Machine learning-driven integration of multimodal data for deciphering breast cancer heterogeneity

by

Qian Liu

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

in the

Individual Interdisciplinary Studies Program

Departments of Biochemistry and Medical Genetics, Computer Science, Statistics
University of Manitoba
Winnipeg

Table of Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

BC	Breast cancer
|logFC|	Absolute value of the log2 transformed folder change
AI	Artificial intelligence
ANOVA	Analysis of Variance
AUC	Area under of the ROC curve
BP	Biological Process
BTF	Bayesian tensor factorization
CC	Cell Component
CDF	Cumulative distribution function
CIT	Cartes d'Identit des Tumeurs
CliXO	Clique Extracted Ontology
CNV	Copy number variation
CoC	Consensus clustering
CP	Canonical decomposition / parallel factors
DCE-MRI	Dynamic contrast enhanced MRI
DepMap	The Cancer Dependency Map
DepOI	dependency of interest
DLR	Deep learning-based radiomics
DSC	Dice similarity coefficient
DTI	Drug-target interaction
eQTL	Expression quantitative trait loci
ER	Estrogen receptor
ESM	Evolutionary Scale Modeling
Fb	BCubed F-score
FDR	False discovery rate
Fp	Pairwise F-score
GENIUS	Gene progNostic Index Using Subtypes
GEO	Gene Expression Omnibus
GGI	Gene expression Grade Index
GNN	Graph neural network
GO	Gene Ontology
GSEA	Gene set enrichment analysis
HER2	Human epidermal growth factor receptor2
HR	Hormone receptor
IHC	Immunohistochemistry
KM	Kaplan-Meier
METABRIC	Taxonomy of Breast Cancer International Consortium
MF	Molecular Function

MPGNN	Message passing GNN
MRI	Magnetic Resonance Imaging
NHS	National Health Service
NMF	Non-negative matrix factorization
NMI	Normalized Mutual Information
NP	non-deterministic polynomial-time
OR	Odds ratio
PCA	Principal component analysis
PCR	Pathological complete response
PPI	Protein-protein interaction
PPR	Pathological partial response
PR	Progesterone receptor
RF	Random forest
RMSE	Root Mean Square Error
ROC	Receiver operating characteristic
ROI	Region of interest
SAR	Semi-auto radiomics
SCM	Subtype Classification Model
SHAP	Shapley Additive Explanation
SL	Synthetic lethality
SNF	Similarity Network Fusion
SSP	Single sample predictor strategy
SVM	Support vector machine
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive
TIL	Tumor-infiltrating lymphocyte
TIMER	Tumor immune estimation resource
TNBC	Triple-negative breast cancer
TNM	Tumor, Node, Metastasis
XGBoost	eXtreme Gradient Boosting

Abstract

Breast cancer (BC) is a complex disease with a high degree of heterogeneity. The heterogeneity of BC could be detected at different biological levels using a variety of modern molecular biological techniques. These biotechniques could generate high-throughput and quantitative measurements, such as gene expression, copy number variation (CNV), DNA methylation, proteomics measurements, and so on. Meanwhile, the tumor morphology information obtained from medical images is also worthy of consideration in evaluating the heterogeneity of BC. Many machine-learning algorithms have been developed to help us to explore the heterogeneity of cancer from the abovementioned high-dimensional measurements. However, there are several challenges for characterizing BC heterogeneity based on the multi-modal biodata using the existing computational data analysis techniques. The first challenge is how to effectively combine the multi-modal biodata and find comprehensive and interpretable representations from them. Another challenge is how to address the execution infeasibility caused by the unpaired data problem (the publicly available datasets have unmatched multi-omics, medical images, and clinical outcome data). Besides, the model interpretability and privacy issues should also be carefully taken into consideration in machine learning-based BC research. This thesis aims to explore the BC heterogeneity using thriving machine-learning algorithms at different data resolutions ranging from single genomics, multi-genomics, to proteomics and radiogenomics. We have four major objectives: 1)human epidermal growth factor receptor2 positive/estrogen receptor positive (HER2+/ER+) BC stratification and prognostic gene signature identification using single genomic data; 2) BC subtyping using multiple genomics data; 3) Graph neural network (GNN) for BC hierarchical biological system mapping using graph structured proteomics data; 4) BC prognostic radiogenomic biomarker identification. This thesis demonstrates the promising applications of

machine learning in deciphering BC heterogeneity at different biological levels. Moreover, the resulting 15-gene HER2+/ER+ BC gene expression signature, multi-omics-based BC subtypes, hierarchical biological systems/protein communities, and prognostic radiogenomic biomarkers have the potential to benefit clinical practice for BC.

Acknowledgement

I would first like to express my sincere gratitude to my supervisor Dr. Pingzhao Hu for his persistent support, encouragement, and guidance in overcoming numerous obstacles I have been facing through my thesis research during the past 6 years I spent at the University of Manitoba. His insightful instructions in the computational biology course also open my mind to the bioinformatics world. Furthermore, his passion and enthusiasm for academic research have encouraged me profoundly, leading me to develop academia as my career.

I would also like to extend my gratitude to my supervisory committee members Dr. Saman Muthukumarana, Dr. Leigh Murphy, and Dr. Yang Wang. Their feedback, suggestions, and encouragement have kept me on track during this thesis research and have been particularly instrumental during key academic milestones including the candidacy exam, thesis proposal, and all the review meetings at both department-level and university-level.

I'm also grateful to all the faculty and staff members in the Department of Biochemistry and Medical Genetics who have supported me in various ways. Special thanks to my colleagues at The Hu Lab and my department for all the communication, cooperation, and of course friendship.

Finally, I would like to dedicate this thesis to my beloved Xingying Zhang for supporting me spiritually throughout my Ph.D. study and my life in general.

Publications and Contributions

1. **Q Liu**, P Hu. 2022. Extendable and explainable deep learning for pan-cancer radiogenomics research. *Current Opinion in Chemical Biology.* 66:102111, https://doi.org/10.1016/j.cbpa.2021.102111. (IF: 8.972). This is a part of Chapter 1 and Chapter 6. I surveyed the literature, wrote the manuscript, and revised the manuscript with the guidance from Dr. Pingzhao Hu.
2. **Q Liu*,** S Huang*, Z Zhang, T Lakowski, W Xu, P Hu. 2022. Multiomics-based tensor decomposition for characterizing breast cancer heterogeneity. This is a book chapter in "Machine Learning Methods for Multi-Omics Data Integration" to be published by Springer Nature. I wrote the manuscript and did the revision with the assistance of Dr. Shujun Huang and Zhongyuan Zhang. Dr. Ted M. Lakowski, Dr. Wei Xu, and Dr. Pingzhao Hu provided guidance and supervision.
3. **Q Liu**, S Huang, D Desautels, KJ McManus, L Murphy, P Hu. Development and validation of a prognostic 15-gene signature for stratifying HER2+/ER+ breast cancer. This manuscript is under review by *Computational and Structural Biotechnology Journal*, which is included in Chapter 2. I conducted the data analysis and wrote the manuscript with guidance from Dr. Pingzhao Hu, Dr. Leigh Murphy, and Dr. Kirk McManus.
4. **Q Liu***, B Cheng*, Y Jin, P Hu. 2022. Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data. *Journal of Biomedical Informatics.* 125:103958, https://doi.org/10.1016/j.jbi.2021.103958. (IF: 8). *These authors contributed equally to the work. The paper is included in Chapter 3. I did the data analysis, manuscript writing, and manuscript revisions with assistance from Bowen Cheng and the guidance from Dr. Pingzhao Hu.

5. **Q Liu,** Y Sun, P Hu. 2022. A novel graph neural network algorithm for interpreting breast cancer mutation and prognostic risk using a hierarchical map of protein communities. The manuscript is under preparation and is included in Chapter 4. I conducted the data analysis with the help of Yan Sun and I wrote the manuscript with guidance from Dr. Pingzhao Hu.
6. **Q Liu**, P Hu. 2022. A novel integrative computational framework for breast cancer radiogenomic biomarker discovery. *Computational and Structural Biotechnology Journal.* 66:102111, https://doi.org/10.1016/j.csbj.2022.05.031. (IF: 6.155). This published paper is included in Chapter 5. I did the data analysis, manuscript writing and revisions with guidance from Dr. Pingzhao Hu.

1 Chapter 1: Introduction

Partial of this chapter has been documented into a review paper published: ***Q Liu****, P Hu. 2022. Extendable and explainable deep learning for pan-cancer radiogenomics research. Current Opinion in Chemical Biology. 66:102111, https://doi.org/10.1016/j.cbpa.2021.102111. (IF: 8.972). I did the literature reviewing, manuscript writing, and manuscript revision with the guidance from Dr. Pingzhao Hu.*

Breast cancer (BC) is the most commonly diagnosed cancer and is one of the leading causes of cancer death for women worldwide (Van Goethem *et al.*, 2006). According to Canadian Cancer Statistics 2022, 28,600 Canadian women will be diagnosed with BC, which represents 25% of all new cancer cases in women. Approximate 5,500 Canadian women will die from BC, which represents 14% of all cancer deaths in women in 2022 (Canadian Cancer Statistics Advisory Committee, 2022). The incidence of new BC cases among Canadian women increased between 1984 and 1991. Part of this increase was due to the more frequent use of mammography, which meant more cases of BC were detected (Canadian Cancer Statistics Advisory Committee, 2022). However, the incidence rate has fluctuated with an overall slight decline since then. The reasons for this fluctuation are unknown, but may involve long-term changes in hormonal factors, such as whether women start menstruating at a young age, breastfeeding, and oral contraceptive utilization (Canadian Cancer Statistics Advisory Committee, 2022). BC mortality peaked in 1986 and has

been declining since then. The decline in mortality likely reflects the impact of improved BC screening and treatment (Canadian Cancer Statistics Advisory Committee, 2022).

Screening for BC can help find cancer earlier and save lives. Screening includes self-exam or having a doctor to exam the breast or having a screening breast mammogram. If a change is detected in the screen mammogram, a diagnostic mammogram, or ultrasound, or breast MRI will be used to check more details about the abnormal. But the golden standard of BC diagnosis is the pathology biopsy. It is the test that removes a small amount of tissue from the suspicious area in the breast. Then some further examinations, such as Immunohistochemistry (IHC) examination or other molecular omics measurements, can be performed on the suspicious tissues. Although often referred to as a single disease, BC is heterogeneous in terms of histology, progression, treatment response and clinical outcome (Dai *et al.*, 2016). Researchers have made significant efforts to find biomarkers that can indicate BC progression, or responses to an exposure or intervention. For example, the hormone receptor (HR) status is the biomarker for BC that is now frequently used in clinical practice. It can indicate the response of BC patients to certain hormone therapies. However, this kind of simple biomarker cannot fully represent the heterogeneity of BC. Sometimes patients with the same HR status still show different prognosis or treatment response. Thus, researchers want to delve more deeply into BC molecular mechanisms and find better biomarkers for the patients. A successful case is the Prediction Analysis of Microarray 50 (PAM50) signature (Parker et al., 2009), which is the gene expression (transcriptome) patterns of several genes, and can help predict the chance of metastasis and survival of BC. Now, with the advent of variety of molecular profiling techniques and machine learning-based data analyzing techniques, people have been investigating better biomarkers for BC by integrating multiple data sources together. This thesis

aims to explore BC heterogeneity based on multi-omics data and uncover biomarkers using machine learning methods.

Chapter 1 aims at reviewing related literatures, which have provided guidance for the research. Section 1.1 introduces BC heterogeneity in perspectives of the subtyping schemes and existing biomarkers for BC. Section 1.2 describes the limitations of current BC subtyping schemes and biomarkers as well as the challenges in using machine learning in BC heterogeneity exploration. Section 1.3 reviews the strategies of solving the challenges mentioned in Section 1.2, including the approaches to acquire BC multi-omics measurements and the machine learning methodologies to analyze them.

1.1 BC heterogeneity

BC is a complex disease and has high inter-tumor and intra-tumor heterogeneity that result from both genetic and non-genetic alterations (Polyak, 2011). Inter-tumor heterogeneity refers to variances between tumors from different people, while intra-tumor heterogeneity represents cellular differences within a single tumor of a patient (Polyak, 2011; Martelotto *et al.*, 2014). Intra-tumor heterogeneity can be further divided into spatial heterogeneity and temporal heterogeneity. Spatial heterogeneity represents regional histological/genetic/epigenetic differences within a tumor, while temporal heterogeneity reflects histological/genetic/epigenetic changes of the tumor at different time points (Martelotto *et al.*, 2014).

BC heterogeneity has become one of the most important and clinically relevant areas of BC research, as in-depth characterization and understanding of BC heterogeneity are critical for improving diagnosis, uncovering prognostic and predictive biomarkers, and designing therapeutic strategies (Koren and Bentires-Alj, 2015). This thesis mainly focuses on exploring BC inter-tumor

heterogeneity. Section 1.1.1 discusses the traditional histopathological staging and grading of BC. Section 1.1.2 introduces the IHC detected HR biomarkers and subtyping schemes of BC. Section 1.1.3 details the intrinsic BC subtyping scheme and other gene expression-based subtype schemes. Section 1.1.4 summarizes BC subtyping schemes that involve other data types such as copy number variations (CNVs). As the intra-tumor heterogeneity is not in the scope of this thesis, it is not discussed in this chapter.

1.1.1 BC histopathological staging and grading

Traditionally, the histopathological findings have been the main indicators for BC inter-tumor heterogeneity, which have also acted as the foundation of different BC stratification schemes. Histologically, BC can arise from epithelial cells in ducts or lobules, leading to ductal or lobular carcinoma, respectively (Malhotra *et al.*, 2010). Depending on whether the cancer cells are still only in the epithelial area or have invaded to surrounding areas, cancers can be *in situ* or invasive (Malhotra *et al.*, 2010). Around 70~80% invasive breast carcinomas are invasive ductal carcinoma (Malhotra *et al.*, 2010). Histopathological staging and grading have a strong impact on current clinical decision-making in BC. The Tumor, Node, Metastasis (TNM) system is the most commonly used staging method in the clinic (Paleri *et al.*, 2010). It was originally developed by Pierre Denoix in the 1940s and 1950s and then edited and organized continuously by the American Joint Committee on Cancer and the Union for International Cancer Control (Kalli *et al.*, 2018). T (Tumor) is used to represent the size of the tumor, N (Node) is used to describe nearby (regional) lymph nodes that are involved, and M (Metastasis) is used to describe whether the cancer has spread from one part of the body to another(Paleri *et al.*, 2010). An overall Stage metric that ranges

from 0 to IV is then used to summarize the T, N, M information. The details of the levels of the metrics T, N, M, and Stage are shown in **Table 1-1**.

Table 1-1. Tumor (T), Node (N), and Metastasis (M) metrics in TNM staging system.

Metric	Category		Meaning
Tumor (T)	T0		No evidence of primary tumor.
	TX		Primary tumor but the size cannot be assessed.
	Tis		Ductal carcinoma in situ.
	T1	T1mi	Tumor size is 0, centimetres or less.
		T1a	Tumor size is more than 0.1 centimetres but not more than 0.5 centimetres.
		T1b	Tumor size is more than 0.5 centimetres but not more than 1 centimetres.
		T1c	Tumor size is more than 1 centimetres but not more than 2 centimetres.
	T2		Tumor size is more than 2 centimetres but no more than 5 centimetres.
	T3		Tumor size is more than 5 centimetres.
	T4	T4a	Tumor has extended to the chest.
		T4b	Tumor has spread into the skin.
		T4c	Both T4a and T4b.
		T4d	Inflammatory carcinoma – the overlying skin is red, swollen, and painful.
Node (N)	N0		There are no cancer cells in any nearby nodes.
	NX		Lymph nodes cannot be assessed (for example, if they were previously removed).

	N2	N2a	Cancer cells have spread to the lymph nodes in the armpit, which are stuck to each other and to other structures.
		N2b	Cancer cells have spread to the lymph nodes behind the breastbone (the internal mammary nodes), which can be seen on a scan or felt by the doctor. There is no evidence of cancer cells in lymph nodes in the armpit.
	N3	N3a	Cancer cells have spread to lymph nodes below the collarbone.
		N3b	Cancer cells have spread to lymph nodes in the armpit and behind the breastbone.
		N3c	Cancer cells have spread to lymph nodes above the collarbone.
Metastasis (M)	M0		No cancer cells are found in the any distant body parts.
		cM0(i+)	Small numbers of cancer cells are found in blood or bone marrow, or cancer cells have spread to lymph nodes away from the armpit, collarbone, and breastbone.
	M1		Cancer cells are found in the distant organs such as bones, lungs, bran, or liver.
Stage	0		Tis, N0, M0
	I	IA	T1, N0, M0
		IB	T0, N1mi, M0
			T1, N1mi, M0
	II	IIA	T0, N1, M0
			T1, N1, M0
			T2, N0, M0
		IIB	T2, N1, M0
			T3, N1, M0
	III	IIIA	T0, N2, M0

		T1, N2, M0
		T2, N2, M0
		T3, N1, M0
		T3, N2, M0
	IIIB	T4, N0, M0
		T4, N1, M0
		T4, N2, M0
	IIIC	T, N3, M0
IV		T, N, M1

*This table is adapted and summarized from Kalli *et al.*'s work (Kalli *et al.*, 2018).

Besides the staging, histopathological grading also has a strong impact on current clinical decision-making in BC (Turashvili and Brogi, 2017). The Nottingham grading is the most widely used grading method in the clinic (Elston and Ellis, 1991). It was modified from the Bloom-Richardson BC grading system which was developed 60 years ago (Elston and Ellis, 1991). Three tumor features are considered to assess the Nottingham grading of BC, namely glandular/tubular differentiation, nuclear pleomorphism, and mitotic counts. The details of each Nottingham grading feature and the grade standard based on them are summarized in **Table 1-2**.

Table 1-2. Summary of the Nottingham grading system for assessing BC histological grade.

Feature	Points/grade	Meaning
Glandular/tubular formation	1	Majority of the tumor (>75%) with glandular/tubular formation.
	2	Moderate degree (10 – 75%) of glandular/tubular differentiation.
	3	Little or none (<10%) glandular/tubular differentiation.
	1	Small, regular uniform nuclear.

Nuclear pleomorphism	2	Moderate increase in nuclear size and variability.
	3	Marked nuclear variability.
Mitotic counts	1	0-9 mitosis observed by a 0.274-millimeter field area Leitz Ortholux microscope
		0-5 mitoses observed by a 0.152-millimeter field area Nikon Labophot microscope
		0-11 mitoses observed by a 0.312-millimeter field area Leitz Diaplan microscope
	2	10-19 mitoses observed by a 0.274-millimeter field area Leitz Ortholux microscope
		6-10 mitoses observed by a 0.152-millimeter field area Nikon Labophot microscope
		12-22 mitoses observed by a 0.312-millimeter field area Leitz Diaplan microscope
	3	More than 20 mitoses observed by a 0.274-millimeter field area Leitz Ortholux microscope
		More than 11 mitoses observed by a 0.152-millimeter field area Nikon Labophot microscope
		More than 23 mitoses observed by a 0.312-millimeter field area Leitz Diaplan microscope
Grade	I	The total points of the above three features are 3-5.
	II	The total points of the above three features are 6-7.
	III	The total points of the above three features are 8-9.

*This table is adapted and summarized from Elston and Ellis's work (Elston and Ellis, 1991).

1.1.2 Receptor expression-based BC subtyping

With the development of modern molecular biology techniques, the heterogeneity in BC has been better recognized through well-established molecular biomarkers, such as estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor2 (HER2) (Harbeck *et al.*, 2019). The expressions of these three receptors are recognized by international guidelines as prognostic and therapeutic biomarkers for invasive BC and are routinely tested through IHC in clinic practice today. A brief illustration of the National Health Service (NHS) clinical management guideline for BC is shown in **Figure 1-1** to demonstrate the clinical

applications of both histopathological biomarkers, staging/grading, and the status of the above mentioned three receptors. According to the standards from American Society of Clinical Oncology/College of American Pathologist, if any nuclear staining is observed in more than 1% of the cancer cells (no matter how strong or weak the signal is), the cancer is considered ER-positive (ER+) and/or PR-positive (PR+). Conversely, if nuclear staining is observed in less than 1% of the cancer cells, the cancer is considered ER-negative (ER-) and PR-negative (PR-) (Weigel and Dowsett, 2010). About 80% of BCs are ER+ and about 75% are PR+. Approximately 65~75% of BCs co-express ER and PR (ER+/PR+), 13% are ER+/PR- and only 2% are ER-/PR+ (Harbeck *et al.*, 2019) (**Figure 1-2**). For the HER2 status, it can be assessed both by IHC and by the *in-situ* hybridization detected amplification of HER2 gene, but the IHC is the most widely used approach due to its cost-effectiveness. The expression of HER2 protein can be quantified and scored on a scale of 0-3. If the HER2 protein expression is scored 0 or 1+, the tumor is considered HER2-negative (HER2-), while if the score is 2+, the tumor is considered equivocal, and if the score is 3+, the tumor is then considered HER2-positive (HER2+) (Wolff *et al.*, 2018). About 15~25% of BCs are HER2+ (Wolff *et al.*, 2018) (**Figure 1-2**).

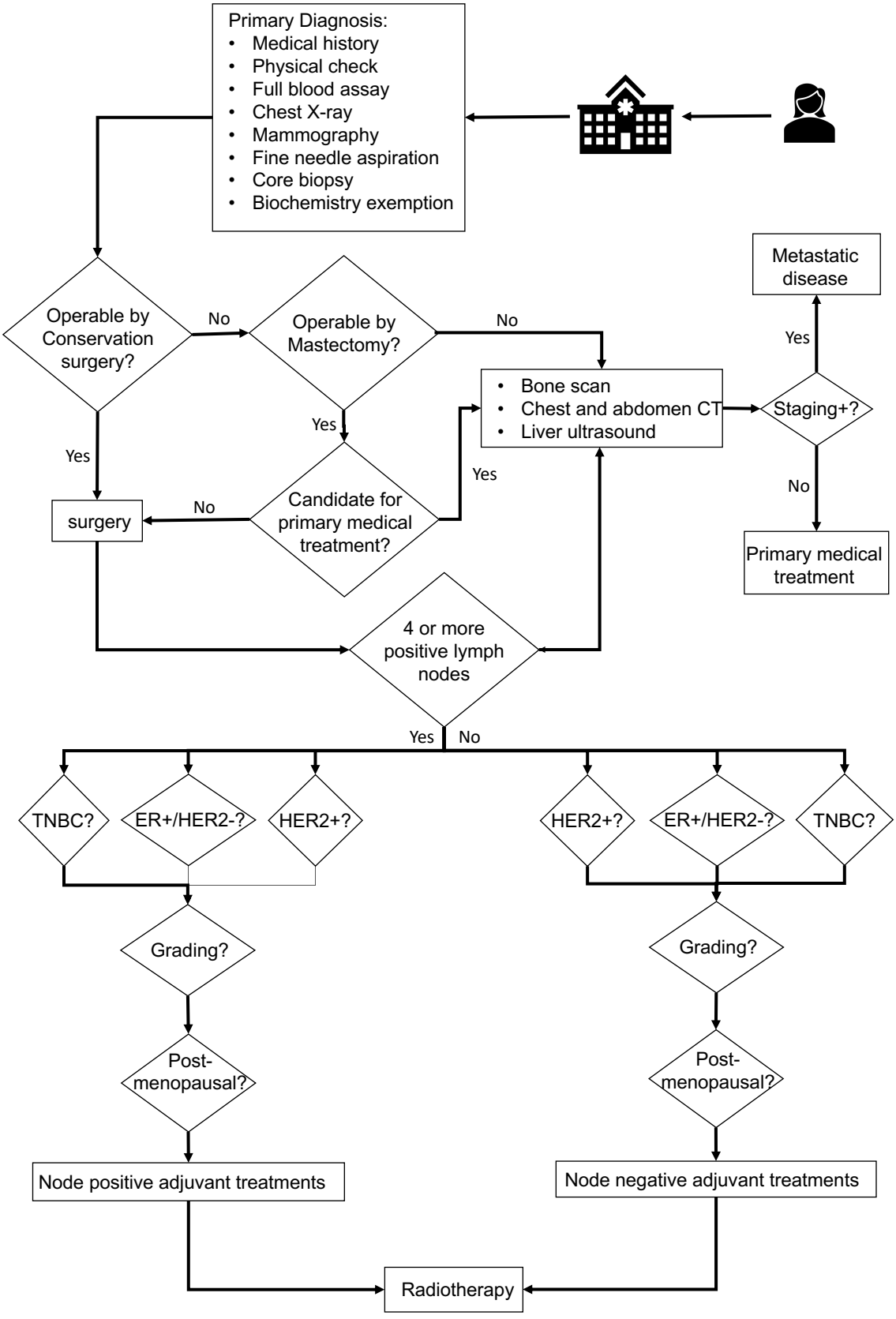

Primary Diagnosis:
• Medical history
• Physical check
• Full blood assay
• Chest X-ray
• Mammography
• Fine needle aspiration
• Core biopsy
• Biochemistry exemption
Operable by Conservation surgery?
No
Operable by Mastectomy?
No
Metastatic disease
Yes
• Bone scan
• Chest and abdomen CT
• Liver ultrasound
Staging+?
Yes
No
Primary medical treatment
Yes
Yes
surgery
No
Candidate for primary medical treatment?
Yes
4 or more positive lymph nodes
Yes
No
TNBC?
ER+/HER2-?
HER2+?
HER2+?
ER+/HER2-?
TNBC?
Grading?
Grading?
Post-menopausal?
Post-menopausal?
Node positive adjuvant treatments
Node negative adjuvant treatments
Radiotherapy

Figure 1-1. A summary of the National Health Service (NHS) clinical management guideline for BC. Source is from the National Health Service (NHS) North Cancer Alliance (NCA) https://www.nhsscotlandnorth.scot/nca/clinical-management-guidelines.

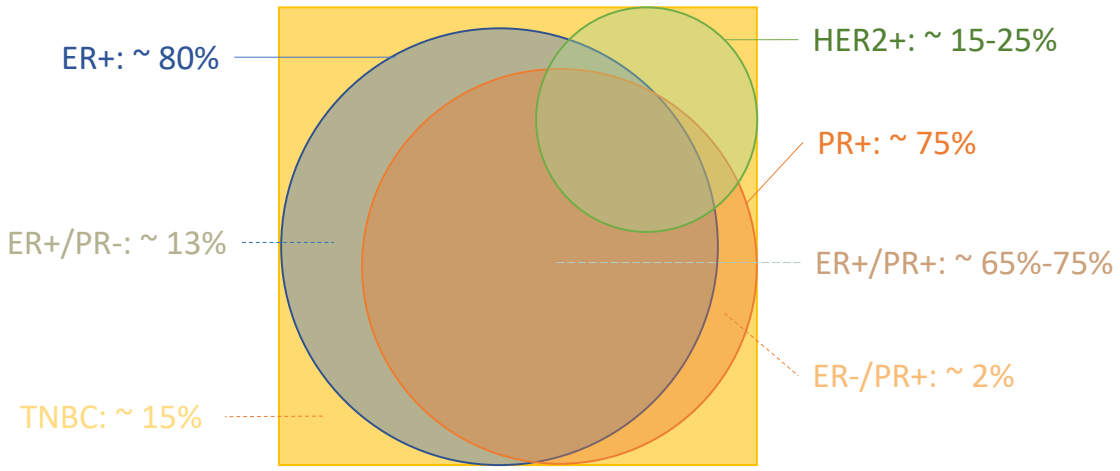


Figure 1-2. The Proportions of BC ER, PR, and HER2 subtypes.

