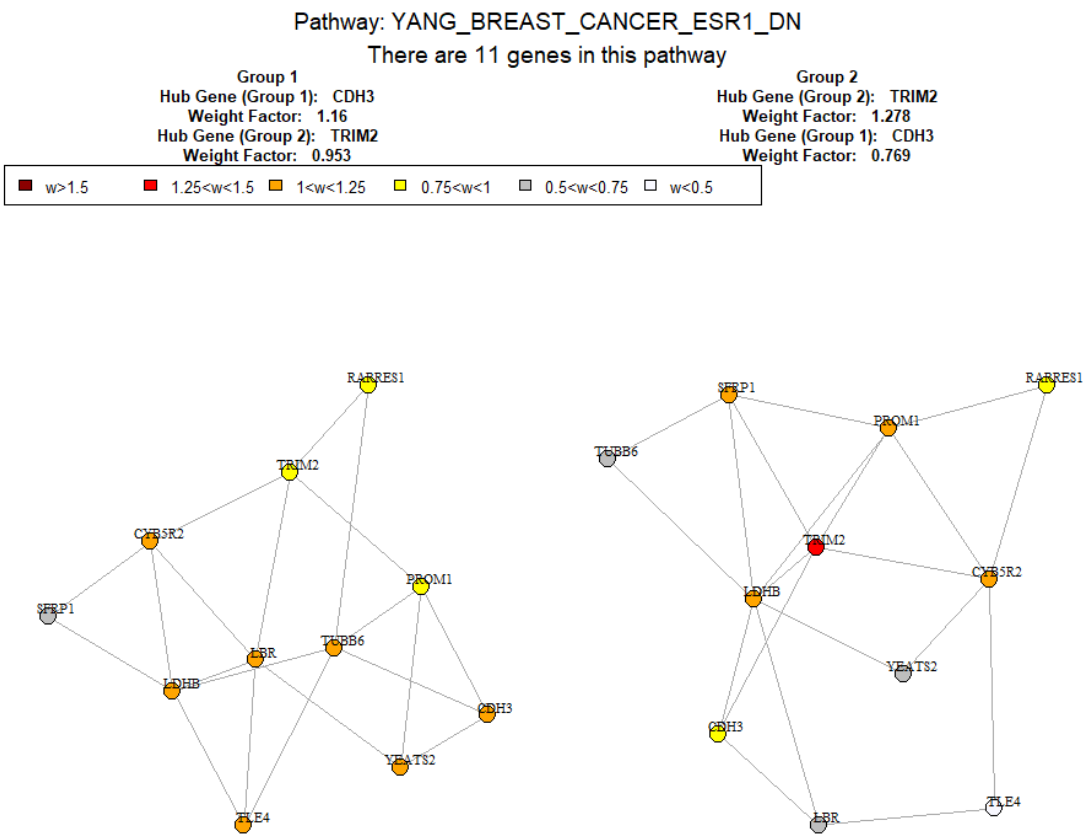

Figure 5. Different co-expression network configuration of Yang breast cancer ESR1 Dn in Luminal B subtype vs other subtypes

The changes in the 'Yang breast cancer ESR1 DN' will be discussed in detail below. The hub gene for the Group 1 of the Luminal A subtype breast cancer gene pathway is *CDH3*. Its product, cadherin 3 (CDH3) plays an important role in cell adhesion, sorting and cell recognition-related signaling. Particularly, in breast cancers the CDH3 gene promotes the cell division and tumor aggressiveness; both factors contribute to a poor prognosis [51]. On the other hand, the hub gene for the co-expression pathway of other three subtypes (group 2) was *TRIM2*. This gene plays an important neuroprotective role, and functions as an E3-ubiquitin ligase in proteasome-mediated

degradation pathways [50]. *TRIM2* product is a direct regulator of Bim degradation in TAM-resistant breast cancer cells [52].