Patients that express ER and PR can benefit from endocrine/hormonal therapy targeting these receptors. For example, ER is a protein inside the cells and can be activated by the hormone estrogen. Once it is activated, the ER will move into nucleus and bind to DNA to regulate activity of different genes and then stimulate the proliferation of mammary cells. Besides, the estrogen metabolism can produce genotoxic waste. These two mechanisms lead to disruption of the cell cycle, apoptosis, and DNA repair, thus leading to tumorigenesis. ER antagonists such as tamoxifen can block estrogen activity by competitively binding to ER, thereby preventing cancer cell proliferation. HER2+ tumors tend to be more aggressive and less responsive to hormonal therapy, and thus have worse prognosis, but they respond to anti-HER2 therapy (e.g., trastuzumab and lapatinib) (Dean-Colomb and Esteva, 2008). BCs that do not express ER, PR, and do not over

express HER2 are called triple-negative BC (TNBC) which account for around 15% of all invasive BCs and carry the worst prognosis and treatment response (Zagami and Carey, 2022).

1.1.3 Gene expression-based BC subtyping

Gene expression has been used in cancer stratification and gene signature identification since 1999 (Qian *et al.*, 2021). Gene expression signatures can help improve patient care by classifying tumors into distinct groups, providing guidance for personalized clinical decisions. In the past two decades, gene expression pattern that takes several transcriptional variations into consideration has been used to study BC inter-tumor heterogeneity. Among them, the most widely established one is the intrinsic classification scheme. It was first proposed in the year of 2000 by Perou *et al.,* where hierarchical clustering was performed on the expression values of 496 genes from 42 BC patients and identified four patient clusters: Luminal-like, HER2-enriched, Basal-like, and Normal-like (Perou *et al.*, 2000). The Luminal-like subtype was named for its high expression of many luminal epithelial cells, which was further evaluated by IHC using antibodies against the luminal-like keratins. In contrast, the HER2-enriched BC subtype underexpresses luminal epithelial genes but overexpresses a subset of genes (including HER2 and GRB7) located in the HER2 amplicon on chromosome 17q12. Similarly, the Basal-like subtype shows low expression of luminal epithelial genes but high expression of breast basal epithelial genes, which can be confirmed by IHC with antibodies against basal keratins. BC samples in Normal-like subtype were clustered close to the normal breast samples, and they overexpress adipose and basal epithelial genes but underexpress luminal epithelial genes (Perou *et al.*, 2000). A follow-up study with an increased sample size (from 42 patients to 84 patients) was published in 2001 by Sørlie *et al*. In this follow-up study, a similar hierarchical clustering method was used to further divide the

previous Luminal-like subtypes into Luminal A, Luminal B, and Luminal C subtypes, resulting in a total of six intrinsic subtypes (Luminal A, Luminal B, Luminal C , HER2-enriched, Basal-like and Normal-like). Luminal A shows the highest expression of luminal epithelial genes, while Luminal B and Luminal C show relatively low or moderate luminal epithelial gene expression. However, Luminal C highly expresses a set of genes of unknown function that are not highly expressed in Luminal A and Luminal B. Furthermore, this follow-up study found that the mutation frequency of the TP53 gene is significantly different among the identified intrinsic subtypes (13% in Luminal A, 71% in HER2-enriched, 82% in Basal-like). The overall survival rates are also significantly different among the identified intrinsic subtypes (Basel-like and HER2-enriched shows the worst survival rates) (Sørlie *et al.*, 2001). Two years after the first follow-up study, Sørlie *et al*. published another follow-up study with a total of 115 BC samples. Similar hierarchical clustering of the expression of 534 genes from these 115 samples revealed five intrinsic subtypes: Luminal A, Luminal B, HER2-enriched, Basal-like and Normal-like (Luminal C in the previous version has been removed in this version) (Sørlie *et al.*, 2003).

To further evaluate the intrinsic subtyping scheme and translate it into clinical practice, more studies were conducted to confirm its reproducibility and develop more precise and concise prediction tools for it (Hu *et al.*, 2006) (Parker *et al.*, 2009). A single sample predictor (SSP) strategy was proposed in 2006 by Hu *et al*. to assign intrinsic subtype labels to incoming new samples based on the similarity of the expression profiles of the 306 intrinsic genes in the samples with the centroid of each subtype expression level of these 306 intrinsic genes (Hu *et al.*, 2006). In 2009, Parker *et al.* were able to reduce the number of intrinsic genes used in the SSP from 306 genes to 50 genes with comparable assignment accuracy (Parker et al., 2009). This set of 50 genes was named PAM50 and significantly reduced the difficulty of the intrinsic subtypes being used in

clinical assay (Parker et al., 2009). As a matter of course, later the PAM50 assay was successfully commercialized and has been commonly used in clinical practice since then. The history of the intrinsic subtyping scheme is shown in **Figure 1-3** timeline.

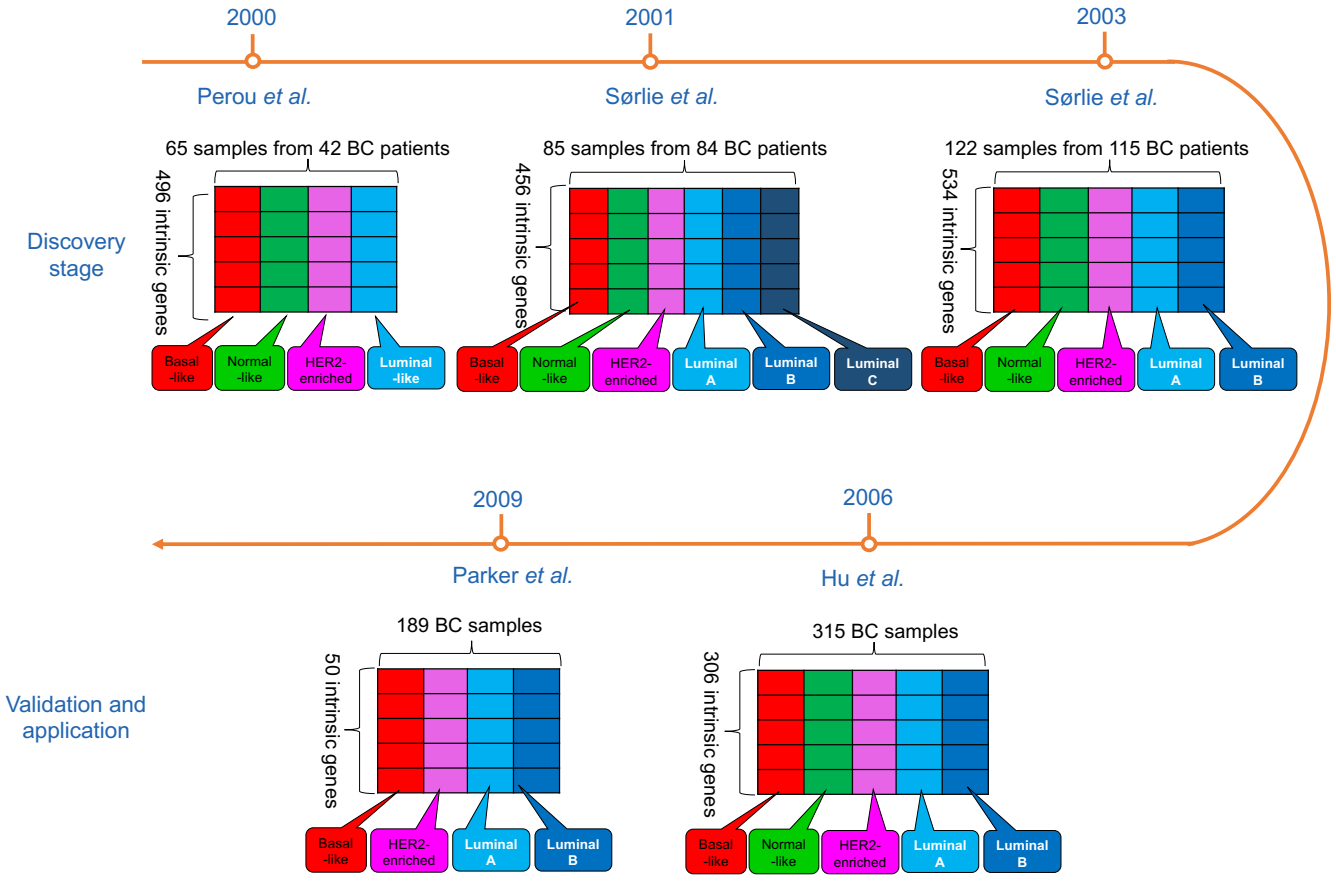


Figure 1-3. The history of the intrinsic subtyping scheme development.

In 2007, Herschkowitz *et al*. identified a new BC subtype, namely Claudin-low, using both 232 human BC samples and 108 mouse BC models (Herschkowitz *et al.*, 2007). The Claudin-low subtype is an extension of above-described intrinsic subtypes and better represents the heterogeneity of some Basal-like and Luminal A/Luminal B BCs (Prat and Perou, 2011). Claudin-low BCs are characterized by the low expression of several Claudin genes involved in tight

junctions and cell-cell adhesion (Herschkowitz *et al.*, 2007) and have poor prognosis and intermediate chemotherapy sensitivity (Prat *et al.*, 2010).

Besides the successful intrinsic subtyping scheme, some other gene expression profile-based BC subtyping methods were proposed in parallel. For example, using the expression of 706 genes from 99 BC samples, Sotiriou *et al.* were able to identify six BC subtypes: Luminal-like 1, Luminal-like 2, Luminal-like 3, Her-2/neu, Basal-like 1, and Basal-like 2 (Sotiriou *et al.*, 2003). Three versions of Subtype Classification Model (SCM), namely SCMOD1, SCMOD2, and SCMGENE were released subsequently to assign subtype labels to incoming new BC samples based on the expression of the 726 genes, 663 genes, and 3 genes, respectively (Haibe-Kains *et al.*, 2012). However, this BC subtyping system has not been applied into clinical practice. Another example is the Cartes d'Identit des Tumeurs (CIT) BC classification project, which identified six BC subtypes (Luminal A, Luminal B, Luminal C, mApo, Basal-like, and Normal-like) based on expression of 256 genes from 537 BC samples (Guedj *et al.*, 2012). Both the SCM and CIT subtyping schemes overlap to some extent with intrinsic subtyping, indicating their similarities and uniqueness.

1.1.4 Copy number alteration-based BC subtyping

The successful BC intrinsic subtyping scheme, as well as the SCM and CIT BC subtyping systems were developed based on the gene expression profiles, thus they can only stratify BC patients into transcriptionally distinct subtypes. However, Guedj *et al*. found that each of those six CIT subtypes shows different CNVs, indicating its ability to capturing BC heterogeneity not only at the transcription level but also at the alterations of the genome (Guedj *et al.*, 2012). This finding

inspired researchers to extend their research scope from the well-studied gene expression profiles to CNV profiles. In 2010, Jönsson and colleagues proposed the first CNV-based BC subtyping scheme which divided BCs into six genomic subtypes: 17q12, Basal-complex, Luminal-simple, Luminal-complex, Amplifier, and Mixed (Jönsson *et al.*, 2010). In the same year, Curtis et al. proposed a subtyping scheme that takes both gene expression and CNV data into consideration (Curtis *et al.*, 2012). Their data came from the Molecular Taxonomy of BC International Consortium (METABRIC) project where the DNA and RNA extracted from a large number of BC samples (997 samples in the discovery set and 995 samples in the validation set) were hybridized onto the Affymetrix SNP 6.0 and Illumina HT-12 v3 platforms to quantify the CNV and gene expression. Expression quantitative trait loci (eQTL) analysis was then performed to identify *cis* genes that are driven by recurrent CNVs. These *cis* genes were finally used to divide BCs in the discovery set into 10 integrative clusters (IntCluster 1 – 10) with different clinical outcomes. And the result can be reproduced in the validation cohort (Curtis *et al.*, 2012). These 10 IntCluster subtypes have the below characteristics. IntClust1, 2, 3, 6, 7, and 8 have a significant number of CNVs, while IntCluster 4 is known as CNV-devoid due to its lower number of CNVs than other IntClusters. The locus of the HER2 genes and the MYC oncogene are amplified in IntCluster 5 and IntCluster 9, respectively. IntClust10 has high genomic instability (Russnes *et al.*, 2017).

Both the single CNV-based genomic subtypes and gene expression & CNV-based IntClust subtypes overlap to some extent with the intrinsic subtypes, as shown in **Table 1-3**.

Table 1-3. The relationship of single CNV-based genomic subtypes and gene expression & CNV-based IntClust subtypes with the single gene expression-based intrinsic subtypes.

Intrinsic subtypes	Genomic subtypes	IntClust subtypes

Luminal A	Luminal-simple, Luminal-complex, Amplifier, Mixed	IntClust 4, IntClust 9
Luminal B	Luminal-simple, Luminal-complex, Mixed, Amplifier	IntClust 5, IntClust 9
HER2-enriched	17q12, Amplifier, Mixed	IntClust 5
Basal-like	Basal-complex, Amplifier, Mixed	IntClust 4, IntClust 10

1.2 Challenges in deciphering BC heterogeneity

Section 1.1 summarizes current research achievements in BC heterogeneity studies. In this section, the limitations of existing BC subtyping schemes and biomarkers and the challenges in this field are briefly described. More details of the challenges are further discussed separately in Chapters 2, 3, 4, and 5.

1.2.1 Substantial heterogeneity in BC subtype

Although existing BC biomarkers and subtyping schemes distinguish the BC well, there is substantial heterogeneity within each subtype. For example, the TNBCs are observed to share the lowest similarity with each other than the other IHC defined BC subtypes (Russnes *et al.*, 2017). To further explore the difference within TNBCs, Lehmann *et al.* studied 587 TNBC cases and identified six TNBC subtypes, namely Basal-like 1, Basal-like 2, Immunomodulatory, Mesenchymal, Mesenchymal stem-like, and Luminal androgen receptor subtypes (Lehmann *et al.*, 2011). Basal-like 1 TNBCs overexpress genes that are involved in cell cycle and cell division. They have a very high mutation rate for tumor suppressor genes such as TP53, BRCA1 and BRCA2. While Basal-like 2 TNBCs overexpress genes that are enriched in growth factor signaling

pathways. As the name suggests, the Immunomodulatory subtype overexpresses genes that are related to immune functions, and the Mesenchymal and Mesenchymal stem-like subtypes overexpress genes that are involved in mesenchymal differentiation and epithelial-mesenchymal transition, etc. Finally, Luminal androgen receptor TNBCs overexpress luminal differentiation and androgen receptor signaling genes (Lehmann *et al.*, 2011). This TNBC subtyping scheme was further evaluated by a study that used the combined gene expression and DNA profiling data from 198 TNBC tumors, where Burstein *et al.* confirmed four distinct TNBC subtypes: Luminal androgen receptor, Mesenchymal, Basal-like immune-suppressed, and Basal-like immune-activated (Burstein *et al.*, 2015). Lehmann *et al.* published a follow-up study that refined their original TNBC molecular subtypes from six to four (Basal-like 1, Basal-like 2, Mesenchymal, and Luminal androgen receptor subtypes) that are the same as the Burstein *et al.*'s result (Lehmann *et al.*, 2016). These four stable TNBC subtypes show distinct prognoses and can represent the differences in TNBC diagnosis age, grade, local/distance disease progression, etc. (Lehmann *et al.*, 2016). They also provided potential subtype-specific treatment targets for TNBC (Burstein *et al.*, 2015).

TNBC has been an active research area because it has the worst survival time and lacks effective therapies. On the contrary, some other subtypes of BC are understudied. For example, as illustrated in **Figure 1-1**, currently all HER2+ BCs are treated in the same way in clinic no matter whether they co-express ER or not. However, the HER2+/ER+ cases usually have better survival in the first 5 years after diagnosis but have higher recurrence risk after 5 years compared to HER2+/ER- (Hwang *et al.*, 2019; Cameron *et al.*, 2017), which may be due to the fact that sustained ER signalling in HER2+ BCs helps escape the HER2 blockade (Bender and Nahta, 2008; Gingras *et al.*, 2017). Currently HER2+/ER+ BC is understudied and lacks biomarkers. A better

understanding of the underlying molecular diversity is important to find new therapy targets for subtypes such as HER2+/ER+ BCs.

As discussed in the previous section, BCs can be divided into 5 intrinsic subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like), but not all HER2+ tumors fall into the HER2-enriched subtype (Koboldt *et al.*, 2012). Based on the expression profiles of the intrinsic genes used for the intrinsic subtyping, Prosigna (rorS) was developed as a gene expression signature that estimates distant recurrence risk of ER+, PR+, hormone-treated, postmenopausal women with BC (Joel S. Parker *et al.*, 2009). Besides rorS, there are several other commercialized BC gene expression signatures, which can be used to estimate different risks for different BC subgroups. Oncotype DX and EndoPredict are gene signatures that estimate the distant recurrence in ER+/HER2- and hormone-treated BC from the expression profiles of 21 genes and 11 genes, separately (Paik *et al.*, 2004; Filipits *et al.*, 2011). PIK3CA-GS was derived from exon 20 (the kinase domain) mutations and is able to predict PIK3CA mutation status and tamoxifen sensitivity of ER+/HER2- BC (Loi *et al.*, 2010). MammaPrint (also called GENE70) is a 70-gene expression signature that could predict the benefit of adjuvant therapy for BC patients under the age of 61 (Joel S. Parker *et al.*, 2009; van 't Veer *et al.*, 2002). Gene progNostic Index Using Subtypes (GENIUS) is a prognostic gene expression signature applicable for any subtype of BC (Bontempi *et al.*, 2010). Gene expression Grade Index (GGI) is a 97-gene signature generated from differentially expressed genes between different histological grades of BC and can estimate the prognostic and recurrence risks of ER+ BC patients (Sotiriou *et al.*, 2006). As can be seen above, there are no such gene signatures that are proposed specifically for HER2+/ER+ BC patients. Thus, exploring the expression and genomic profiles of HER2+/ER+ BC and generating prognostic and predictive gene signatures are critical for better HER2+/ER+ BC clinical guidance.

1.2.2 Integration of BC multi-omics data for understanding BC heterogeneity

In biology, the central dogma model is that DNA is transcribed into mRNA, then translated to proteins, and finally forms the tumor's phenotype. Epigenetic changes can boost or interfere with the transcription of genes. The studies of the whole set of DNAs, epigenetic changes, mRNAs, and proteins are called genomics, epigenomics, transcriptomics, and proteomics, respectively. The medical image features extracted from patients' medical images can reflect phenotypic information such as morphological or intensity difference of the tumor, which is called radiomics (Lambina *et al.*, 2012). Multi-omics layers these omics on top of each other and gives us a multidimensional view of human biology, which can help us discover biomarkers that may be missed if we only focus on a single view of these omics. **Figure 1-4** shows a simplified overview of the multi-omics.

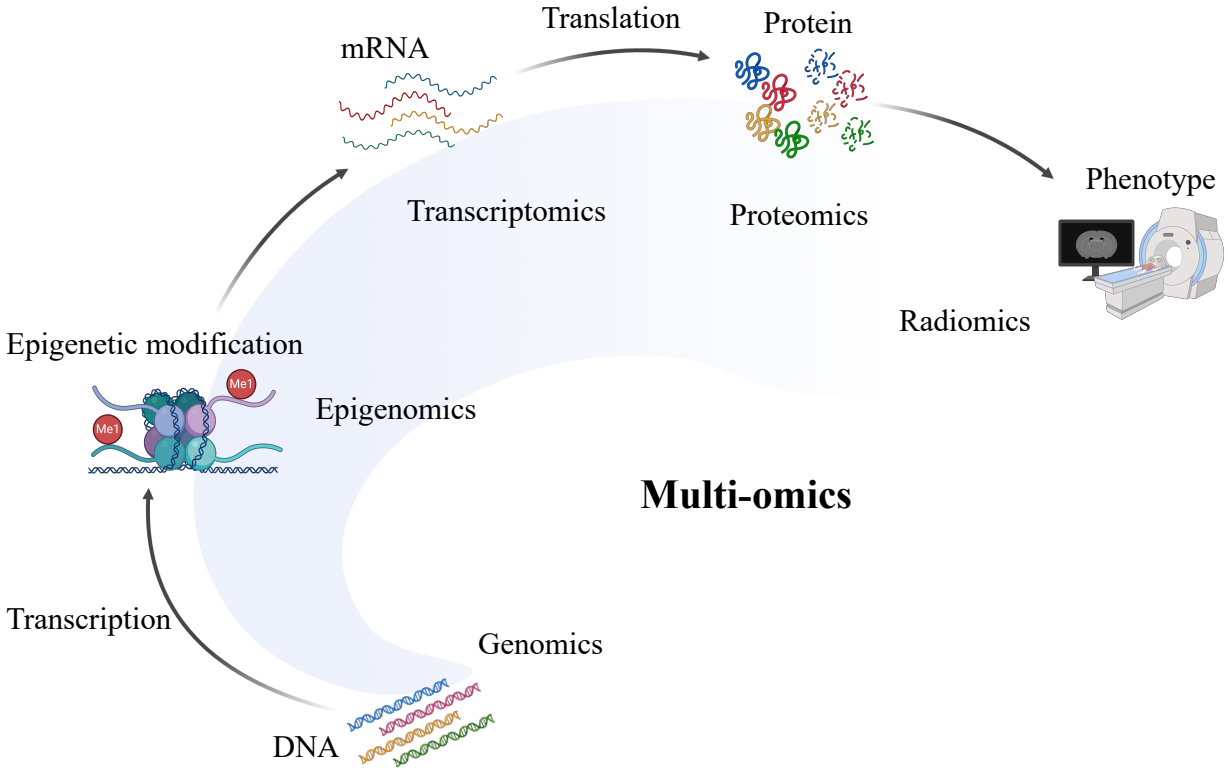


Figure 1-4. The overview of multi-omics.

As shown by Curtis *et al.* in their proposed subtyping scheme that takes both gene expression and CNV data into consideration (Curtis *et al.*, 2012), the heterogeneity of BC could be better detected using data from different biological levels such as gene expressions, CNVs, DNA methylations, proteomics measurements, radiomics, and so on. However, majority of research in BC subtyping are still based on single level of genomic data (Dai *et al.*, 2015) which may fail to synthesize the inherent information across different levels of genomic data. For this reason, cancer subtyping using multi-omics data has its advantage in integrating information from different aspects, leading to a more comprehensive subtyping (Huang *et al.*, 2017). The advancement of a variety of modern molecular biology techniques has made it possible to build an integrated framework for BC heterogeneity study (Rappoport *et al.*, 2019), but the problem is how to effectively integrate/mine multimodal data and preserve cross-level information simultaneously.

According to Fang et al., there exist two kinds of data integration approaches: loose approach and tight approach (Fang, 2019). For the loose approach, different types of datasets are first used to build separate models. Then the results from these models are combined in either a hierarchical, ensemble, or a linear fashion (Thingholm *et al.*, 2016; M. D. Ritchie *et al.*, 2015; Huang *et al.*, 2017). The advantage of this multi-staged method is its simplicity. However, loose integration is ineffective in exploring the complicated patterns that arise from combining distinct types of data (Fang, 2019). Therefore, the tight approach is considered to be more advanced than the loose method. For the tight approach, all different types of data are combined to fit a single subtyping model. The easiest way of tight integration is to concatenate all types of normalized data to establish a large matrix, then input the large matrix directly into the subtyping model (Fang,

2019; Huang *et al.*, 2017). However, this concatenation method usually ignores the fundamental complementarity of different types of data (Wang *et al.*, 2014). Therefore, it is suggested to combine different data sources in a more abstract and practical way (Huang *et al.*, 2017). For example, we can first extract latent features from different data types and then use classic clustering algorithms to do subtyping. Examples include iClusterPlus developed by Mo *et al*. (Mo *et al.*, 2018) and LRAcluster (low-rank approximation based multi-omics data clustering) developed by Wu *et al*. (Wu *et al.*, 2015). Nguyen *et al*. proposed PINSCluster (Nguyen *et al.*, 2020) to integrate multi-omics data using cross-type connectivity, which is also a good example of the abstract and tight data integration methods. A more detailed description and comparison of these methods can be found in a recent review paper on comprehensively evaluating multi-omics-based cancer subtyping methods (Duan *et al.*, 2021). Basically, the review concluded that iClusterPlus, LRAcluster, and PINSCluster show decent and comparable performance but the robustness and computational efficacy are worse than classical cancer subtyping algorithm Similarity Network Fusion (SNF) (Wang *et al.*, 2014).

Some mathematic and computational methods were proposed to address this data combination problem in a more abstract way. The most popular one is the tensor factorization (Kolda and Bader, 2009). Instead of simply concatenating different omics data to a large two-dimensional matrix, we can combine them into a three-dimensional tensor with the new dimension representing data type. This three-dimensional tensor can thus be processed by traditional tensor factorization algorithms for latent factor extraction. In this way, data across different levels are integrated and the cross-level information is retained. These integrated factors are then passed on to downstream subtyping models. Tensor is defined as a high-dimensional data array (Kolda and Bader, 2009). Tensor decomposition or factorization problem has been studied for many years and

the algorithms that emerged to solve the problem have been applied in many domains of research (Kolda and Bader, 2009; Harshman, 1970; Carroll and Chang, 1970). Among those mature tensor factorization algorithms, canonical decomposition (CANDECOM) (Carroll and Chang, 1970) and parallel factors (PARAFAC) (Harshman, 1970) are always referred together as CP (CANDECOM/ PARAFAC) because they both decompose a tensor as a sum of rank-one tensors (Kiers, 2000; Hitchcock, 1927; MÖcks, 1988). A drawback of CP-based tensor factorization is that a predefined hyperparameter, the rank of the tensor, is needed. However, to determine the rank of a given tensor is known as a non-deterministic polynomial-time (NP) -hard problem (Håstad, 1990; Hillar and Lim, 2013). There has been no straightforward algorithm for solving this problem for a long time (Kolda and Bader, 2009) until recently, a Bayesian tensor factorization (BTF) model was proposed to overcome this difficulty (Tang *et al.*, 2018). BTF takes three steps to determine the rank of a tensor: a multi-linear model to decompose the given tensor to latent factors; a high-dimensional variational Bayesian inference method to estimate the posterior distribution of the decomposed latent factors; and finally a filtering procedure for removing the redundancy factors (Tang *et al.*, 2018). Factors extracted by BTF have been proved to have better performance in electroencephalography-based sleep stage classification and seizure detection (Tang *et al.*, 2018).

1.2.3 Complexity in BC multimodal data analyses

Among the multimodal data in different biological level, the graph structured proteomics data and medical image data are significantly different from the matrix format genomics, epigenomics, and transcriptomics data. Thus, new analyzing strategies are needed. Proteomics data could be modeled as a graph where vertices are proteins, and edges are undirected and weighted

connections between the interacting proteins. Graphs have better ability to modelling real-word entities and their relations. A basic graph that only describes the structural information could be noted as $G(V, E)$, where V represents the vertex set and E is the edge set, while an attributed graph $G(V, E, X_V, X_E)$ could house the node features X_V and edge features X_E as well. CComprehensively mapping the structure of BC protein communities using this kind of graph structured proteomics data and identifying potential biomarkers from them are promising approaches for the study of BC heterogeneity, as it can provide a systematic view of BC proteomics for better understanding disease mechanisms and can also help identify cumulative mutation/outcome effects on clusters of proteins/genes rather than individual proteins/genes (Zheng *et al.*, 2021). Currently, the Clique Extracted Ontology (CliXO) algorithm is widely used in detecting hierarchical communities of a network (Zheng et al., 2021; Qin et al., 2021). CliXO takes the weighted graph as input, calculates the maximal cliques, and adds the identified maximal cliques to the ontology as the threshold of edge weight of this graph (here, the combined GO proximity) is lowered (Kramer et al., 2014). However, CliXO only considers the edge weights, and it has several hyperparameters that control the threshold decreasing strategies which need to be pre-specified and tuned manually to achieve a propitiate decision. Furthermore, CliXO is an unsupervised clustering method that cannot take critical clinical information such as patients' prognosis into consideration. Thus, more efficient and automatic algorithms are needed to address these drawbacks.

Medical images have been extensively used in different phases of BC management due to their interpretability, real-time monitoring, noninvasive nature (Fass, 2008). In addition, medical images can capture dynamic and macroscopic information of the entire tumor, which cannot be captured by molecular omics data because biopsies are usually performed at a specific time on a small part of tumor tissues. Once its relationship with other molecular omics data and its role in

BC heterogeneity are studied thoroughly, medical image can serve as both a great resource of information itself and an important surrogate for molecular information. This brings up two relatively fresh but important research fields, namely radiomics and radiogenomics.

Radiomics is a research field where high-throughput medical image features are used to describe disease phenotypes (Lambina *et al.*, 2012). Current radiomic studies face feature subjectivity and interpretability trade-off issues. Traditional computational engineering methods usually involve human experts' pre-processing which introduces subjectivity. Although human understandable image features such as tumor morphological features and the first-order, second-order statistic features of the image pixel distribution (Van Griethuysen *et al.*, 2017) could be generated using the traditional feature engineering methods, they are pre-defined and limited by human knowledge. Therefore, they may not be able to fully represent the image heterogeneity. Recently, with the fast development of deep learning techniques, deep learning-based feature extraction approaches have been widely used in radiomics (Visvikis *et al.*, 2019). Deep learning is highly flexible and accurate in analyzing multimodal volumetric and dynamic medical images in a fully automatic and non-linear manner (Nie *et al.*, 2019). but image features extracted by sophisticated deep learning models are considered to be not human understandable. Therefore, it is critical to explore potential tools to increase their explainability (Barredo Arrieta *et al.*, 2020).

Combining radiomics with genomics leads to the field of radiogenomics, which has a goal of noninvasively uncovering the radiogenomic biomarkers that could indicate the heterogeneity of the disease (Rutman and Kuo, 2009). Besides the challenges in radiomics and genomics, radiogenomics also faces the unpaired data problem. Most advanced complex models usually have a huge number of parameters to be optimized. Hence, they require comparable large datasets for training the parameters (Mikołajczyk and Grochowski, 2018). A small dataset can not only lead

to inadequate stratification among training, validation, and testing, but also easily encounter unpaired data problems. The inadequate stratification issue might be solved by data augmentation (Hussain *et al.*, 2017), but the unpaired data problem might make the entire study infeasible. Scientific research often compares or associates two or more sets of measured data from the same set of samples. The data are defined as unpaired data if the sets of data arise from separate individuals (Jin *et al.*, 2019; Almahairi *et al.*, 2018). For example, a dataset may contain genomic and survival information for a large number of patients, but only a subset of the patients have medical image data. Due to the small size of the image subset, the patients' prognostic information may not be adequate to perform survival analysis (e.g., patients with medical images might be all alive so there is no death event observed in the data). Hence the prognostic values of the image features could not be evaluated (i.e. the image features cannot be identified as prognostic biomarkers) (Burnside *et al.*, 2016a). Effective utilization of the unpaired imaging, genomic, and clinical data should be considered wisely. To solve the challenge, Gevaert et al. proposed a leverage strategy (O. Gevaert *et al.*, 2012), where we could first generate radiogenomic correlation maps between the radiomic features and the genomic features, then the correlation map can be checked to select correlated genomic features for predicting individual radiomic features. The well-trained predictive model can be applied to a genomic dataset without paired images but with sufficient long-term clinical outcome data. This leveraging strategy could be used together with explainable deep learning techniques to identify deep radiogenomic biomarkers from unpaired deep image features and clinical outcomes by leveraging the genomic features.

1.3 Solutions

Section 1.2 discusses some of the current challenges in BC heterogeneity studies. This section intends to summarize some previous works, which lead to potential strategies in solving those challenges. **Table 1-4** is a high-level summary of the contents to be discussed in this section.

Table 1-4. Current challenges and potential solutions in deciphering BC heterogeneity.

Challenges		Heterogeneity within BC subtypes	Integration of BC multi-omics data	complexity of BC multimodal data	
Solutions	Single/multi-omics data	Gene expressions	CNVs, Gene expressions, DNA methylations	Medical images, CNVs, Gene expressions, DNA methylations	Graph structured proteomics
	Machine learning/ deep learning methods	Unsupervised clustering	Tensor Factorization	Deep learning-based radiogenomics framework	Graph neural network

1.3.1 BC multimodal data

As shown in **Figure 1-4,** the biological information of BC can be captured at different omics levels. Starting from the genomic level, CNVs are structural variants in the genome that involve changes of DNA copies in the certain regions (Feuk *et al.*, 2006) (Shlien and Malkin, 2009). These DNA regions are normally larger than one kilobase pairs and may span many different genes. CNVs can arise by deletions (which can lead to gene copy number loss) or duplications (which can lead to gene copy number gain). In principle, a gene has two copies in the

diploid genome of normal human cells, with each copy inherited from each parent. But in cancer cells, the number of copies of some genes becomes smaller (gene copy number loss) or greater (gene copy number gain) than two and thus decreases or increases expression of these otherwise normal genes, respectively. CNVs are considered to be detrimental because they alter gene expression patterns and thus lead to an imbalance in cell growth and differentiation (Gonçalves *et al.*, 2017). For a long time, microarray-based methods (such as Array Comparative Genomic Hybridization and Single Nucleotide Polymorphism genotyping arrays) have been used to detect genome-wide CNV, which are efficient for large CNV detection (Fang and Wang, 2018). As next-generation sequencing is widely used due to the rapid price decreases, CNV studies are increasingly using sequencing for detecting small or novel CNVs that array-based methods often miss (Shen *et al.*, 2019) (Abel and Duncavage, 2013) (Li *et al.*, 2015).

Moving from genomics to epigenomics, DNA methylation is epigenetic modification that plays an important regulatory role in downstream gene expression (Robertson, 2005). In the human genome, the promoter regions of many genes contain CpG islands, and these regions contain a high frequency of CpG sites in which cytosine nucleotides are followed by guanine nucleotides (Smith and Meissner, 2013). DNA methylation can occur at CpG sites by adding a methyl group to the C5 position of cytosine in a CpG dinucleotide, forming 5-methylcytosine, a reaction catalyzed by DNA methyltransferases (Bird, 2002). Methylation of multiple sites within a CpG island is called hypermethylation. If hypermethylation occurs in the promoter, it eventually leads to transcriptional silencing of the gene (Jones, 2012) (Domcke *et al.*, 2015) (Blattler and Farnham, 2013). In normal cells, only a few promoter CpG islands are hypermethylated (Robertson, 2005). However, some cancers may have aberrant promoter CpG island hypermethylation in certain genes. This aberrant promoter CpG island hypermethylation can lead to inappropriate gene silencing

(Jones, 2012) (Chatterjee and Vinson, 2012). If it occurs in a tumor suppressor gene, cancer may occur. In the past two decades, more than 100 genes have been reported to be transcriptionally silenced in BC by aberrant CpG island hypermethylation (Khatri *et al.*, 2012; Jovanovic *et al.*, 2010; Basse and Arock, 2015; Pasculli *et al.*, 2018). Bisulfite conversion-based methods, such as Illumina methylation arrays (HumanMethylation450 and HumanMethylation850 arrays), and whole-genome bisulfite sequencing have been used to detect DNA methylation across the genome, with the latter being able to assess nearly every methylation status of CpG sites (Gupta *et al.*, 2010; Wang *et al.*, 2018).

As the result of genetic and epigenetic alterations, abnormal gene expression is one of the most important cancer mechanisms (Vogelstein *et al.*, 2013). As described in Section 1.1, gene expression is the most commonly used data type in the study of BC heterogeneity among other omics data. As an intermediate stage between DNA and proteins, transcriptome enables researchers to link a cell's phenotype (or proteome) to its molecular machinery (genome). When it comes to cancer, studying the transcriptome provides the link between the genetic causes of cancer and its phenotypic consequences. In just a few years, the key technology for profiling cancer transcriptomes has convincingly shifted from microarray-based to RNA sequencing-based technologies (Van den Berge *et al.*, 2019). For RNA- sequencing, RNA molecules are first extracted from a sample of interest and then reverse transcribed into cDNA molecules. The cDNA library is then sequenced to generate millions of short reads. The generated reads can be aligned by mapping to a reference genome or by de novo assembly (Van den Berge *et al.*, 2019). After mapping sequence reads to a reference genome, the next step is to quantify transcript-level or gene-level abundance by assigning mapped reads to specific transcripts or genes (Van den Berge *et al.*, 2019). The quantified results are combined into an expression matrix with each feature (gene or

transcript) in a row and each sample in a column, where the values are actual read counts or estimated abundances. The CNV/DNA methylation/gene expression data are usually in a matrix format where rows are patients and columns are genes (**Figure 1-5**).

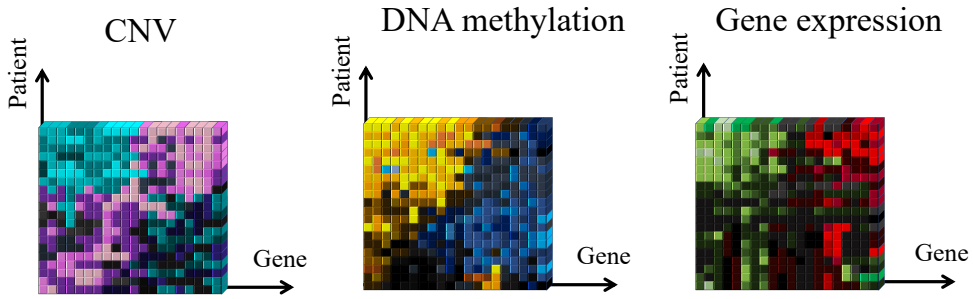


Figure 1-5. An illustration of the matrix format CNV/DNA methylation/gene expression data.

As the products of gene expression, proteins are the building blocks of all cells in our bodies. Protein can be represented by its amino acid sequence which can be determined by the mass spectrometry techniques (Zhang, 2009). However, the function of a protein is largely determined by its three-dimensional (3D) conformation, thus extracting numeric features that can represent a protein's 3D structure is of great value (Fasoulis, 2021). Besides the structure, proteins also achieve phenotypic function by interactions and associations (Rao *et al.*, 2014). The physical protein-protein interactions (PPI) are often quantified from purified proteins using techniques such as two-hybrid systems, mass spectrometry, phage display, and protein chip technology (Zhang, 2009). The purification can be affinity purification, Y2H (yeast 2 hybrid), TAP (tandem affinity purification) and so on (Zhang, 2009). Proteomic information is essential for deciphering BC heterogeneity from a systematic perspective. It has become more and more important in BC functional subtyping, staging, recurrence prediction, and resistance assessment (Neagu *et al.*, 2021). As mentioned in Section 1.2.3, unlike the matrix format CNV/DNA methylation/gene expression data (**Figure 1-5**), proteomic data could be modeled as a graph where vertices are

proteins, and edges are undirected and weighted connections between the interacting proteins. **Figure 1-6** shows an example of proteomic data in a graph structure.

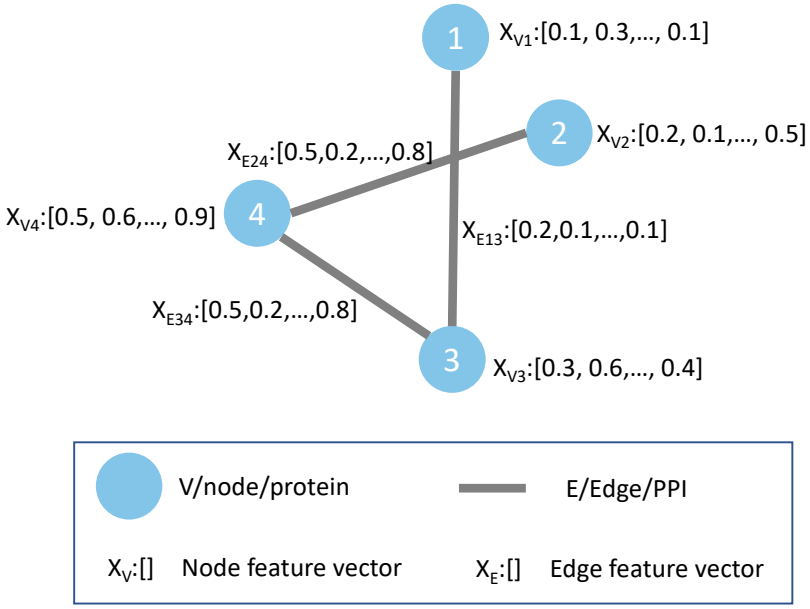


Figure 1-6. An example of attributed graph structure of proteomic data. An attributed graph *G(V, E, X_V, X_E)* could represent both the structure information and the node/edge features (X_V and X_E).

Medical image data are also quite different from the matrix format CNV/DNA methylation/gene expression data and graph format proteomic data. A two-dimensional (2D) image is also a matrix but with the rows and columns representing the x-axis (horizontal) and y-axis (vertical) plane coordinates. Each element in the plane is called a pixel, and the value of the pixel in position (*x, y*) represents the intensity of the information captured at the specific coordinate (*x, y*). Similar to the 2D images, a 3D image has the same format but with an additional axis z. **Figure 1-7** shows examples of a 2D image and a 3D image.

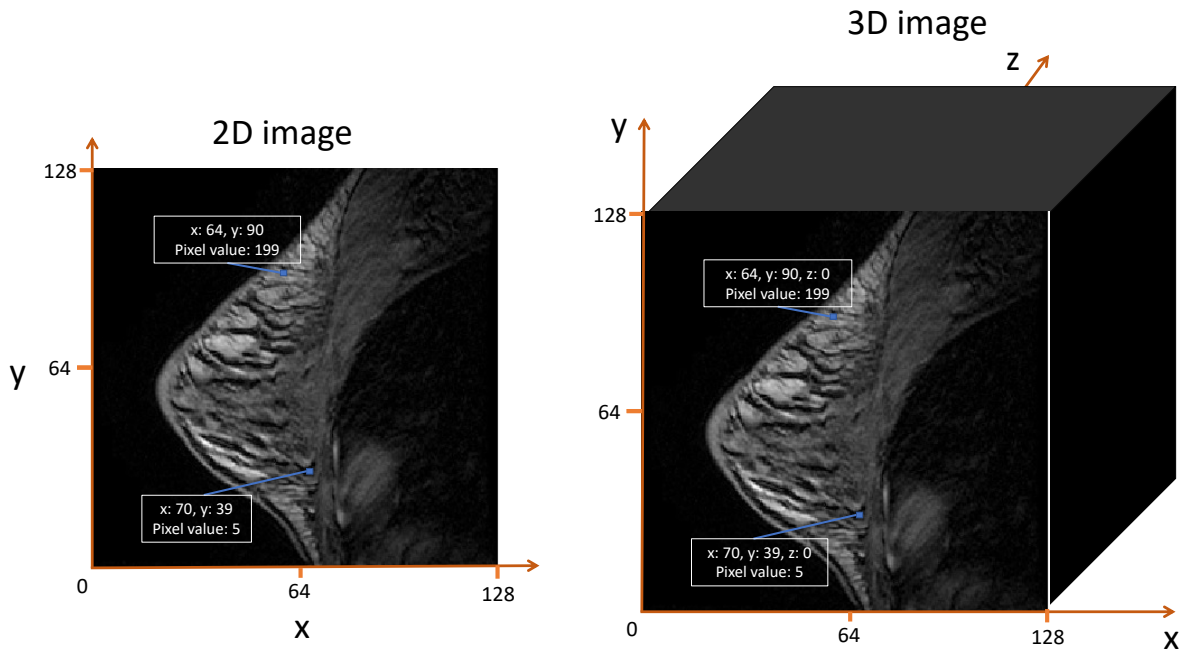


Figure 1-7. Examples of a 128 × 128 size 2D image and a 128 × 128 × 128 size 3D image.

Currently, there are many medical imaging technologies such as X-ray, Computed Tomography, Ultrasound, Magnetic Resonance Imaging (MRI), etc. Each type of imaging technology gives different information about the area of the body being studied. Among these imaging types, MRI is considered the most complex one (Runge, 2015) and is widely used in BC radiogenomic studies. MRI machines use very strong magnetic fields and radio waves to make pictures. During an MRI exam, an electric current is passed through a coil, creating a temporary change in the magnetic field in the patient. While recovering the changes, radio waves are received from a transmitter/receiver in the machine and these signals are used to make a digital image of the area of the body being scanned (Brown *et al.*, 2014). These signals mainly come from protons in body fat and water molecules (Brown *et al.*, 2014). For some MRI exams, a contrast agent (such

as gadolinium) is injected into a vein in the patient's arm to change the contrast of the MRI. The resulting MRI is called dynamic contrast enhanced MRI (DCE-MRI).

1.3.2 Machine learning for understanding BC heterogeneity

Artificial intelligence (AI) is a computer science technique used to train a computer how to learn, reason, perceive, infer, communicate and make decisions to simulate what humans do with their intelligence (Moor, 2006). Machine learning is a subfield of AI and is the process by which a computer learns from its experience and improves its performance over time without explicitly being programmed (Moor, 2006). A typical machine learning project has four main steps: data collection, data preparation, model training, and model testing. There are two major subcategories of machine learning: unsupervised and supervised. Unsupervised machine learning models look for patterns or trends in unlabeled data. Principal component analysis (PCA), matrix decomposition (Brunet *et al.*, 2004), and tensor factorization (Kolda and Bader, 2009; Khan and Ammad-ud-din, 2016) fall into the unsupervised machine learning category. Supervised machine learning models are trained with labeled data, which allow the models to learn and become more and more accurate over time. They are carried out frequently in solving data science problems (Kotsiantis *et al.*, 2006). Examples of supervised machine learning algorithms are linear/logistic regression, support vector machine (SVM), random forest (RF), supervised deep artificial neural networks, etc.

Machine learning has been applied to BC studies for decades (Tang *et al.*, 2009). Unsupervised machine learning methods such as consensus clustering has been used in omics-based cancer subtyping (Monti *et al.*, 2003). Traditional subtyping methods such as K-mean (Lloyd, 1982) and hierarchical clustering have some limitations. For K-mean clustering, a

predefine K, which is the number of clusters, is needed. However, for most of the unsupervised clustering problems, the number of clusters is unknown. Although hierarchical clustering could provide us tree-based results, a predefined cut-off point is still needed to decide the number of clusters. However, CoC is a resampling-based clustering algorithm which could estimate the number of clusters and obtain robust clustering results according to the consensus among several clustering runs (Monti *et al.*, 2003; Wilkerson and Hayes, 2010). Several multi-omics data integration models introduced in Section 1.2.2 such as SNF, CP, BTF, etc. are actually machine learning based.

Among different types of machine learning, deep learning has become the most popular learning strategy in recent years due to its excellent performance in many fields such as computer vision and natural language processing (Lecun *et al.*, 2015). Deep learning is inspired by the human brain, in which tremendous amounts of processing nodes are interconnected and organized into many layers. Each node processes inputs and produces an output that is mapped to the nodes in next layer. Different nodes would assess the information layer by layer and make a final output that indicates the label information. Deep learning models vary in terms of input data formats, processing pipelines, network structures, and learning targets, which result in a series of variations of deep learning such as deep convolutional neural networks (DCNNs) (Krizhevsky *et al.*, 2017), deep recurrent neural networks (DRNNs) (Jozefowicz *et al.*, 2015), deep graph neural networks (GNNs) (Gori *et al.*, 2005), deep reinforcement learning (Mnih *et al.*, 2015), generative adversarial networks (GANs) (Goodfellow *et al.*, 2014), etc.

Deep learning has been introduced to BC single omics and multi-omics studies, where the input of the models could be the omics information, and the output could be a variety of clinical or biological outcomes depending on the objectives of the studies. Due to the extraordinary success

of deep learning in solving computer vision tasks, it has been involved in many BC radiogenomics studies. The first deep learning radiogenomic study of BC was published in 2018 by Yoon *et al* (Yoon *et al.*, 2018). The study objective was to predict clinical characteristics such as the tumor stage, ER, PR, and HER2 status. They used multimodal three-dimensional MRI data (T1 and T1-contrast) and 70 risk gene expressions as input. A deep learning model was used to extract radiomic features and map the extracted radiomic features along with gene expressions to clinical characteristics (Yoon *et al.*, 2018). Ha and his team published two studies about BC radiogenomics using deep learning (Ha, Chang, *et al.*, 2019; Ha, Mutasa, *et al.*, 2019). Manually generated MRI tumor segments were obtained first, then deep learning was used to map tumor segments to BC intrinsic subtypes (Luminal A, Luminal B, HER2-enriched, Basal-like) and Oncotype Dx recurrence risk scores, respectively (Ha, Chang, *et al.*, 2019; Ha, Mutasa, *et al.*, 2019). Zhu *et al.* used pre-trained deep learning models based on ImageNet dataset (Deng *et al.*, 2010) to extract radiomic features from breast MRI, then input them into a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) model for final molecular subtypes classification (Zhu *et al.*, 2017). Zhang *et al*. applied advanced deep learning techniques to improve breast tumor auto-segmentation, then extracted several handcraft features from the segments. They claimed these handcraft features could be applied to classify patients' intrinsic subtypes (Zhang *et al.*, 2019).

Deep learning is also a good way to analyze the complex graph structured proteomic data. Recently, graph neural network (GNN) (Gori *et al.*, 2005) has shown great capability in automatically analyzing biological graphs, and it is also powerful in combining other auxiliary data sources to achieve best model performance (Long *et al.*, 2020, 2022, 2021). For example, Yang *et al.* proposed to use a GNN called signed variational graph autoencoder (S-VGAE) to predict the interactions between proteins using the protein-protein interaction (PPI) graph (F. Yang

et al., 2020). This model can effectively take both the graph structure and protein sequence information into consideration. Similarly, Jha et al. constructed a GNN with graph convolutional network (GCN) and graph attention network (GAT) techniques to learn integrated protein features from both the graph structure and the protein sequence information in a PPI graph (Jha *et al.*, 2022).

1.4 Rationale and objectives

Understanding BC heterogeneity is critical to BC management. Although previous studies have made extraordinary contributions in exploring BC heterogeneity and translating the results into BC clinic, there are still some key gaps, such as the BC "sub-heterogeneity" (the heterogeneity within a specific BC subtype), BC cross-level heterogeneity, BC high-throughput data complexity, etc. Machine learning/deep learning has achieved unprecedented success in many fields, including human healthcare.

To help human experts explore BC heterogeneity from multi-genomics data and medical images using machine learning/deep learning, several challenges need to be addressed. The first challenge is how to effectively combine the multi-modal biodata and find comprehensive representatives from them. Another challenge is how to address the executional infeasibility caused by the problem of unpaired data because different subjects may have different modality data. Currently, the publicly available datasets might be incomplete for us to do an imaging biomarker-oriented multi-omics study. Besides, the model interpretability and privacy issues should also be carefully taken into consideration in machine learning-based BC research.

This thesis aims to explore the BC heterogeneity using thriving machine learning/deep learning algorithms at different levels ranging from single genomics, multi-genomics, to proteomics and radiogenomics. Chapter 2 aims to stratify HER2+/ER+ BCs and create a

prognostic and prediction signature using single genomics data (gene expression). We hypothesized that there are distinct subtypes identifiable in gene expressions in HER2+/ER+ BCs. We also hypothesis that there are representative genes that can be organized into a gene signature with prognostic and predictive significance. Chapter 3 aims to stratify BC using multiple genomics data. We hypothesis that BTF algorithm could learn representations from both gene expression, CNV, and DNA methylation data which maintains the cross-omics information. We also hypothesis that the learned features can be used to stratify BC into prognostic significant subtypes which outperforms single genomics-based BC subtyping. Chapter 4 aims to construct a biological system hierarchy based on graph structure proteomics data and then identify clinical hotspots from the hierarchy. We hypothesis that GNN could generate the hierarchical tree with state-of-the-art performance. We also hypothesis that the minor survival hazard and rare mutation of each gene in the human genome could be well-studied in biological system levels. Chapter 5 aims to build a framework for BC prognostic radiogenomic biomarker identification. We hypothesis that the deep learning-based BC DCE-MRI features can be good surrogates for integrated multiple genomic information of BC. We also hypothesis that the mediation analysis among multi-omics features, DCE-MRI features, and clinical outcomes could improve the explainability of the deep learning-based radiogenomic biomarker identification framework.

2 Chapter 2: Gene expression based HER2+/ER+ BC heterogeneity study

2.1 Introduction

In the past, patients with HER2+ tumors had a poor prognosis among all subtypes of invasive BC. With the development of targeted anti-HER2 therapy, such as trastuzumab and lapatinib, survival rates for both early- and late-stage HER2+ disease have increased (Rye *et al.*, 2018) with varying degrees. Though HER2+ BC is a heterogeneous disease showing different relapse, survival outcomes, and treatment responses, there are currently no biomarkers to predict their clinical outcomes (Brandão *et al.*, 2020). Recently, ER status is emerging as a robust predictive marker within HER2+ BCs. Approximately half of all HER2+ BCs co-express HRs, namely ER+ and/or PR+ (Gingras *et al.*, 2017). Of HER2+/HR+ BC patients, over 95% are HER2+/ER+, who usually have better survival in the first 5 years following a diagnosis but have higher recurrence risk after 5 years compared to those with HER2+/ER- tumors (Hwang *et al.*, 2019; Cameron *et al.*, 2017), possibly due to the fact that sustained HR (mainly ER) signalling helps tumor escape from HER2 blockade (Bender and Nahta, 2008; Gingras *et al.*, 2017). Multiple studies found that HER2+/ER+ patients treated with anti-HER2 therapy showed lower pathological complete response (PCR) rates than patients with other types of BCs (Gianni *et al.*, 2012; Baselga *et al.*, 2012; Carey *et al.*, 2016). Given the predictive value of ER status, it would be reasonable for the next wave of clinical trials to target HER2+/ER+ and HER2+/ER- patients separately. As well, the development of new therapeutic strategies is of utmost importance to overcome the limitations to targeted therapies and improve treatment for HER2+/ER+ BC. Currently, there is no comprehensive study that focused specifically on HER2+/ER+ BC. Thus, a

better understanding of the molecular diversity of the ER+/HER2+ BC could pave the way to breakthroughs in HER2+/ER+ treatments.