The Cadherin 3 (CDH3) hub gene of group 1 expression pathway (fig 5) interacts with *YEATS2, TUBB6* and *PROM1* genes. There were several genes upregulated in the Group 1 of the co-expression gene network which are not upregulated in the gene network in Group 2 (**Fig 5**). These genes were *LBR, TUBB6, TLE4, CDH3* and *YEATS2*, which implies that the phenotype of Group 1 is Luminal B subtype. *TUBB6* (Tubulin Beta 6 Class V) protein confers to the breast tumors an increased sensitive to taxane-based chemotherapy. Studies have observed a downregulation of the *TUBB6* in the breast cancers [53]. Nevertheless, when we compare the group 1 (Luminal B) network to the networks of all other breast cancers, relative upregulation of the *TUBB6* gene was observed. *YEATS2* product is the histone acetyltransferase involved in the non-small cell lung carcinoma [54]. The *TLE4* gene may behave as either a tumor suppressor gene or as a facilitator of oncogenesis in invasive breast cancer. *TLE4* gene promotes cell proliferation and invasion via activation of a JNK-c-Jun pathway and leads to increased expression in cyclin D1 and decreased expression in P27Kip1 [55]. Upregulation of the laminin (*LBR*) in the breast cancer declines with tumor grade, and these decreases reflect increased probability of dying from cancer [56].

Overall, a set of proteome derived pathways deregulated in the Luminal B subtype breast cancer was not exactly aligned with the already discovered transcriptional characteristics of the Luminal B subtype breast cancer.

DISCUSSION

There are several computational methods developed exclusively for the analysis of proteomics data for normalizing and preprocessing, the identification and quantification of protein complexes [57], protein-protein interaction network analysis and visualization [58]. Oftentimes, these tools are built upon special dedicated software platforms, where each technique is supported by a group of statistical tools compatible with high dimensional proteomic data analysis [59]. A number of competitive GSA tests were specifically developed for proteomics data [8, 9]. Surprisingly, no self-contained GSA approaches were specially developed for proteomics data analysis, with no significant efforts put to repurposing of existing self-contained GSA tests previously developed for transcriptomics, and to their application to the proteomics data. As self-contained GSA approaches have more power than competitive GSA [2], it is likely that proteomics data analysis maybe aided by adoption of self-contained GSA tests previously developed for transcriptomics [2].

To investigate the possibility of the alignment of the previously known transcriptomic characteristics of breast cancer subtype with the proteomics data I conducted a self-contained GSA analysis, specifically, the gene sets network correlation analysis (GSNCA) of the breast cancer subtype proteome profiles. In total, I have analyzed 77 breast cancer patient samples and 3 biological replicates belonging to four

molecular subtypes. These samples were already analyzed by high-resolution mass spectrometry (MS/MS), thus providing an opportunity for cross-omics comparison.

Typically, experimental proteomics and phosphoproteomic data are plagued by large amounts of missing values, partly because of particular protein expression in a particular sample not recorded due to sample complexity and variation in sampling from one specimen to another. To overcome the problem of missing value, I have removed the proteins that were quantified sparingly. Then, I have found that the rest of the missing data values are at random (MAR) in the dataset, and implemented the MICE method of data imputation. Since the missing values of the proteome data mostly correspond to the proteins with low levels of expression, these missing values are replaceable with the calculated number. In a nutshell, the probability that a value is missing depends only on observed values and is predicted using them. The missing datapoints are imputed on a variable basis by specific imputation model per variable.

In order to understand the biological mechanism and the associated network, we need to have a thorough understanding of how proteins and mRNAs respond to external stimulations and how gene commands are neglected, leading to a decrease in a correlation between transcriptomic and proteomic details [60]. Exact congruency of the transcriptome and proteome-based portraits of the breast cancer subtypes may not be anticipated due to several reasons. Due to a combination of the molecular properties of proteins and the technical difficulties in assaying, proteomics data differ from transcriptomic data. For example, protein half-lives are defined by post transcription

machinery [60], not by the state of its own gene. At post-transcriptomic level, both translation and protein degradation contribute to the steady state abundance of each proteins [61]. In addition, despite the power of predominant MS-based technologies, some parts of the proteome remain secret due to technical limitations of the extraction and solubilization techniques [62]. The correlations between protein and mRNA abundances in both bacteria and eukaryotes are estimated at approximately a square Pearson association coefficient of ~0.40, i.e. only 40 percent of the variance in protein abundance can be explained by the abundance of respective mRNAs [61].

Given all these difficulties, it is surprising how well the self-contained GSA test was able to align many characteristic features found at mRNA level with the protein level. It should be noted that the competitive GSA assessments at the specified level of significance may not have sufficient power to identify differentially expressed pathways, while self-contained GSA test are capable of this feat.

In conclusion, our work points that the use of self-contained GSA methods makes it possible to integrate observations derived from molecular profiles of breast cancer tumor subtypes independently based on transcriptomics and proteomics. In addition, applying self-contained GSA methods at the protein level allow one to extract additional insights concerning molecular mechanisms and actionable targets, which are only accessible at the protein level. Complementing proteomics data analysis with self-contained GSA tests, in addition to competitive tests, explicitly designed for proteomics data, will find it use in the future.

APPENDIX I

The R script that I used in this study is mentioned here.

```
library(GSAR)

library(GSVAdata)

proteomes <- read.csv("C:/Users/koush/OneDrive/Desktop/Thesis/other
datasets/basal_proteomes.csv")

data("c2BroadSets")

library(org.Hs.eg.db)

library(GSEABase)

#proteomes2 <- proteomes[c(1:len), c(1,2:20,34:56)] (Basal & Luminal-A)

rownames(proteomes) <- proteomes[,1]

proteomes <- proteomes[,-1]

View(proteomes)

dim(proteomes)

proteomes <- data.matrix(proteomes)

C2 <- as.list(geneIds(c2BroadSets))

len <- length(C2)

genes.entrez <- unique(unlist(C2))

genes.symbol <- array("",c(length(genes.entrez),1))

x <- org.Hs.egSYMBOL

mapped_genes <- mappedkeys(x)
```

```
xx <- as.list(x[mapped_genes])
for (ind in 1:length(genes.entrez)){
  if (length(xx[[genes.entrez[ind]]])!=0)
    genes.symbol[ind] <- xx[[genes.entrez[ind]]]
}
## discard genes with no mapping to gene symbol identifiers
genes.no.mapping <- which(genes.symbol == "")
if(length(genes.no.mapping) > 0){
  genes.entrez <- genes.entrez[-genes.no.mapping]
  genes.symbol <- genes.symbol[-genes.no.mapping]
}
names(genes.symbol) <- genes.entrez
##discard genes in C2 pathways which do not exist in proteomes dataset
overlap <- rownames(proteomes)
remained <- array(0,c(1,len))
for (k in seq(1, len, by=1)) {
  remained[k] <- sum((genes.symbol[C2[[k]]] %in% overlap) &
                     (C2[[k]] %in% genes.entrez))
}
## discard C2 pathways which have less than 10 or more than 500 genes
C2 <- C2[(remained>=10)&&(remained<=500)]
pathway.names <- names(C2)
c2.pathways <- list()
for (k in seq(1, length(C2), by=1)){
```

```
  selected.genes <- which(overlap %in% genes.symbol[C2[[k]]])

  c2.pathways[[length(c2.pathways)+1]] <- overlap[selected.genes]

}

names(c2.pathways) <- pathway.names

path.index <- which(names(c2.pathways) ==
"DOANE_BREAST_CANCER_CLASSES_DN")

target.pathway <-
proteomes[c2.pathways[["DOANE_BREAST_CANCER_CLASSES_DN"]],]

target.pathway <- data.matrix(target.pathway)

View(target.pathway)

group.label <- c(rep(1,19), rep(2,61))

#dim(target.pathway)

WW_pvalue <- WWtest(target.pathway, group.label)

KS_pvalue <- KStest(target.pathway, group.label)

MD_pvalue <- MDtest(target.pathway, group.label)

RKS_pvalue <- RKStest(target.pathway, group.label)

RMD_pvalue <- RMDtest(target.pathway, group.label)

F_pvalue <- AggrFtest(target.pathway, group.label)

GSNCA_pvalue <- GSNCAtest(target.pathway, group.label)

WW_pvalue

KS_pvalue

MD_pvalue

RKS_pvalue

RMD_pvalue

F_pvalue
```

```
GSNCA_pvalue

plotMST2.pathway(object=proteomes[c2.pathways[[path.index]],],

        group=c(rep(1,19), rep(2,61)),
name="DOANE_BREAST_CANCER_CLASSES_DN",

        legend.size=1.2, leg.x=-1.2, leg.y=2,

        label.size=1, label.dist=0.8, cor.method="pearson")

results <- TestGeneSets(object=proteomes, group=group.label,

            geneSets=c2.pathways[1:3272], min.size=10, max.size=100,
test="GSNCAtest")

results

#write.csv(results, file = "basal_pathways.csv")
```