Gene expression has been used in cancer stratification and gene signature identification since 1999 (Qian *et al.*, 2021). Gene expression signatures have been very successful in helping improve patient care by classifying tumors into distinct groups, providing guidance for personalized clinical decisions. Molecular classification of BC based on gene expression profiles has been extensively explored in the past and has been approved to be a good strategy. Several gene expression based signatures are successfully commercialized such as Prosigna (rorS) (Joel S. Parker *et al.*, 2009), Oncotype DX, EndoPredict (Paik *et al.*, 2004; Filipits *et al.*, 2011), PIK3CA-GS (Loi *et al.*, 2010), MammaPrint/GENE70 (Joel S. Parker *et al.*, 2009; van 't Veer *et al.*, 2002), GENIUS (Bontempi *et al.*, 2010), GGI (Sotiriou *et al.*, 2006), and so on. However, these gene signatures were not designed for HER2+/ER+ BC patients. Currently, there is no such gene signature for HER2+/ER+ BC patients. Thus, exploring the gene expression profiles of HER2+/ER+ BC and generating prognostic and predictive gene signatures is critical for better HER2+/ER+ BC clinical guidance.

To better understand the heterogeneity of HER2+/ER+ breast tumors, we sought to improve the current HER2+/ER+ BC stratification using gene expression profiles, aiming to identify reproducible HER2+/ER+ BC subgroups with well-established prognostic gene signatures. We applied unsupervised clustering on Cox regression filtered genes for HER2+/ER+ BC stratification and subsequently performed genome-wide expression differential analysis between the identified HER2+/ER+ BC subgroups. A gene expression signature was generated based on the significant genes in both Cox regression analysis and gene expression differential analysis. A

supervised subgroup classifier was then trained based on the expression of the proposed gene signature and validated in two independent HER2+/ER+ BC cohorts.

2.2 Methods

2.2.1 Data sources

An overall workflow of this study is shown in **Figure 2-1**. The raw read counts of RNA-sequencing data and clinical information of The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) cohort (Liu *et al.*, 2018) were downloaded through R package "TCGAbiolinks" (Colaprico *et al.*, 2016). TCGA-BRCA has 1098 patients, with each having more than 55,368 gene expression values. After filtering, there are a total of 123 HER2+/ER+ BC patients. The microarray-based gene expression data of METABRIC cohort (Curtis *et al.*, 2012) and Gene Expression Omnibus (GEO) (accession number GSE149283) BC cohort (Cecco, 2021) were also collected. METABRIC provides more than 24,368 gene expression patterns for all 1904 BC patients, with 104 patients being HER2+/ER+. GSE149283 provides neoadjuvant trastuzumab therapy response information and 24,352 gene expressions for 18 BC samples. Among these 18 samples, 14 are HER2+/ER+ BCs.

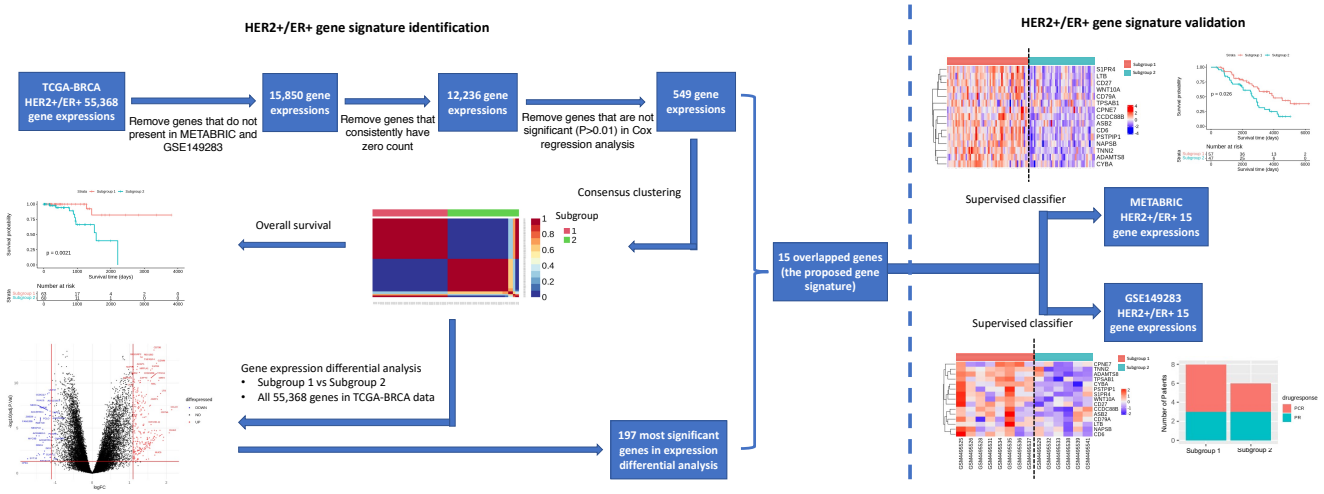


Figure 2-1. Overall workflow of this study. 15,850 genes are in common among TCGA-BRCA, METABRIC, and GSE149283 HER2+/ER+ patients. 12,236 of them with at least one count in one sample are kept and inputted into a Cox regression-based feature selection step which results in 549 significant genes based on the criteria of *p-value* < 0.01. Consensus clustering is then performed to stratify TCGA-BRCA HER2+/ER+ patients based on gene expression profile of these 549 significant genes. Gene differential analyses are done among the identified subtypes to identify most differentially expressed genes. Genes that are significant in both Cox regression analysis and gene expression differential analysis are selected to form the proposed gene signature. Validations of this gene signature are performed on METABRIC and GSE149283 HER2+/ER+ cohorts. An eXtreme Gradient Boosting (XGBoost) classifier is trained using the proposed gene signature on TCGA-BRCA data, then applied to assign METABRIC and GSE149283 BCs into two subgroups. For METABRIC, survival difference of the predicted subgroups is tested. For GSE149283, the drug response difference between the predicted subgroups is tested.

TCGA-BRCA was used as the discovery data in this study, while METABRIC and GSE149283 were used for validation. A total of 15,850 common genes present in all three cohorts was kept. Batch effect removal was performed among TCGA-BRCA, METABRIC and GSE149283 data using "ComBat_seq()" function in the "sva" R package (Leek JT, Johnson WE,

Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD, 2022). The gene expression data of TCGA-BRCA and normal samples before and after batch effect removal were visualized using principal component analysis (PCA) plots (**Figure 2-2A**). The PCA plots of the expression data from the three data sets before and after batch effect removal are shown in **Figure 2-2B.**

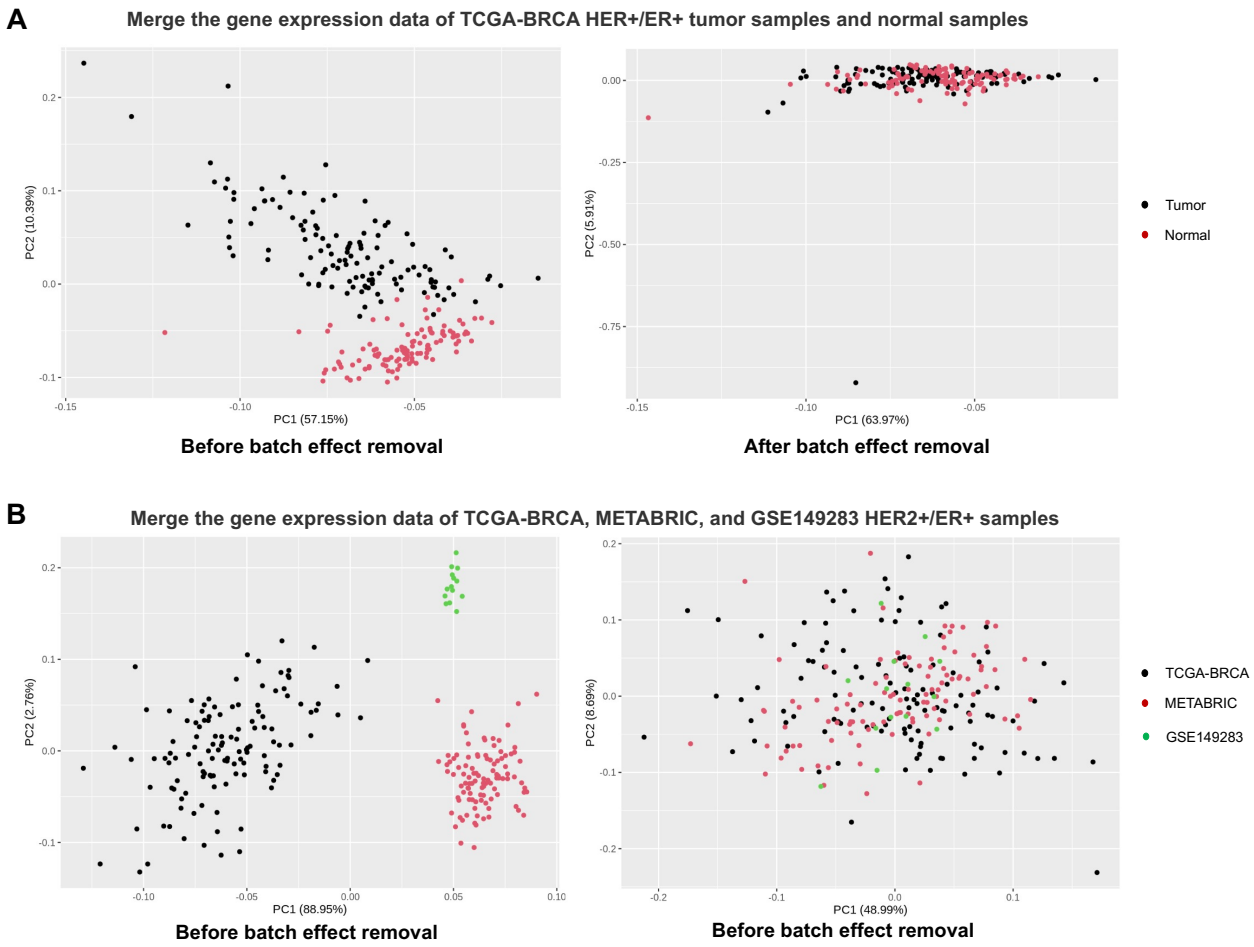


Figure 2-2. PCA plots of the data used in this study. A: The PCA plots of the TCGA-BRCA HER2+/ER+ tumor and normal gene expression data with and without batch effect removal. The batch effect removed data was used in the gene expression differential analysis between TCGA-BRCA HER2+/ER+ subgroups. **B:** The PCA plots of the three HER2+/ER+ datasets with and without batch effect removal. The batch effect

removed data were used in training the supervised subgroup classifier and evaluating the proposed 15-gene signature.

For the TCGA-BRCA data, we removed the genes that consistently have zero counts, which resulted 12,236 genes left for the following analysis. We first normalized the raw count data based on Trimmed Mean of M values (TMM) (Robinson and Oshlack, 2010) and then calculated the log transformed count per million (LogCPM) value using R package "edgR". We subsequently performed feature selection using Cox regression model with a cut-off *p-value* of 0.01. Genes significant in univariate Cox regression analysis were selected, resulting in 549 survival-associated genes. TCGA-BRCA also provides the gene expression date of 113 normal adjacent breast tissue samples. We used this normal expression data as reference for the later gene expression differential analysis. For GSE149283, the drug response difference between the predicted subgroups is tested.

2.2.2 Unsupervised clustering on discovery data

We used the R package "CancerSubtypes" (Xu *et al.*, 2017) to run the CoC algorithm (Monti *et al.*, 2003) on the TCGA-BRCA data matrix (549 genes × 123 patients). Traditional subtyping methods such as K-mean (Lloyd, 1982) and hierarchical clustering have some limitations. For K-mean clustering, a predefine K, which is the number of clusters, is needed. However, for most of the unsupervised clustering problems, the number of clusters is unknown. Although hierarchical clustering could provide us a tree-based results, a predefined cut-off point is still needed to decide the number of clusters. However, CoC is a resampling-based clustering algorithm which could estimate the number of clusters and obtain robust clustering results

according to the consensus among several clustering runs (Monti *et al.*, 2003; Wilkerson and Hayes, 2010).

In our case, CoC first subsampled the gene expression data matrix for 30 times. Then, non-negative matrix factorization (NMF) was applied on these 30 sample sets to obtain 30 clustering results. The maximum number of subtypes (k) was set as 10, which means we could expect up to 10 subtypes during each of the 30 clustering runs. The 30 clustering results for each subtype number (from 1 to k, which is 10 in our case) were then used to calculate the pairwise consensus value, which is defined as the probability of two items being clustered together (Monti *et al.*, 2003). These consensus values formed a consensus matrix for each cluster number. Which means, we obtained 10 consensus matrices, each one corresponding to a specific number of clusters. In the end, an agglomerative hierarchical consensus clustering was applied to each of the 10 consensus matrices, to obtain the final 10 clustering results.

Next, we calculated the silhouette value for each patient, which measures how similar a sample is to its own cluster compared to other clusters. The silhouette value ranges from -1 to 1 with a high value indicating that the sample is well matched to its own cluster and poorly matched to other clusters. If most samples have a high positive value, then the clustering configuration is appropriate. We then did survival analysis to test whether there are survival differences between the identified subtypes. The result was visualized using the Kaplan-Meier (KM) curve (Kaplan and Meier, 1958), which is widely used in clinical and healthcare fundamental research. It shows what the probability of an event (survival) is at a certain time interval.

2.2.3 Gene expression differential analysis for identified subgroups

We performed the differential analyses of each of the identified subgroups relative to the normal tissue samples in TCGA-BRCA cohort using R package "limma" (M. E. Ritchie *et al.*, 2015). We performed differential analyses between the identified HER2+/ER+ subgroups. The differentially expressed genes of each HER2+/ER+ subgroup versus the normal samples were then used to perform gene set enrichment analysis (GSEA; detailed below) to explore the enriched biological pathways. The differentially expressed genes between the identified subgroups were further used to filter for the genes for gene signature construction.

2.2.4 Gene set enrichment analysis for identified subgroups

GSEA was applied on the differentially expressed genes identified in the previous step, and only included genes with adjusted *p-value* less than 0.05 and |logFC| (absolute value of the log2 transformed fold change) larger than 0.5. We pre-ranked the genes based on their adjusted *p-values* in the differential analysis, then inputted them into the GSEA software (Subramanian *et al.*, 2005).

2.2.5 Gene signature creation and validation

Using eXtreme Gradient Boosting (XGBoost)-based supervised classification, we built a gene signature from the genes that are significant in both Cox regression analysis and the expression differential analysis between the identified HER2+/ER+ subgroups. The TCGA-BRCA data were then used to train and evaluate the XGBoost classifier in a 5-fold cross validation way using the R package "caret" (Kuhn, 2008). The model's performance was measured by the area under of the curve (AUC) of receiver operating characteristic (ROC), accuracy, sensitivity, and specificity. Shapley Additive Explanation (SHAP) values were calculated to increase the

interpretability of the XGBoost model (Lundberg and Lee, 2017). A higher SHAP value of a given feature in the model refers to stronger influence on the model's decision.

We applied the well-trained XGBoost model to classify the METABRIC and GSE149283 cohorts using the expression profile of the proposed gene signature. Survival analysis and hierarchical heatmap were also performed to visualize the predicted subgroup on METABRIC. For GSE149283, we visualized the difference of the neoadjuvant trastuzumab therapy response between each subgroups using bar plot.

We further checked the uniquely mutated genes in each subgroup for both TCGA-BRCA cohort and METABRIC cohort using the CBioPortal OncoPrint function (Gao *et al.*, 2013). The tumor immune estimation resource (TIMER) (Li *et al.*, 2016) quantified abundance of the tumor-infiltrating lymphocytes (TILs) (B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells), the PAM50 intrinsic subtypes and rorS, GENIUS, GENE70, GGI scores of each sample in both TCGA-BRCA and METABRIC cohorts were also tested.

The Cancer Dependency Map (DepMap) project provides systematically identified genetic and pharmacologic dependencies that were measured on CRISPR-Cas9 knockout screened cancer cell lines (Tsherniak *et al.*, 2017). To extend the cell line DepMap to tumors, Chiu et al. developed a deep learning model named DeepDep to predict the effect scores of the dependency of interest (DepOI) from genomics data (Chiu *et al.*, 2021). We applied the well-trained DeepDep on our HER2+/ER+ BC gene expression data to get the predicted effect score of the DepOIs for TCGA-BRCA and METABRIC cohorts, then checked the difference between the two HER2+/ER+ BC subtypes.

2.3 Results

2.3.1 Two distinct subgroups within HER2+/ER+ identified

To stratify TCGA-BRCA HER2+/ER+ patients, we performed CoC on survival significant genes. We first obtained 549 genes associated with survival outcome according to the genome-wide univariate Cox regression analysis. When the unsupervised CoC cluster number equals to 2, we obtained the most significant survival difference (*p-value* = 0.0021). The consensus matrix heatmap, silhouette plot, gene expression heatmap, and KM plot are shown in **Figure 2-3**. We observed a clear two-cluster pattern on both consensus matrix heatmap (**Figure 2-3A**) and gene expression heatmap (**Fig. 2-4**). The average silhouette value is 0.91, indicating two robust different subgroups existing in the TCGA-BRCA HER2+/ER+ cohort (**Figure 2-3B**). Sixty-three patients were assigned to Subgroup 1, while the other sixty patients were assigned to Subgroup 2. The survival difference between these two subgroups is significant (*p-value* = 0.0021). Patients in Subgroup 2 suffered poor prognosis (**Figure 2-3C**).

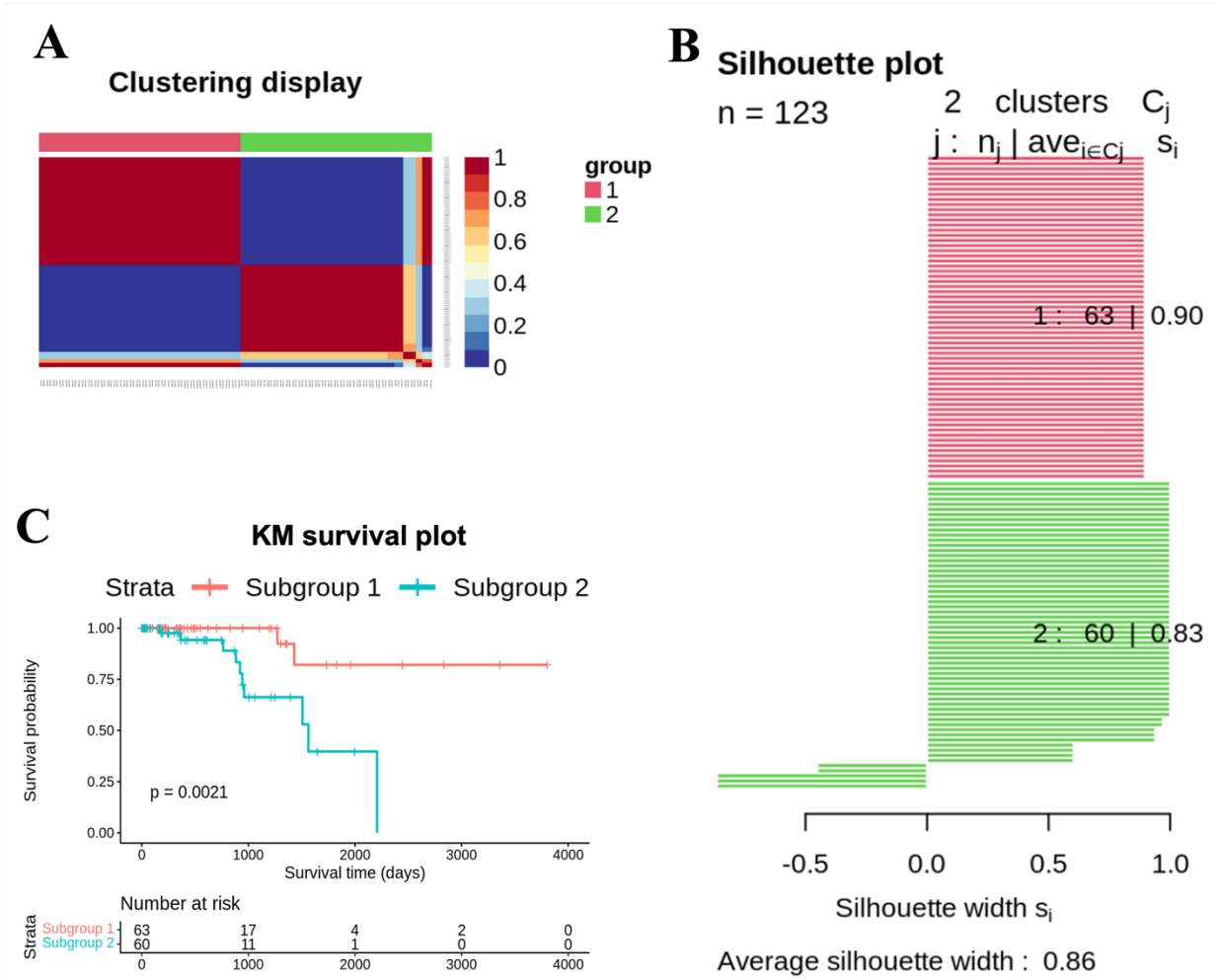


Figure 2-3. Results of consensus clustering on TCGA-BRCA data. A: Symmetric consensus matrix hierarchical clustering heatmap for TCGA-BRCA data. Columns and rows are patients. The color represents the probability that two patients were clustered together. **B:** Silhouette plot for the TCGA-BRCA data. Each horizontal line represents a sample, and the length of the line is the silhouette value for the sample. The color represents different subtypes: red ones are in Subgroup 1, while green ones are in Subgroup 2. A high value indicates that the sample is well matched to its own cluster and poorly matched to other clusters. If most samples have a high positive value, then the clustering configuration is appropriate.

The overall silhouette value is 0.91, which means the clustering is appropriate. C: KM plot of two subgroups identified by CoC.

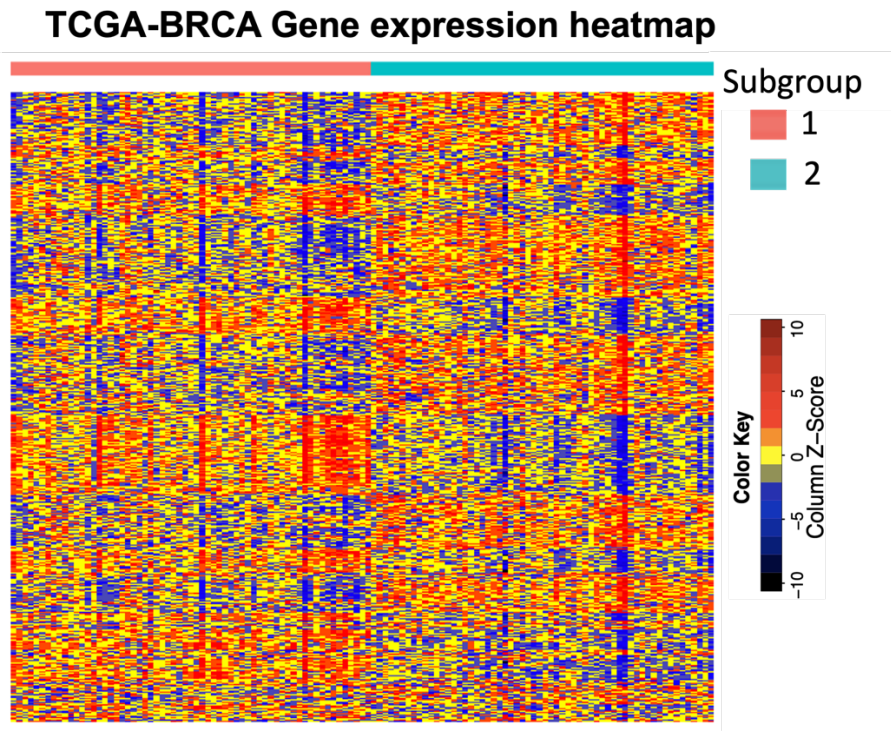
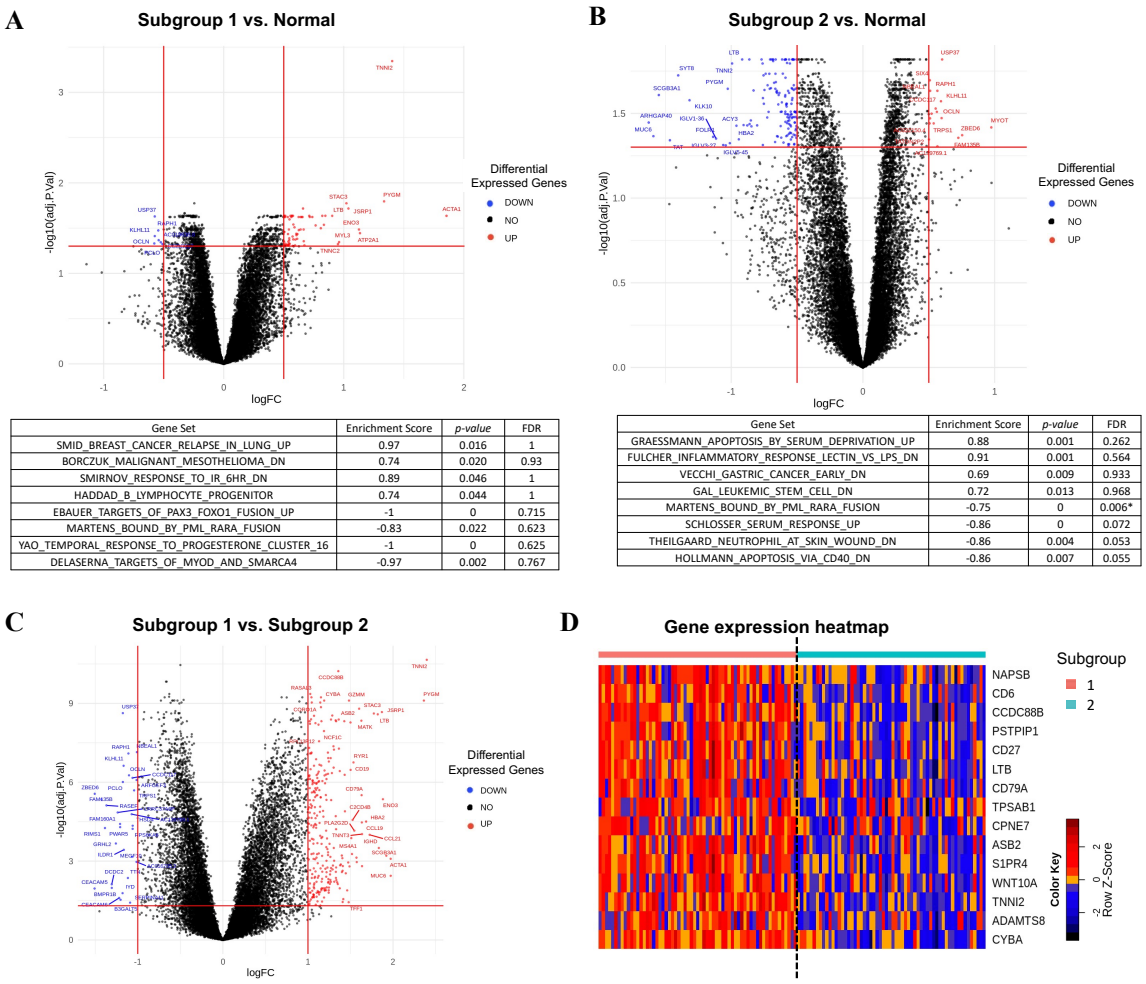


Figure 2-4. TCGA-BRCA gene expression visualization using heatmap. Columns are 123 patients, while rows are 549 genes. Unsupervised CoC results are shown in the topside bar.

In addition, from the differential analyses between Subgroup 1 versus normal and Subgroup 2 versus normal (**Figure 2-5**). Subgroup 2 has more differentially expressed genes than Subgroup 1 relative to normal. GSEA results (**Figure 2-5**) based on the two pre-ranked lists of differentially expressed genes (one list each for Subgroup 1 or Subgroup 2 versus normal) found that only the Martens Bound by Promyelocytic leukemia (PML) retinoic acid receptor alpha (RARA) Fusion pathway is significantly downregulated in Subgroup 2 relative to normal (false positive rate (FDR) = 0.006).



Gene Set	Enrichment Score	*p-value*	FDR
SMID_BREAST_CANCER_RELAPSE_IN_LUNG_UP	0.97	0.016	1
BORCZUK_MALIGNANT_MESOTHELIOMA_DN	0.74	0.020	0.93
SMIRNOV_RESPONSE_TO_IR_6HR_DN	0.89	0.046	1
HADDAD_B_LYMPHOCYTE_PROGENITOR	0.74	0.044	1
EBAUER_TARGETS_OF_PAX3_FOXO1_FUSION_UP	-1	0	0.715
MARTENS_BOUND_BY_PML_RARA_FUSION	-0.83	0.022	0.623
YAO_TEMPORAL_RESPONSE_TO_PROGESTERONE_CLUSTER_16	-1	0	0.625
DELASERNA_TARGETS_OF_MYOD_AND_SMARCA4	-0.97	0.002	0.767

Gene Set	Enrichment Score	*p-value*	FDR
GRAESSMANN_APOPTOSIS_BY_SERUM_DEPRIVATION_UP	0.88	0.001	0.262
FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS_DN	0.91	0.001	0.564
VECCHI_GASTRIC_CANCER_EARLY_DN	0.69	0.009	0.933
GAL_LEUKEMIC_STEM_CELL_DN	0.72	0.013	0.968
MARTENS_BOUND_BY_PML_RARA_FUSION	-0.75	0	0.006*
SCHLOSSER_SERUM_RESPONSE_UP	-0.86	0	0.072
THEILGAARD_NEUTROPHIL_AT_SKIN_WOUND_DN	-0.86	0.004	0.053
HOLLMANN_APOPTOSIS_VIA_CD40_DN	-0.86	0.007	0.055

Figure 2-5. Gene expression differential analysis results of the identified HER2+/ER+ subgroups. A top: Volcano plot of the gene expression differential analysis between TCGA-BRCA HER2+/ER+ Subgroup 1 and normal samples. The thresholds are highlighted in red lines (|logFC| > 0.5 and adjusted *p-value* < 0.05). **A bottom:** Gene set enrichment analysis (GSEA) result of the differentially expressed genes in TCGA-BRCA HER2+/ER+ Subgroup 1**. B top:** Volcano plot of the gene expression differential analysis between TCGA-BRCA HER2+/ER+ Subgroup 2 and normal samples. The thresholds are highlighted in red lines (|logFC| > 0.5 and adjusted *p-value* < 0.05). **B bottom:** GSEA result of the differentially expressed genes in TCGA-BRCA HER2+/ER+ Subgroup 2. **C:** Volcano plot of the gene expression differential

analysis between TCGA-BRCA HER2+/ER+ Subgroup 1 and Subgroup 2. There are 197 genes that are differentially expressed based on the significance thresholds highlighting in red lines (|logFC| > 1 and adjusted *p-value* < 0.05). **D:** The expression profile of the proposed 15-gene signature visualized using heatmap. Columns are 123 patients, while rows are 15 genes. Subgroup labels are showed on the top bar.

The demographic information and general BC clinical information of the two identified subgroups for TCGA-BRCA HER2+/ER+ cohort are shown in **Table 2-1**. Except for age (*p-value* = 0.0258), no other demographics or clinical differences exist between these two subgroups as the *p-values* (student t-test or chi-square test) are > 0.05.

Table 2-1. Demographic and clinical information of the identified HER2+/ER+ subgroups in TCGA-BRCA cohort.

		Subgroup 1	Subgroup 2	*p-value*
No. patients		63	60	
Age	min	34	29	0.02581*
	max	88	90	
	mean	57.13	62.65	
	Standard deviation	13.67	13.46	
T	T1, T1b, T1c	14	10	0.2133
	T2	41	39	
	T3	7	7	
	T4 (T4, T4b)	1	4	
N	N0, N0 (i-), N0 (i+)	31	23	0.2202
	N1, N1a, N1b	23	19	
	N2, N2a	6	11	
	N3, N3a	3	6	
	NX	0	1	

M	cM0 (i+), M0	50	50	0.1991
	M1	1	1	
	MX	12	9	
Stage	I, IA	7	9	0.2414
	II, IIA, IIB	43	28	
	IIIA, IIIB, IIIC	12	21	
	IV	1	1	
	X	0	1	
Surgery	Lumpectomy	8	4	0.2650
	Modified Radical Mastectomy	17	20	
	Simple Mastectomy	12	4	
	Other	19	28	
	Not Available	7	4	

The differential analyses of Subgroup 1 versus Subgroup 2 identified 197 genes differentially expressed (|logFC| > 1 and adjusted *p-value* < 0.05) (**Figure 2-5**). Among these 197 differentially expressed genes, 15 overlapped with the 549 survival-associated genes from Cox regression analyses: *TNNI2*, *CCDC88B*, *CYBA*, *ASB2*, *LTB*, *S1PR4*, *PSTPIP1*, *CD6*, *CD27*, *WNT10A*, *NAPSB*, *CD79A*, *ADAMTS8*, *CPNE7*, and *TPSAB1*. These 15 genes have both survival significance and subgroup distinguish significance and constitute the proposed gene signature. In this way, we decreased the gene number from 549 to 15 for easier application in other datasets.

2.3.2 HER2+/ER2+ BC gene signature development

A supervised classification model (XGBoost) was trained to classify the TCGA-BRCA HER2+/ER+ patients into the CoC identified two subgroups. The fine-tuning process and 5-fold cross validated performance of the XGBoost classifier are shown in **Table 2-2**. The 5-fold cross

validated AUC, sensitivity, and specificity on TCGA-BRCA data are 0.85, 0.76, and 0.77, respectively. The importance score of each gene in the XGBoost model are shown in **Figure 2-6**. TNNI2 is the most important feature for the model to make the prediction decision.

Table 2-2. Hyperparameter finetuning process and 5-fold cross validated model performance of the XGBoost classifier.

Learning rate	Max depth	AUC	Sensitivity	Specificity
0.005	6	0.84	0.78	0.73
	8	0.82	0.78	0.75
	10	0.84	0.78	0.75
	12	0.82	0.73	0.75
0.01	6	0.81	0.76	0.73
	8	0.83	0.81	0.75
	10	**0.85**	**0.76**	**0.77**
	12	0.83	0.79	0.77
0.05	6	0.81	0.79	0.77
	8	0.81	0.79	0.73
	10	0.82	0.78	0.72
	12	0.79	0.79	0.73
0.07	6	0.80	0.78	0.72
	8	0.82	0.76	0.75
	10	0.80	0.76	0.75
	12	0.80	0.79	0.72
0.1	6	0.80	0.78	0.72
	8	0.82	0.79	0.72
	10	0.80	0.81	0.75
	12	0.79	0.81	0.73

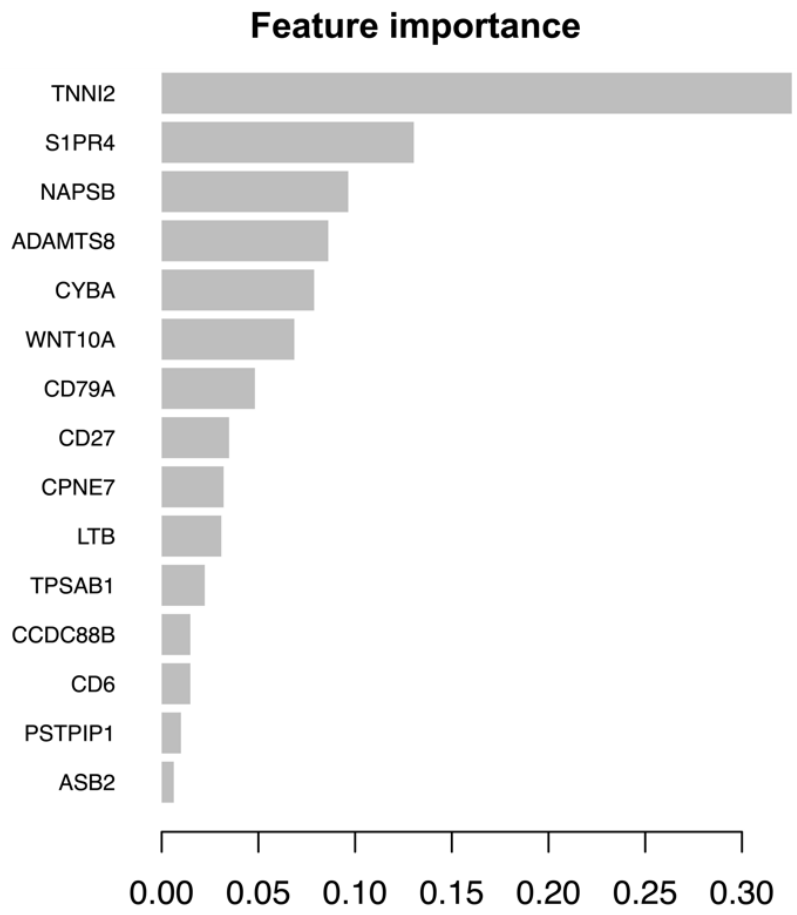


Figure 2-6. The SHAP importance scores of each gene in the XGBoost classifier.

2.3.3 HER2+/ER2+ BC gene signature validation

The well-trained XGBoost classifier was applied to two external datasets: METABRIC and GSE149283 for validation. The expression profiles of the 15-gene signature on METABRIC cohort and GSE149283 cohort are presented in **Figure 2-7A** and **Figure 2-7B,** respectively. Both METABRIC and GSE149283 patients assigned in Subgroup 2 showed overall lower expression values of the 15 gene signature than Subgroup 1, which is similar to that observed in the TCGA-BRCA cohort.

For the 104 HER2+/ER+ patients from the METABRIC cohort, 57 were assigned to Subgroup 1, while 47 were assigned to Subgroup 2 by the XGBoost model. The survival difference of these two subgroups was significant with Subgroup 2 showing worse survival than Subgroup 1 (**Figure 2-7C**), which is similar to that observed in the TCGA-BRCA cohort (**Figure 2-7C**). However, unlike the TCGA-BRCA cohort findings, there were no significant differences in demographic or clinical characteristics between these two subgroups (**Table 2-3**). For the 14 patients in GSE149283, eight patients were in Subgroup 1 and six patients were assigned to Subgroup 2. According to **Figure 2-7D**, there is a higher proportion of patients in Subgroup 2 (three out of six) showed partial response to trastuzumab than Subgroup 1 (three out of eight patients). However, the difference was not significant according to the Fisher's exact test (*p-value* = 1, odds ratio = 1.61).

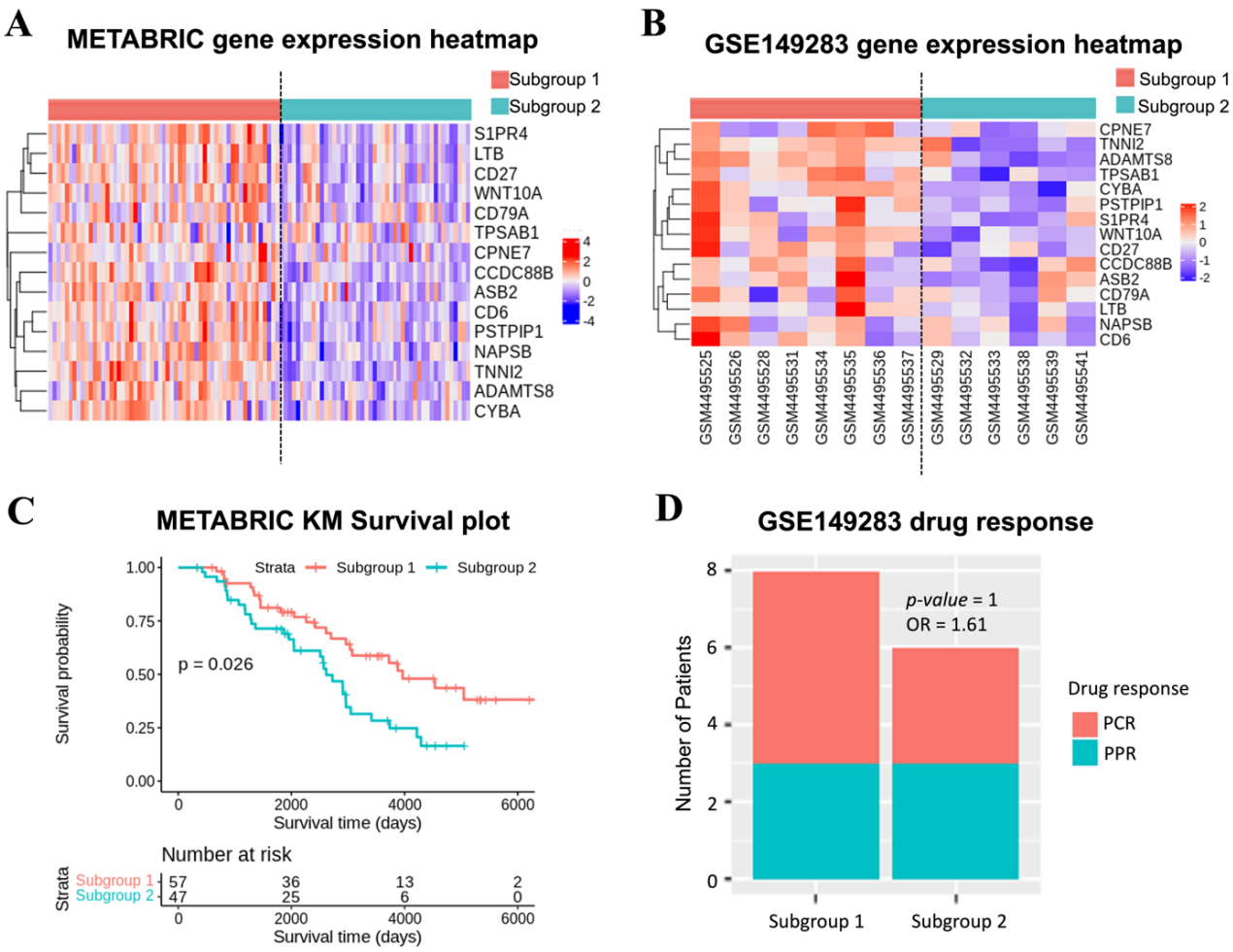


Figure 2-7. Predicted subgroups of external validation HER2+/ER+ BC cohorts. A: The expression profile of the proposed 15-gene signature on METABRIC HER2+/ER+ BC cohort. Columns are 104 patients, while rows are 15 genes. The XGBoost predicted subgroup labels are shown in the top side bar. **B:** The expression profile of the proposed 15-gene signature on GSE149283 HER2+/ER+ BC cohort. Columns are 14 patients, and rows are 15 genes. The XGBoost predicted subgroup labels are shown in the topside bar. **C:** KM plot of the two subgroups of METABRIC cohort predicted by XGboost. **D:** The stacked histogram of the trastuzumab therapy response for the XGBoost predicted subgroups (PCR, pathological complete response; PPR, pathological partial response; OR, odds ratio).

Table 2-3. METABRIC demographic information.

	Subgroup 1	Subgroup 2	*p-value*	
No. patients	57	47		
Age	Min	29	39	0.1035
	Max	87	87	
	Mean	60	64	
	Standard deviation	13.72	12.20	
Tumor size	Min	5	10	0.1074
	Max	70	50	
	Mean	27	23.17	
	Standard deviation	13.60	7.60	
Lymph nodes positive	Min	0	0	0.0894
	Max	25	25	
	Mean	1.83	3.68	
	Standard deviation	4.61	6.11	
Grade	1	2	1	0.2318
	2	18	14	
	3	35	32	
	Null	2	0	
Stage	0	15	9	0.2424
	1	14	8	
	2	13	19	
	3	2	2	
	4	1	0	
	Null	12	9	
Treatment	Chemotherapy (CT)	2	1	0.2578
	Radiotherapy (RT)	4	2	
	Hormonotherapy (HT)	12	9	
	CT/RT	3	2	
	CT/HT	0	3	

	CT/HT/RT	5	6	
	HT/RT	23	21	
	None	8	3	

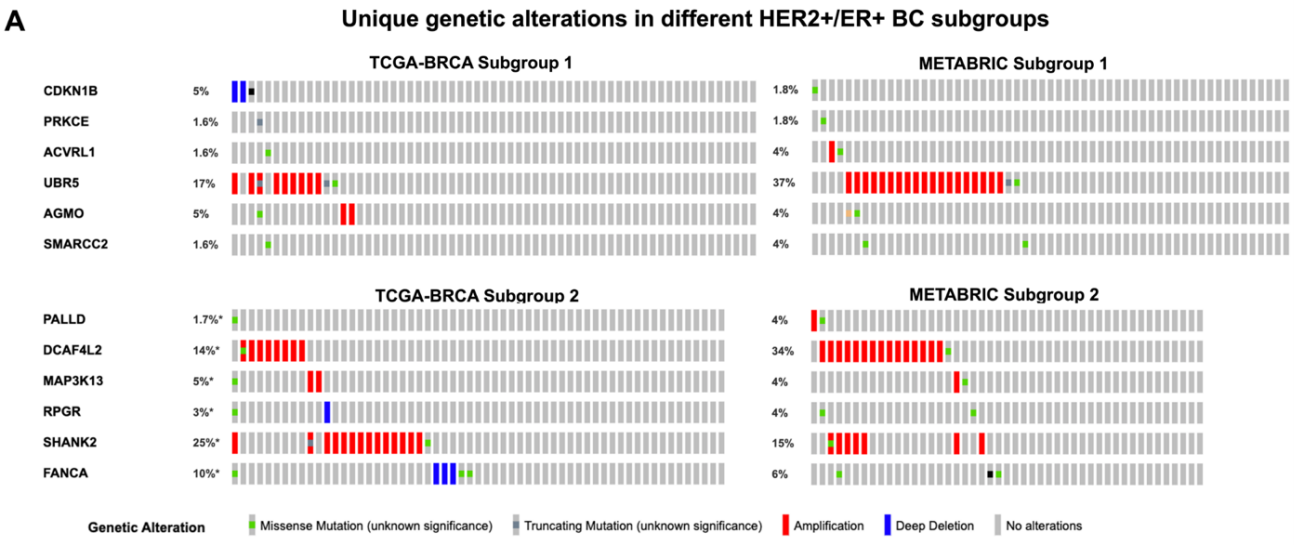
2.3.4 Computational characterization of the two subgroups (external validation)

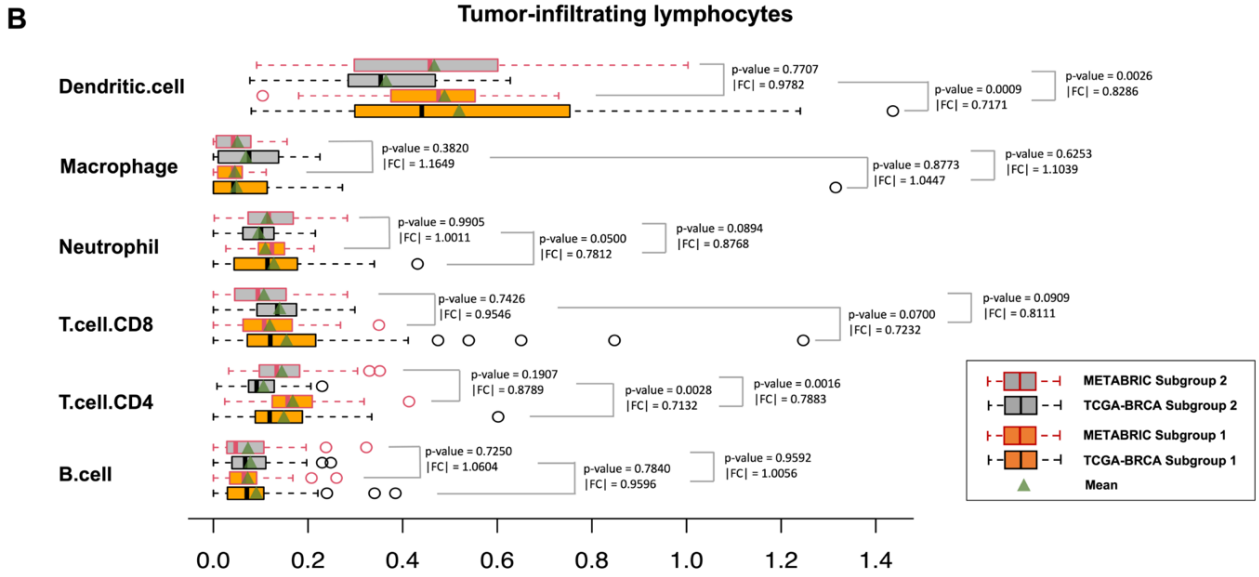
Uniquely mutated genes, TILs, PAM50 subtypes, some other published gene signatures such as rorS, GENIUS, GENE70, GGI scores, and DepMap dependency were calculated to characterize the identified two HER2+/ER+ subgroups. We found that there are 1914 mutated genes in the genome of TCGA-BRCA Subgroup 1 patients that were not observed within Subgroup 2 patients' genome, while TCGA-BRCA Subgroup 2 patients' genome have 3,293 mutated genes that are absent in Subgroup 1's genome. There are six genes (*CDKN1B, PRKCE, ACVRL1, UBR5, AGMO, SMARCC2*) commonly mutated in both TCGA-BRCA Subgroup 1 and METABRIC Subgroup 1 (**Figure 2-8A** top left). While another six common genes (*PALLD, DCAF4L2, MAP3K13, RPGR, SHANK2, FANCA*) are altered in both TCGA-BRCA Subgroup 2 and METABRIC Subgroup 2 (**Figure 2-8A** bottom left).

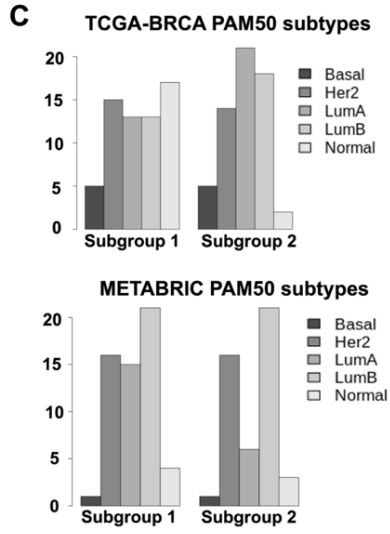
The quantified abundance of six immune cell types were estimated using TIMER on both TCGA-BRCA and METABRIC cohorts to check the TILs difference in two HER2+/ER+ subgroups (**Figure 2-8B)**. The infiltration of dendritic cells, neutrophils, and CD4+ T cells are significantly lower in TCGA-BRCA Subgroup 2 than in Subgroup 1 (*p-values* are 0.0009, 0.0500, and 0.0028, while absolute fold changes are 0.7171, 0.7812, and 0.7132). The combined Subgroup 2 patients (TCGA-BRCA Subgroup 2 plus METABRIC Subgroup 2) also show fewer dendritic cell (*p-value* = 0.0026, absolute fold change = 0.8286) and CD4+ T cell infiltrations (*p-value* =

0.0016, absolute fold change = 0.7883) than the combined Subgroup 1 patients (TCGA-BRCA Subgroup 1 plus METABRIC Subgroup 1).

Figure 2-8C and **Figure 2-8D** are showing the PAM50 intrinsic subtypes and the published gene signatures (rorS and GENIUS) of different subgroups in both TCGA-BRCA and METABRIC cohorts. There is a lower proportion of Normal PAM50 subtype in TCGA-BRCA HER2+/ER+ Subgroup 2 compared with TCGA-BRCA HER2+/ER+ Subgroup 1. For METABRIC, there is a lower proportion of Luminal A type in the HER2+/ER+ Subgroup 2 than Subgroup 1. TCGA-BRCA HER2+/ER+ Subgroup 2 showed significantly higher intrinsic rorS score and GENIUS score than Subgroup 1. However, other published gene signatures didn't show any difference between the two HER2+/ER+ subgroups in both cohorts. The predicted effect scores of DepOIs are visualized in **Figure 2-9**. Please be noted that only top 15 DepOIs with the most significant differences between HER2+/ER+ Subgroup 1 and Subgroup 2 are shown in the heatmaps. As can be seen, there are visible differences in TCGA-BRCA cohort, but not in METABRIC cohort. TCGA-BRCA HER2+/ER+ Subgroup 1 show lower dependency effect scores of the top 15 DepOIs.