REFERENCES

1. Gulcicek, E. E. *et al.* Proteomics and the Analysis of Proteomic Data: An Overview of Current Protein-Profiling Technologies. *Curr Protoc Bioinformatics* **0 13**, (2005).

2. Glazko, G., Zybailov, B., Emmert-Streib, F., Baranova, A. & Rahmatallah, Y. Proteome-transcriptome alignment of molecular portraits achieved by self-contained gene set analysis: Consensus colon cancer subtypes case study. *PLOS ONE* **14**, e0221444 (2019).

3. Rahmatallah, Y., Emmert-Streib, F. & Glazko, G. Gene set analysis approaches for RNA-seq data: performance evaluation and application guideline. *Brief Bioinform* **17**, 393–407 (2016).

4. Set-Based Test Procedures for the Functional Analysis of Protein Lists from Differential Analysis. - Abstract - Europe PMC. Available at: https://europepmc.org/article/med/26519175. (Accessed: 10th April 2020)

5. Ahrens, M. *et al.* Detection of patient subgroups with differential expression in omics data: a comprehensive comparison of univariate measures. *PLoS ONE* **8**, e79380 (2013).

6. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).

7. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987 (2007).

8. Cha, S. *et al.* In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology. *Mol. Cell Proteomics* **9**, 2529–2544 (2010).

9. Lavallée-Adam, M., Rauniyar, N., McClatchy, D. B. & Yates, J. R. PSEA-Quant: a protein set enrichment analysis on label-free and label-based protein quantification data. *J. Proteome Res.* **13**, 5496–5509 (2014).

10. Rahmatallah, Y., Emmert-Streib, F. & Glazko, G. Comparative evaluation of gene set analysis approaches for RNA-Seq data. *BMC Bioinformatics* **15**, 397 (2014).

11. DeSantis, C., Ma, J., Bryan, L. & Jemal, A. Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians* **64**, 52–62 (2014).

12. Weiss, A. *et al.* Validation Study of the American Joint Committee on Cancer Eighth Edition Prognostic Stage Compared With the Anatomic Stage in Breast Cancer. *JAMA Oncol* **4**, 203–209 (2018).

13. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).

14. Wang, X., Markowetz, F., De Sousa E Melo, F., Medema, J. P. & Vermeulen, L. Dissecting cancer heterogeneity--an unsupervised classification approach. *Int. J. Biochem. Cell Biol.* **45**, 2574–2579 (2013).

15. Zhao, L., Lee, V. H. F., Ng, M. K., Yan, H. & Bijlsma, M. F. Molecular subtyping of cancer: current status and moving toward clinical applications. *Brief. Bioinformatics* **20**, 572–584 (2019).

16. Finnegan, T. J. & Carey, L. A. Gene-expression analysis and the basal-like breast cancer subtype. *Future Oncology* **3**, 55–63 (2007).

17. Lesurf, R. *et al.* Genomic characterization of HER2-positive breast cancer and response to neoadjuvant trastuzumab and chemotherapy—results from the ACOSOG Z1041 (Alliance) trial. *Ann Oncol* **28**, 1070–1077 (2017).