A
Unique genetic alterations in different HER2+/ER+ BC subgroups
TCGA-BRCA Subgroup 1
METABRIC Subgroup 1
CDKN1B 5% 1.8%
PRKCE 1.6% 1.8%
ACVRL1 1.6% 4%
UBR5 17% 37%
AGMO 5% 4%
SMARCC2 1.6% 4%
TCGA-BRCA Subgroup 2
METABRIC Subgroup 2
PALLD 1.7%* 4%
DCAF4L2 14%* 34%
MAP3K13 5%* 4%
RPGR 3%* 4%
SHANK2 25%* 15%
FANCA 10%* 6%
Genetic Alteration
Missense Mutation (unknown significance)
Truncating Mutation (unknown significance)
Amplification
Deep Deletion
No alterations


B
Tumor-infiltrating lymphocytes
Dendritic.cell
Macrophage
Neutrophil
T.cell.CD8
T.cell.CD4
B.cell
p-value = 0.7707 |FC| = 0.9782
p-value = 0.0009 |FC| = 0.7171
p-value = 0.0026 |FC| = 0.8286
p-value = 0.3820 |FC| = 1.1649
p-value = 0.8773 |FC| = 1.0447
p-value = 0.6253 |FC| = 1.1039
p-value = 0.9905 |FC| = 1.0011
p-value = 0.0500 |FC| = 0.7812
p-value = 0.0894 |FC| = 0.8768
p-value = 0.7426 |FC| = 0.9546
p-value = 0.0700 |FC| = 0.7232
p-value = 0.0909 |FC| = 0.8111
p-value = 0.1907 |FC| = 0.8789
p-value = 0.0028 |FC| = 0.7132
p-value = 0.0016 |FC| = 0.7883
p-value = 0.7250 |FC| = 1.0604
p-value = 0.7840 |FC| = 0.9596
p-value = 0.9592 |FC| = 1.0056
METABRIC Subgroup 2
TCGA-BRCA Subgroup 2
METABRIC Subgroup 1
TCGA-BRCA Subgroup 1
Mean
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4


C
TCGA-BRCA PAM50 subtypes
Basal
Her2
LumA
LumB
Normal
Subgroup 1
Subgroup 2
METABRIC PAM50 subtypes
Basal
Her2
LumA
LumB
Normal
Subgroup 1
Subgroup 2

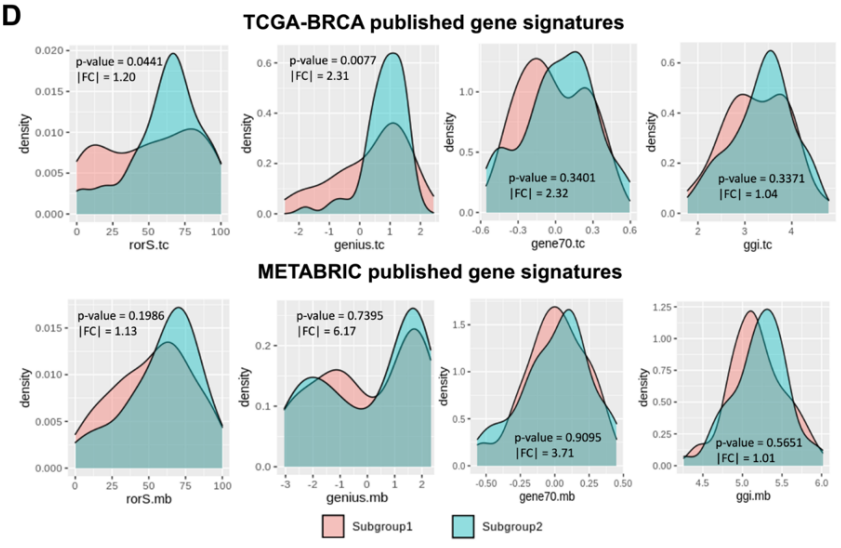

D
TCGA-BRCA published gene signatures
p-value = 0.0441 |FC| = 1.20
p-value = 0.0077 |FC| = 2.31
p-value = 0.3401 |FC| = 2.32
p-value = 0.3371 |FC| = 1.04
rorS.tc
genius.tc
gene70.tc
ggi.tc
METABRIC published gene signatures
p-value = 0.1986 |FC| = 1.13
p-value = 0.7395 |FC| = 6.17
p-value = 0.9095 |FC| = 3.71
p-value = 0.5651 |FC| = 1.01
rorS.mb
genius.mb
gene70.mb
ggi.mb
density
Subgroup1
Subgroup2

Figure 2-8. Computational characterization of the HER2+/ER+ subgroups for both TCGA-BRCA cohort and METABRIC cohort. A top panel: The common genes that are mutated in both TCGA-BRCA Subgroup 1 and METABRIC Subgroup 1. **A bottom panel:** The common genes that are altered in both TCGA-BRCA Subgroup 2 and METABRIC Subgroup 2. **B:** The TIMER quantified abundances of tumor-infiltrating lymphocytes for both TCGA-BRCA and METABRIC cohorts. T-test were used to test the significance of the differences. **C:** Histograms of the PAM50 intrinsic subtypes distributions for two subgroups. **D:** Density plots of the published gene signatures (rorS, GENIUS, GENE70, GGI) of different subgroups in both TCGA-BRCA and METABRIC cohorts.

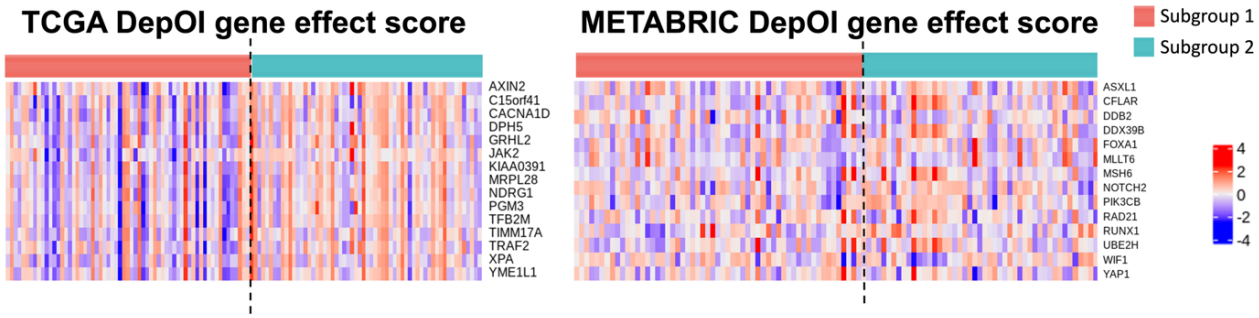


Figure 2-9. Visualization of the predicted effect scores of the DepOIs for TCGA-BRCA and METABRIC cohorts. Only top 15 DepOIs with most significant differences between HER2+/ER+ Subgroup 1 and Subgroup 2 are shown in the heatmaps.

2.4 Discussion

We have identified a 15-gene expression signature which could stratify HER2+/ER+ BC patients into two prognostically different subgroups in both unsupervised and supervised manners. The prognostic difference between the two HER2+/ER+ subgroups was observed in both TCGA-BRCA and METABRIC cohorts, not confounded by other clinical characteristics, including tumor

size, grade, or stage. The two subgroups also tend to exhibit difference in terms of their response to trastuzumab in GSE149283 with 14 samples, suggesting the predictive potential of the proposed 15-gene signature. However, no statistical significance was observed in the GSE149283 cohort, which might be due to the small sample size (n = 14). Thus, further validation is required once a large drug response dataset becomes available. According to the GSEA results of the differentially expressed genes, Martens Bound by PML RARA Fusion pathway was significantly enriched in Subgroup 2 but not in Subgroup 1. This pathway is a diagnostic marker of the acute promyelocytic leukemia (Martens *et al.*, 2010) and maybe used as a diagnostic marker for HER2+/ER+ Subgroup 2.

Six genetic alterations were found in Subgroup 1 that were not seen in Subgroup 2. Among them, *UBR5* amplification was observed in 17% TCGA-BRCA Subgroup 1 patients' genome and 37% METABRIC Subgroup 1 patients' genome. *UBR5* encodes a HECT-domain containing E3 ubiquitin ligase that is involved in regulating DNA damage response, cell cycle, metabolism, transcription, and apoptosis (Xiang *et al.*, 2022). Multiple studies have demonstrated that elevated expression of *UBR5* is implicated in different cancers, including breast and ovarian cancers, and is closely associated with advanced clinical stage, distant metastasis, and shorter overall survival in patients (Xiang *et al.*, 2022). *UBR5* exhibits oncogene-like characteristics as it is proposed to promote breast and ovarian cancer growth and metastasis, which makes it an attractive therapeutic target for aggressive BC (Song *et al.*, 2020). Our study further suggests that UBR5 might offer a potential way to target HER2+/ER+ Subgroup 1 as it is in Subgroup 1 (17% TCGA-BRCA HER2+/ER+ BC Subgroup 1 patients; 37% METABRIC HER2+/ER+ BC Subgroup 1 patients) but not in Subgroup 2. Similarly, several other genetic alterations were unique to Subgroup 2, including *DCAF4L2* amplification (14% TCGA-BRCA HER2+/ER+ BC Subgroup 2 patients; 34%

METABRIC HER2+/ER+ BC Subgroup 2 patients) and *SHANK2* amplification (25% TCGA-BRCA HER2+/ER+ BC Subgroup 2 patients; 15% METABRIC HER2+/ER+ BC Subgroup 2 patients). DCAF4L2 belongs to the WD-repeat domain (WDR) protein family, which commonly functions by mediating protein-protein interactions (Hu *et al.*, 2019). *DCAF4L2* overexpression in human colorectal cancer is associated with a more advanced clinical stage as in lymphatic and distant metastasis. Moreover, overexpression was also found to promote cell migration, invasion, and epithelial-mesenchymal-transition (EMT) through activating NFκB signal pathway (Wang *et al.*, 2016). High expression of *DCAF4L2* may be positively associated with poor overall survival of BC (Hu *et al.*, 2019). However, it remains to be determined how *DCAF4L2* is implicated in HER2+/ER+ BC pathology and whether it could be used as a novel candidate target for HER2+/ER+ treatment. *SHANK2* is one of the *SHANK* family of master scaffolding proteins and plays important roles in regulating synapse plasticity. The *SHANK* family proteins were recently found to be involved in cancer cell invasion (Liu *et al.*, 2021). Methylation of *SHANK2* could promote BC cell migration through activating endosome focal adhesion kinase FAK signalling (Liu *et al.*, 2021). FAK is a cytoplasmic protein-tyrosine kinase which is important in cell adhesion, survival and migration (Alanko and Ivaska, 2016). *SHANK2* methylation was also identified as a potential biomarker of BC metastasis (Liu *et al.*, 2021). However, the precise roles of *SHANK2* in HER2+/ER+ BC have yet to be determined.

The TIMER estimated that tumor infiltrations of two types of immune cells (dendritic cells and CD4+ T cells) were significantly lower in HER2+/ER+ BC Subgroup 2 than in Subgroup 1. According to Jin et al., lower tumor infiltration of these two types of immune cells were associated with worse prognosis (Jin and Hu, 2020), which is consistent with our findings that Subgroup 2 patients have lower TILs and worse prognosis. PAM50 intrinsic subtypes, the related rorS score,

and the Genius score also showed different distributions between the two HER2+/ER+ BC subgroups in TCGA-BRCA cohort, indicating a higher survival risk of Subgroup 2 than Subgroup 1. However, these differences were not observed in METABRIC HER2+/ER+ BC cohort. Other published gene signatures such as GENE70 and GGI were not significantly different between two subgroups in both TCGA-BRCA HER2+/ER+ and METABRIC cohorts, suggesting their unsuitability for HER2+/ER+ BC. DeepDep predicted DepMap dependency scores showed visible different patterns between the two subgroups in TCGA-BRCA HER2+/ER+ cohort according to our results, which further confirmed the difference between these two subgroups. However, this pattern cannot be reproduced in METABRIC HER2+/ER+ cohort.

In summary, we found that some of the differences between the proposed two HER2+/ER+ subgroups were observed in TCGA-BRCA cohort but not in METABRIC cohort, such as the tumor infiltrations of dendritic and CD4+ T cells, two published gene signatures (rorS and GENIUS), DeepDep predicted gene dependency scores, etc., possibly due to the different acquisition technologies used to obtain the raw gene expression data of these two cohorts (Bismeijer *et al.*, 2018). TCGA-BRCA gene expression data were obtained through RNA sequencing, while METABRIC used microarray technology. Another possible reason is that TCGA-BRCA biospecimens were collected from newly diagnosed patients who had received no prior treatment, while lymph node-positive METABRIC patients received chemotherapy before the biopsy (Liu *et al.*, 2018) (Curtis *et al.*, 2012).

2.5 Conclusion

In conclusion, our study is the first to explore the molecular heterogeneity associated with HER2+/ER+ BCs. We identified and validated the potential subgroups of HER2+/ER+ breast

tumors with reproducible prognostic and other properties. We provided a well-trained HER2+/ER+ subgroup classifier to assign new patients to a specific subgroup. We also discussed potential biological explanations of the identified subgroups and linked it with existing knowledge of BC. Most important, our findings may provide guidance for future target therapy of the HER2+/ER+ BC patients.

3 Chapter 3: Multi-omics-based tensor subtyping for BC

This chapter is modified from the published paper below. I performed all data analyses, paper writing and revision.

Q Liu*, B Cheng*, Y Jin, P Hu. 2022. Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data. Journal of Biomedical Informatics. 125:103958, https://doi.org/10.1016/j.jbi.2021.103958. (IF: 8). *These authors contributed equally to the work. I did the data analysis, manuscript writing, and manuscript revision with the assistant from Bowen Cheng and the guidance from Dr. Pingzhao Hu.*

3.1 Introduction

Currently, the majority of research in BC subtyping is still based on individual genomic datasets, especially gene expression data (Dai *et al.*, 2015). Fruitful achievements are obtained in gene expression-based subtyping such as the intrinsic classification scheme (Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like) (Joel S. Parker *et al.*, 2009), claudin-low extension of the intrinsic subtypes (Herschkowitz *et al.*, 2007), CIT subtyping scheme (Luminal A, Luminal B, Basal-like, and Normal-like) (Guedj *et al.*, 2012), and so on. Other than gene expression data, Jönsson et al. proposed an exclusively CNV-based classification scheme to distinguish six genomic subtypes in BC: 17q12, Basal-complex, Luminal-simple, Luminal-complex, Amplifier, and Mixed subtypes (Jönsson *et al.*, 2010). BC subtyping systems like the above-mentioned ones based on only one data type may fail to synthesize the inherent information across different levels of genomic data. For this reason, cancer subtyping using multi-omics data has its advantage in integrating information from different aspects, leading to a more

comprehensive subtyping (Huang *et al.*, 2017). For example, Tao et al. developed network-based algorithms to take advantage of multi-genomic information for studying a variety of cancer types, such as lung cancer, colon cancer, and rectal cancer (Huang *et al.*, 2015; Huang and Cai, 2016; Huang *et al.*, 2012). With thousands of BC samples having been extensively studied by a variety of molecular analyses during the last few decades, it is expected that integration of knowledge from multiple data types measured from the same individuals could further improve the performance of BC subtyping in predicting patient outcomes (Rappoport *et al.*, 2019).

According to Fang et al., there exist two kinds of data integration approaches: loose approach and tight approach (Fang, 2019). For the loose approach, different types of datasets are used to build separate models. Then the results from these models are combined in either a hierarchical, ensemble, or a linear fashion (Thingholm *et al.*, 2016; M. D. Ritchie *et al.*, 2015; Huang *et al.*, 2017). The advantage of this multi-staged method is its simplicity. However, loose integration is ineffective in exploring the complicated patterns that arise from combining distinct types of data (Fang, 2019). Therefore, the tight approach is considered to be more advanced than the loose method. For the tight approach, all different types of data are combined to fit a single subtyping model. The easiest way of tight integration is to concatenate all types of normalized data to establish a large matrix, then input the large matrix directly into the subtyping model (Fang, 2019; Huang *et al.*, 2017). However, this concatenation method usually ignores the fundamental complementarity of different types of data (Wang *et al.*, 2014). Therefore, it is suggested to combine different data sources in a more abstract and practical way (Huang *et al.*, 2017). A sophisticated example of combining multiple data in a practical way is the METABRIC project (Curtis *et al.*, 2012). It proposed a subtyping scheme by combining gene expression and CNV data. The METABRIC study first identified genes whose expression across breast tumors were driven

by recurrent CNVs that were cis (within a 3-megabase window surrounding the gene of interest) using an expression quantitative trait loci analysis. These cis-driven genes were then used to classify the tumors into 10 integrative cluster subtypes (IntClust1-10) with distinct clinical outcomes and responses to therapy (Curtis *et al.*, 2012). Another way of integrating multiple data in an abstract way is to first extract latent features from different data types and then use classic clustering algorithms to do subtyping. Examples include SNF (Wang *et al.*, 2014), iClusterPlus (Mo *et al.*, 2018), LRAcluster (Wu *et al.*, 2015), PINSCluster (Nguyen *et al.*, 2020), etc.

Some mathematic and computational methods were proposed to address this data combination problem in a more abstract way. The most popular one is the tensor factorization (Kolda and Bader, 2009). Instead of simply concatenating different omics data to a large two-dimensional matrix, we can combine them into a three-dimensional tensor with the new dimension representing data type. This three-dimensional tensor can thus be processed by traditional tensor factorization algorithms for latent factor extraction. In this way, data across different levels are integrated and the cross-level information is retained. These integrated factors are then passed on to downstream subtyping models.

Tensor is defined as a high-dimensional data array (Kolda and Bader, 2009). Tensor decomposition or factorization problem has been studied for many years and the algorithms that emerged to solve the problem have been applied in many domains of research (Kolda and Bader, 2009; Harshman, 1970; Carroll and Chang, 1970). Among those mature tensor factorization algorithms, CP (Carroll and Chang, 1970) (Harshman, 1970) perhaps is the simplest one. It decomposes a tensor as a sum of rank-one tensors (Kiers, 2000; Hitchcock, 1927; MÖcks, 1988). A drawback of CP-based tensor factorization is that pre-defining the rank of the tensor is a NP-hard problem (Håstad, 1990; Hillar and Lim, 2013). There was no straightforward algorithm for

solving this problem for a long time (Kolda and Bader, 2009) until recently, the BTF algorithm was proposed to overcome this difficulty (Tang *et al.*, 2018). BTF takes three steps to determine the rank of a tensor: a multi-linear model to decompose the given tensor to latent factors; a high-dimensional variational Bayesian inference method to estimate the posterior distribution of the decomposed latent factors; and finally a filtering procedure for removing the redundancy factors (Tang *et al.*, 2018).

After different levels of omics data are integrated into latent factors, a subtyping model is needed to identify BC intrinsic subtypes using these latent factors. The problem then becomes given a matrix with patients in rows and latent factors in columns, to identify the intrinsic clusters of these patients. CancerSubtypes (Xu *et al.*, 2017) is an R package for cancer subtyping and is widely accepted by omics data-based cancer stratifications. It envelopes some classical cancer subtyping algorithms such as CoC (Monti *et al.*, 2003), SNF (Wang *et al.*, 2014), SNF.CC (an ensemble method which combines SNF and CoC) (Xu *et al.*, 2017), Consensus NMF (CNMF) (Brunet *et al.*, 2004), and so on. CoC is a resampling-based clustering algorithm that intends to obtain robust data partitions according to the consensus among several clustering runs. It has the ability of intuitively estimating the number of clusters in a dataset (Monti *et al.*, 2003; Wilkerson and Hayes, 2010). SNF constructs fused sample-level similarity networks for patient subgrouping and has been proved to perform well in subclassification of various cancers (Wang *et al.*, 2014). CNMF incorporates CoC into the traditional NMF (Lee and Seung, 1999) method to stratify patients in a dimension reduction approach (Brunet *et al.*, 2004). NMF, in its core algorithm, is similar to the tensor factorization mentioned in the previous paragraph because a matrix can be considered as a 2D tensor. To this end, involving CNMF as the downstream subtyping method

after BTF can keep the entire workflow of multi-omics-based cancer subtyping under a uniform set of theoretical assumptions.

In this study, we propose a new computational subtyping strategy for BC that effectively interlocks multi-omics, BTF, and CNMF (**Figure 3-1**). A three-dimensional tensor is constructed first using the gene expression data, CNV data, and methylation data measured on 762 BC patients from TCGA-BRCA cohort (Liu *et al.*, 2018). This three-dimensional tensor is then passed on to BTF for latent factor decomposition. The decomposed factors, which contain the comprehensive information of the multi-omics data, are assigned to CoC wrapped K-mean clustering for subtype number determination. Then CNMF is applied for the patient subtype membership assignment. The identified BC subtypes are evaluated by statistical tests and survival analysis. The proposed approach is compared with other state-of-the-art methods. More details of these algorithms can be found in the method section.

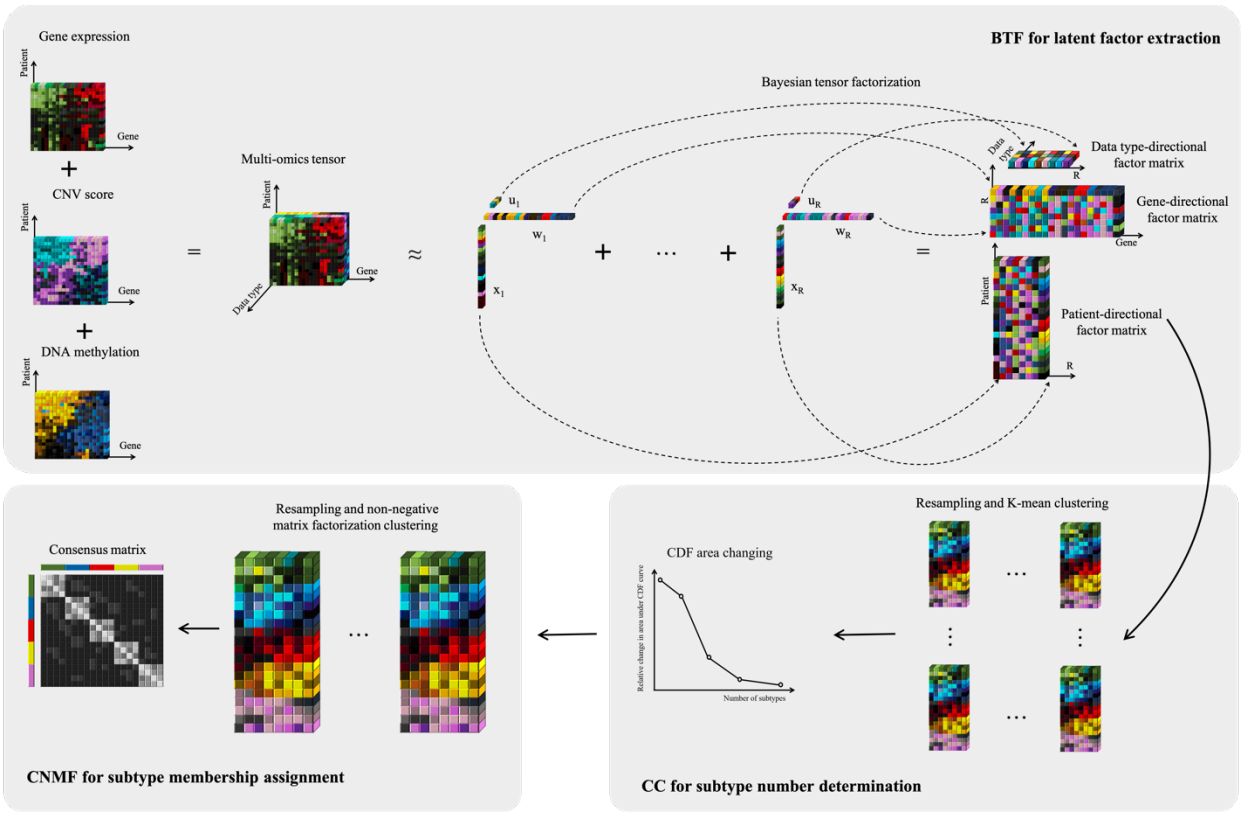


Figure 3-1. The overall workflow of the proposed BTF-CNMF method for multi-omics-based BC subtyping. A three-dimensional multi-omics tensor is constructed by attaching the gene expression data matrix, CNV data matrix, and methylation data matrix together. This three-dimensional tensor is then passed on to BTF for latent factors decomposition. BTF automatically choose the best rank (R) of the tensor, which is controlling the number of latent factors. The patient-directional decomposed factors, which contain the comprehensive information of multi-omics data, are assigned to CoC algorithm wrapped K-mean clustering for subtype number determination. The optimum number of subtypes is determined by observing the dynamic cumulative distribution function (CDF) of the consensus value. The optimum number of subtypes occurred at the saturation point of CDF area changing plot. Once the number of subtypes is established, the CNMF is applied for patient subtype membership assignment.

3.2 Materials and method

3.2.1 Data collection and preprocessing

Three gene-level omics data matrices, copy number variation (~C), gene expression profiles (~G) and methylation profiles (~M), were downloaded from Xena platform (Goldman *et al.*, 2020). The upstream bioinformatics pipeline of data retrieval and preprocessing can be found in TCGA-BRCA (https://www.cancer.gov/tcga.) and Xena (http://xena.ucsc.edu/). The ~C is the gene-level CNV estimated using gistic2 (Mermel *et al.*, 2011) method, which contains CNV information for 1080 patients and 24774 genes. The ~G contains gene expressions for 1104 patients and 20530 genes. The ~M contains averaged methylation scores for 790 patients and 24774 genes. The clinical data (~CD) was downloaded from National Cancer Institute Genomic Data Commons (NCI GDC) (Grossman *et al.*, 2016) and some useful demographical and clinical information were carefully selected such as patients' age, tumor stage, IHC biomarker status (ER, PR, HER2), PAM50 subtype, and survival information (time to event and event indicator). After matching all sample IDs in the data sets, removing all-0 samples (the sample with all feature values are 0) and all-0 features (all samples have 0 value in this specific feature), selecting only top 10 percent of genes with largest coefficient of variation, 4515 genes for 762 patients were kept in the final matrices (~C), (~G) and (~M). The above-mentioned clinical information of these 762 patients was extracted for further analysis.

3.2.2 BTF for multi-omics data integration

BTF introduces Bayesian inference theory to CP algorithm. Objective of CP, as mentioned in the introduction, is to decompose a tensor to a sum of several rank-one tensors. In our implementation the algorithm, CP can be explained by the formula below:

Equation 3-1

$$\min_{(\widetilde{\mathbf{T}})} \left\| (\sim\mathbf{T}) - \widehat{(\sim\mathbf{T})} \right\|$$

Where $\widehat{(\sim T)} = \sum_{r=1}^{R} \widehat{u_r} \otimes \widehat{w_r} \otimes \widehat{x_r} = [\![\hat{U}, \hat{W}, \hat{X}]\!]$. R is the rank of tensor (~T) built upon the three processed omics data matrices (~C), (~G) and (~M). (~T) is a three-dimensional tensor with the size of 762 (number of patients) × 4515 (number of genes) × 3 (number of omics data types). $\widehat{(\sim T)}$ is the estimated low rank approximation of (~T). In other words, $\widehat{(\sim T)}$ approximates (~T) with a sum of R rank-one tensors (denoted as $\widehat{u_r} \otimes \widehat{w_r} \otimes \widehat{x_r}$, where $\widehat{u_r}, \widehat{w_r}, \widehat{x_r}$ (r = 1, 2, … , R) stand for non-zero vectors, and the symbol $\otimes$ denotes outer product calculation. These rank-one tensors can form 3 matrices $(\hat{U}, \hat{W}, \hat{X})$ which act as the extracted latent factors along three dimensions of the tensor (~T). Among these three latent factor matrices, $\hat{X}$ is in patient-level and has a size of 762 (number of patients) × R (rank of the tensor).

As we can see, the rank R is the exact number of latent factors we can extract. If the R is too small, too much information might be lost during tensor factorization. Whereas if R is too large, redundant factors might be generated. To overcome this difficulty, Bayesian inference was introduced to CP for effectively determining a proper rank for a given tensor. The following formula explains the CP-based BTF used in this study.

Equation 3-2

$$(\sim\mathbf{T}) = (\sim\mathbf{T_{true}}) + \in$$

Equation 3-3

$$(\sim\mathbf{T_{true}}) = \sum_{\mathbf{r=1}}^{\mathbf{R}} \mathbf{u_r} \otimes \mathbf{w_r} \otimes \mathbf{x_r} = [\![\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X}]\!]$$

where $(\sim\mathrm{T})$ is assumed to be composed of the true tensor $(\sim T_{\mathrm{true}})$, and the noise tensor $\in$; U represents a latent factor matrix factorized along omics data type direction, which has a size of 3 (number of omics data types) × R, and its r^{th} column is u_r. W and X are the latent factor matrices factorized along gene-direction and patient-direction, respectively. W has a size of 4515 (number of genes) × R, while X has a size of 762 (number of patients) × R. They have w_r and x_r in their r^{th} column, respectively.

Since the prior knowledge of $(\sim\mathrm{T})$ is lacking, the non-informative prior distribution of u_r, w_r, x_r, and the elements of $\in$ are assumed to be i.i.d. Gaussian distribution.

$$u_r \sim \mathrm{N}\ (0, \sigma_u)$$

$$w_r \sim \mathrm{N}\ (0, \sigma_w)$$

$$x_r \sim \mathrm{N}\ (0, \sigma_x)$$

$$\epsilon \sim \mathrm{N}\ (0, \lambda)$$

where ϵ is the element in noise tensor $\in$; λ (scalar), σ_u, σ_w, and σ_x (matrices with a size of R × R) serve as hyperparameters.

Under these assumptions, the conditional probability and the joint distribution of the model can be derived. Then the Variational Bayesian Inference (Zhao *et al.*, 2015) is incorporated to iteratively deduce the posterior distribution of latent factor matrices and the hyperparameters using

the prior distribution and the observed value in (∼T). The obtained posterior factor vectors with small values will be considered as redundant and thus be excluded. In this way, BTF can make attempts to iteratively optimize the rank when a prior rank is given, while the traditional CP takes the given initial rank as the final rank (Zhao *et al.*, 2015; Xiong *et al.*, 2010; Harshman, 1970; Tang *et al.*, 2018). The algorithmic details of the Variational Bayesian Inference technique can be found in other studies (Zhao *et al.*, 2015; Tang *et al.*, 2018).

TensorBF (Khan and Ammad-ud-din, 2016), which was the first package that enabled implementation of BTF using the R programming language platform, was used to extract the proposed multi-omics latent factor matrices. In addition to the original BTF algorithm (Tang *et al.*, 2018) mentioned above, tensorBF introduces a sparsity parameter to remove redundant factors, and thus, it can achieve automatic rank optimization. A noiseProb argument of the tensorBF function is needed to claim the proportion of variance that is expected to be explained with the factors. In our experiment, the noiseProb was set to (0.5,0.5) which meant 50% of the variance could be explained by the factors. Another argument is the initial rank K, which influences the inferenced final rank. We considered the K values of 10, 50, 75, 100, 150, and 200 to fine-tune this parameter.

3.2.3 BC subtyping problem setting and solving strategy

We considered BC subtyping as a two-stage problem. The first stage was to find out how many subtypes are in the BC patients, and the second stage was to assign the membership to each subtype. Only the patient-directional factorized latent matrix X was passed on to the downstream BC subtyping analysis.

3.2.3.1 Determine the number of BC subtypes

The number of BC subtypes is always the primary concern in BC subtyping problem. CoC algorithm (Monti *et al.*, 2003) was used in our study to first determine the number of subtypes after the integrated multi-omics latent factors were extracted. In our case, CoC first subsampled the input data matrix X for 50 times. Then, the fast clustering algorithm, K-means (Lloyd, 1982), was applied on these 50 sample sets to obtain 50 clustering results. The maximum number of subtypes (k) was set as 10, which meant we could expect up to 10 subtypes during each of the 50 clustering runs. The 50 clustering results for each subtype number (from 1 to k, which is 10 in our case) were then used to calculate the pairwise consensus values. The consensus value is defined as the probability of two items being clustered together(Monti *et al.*, 2003). These consensus values then formed a consensus matrix for each subtype number. Therefore, we obtained 10 consensus matrices, each one corresponding to a specific number of subtypes being clustered. In the end, an agglomerative hierarchical consensus clustering was applied to each of the 10 consensus matrices, to obtain the final 10 clustering results. Since every number of subtypes between 1 to k were tested during this procedure, we could use the Consensus Cumulative Distribution Function (CDF) to show the cumulative distribution of the consensus matrix for each run of clustering algorithm to decide which number of subtypes is the best. We used ConsensusClusterPlus (Wilkerson and Hayes, 2010), which is an R package for CoC and its visualization, to plot the CDF against each number of subtypes. ConsensusClusterPlus's CDF plot function shows the relative change in area under the CDF curve comparing the scenarios with increasing numbers of subtypes. If the area under the CDF curve stops significantly increasing after a certain number of subtypes, we selected that number of subtypes as the optimum number.

3.2.3.2 CNMF for BC subtyping

After the number of subtypes was decided, we used CNMF (Brunet *et al.*, 2004) clustering function in the R's package "CancerSubtypes" (Xu *et al.*, 2017) to assign the patients into each subtype class. Since CoC can wrap any clustering method with its resampling procedure, CNMF is such an algorithm that wraps NMF with CoC. The NMF algorithm is a low dimensional form of BTF which can be written as:

Equation 3-4

$$\mathbf{X} \cong [\![\boldsymbol{S}, \boldsymbol{H}]\!] = \sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \mathbf{s_k} \otimes \mathbf{h_k}$$

Where X is the multi-omics latent factor matrix with a size of 762 (number of patients) × R (rank of the tensor). S and H are two non-negative matrices with sizes of 762 × K and K × R, respectively. Here K is the predetermined optimum number of subtypes. By updating the S and H iteratively, the best value for S was obtained which was the subtype indicator we desired to have. Furthermore, CoC was wrapped outside of the NMF algorithm which made the clustering even more robust. More details of the optimization and iteration can be found in the original paper of NMF (Lee and Seung, 1999) and CNMF (Brunet *et al.*, 2004). We name this method as BTF-CNMF. The implementation of this method can be accessed at https://github.com/bowencheng-create/tensorBTF.

3.2.4 Baseline comparisons

The proposed multi-omics BTF-CNMF approach for BC subtyping was compared with other state-of-the-art work in two ways. The first series of comparisons was the proposed method compared with the multi-omics-based subtyping models including iClusterPlus (Mo *et al.*, 2018),

LRAcluster (Wu *et al.*, 2015), PINSCluster (Nguyen *et al.*, 2020), SNF (Wang *et al.*, 2014), and SNF.CC (Xu *et al.*, 2017). The second series of comparisons was the proposed method compared with each single omics-based subtyping model CNMF. The iClusterPlus, LRAcluster, and PINSCluster were implemented using the packages provided by the original studies. The implementations of SNF, SNF.CC and all single omics-based subtyping CNMF models were based on the R package "CancerSubtypes" (Xu *et al.*, 2017).

3.2.5 Subtyping validation

3.2.5.1 Simulation analysis

We simulated three omics data for an artificial cohort with 500 subjects using R package "InterSIM" (Chalise *et al.*, 2016), which is closer to real multi-omics dataset since this package simulates multiple interrelated data types with realistic intra-and inter-relationships based on the DNA methylation, mRNA gene expression, and protein expression data from the TCGA ovarian cancer study. InterSIM generates a predefined number of clusters and associates features to these clusters by shifting their mean values with a fixed amount. Here, we simulated the three omics data sets with 2 clusters. The number of features for the simulated methylation, gene expression and protein expression data are 367, 131 and 160, respectively, which are matched to 130 common genes in the three omics data sets to build our tensor. Other parameters we set are: the proportion of samples in the clusters are 0.6 and 0.4; cluster mean shift for gene expression, methylation and protein expression is 0.2. We also tested this for the cases 0.3, 0.4 and 0.6; proportion of differentially expressed genes, CpG sites and proteins is 0.2. The rest of the parameters are set as NULL according to the package's default.

We then executed the proposed multi-omics BTF-CNMF model and all the baseline models including multi-omics-based iClusterPlus, PINSPlus, LRAcluster, SNF, and SNF.CC models, as

well as the three single-data type-based CNMF models using each single omics data as an input separately. The performance metrics included precision, recall, balanced accuracy, specificity and silhouette score (Rousseeuw, 1987). The formulas of calculating these metrics were listed in **Table 3-1**.

Table 3-1. The definition and formula of the model performance metrics used in the simulation analysis.

Metrics	Formula
Precision	TP[1] / (TP + FP[2])
Recall	TP / (TP + FN[3])
Specificity	TN[4] / (TN + FP)
balanced.accuracy	(TP / (TP + FN) + TN / (TN + FP)) / 2 = (Recall + Specificity) / 2
Silhouette	(b[5]-a[6]) /max (a,b)

[1]TP: True positive;

[2]FP: False positive;

[3]FN: False negative;

[4]TN: True negative.

[5]a: average intra-cluster distance i.e the average distance between each point within a cluster.

[6]b: average inter-cluster distance i.e the average distance between all clusters.

3.2.5.2 Survival analysis

Survival analysis was applied to explore the survival differences among the identified subtypes for the proposed method and the baselines. We performed survival analyses using R's "survival" package (Therneau and Grambsch, 2000) and plotted the KM curve (Kaplan and Meier, 1958) using R's package "ggplot2" (Wickham, 2016). We performed this for overall survival (the event can be caused by any disease) and disease specific survival (the event is caused by BC), separately.

3.2.5.3 Demographical and clinical significance of the identified BC subtypes

The demographical and clinical information includes patients' age of diagnosis, ER status, PR status, HER2 status, tumor stage, and PAM50 subtypes. The differences among different subgroups were tested using either one-way analysis of variance (ANOVA) (for continuous variable such as age) or chi-square test (for categorical variable such as tumor stage).

3.2.6 Subtype interpretation

3.2.6.1 Gene expression visualization and differential analysis

The full gene expression data including the gene expressions of normal tissue samples were visualized by t-SNE using R's "tsne" (Justin Donaldson, 2016) and "ggplot2" package. The differential analyses were performed using the R's package "limma". Six contrasts corresponding to the six identified subtypes versus the normal tissue samples were created and fitted. The results were visualized using volcano plots generated by ggplot2.

3.2.6.2 Gene set enrichment analysis

The identified significant and unique genes in abovementioned six differential analyses were extracted and ranked according to their adjusted p-values. The ranked gene lists were then input into the Enrichr software (Chen *et al.*, 2013) for enriched pathway identification. The significantly enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were visualized using bar plots.

3.3 Results

3.3.1 Data collection and preprocessing

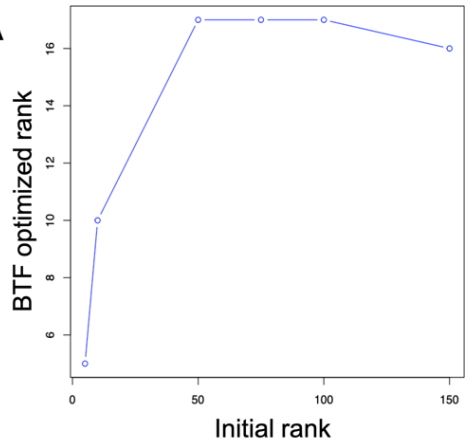
Three gene-level omics data: gene expression, CNV, and gene methylation, were downloaded from Xena platform (Goldman *et al.*, 2020). The clinical data were downloaded from National Cancer Institute Genomic Data Commons (NCI GDC) (Grossman *et al.*, 2016), which included patients' demographic information, tumor stage, IHC marker status (ER, PR, HER2), PAM50 subtype, and survival information (time to event and event indicator). After samples from different sources were matched, all-0 samples (the sample with all feature values are 0) and all-0 features (all samples have 0 value in this specific feature) were removed, and the top 10 percent of genes with the largest coefficient of variation were selected. After filtering and feature selection, only 4515 genes for 762 patients remained in the final data.

3.3.2 BTF for multi-omics data integration

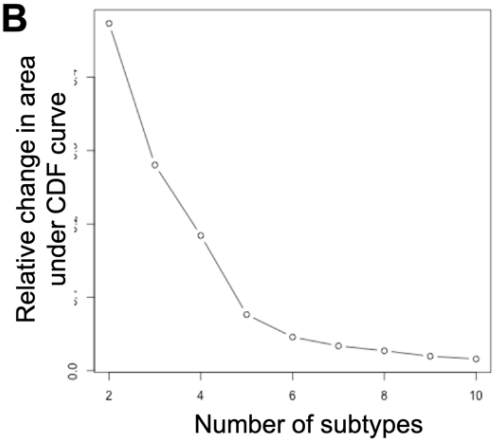
BTF introduces Bayesian inference theory to the CP algorithm. The objective of CP, as mentioned in the introduction, is to decompose a tensor to a sum of R rank-one tensors. R is the

rank of the multi-omics tensor built upon the three processed omics data matrices (**Figure 3-1**). The multi-omics tensor is a three-dimensional tensor with the size of 762 (number of patients) × 4515 (Number of genes) × 3 (number of omics data types). Traditional CP algorithm cannot estimate the rank R, while BTF can find the optimum R when given an initial R to optimize on. Several initial rank R values were explored, and the optimum posterior rank was found to be 17 as shown in **Figure 3-2A**. The decomposed rank-one tensors form 3 new matrices: the data type-directional factor matrix with the size of 3 (number of data types) × 17 (rank); the gene-directional factor matrix with the size of 4515 (genes) × 17 (rank); and the patient-directional factor matrix with the size of 762 (patients) × 17 (rank). The patient-directional factor matrix was interpreted as the extracted multi-omics latent factors for subsequent clustering.

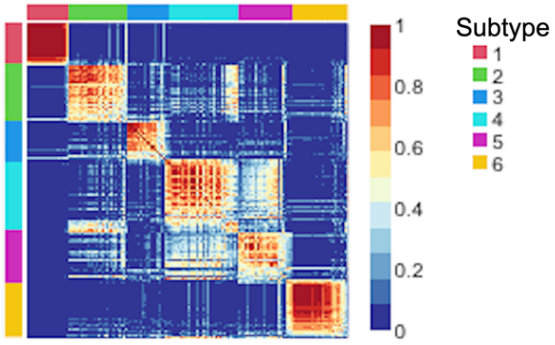
A



B



C



D

	Subtype 1	Subtype 2	Subtype 3	Subtype 4	Subtype 5	Subtype 6
Subtype 1	1	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Subtype 2	< 0.001	1	< 0.001	< 0.001	< 0.001	< 0.001
Subtype 3	< 0.001	< 0.001	1	< 0.001	< 0.001	< 0.001
Subtype 4	< 0.001	< 0.001	< 0.001	1	< 0.001	< 0.001
Subtype 5	< 0.001	< 0.001	< 0.001	< 0.001	1	< 0.001
Subtype 6	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	1

E

		Subtype 1	Subtype 2	Subtype 3	Subtype 4	Subtype 5	Subtype 6	p-value
Number of Patients		101	138	102	164	127	130	
Age at diagnosis	≥65	25	51	16	60	33	38	0.000725***
	<65	76	87	86	104	94	92	
	mean	57	59	54	61	57	58	
	min	34	27	26	29	28	29	
	max	90	90	76	90	90	90	
	SD	11.95	14.26	10.50	14.33	12.61	14.32	
Stage	I	14	10	17	32	24	24	6.318e-05***
	II	58	82	47	85	77	76	
	III	24	44	31	44	25	27	
	X or IV	5	1	5	1	0	3	
	Other	0	1	2	2	1	0	
ER Status	Positive	77	109	77	127	71	91	0.683
	Negative	24	28	25	32	23	35	
	Not Evaluated	0	1	0	5	33	4	
PR Status	Positive	61	89	67	118	67	80	0.1157
	Negative	40	48	35	39	26	46	
	Indeterminate	0	0	0	2	0	0	
	Not Evaluated	0	1	0	5	34	4	
HER2 Status	Positive	7	21	9	20	18	13	0.1426
	Negative	58	80	44	101	43	61	
	Indeterminate	1	5	0	4	2	0	
	Equivocal	25	19	20	22	28	22	
	Not Evaluated/Available	10	13	29	17	36	34	
Pam50 subtype	LumA	47	83	54	80	67	78	3.889e-05***
	LumB	19	16	21	44	33	8	
	Her2	4	6	5	15	6	9	
	Basal	25	27	20	16	21	24	
	Normal	6	6	2	9	0	11	

F

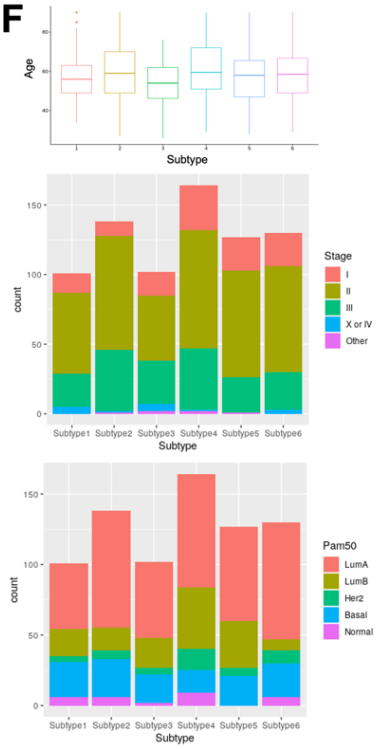


Figure 3-2. The identified six BC subtypes using proposed method. Panel A shows the Bayesian posterior optimized tensor rank against the initial rank. The optimized rank is 17, which means the extracted patient-directional factor matrix has a size of 762 (patients) × 17. After several resampling and K-mean clustering for different number of subtypes (2 to10), the probability of two items being clustered together is stored in a consensus matrix. The relative change in area under CDF curve of the consensus values against these number of subtypes is shown on **Panel B**. By observing the saturate point of the plot, we decide to select 6 as the optimum number of subtypes because the saturation occurs when the number of subtypes equates to 6. This information acts as a hyperparameter of the following CNMF clustering model for subtype membership assignment. The clustered CNMF consensus matrix is shown in **Panel C** with very clear 6-cluster pattern. **Panel D** indicates the pair-wise statistical significance of the 6 clusters. Since the p-values between different pairs of the identified subtypes approach zero (far below the significant threshold 0.05), there is strong evidence to say that the identified subtypes are from different Gaussian distributions. **Panel E** shows the clinical characteristic differences of the identified subtypes. The identified subtypes are different in age, tumor stage, and PAM50 subtypes. **Panel F** is the boxplot and barplots for the abovementioned three significantly differed clinical variables (age, tumor stage, and PAM50).

3.3.3 BC subtyping

We considered BC subtyping as a two-stage problem. The first stage was to find out how many subtypes exist in BC patients, and the second stage was to assign the patients to each subtype. Only the patient-directional latent factor matrix was passed on to the downstream BC subtyping.

The number of subtypes is always the primary concern in the BC subclassification problem. CoC algorithm (Monti *et al.*, 2003) was used in our study to determine the number of subtypes after the integrated multi-omics latent factors are extracted. In our case, CoC first subsampled the above-mentioned patient-directional factor matrix, then applied K-means (Lloyd, 1982) clustering on these sample sets. A series of subtype numbers were tried during these clustering runs. The

clustering results for each subtype number were then used to calculate the pairwise consensus values. The consensus value was defined as the probability of two patients being clustered together (Monti *et al.*, 2003). These consensus values then form a consensus matrix for each subtype number. In the end, an agglomerative hierarchical clustering was applied to each of the consensus matrices to get the final clustering results. Since different numbers of subtypes were tried during this procedure, the CDF was used to show the cumulative distribution of the consensus matrix at each run of clustering to decide which number of subtypes is the best (**Figure 3-2B**). CDF plot function shows the relative change in area under the CDF curve comparing the scenarios with increasing number of subtypes. If the area under the CDF curve stops significantly increasing after a certain number of subtypes, we can say that number of subtypes is the optimum number. In our case, the optimum number of subtypes was 6 (**Figure 3-2B**).

After the number of subtypes was decided, we moved on to the next stage—subtype membership assignment. CNMF (Brunet *et al.*, 2004) clustering combines NMF and CoC. The NMF algorithm is a low dimensional form of BTF which further decomposes the patient-directional factor matrix into two non-negative matrices with sizes of 762 (patients) × 6 (subtypes) and 6 × 17 (tensor rank), respectively. The first non-negative matrix is the subtype membership indicator. By updating these two non-negative matrices iteratively, the best subtype indicator is obtained. Besides, the resampling-based consensus clustering built on the NMF algorithm will make the clustering result even more robust. The clustered consensus matrix of CNMF can be found in **Figure 3-2C**. And the unsupervised gene expression t-SNE visualization is shown in **Figure 3-3**.

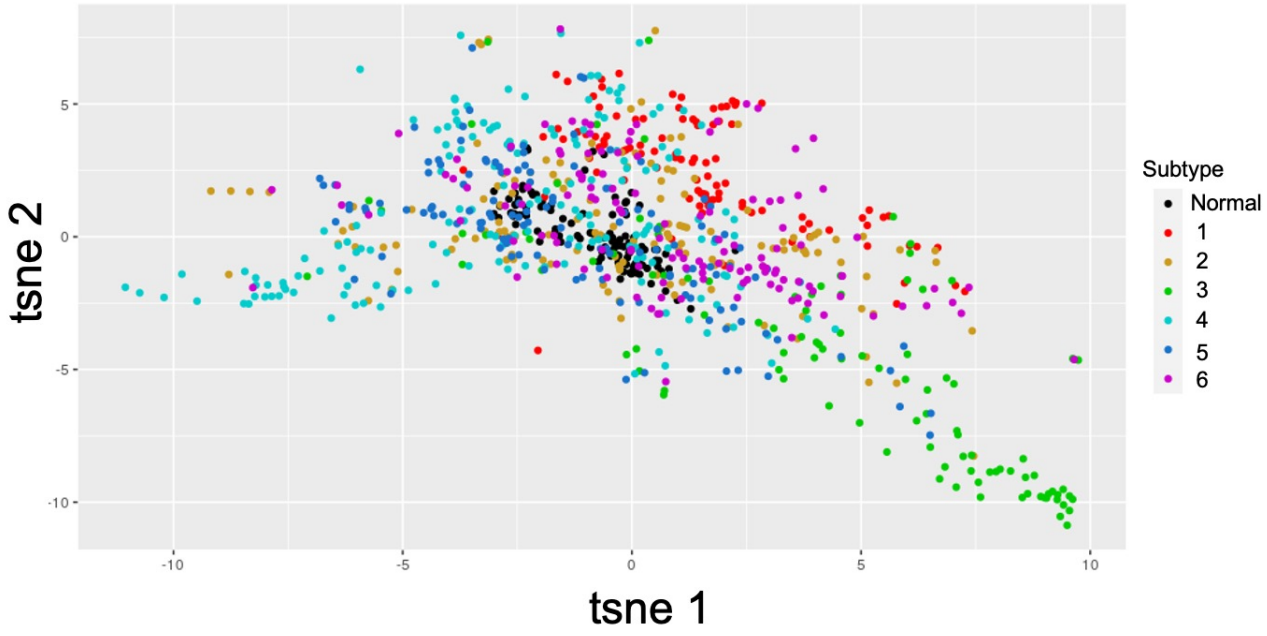


Figure 3-3. The t-SNE map for gene expressions.

Statistical significance of the identified subtypes was tested using the hypothesis assumption and the ratio between the within-cluster sum of squares and the total sum of squares act as the statistic. Since the *p-values* between different pairs of the identified subtypes approached zero (**Figure 3-2D**), there was strong evidence to reject H0 (i.e., the data of the identified subtypes is from a single Gaussian distribution) and accept H1 (H1: The data of the identified subtypes is from a mixture of Gaussian distributions). Therefore, we can conclude that the identified subtypes are valid and real.

The clinical information we collected included patients' age of diagnosis, ER status, PR status, HER2 status, tumor stage, and PAM50 subtypes. These clinical differences among different subtypes were tested using either one-way Analysis of Variance (ANOVA) (for continuous variable such as age) or chi-square test (for categorical variable such as tumor stage). As shown in **Figure 3-2E**, three significant clinical variables are identified, age, stage, and PAM50 subtypes.

The distributions of these three clinical variables are further emphasized in **Figure 3-2F**. Patients in subtype 3 have the lowest average age, the highest proportion of X or IV tumor stage, and the lower proportion of Normal PAM50 subtype (**Figure 3-2F**).

3.3.4 Baseline comparison

The proposed multi-omics BTF-CNMF approach for BC subtyping was compared with other works in two ways. The first series of the comparison is the proposed method versus multi-omics-based iClusterPlus, PINSPlus, LRAcluster, SNF and SNF.CC. The second series of the comparison is the proposed method compared with the CNMF model using each single omics data as an input. Survival analysis is applied to explore the survival differences among the identified subtypes for the proposed method and the baselines. We do this for overall survival (the event can be caused by any disease) and disease-specific survival (the event is caused by BC), separately. Since Duan *et al.* have shown that the Similarity Network Fusion (SNF) method (Wang *et al.*, 2014) has relatively more robust and efficient performance than other methods (Duan *et al.*, 2021), and the real data have no ground truth of the cluster labels, our real data analysis was just focused on our proposed method and the SNF and SNF.CC methods. For the simulated data analysis, since we have the ground truth of the cluster labels, we compared the performance of our proposed method with all suggested methods.

As shown in **Table 3-2**, generally speaking, the multi-omics -based models (such as BTF-CNMF, SNF) show better performance to identify the predefined true clusters than those based on individual omics data. The proposed multi-omics BTF-CNMF approach shows better performance to identify the predefined true clusters than the baseline methods using either multi-omics or

single-omics data. We also observed that the performance of the SNF model is better than that of the iClusterPlus, PINSPlus, LRAcluster methods, which is similar to what Duan *et al* observed.

Table 3-2. The simulation analysis performance of the proposed method and the baseline methods.