18. Roy, V. & Perez, E. A. Beyond Trastuzumab: Small Molecule Tyrosine Kinase Inhibitors in HER-2–Positive Breast Cancer. *The Oncologist* **14**, 1061–1069 (2009).

19. Prat, A. *et al.* Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. *J. Clin. Oncol.* **31**, 203–209 (2013).

20. Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55–62 (2016).

21. Proteomics Data Analysis (2/3): Data Filtering and Missing Value Imputation. *DataScience+* Available at: https://datascienceplus.com/proteomics-data-analysis-2-3-data-filtering-and-missing-value-imputation/. (Accessed: 6th December 2019)

22. Berg, P., McConnell, E. W., Hicks, L. M., Popescu, S. C. & Popescu, G. V. Evaluation of linear models and missing value imputation for the analysis of peptide-centric proteomics. *BMC Bioinformatics* **20**, 102 (2019).

23. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* **20**, 40–49 (2011).

24. Rahmatallah, Y., Zybailov, B., Emmert-Streib, F. & Glazko, G. GSAR: Bioconductor package for Gene Set analysis in R. *BMC Bioinformatics* **18**, 61 (2017).

25. Wu, D. *et al.* ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* **26**, 2176–2182 (2010).

26. Rahmatallah, Y., Emmert-Streib, F. & Glazko, G. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics* **30**, 360–368 (2014).

27. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267–273 (2003).

28. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

29. Bertucci, F., Finetti, P. & Birnbaum, D. Basal Breast Cancer: A Complex and Deadly Molecular Subtype. *Curr Mol Med* **12**, 96–110 (2012).

30. Barnett, D. H. *et al.* Estrogen receptor regulation of carbonic anhydrase XII through a distal enhancer in breast cancer. *Cancer Res.* **68**, 3505–3515 (2008).

31. Fedele, C. G. *et al.* Inositol polyphosphate 4-phosphatase II regulates PI3K/Akt signaling and is lost in human basal-like breast cancers. *PNAS* **107**, 22231–22236 (2010).

32. Shi, L., He, C., Li, Z., Wang, Z. & Zhang, Q. FBP1 modulates cell metabolism of breast cancer cells by inhibiting the expression of HIF-1α. *Neoplasma* **64**, 535–542 (2017).

33. Kalari, K. R. *et al.* An Integrated Model of the Transcriptome of HER2-Positive Breast Cancer. *PLoS One* **8**, (2013).

34. Guo, Y., Kim, C., Ahmad, S., Zhang, J. & Mao, Y. CENP-E–dependent BubR1 autophosphorylation enhances chromosome alignment and the mitotic checkpoint. *J Cell Biol* **198**, 205–217 (2012).

35. Chemogenetic Evaluation of the Mitotic Kinesin CENP-E Reveals a Critical Role in Triple-Negative Breast Cancer | Molecular Cancer Therapeutics. Available at: https://mct.aacrjournals.org/content/13/8/2104. (Accessed: 10th April 2020)

36. Zhou, J. *et al.* KIF11 Functions as an Oncogene and Is Associated with Poor Outcomes from Breast Cancer. *Cancer Res Treat* **51**, 1207–1221 (2019).

37. Meng, Q.-C. *et al.* Overexpression of NDC80 is correlated with prognosis of pancreatic cancer and regulates cell proliferation. *Am J Cancer Res* **5**, 1730–1740 (2015).

38. Zhang, Y. *et al.* High expression of PRKDC promotes breast cancer cell growth via p38 MAPK signaling and is associated with poor survival. *Mol Genet Genomic Med* **7**, (2019).

39. Ueda, H., Abbi, S., Zheng, C. & Guan, J. L. Suppression of Pyk2 kinase and cellular activities by FIP200. *J. Cell Biol.* **149**, 423–430 (2000).

40. Abbi, S. *et al.* Regulation of focal adhesion kinase by a novel protein inhibitor FIP200. *Mol. Biol. Cell* **13**, 3178–3191 (2002).

41. Alsiary, R. *et al.* Deregulation of Microcephalin and ASPM Expression Are Correlated with Epithelial Ovarian Cancer Progression. *PLoS One* **9**, (2014).

42. Shubbar, E. *et al.* Elevated cyclin B2 expression in invasive breast carcinoma is associated with unfavorable clinical outcome. *BMC Cancer* **13**, 1 (2013).

43. Roy, V. & Perez, E. A. Beyond Trastuzumab: Small Molecule Tyrosine Kinase Inhibitors in HER-2–Positive Breast Cancer. *The Oncologist* **14**, 1061–1069 (2009).

44. Yersal, O. & Barutca, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol* **5**, 412–424 (2014).

45. Sasaki, T. *et al.* Genome-Wide Gene Expression Profiling Revealed a Critical Role for GATA3 in the Maintenance of the Th2 Cell Identity. *PLoS ONE* **8**, e66468 (2013).

46. Cohen, H. *et al.* Shift in GATA3 functions, and GATA3 mutations, control progression and clinical presentation in breast cancer. *Breast Cancer Research* **16**, 464 (2014).

47. Trock, B. J., Leonessa, F. & Clarke, R. Multidrug resistance in breast cancer: a meta-analysis of MDR1/gp170 expression and its possible functional significance. *J. Natl. Cancer Inst.* **89**, 917–931 (1997).

48. Takaku, M., Grimm, S. A. & Wade, P. A. GATA3 in breast cancer: tumor suppressor or oncogene? *Gene Expr* **16**, 163–168 (2015).

49. Wu, K. *et al.* DACH1 Is a Cell Fate Determination Factor That Inhibits Cyclin D1 and Breast Tumor Growth. *Mol Cell Biol* **26**, 7116–7129 (2006).

50. Reis-Filho, J. S., Weigelt, B., Fumagalli, D. & Sotiriou, C. Molecular profiling: moving away from tumor philately. *Sci Transl Med* **2**, 47ps43 (2010).

51. Paredes, J. *et al.* P-cadherin expression in breast cancer: a review. *Breast Cancer Res* **9**, 214 (2007).

52. Yin, H. *et al.* GPER promotes tamoxifen-resistance in ER+ breast cancer cells by reduced Bim proteins through MAPK/Erk-TRIM2 signaling axis. *International Journal of Oncology* **51**, 1191–1198 (2017).

53. Nami, B. & Wang, Z. Genetics and Expression Profile of the Tubulin Gene Superfamily in Breast Cancer Subtypes and Its Relation to Taxane Resistance. *Cancers (Basel)* **10**, (2018).

54. Mi, W. *et al.* YEATS2 links histone acetylation to tumorigenesis of non-small cell lung cancer. *Nat Commun* **8**, 1088 (2017).

55. Yuan, D., Yang, X., Yuan, Z., Zhao, Y. & Guo, J. TLE1 function and therapeutic potential in cancer. *Oncotarget* **8**, 15971–15976 (2016).

56. Sakthivel, K. M. & Sehgal, P. A Novel Role of Lamins from Genetic Disease to Cancer Biomarkers. *Oncol Rev* **10**, (2016).

57. Heusel, M. *et al.* Complex-centric proteome profiling by SEC-SWATH-MS. *Molecular Systems Biology* **15**, e8438 (2019).

58. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

59. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**, 731–740 (2016).

60. Haider, S. & Pal, R. Integrated Analysis of Transcriptomic and Proteomic Data. *Curr Genomics* **14**, 91–110 (2013).

61. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* **13**, 227–232 (2012).

62. Laukens, K., Naulaerts, S. & Berghe, W. V. Bioinformatics approaches for the functional interpretation of protein lists: From ontology term enrichment to network analysis. *PROTEOMICS* **15**, 981–996 (2015).

BIOGRAPHY

Koushik Ayaluri is a master's student at George Mason University in Bioinformatics and Computational Biology department. He received his Bachelor of Technology in Bioinformatics from Amity Institute of Biotechnology, Amity University, Noida, India in 2017. His research interests are in cancer genomics, Next Generation sequencing and epigenetics.