		Precision	Recall	Accuracy	Specificity	Silhouette (similarity)[5]	Silhouette (distance)[6]
Delta=0.2 Cluster=2 Sample=500	**Multi BTF-CNMF**[1]	**0.75**	**0.76**	**0.76**	**0.75**	**0.86**	**0.05**
	Multi iClusterPlus	0.52	0.52	0.52	0.52	NA[7]	0.01
	Multi PINSPlus	0.52	0.52	0.52	0.52	NA	0.02
	Multi LRAcluster	0.55	0.54	0.54	0.54	NA	0.01
	Multi SNF	0.51	0.51	0.51	0.51	0.23	0.01
	Multi SNF.CC	0.51	0.51	0.51	0.51	0.61	0.01
	Single CNMF 1[2]	0.52	0.52	0.52	0.52	0.99	0.03
	Single CNMF 2[3]	0.50	0.50	0.50	0.50	0.90	0.02
	Single CNMF 3[4]	0.52	0.53	0.53	0.53	0.98	0.01
Delta=0.3 Cluster=2 Sample=500	**Multi BTF-CNMF**	**0.90**	**0.91**	**0.91**	**0.91**	**0.86**	**0.09**
	Multi iClusterPlus	0.57	0.57	0.57	0.57	NA	0.03
	Multi PINSPlus	0.52	0.52	0.52	0.52	NA	0.01
	Multi LRAcluster	0.55	0.55	0.55	0.55	NA	0.01
	Multi SNF	0.87	0.88	0.88	0.88	0.8	0.08
	Multi SNF.CC	0.52	0.52	0.52	0.52	0.66	0.01
	Single CNMF 1	0.51	0.51	0.51	0.51	0.97	0.02
	Single CNMF 2	0.52	0.52	0.52	0.52	0.95	0.02

	Single CNMF 3	0.52	0.52	0.52	0.52	0	0.01
Delta=0.4 Cluster=2 Sample=500	**Multi BTF-CNMF**	**0.97**	**0.98**	**0.54**	**0.98**	**0.97**	**0.13**
	Multi iClusterPlus	0.58	0.57	0.58	0.57	NA	0.02
	Multi PINSPlus	0.53	0.53	0.53	0.53	NA	0.03
	Multi LRAcluster	0.55	0.54	0.54	0.54	NA	0.01
	Multi SNF	0.95	0.95	0.95	0.95	0.32	0.11
	Multi SNF.CC	0.96	0.96	0.96	0.96	0.33	0.08
	Single CNMF 1	0.55	0.55	0.55	0.55	0.97	0.01
	Single CNMF 2	0.52	0.51	0.51	0.51	0.91	0.01
	Single CNMF 3	0.55	0.55	0.55	0.55	0.98	0.03
Delta=0.6 Cluster=2 Sample=500	**Multi BTF-CNMF**	**1**	**1**	**1**	**1**	**1**	**0.13**
	Multi iClusterPlus	0.87	0.87	0.87	0.87	NA	0.08
	Multi PINSPlus	1	1	1	1	NA	0.13
	Multi LRAcluster	1	1	1	1	NA	0.11
	Multi SNF	1	1	1	1	0.53	0.11
	Multi SNF.CC	1	1	1	1	0.90	0.11
	Single CNMF 1	0.54	0.54	0.54	0.54	0.25	0.01
	Single CNMF 2	0.56	0.56	0.56	0.56	0.94	0.03
	Single CNMF 3	1	1	1	1	1	0.11

[1]Multi BTF CNMF: The proposed new approach using the multi-omics (the simulated gene expression data, DNA methylation data and proteomics data) as input.
[2]Single CNMF 1: The CNMF approach using only simulated gene expression data as input.
[3]Single CNMF 2: The CNMF approach using only simulated DNA methylation data as input.
[4]Single CNMF 3: The CNMF approach using only simulated proteomics data as input.
[5]Silhouette (similarity): The Silhouette metric calculated based on the similarity/consensus matrix.
[6]Silhouette (distance): The Silhouette metric calculated based on the distance matrix.
[7]Not applicable (the number cannot be estimated using the packages).

Figure 3-4 shows the disease-specific survival analysis results of the proposed method and the above-mentioned baseline methods. There were significant disease-specific survival differences among the subtypes identified by the proposed method as shown in **Figure 3-4A**. Furthermore, patients in Subtype 3 suffered a worse prognosis compared to other subtypes. While the subtypes identified by the baseline models showed no significant survival differences (**Figure 3-4B-F**). The overall survival analysis results showed a similar trend (**Figure 3-5)**.

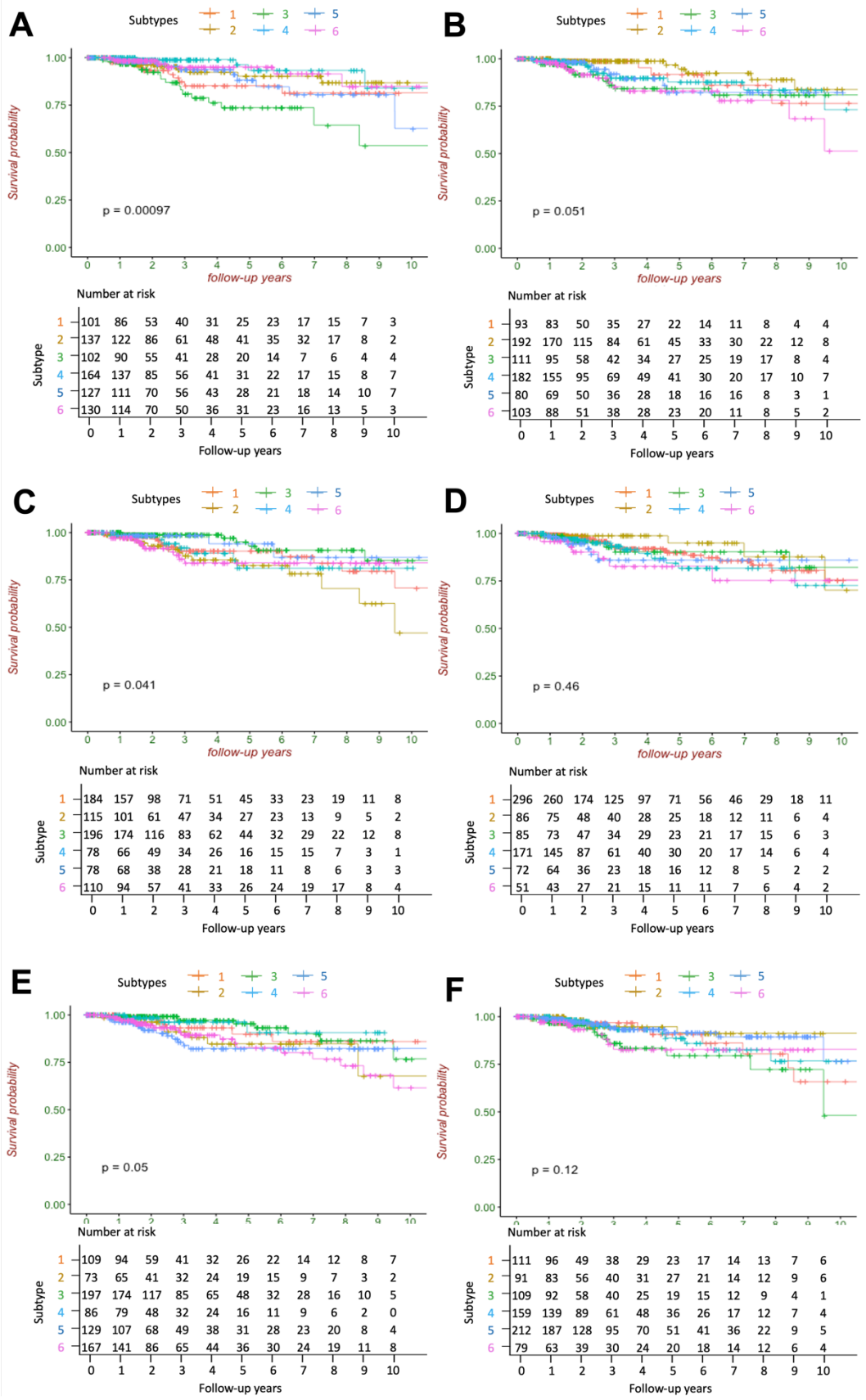

A
Subtypes 1 2 3 4 5 6
Survival probability
p = 0.00097
follow-up years
Number at risk
Subtype
1 101 86 53 40 31 25 23 17 15 7 3
2 137 122 86 61 48 41 35 32 17 8 2
3 102 90 55 41 28 20 14 7 6 4 4
4 164 137 85 56 41 31 22 17 15 8 7
5 127 111 70 56 43 28 21 18 14 10 7
6 130 114 70 50 36 31 23 16 13 5 3
0 1 2 3 4 5 6 7 8 9 10
Follow-up years
B
Subtypes 1 2 3 4 5 6
Survival probability
p = 0.051
follow-up years
Number at risk
Subtype
1 93 83 50 35 27 22 14 11 8 4 4
2 192 170 115 84 61 45 33 30 22 12 8
3 111 95 58 42 34 27 25 19 17 8 4
4 182 155 95 69 49 41 30 20 17 10 7
5 80 69 50 36 28 18 16 16 8 3 1
6 103 88 51 38 28 23 20 11 8 5 2
0 1 2 3 4 5 6 7 8 9 10
Follow-up years
C
Subtypes 1 2 3 4 5 6
Survival probability
p = 0.041
follow-up years
Number at risk
Subtype
1 184 157 98 71 51 45 33 23 19 11 8
2 115 101 61 47 34 27 23 13 9 5 2
3 196 174 116 83 62 44 32 29 22 12 8
4 78 66 49 34 26 16 15 15 7 3 1
5 78 68 38 28 21 18 11 8 6 3 3
6 110 94 57 41 33 26 24 19 17 8 4
0 1 2 3 4 5 6 7 8 9 10
Follow-up years
D
Subtypes 1 2 3 4 5 6
Survival probability
p = 0.46
follow-up years
Number at risk
Subtype
1 296 260 174 125 97 71 56 46 29 18 11
2 86 75 48 40 28 25 18 12 11 6 4
3 85 73 47 34 29 23 21 17 15 6 3
4 171 145 87 61 40 30 20 17 14 6 4
5 72 64 36 23 18 16 12 8 5 2 2
6 51 43 27 21 15 11 11 7 6 4 2
0 1 2 3 4 5 6 7 8 9 10
Follow-up years
E
Subtypes 1 2 3 4 5 6
Survival probability
p = 0.05
Number at risk
Subtype
1 109 94 59 41 32 26 22 14 12 8 7
2 73 65 41 32 24 19 15 9 7 3 2
3 197 174 117 85 65 48 32 28 16 10 5
4 86 79 48 32 24 16 11 9 6 2 0
5 129 107 68 49 38 31 28 23 20 8 4
6 167 141 86 65 44 36 30 24 19 11 8
0 1 2 3 4 5 6 7 8 9 10
Follow-up years
F
Subtypes 1 2 3 4 5 6
Survival probability
p = 0.12
Number at risk
Subtype
1 111 96 49 38 29 23 17 14 13 7 6
2 91 83 56 40 31 27 21 14 12 9 6
3 109 92 58 40 25 19 15 12 9 4 1
4 159 139 89 61 48 36 26 17 12 7 4
5 212 187 128 95 70 51 41 36 22 9 5
6 79 63 39 30 24 20 18 14 12 6 4
0 1 2 3 4 5 6 7 8 9 10
Follow-up years

Figure 3-4. The disease specific 10-year survival comparison of the proposed method with baseline methods. **Panel A** is the disease specific survival plot of the subtypes identified by the proposed multi-omics BTF-CNMF method. **Panel B** is the disease specific survival plot of the subtypes identified by the multi-omics plus BTF plus SNF. **Panel C** is the disease specific survival plot of the subtypes identified by the multi-omics plus BTF plus SNF.CC. **Panel D** is the disease specific survival plot of the subtypes identified by the BTF-CNMF method with only CNV data as input. **Panel E** is the disease specific survival plot of the subtypes identified by the BTF-CNMF method with only gene expression data as input. **Panel F** is the disease specific survival plot of the subtypes identified by the BTF-CNMF method with only gene methylation data as input.

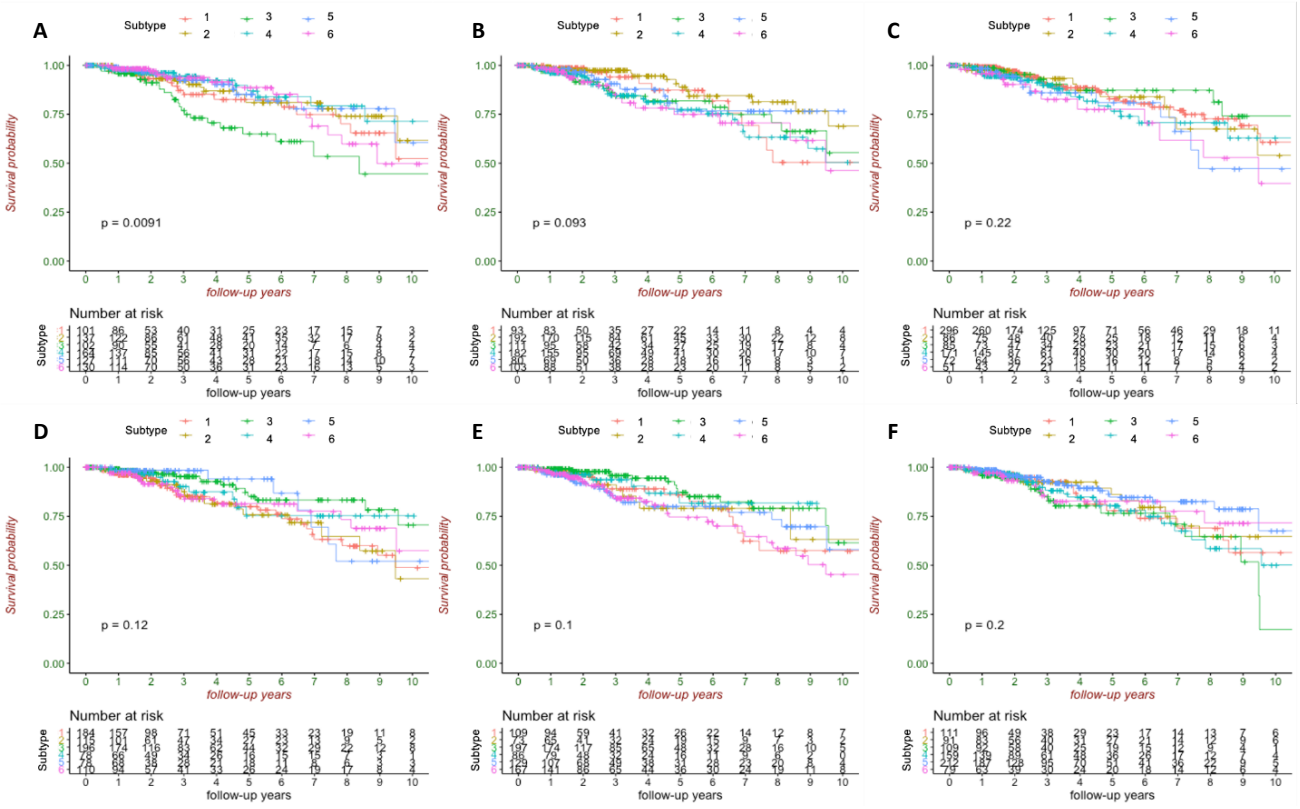


Figure 3-5. The overall survival comparison of proposed method with baseline methods. **Panel A** is the overall survival plot of the subtypes identified by the proposed multi-omics BTF-CNMF method. **Panel B** is the overall survival plot of the subtypes identified by the multi-omics plus BTF plus SNF. **Panel C** is

the overall survival plot of the subtypes identified by the multi-omics plus BTF plus SNF.CC. **Panel D** is the overall survival plot of the subtypes identified by the CNMF method with only CNV data as input. Panel E is overall survival plot of the subtypes identified by the CNMF method with only gene expression data as input. **Panel F** is the overall survival plot of the subtypes identified by the CNMF method with only gene methylation data as input. Note: the top part from left to right is Panels A, B and C and the bottom part from left to right is Panels D, E and F.

3.3.5 Subtype interpretation

The identified subtypes were further explored in terms of their gene expression differences and gene set enrichment differences. We included the gene expressions of 114 normal tissue samples as a control group. The differential analyses were done between each subtype with the normal tissue sample group, respectively. The volcano plots of the differential analysis corresponding to each of the six subtypes compared to the normal samples are shown in **Figure 3-6**. We highlighted those genes with the adjusted *p-value* < 1e-100 and the absolute log2(fold change) >= 0.585 (1.5-fold change). The details of these highlighted genes for each of the 6 subtypes or clusters (c1- c6) and their comparison (e.g., common ones, unique ones, and the percentage of unique ones among total) can be found in **Appendix A**.

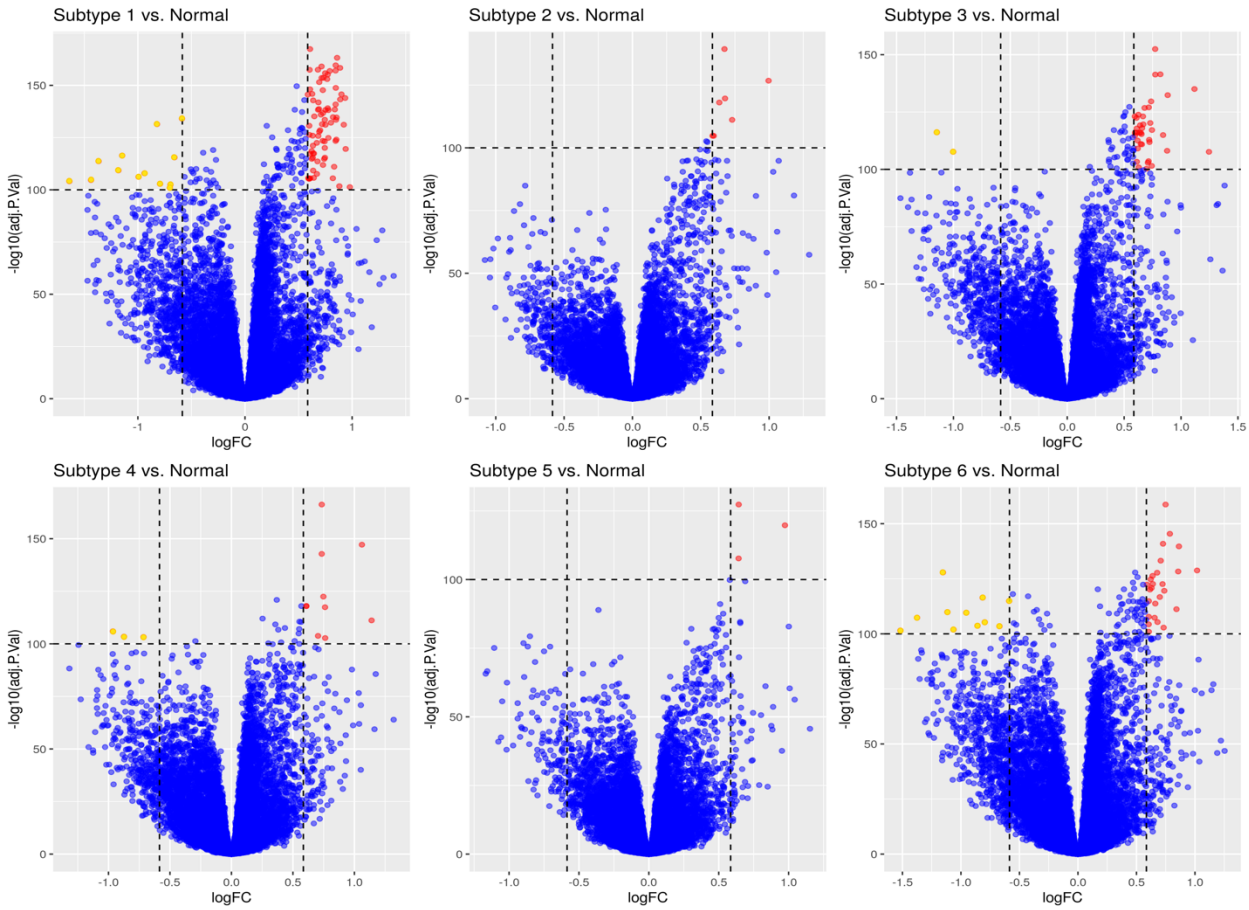


Figure 3-6. Gene expression differential analysis of the identified six BC subtypes. The results of the six gene expression differential analyses corresponding to the six subtypes compared to the normal samples, respectively. Red dots represent the genes with the adjusted *p-value* < 1e-100 and the log2(fold change) >= 0.585 (1.5-fold change). Yellow dots represent the genes with the adjusted *p-value* < 1e-100 and the log2(fold change) <= -0.585 (1.5-fold change).

Differentially expressed genes that were both statistically significant (adjusted p-value < 1e-100 and log2(fold change) >= 0.585) and unique to each of the six subtypes were then passed to GSEA (Chen *et al.*, 2013) to identify which of the KEGG (Ogata *et al.*, 1999) pathways, if any, the gene sets from different subtypes were enriched in. Please be noted that Subtype 2 and Subtype

5 did not have unique genes, thus the full gene lists were used for the GSEA for these two subtypes. The identified pathways significantly enriched in the subtype-specific gene sets are shown in **Figure 3-7**.

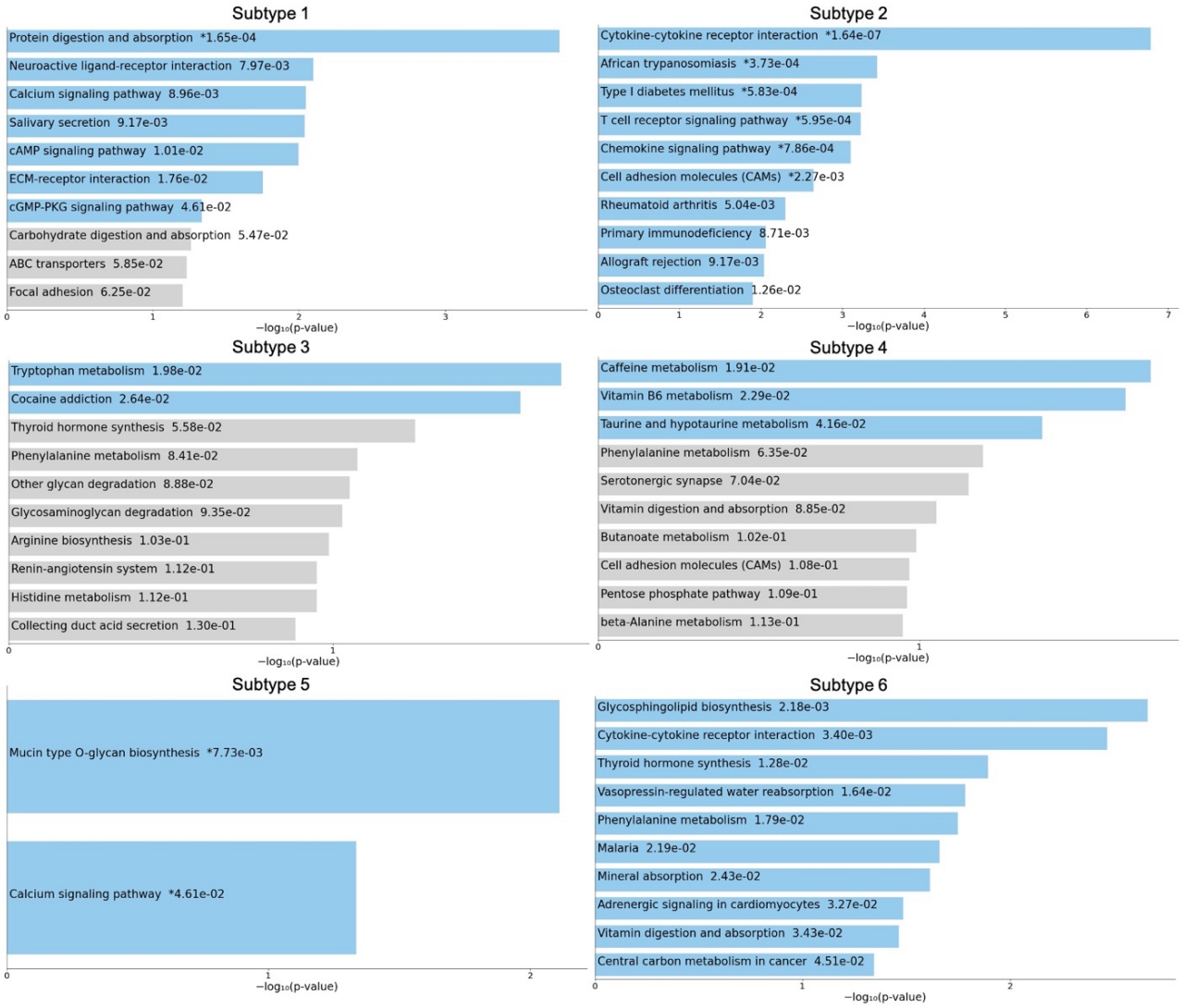


Figure 3-7. The bar chart of the top enriched terms in the KEGG library. Colored bars correspond to enriched KEGG pathways with significant *p-values* (<0.05). An asterisk (*) next to a *p-value* indicates the term also has a significant adjusted *p-value* (<0.05).

3.4 Discussion

The proposed BTF-CNMF method is a novel tensor-based integrative clustering approach and in this study, it successfully identified 6 BC subtypes. These subtypes were significant in terms of statistical tests, age-specific ANOVA test, tumor stage and PAM50 subtype-specific chi-square tests, survival test, and gene expression t-SNE clustering test. On the contrary, all other baseline methods could not identify subtypes with statistically significant survival patterns. In this sense, we have shown the advantages of the proposed BTF-CNMF over other baselines.

Patients with Subtype 3 as classified by BTF-CNMF suffered the worst prognosis. Deeper insights into this observation are provided in this study. From the age-specific ANOVA test and boxplot, we can see that Subtype 3 has the lowest average age at diagnosis. For BC, young age at diagnosis is known to be associated with poor prognosis, which is supported by previous research (Martínez *et al.*, 2019; Johnson *et al.*, 2015; Anders *et al.*, 2008). Our gene expression t-SNE map showed that patients with Subtype 3 are clustered and present far away from other subtypes, which also indicated the distinctiveness of Subtype 3. Differentially expressed genes were identified for each subtype compared to the normal tissue (**Appendix A1-6**). The significant genes unique to Subtype 1 were enriched in protein and carbohydrate digestion and absorption pathways as well as several signalling pathways (**Appendix A7**). More importantly however, Subtype 1 was enriched in ABC transporters, which is commonly associated with multi-drug resistant phenotype in cancer. Subtype 2-associated genes were enriched in several regulatory pathways, such as Cytokine-cytokine receptor interaction pathway, T cell signaling pathway, Chemokine signaling pathway. Rheumatoid arthritis, Primary immunodeficiency, and Allograft rejection. Thus, Subtype 2 may be best described as “immune-related” or “inflammatory” subtype. The significantly enriched pathways for Subtype 3 were the Tryptophan metabolism pathway and Cocaine addiction pathway. Interestingly, Subtype 3 was also enriched in Glycosaminoglycan and other glycan

degradation pathways and in addition to being absent in enrichment of adhesion-related pathways unlike Subtype 1 (Focal adhesion), Subtype 2 (Cell adhesion molecules (CAMs)), and Subtype 4 (CAMs). This suggests that patients classified as Subtype 3 may suffer from metastatic phenotype more so than other patient groups, due to degradation of extracellular matrix and loss of adhesion. Many of the pathways enriched in Subtype 4-specific genes were related to metabolism; Hence Subtype 4 may be best described as the "metabolic" subtype. Some biosynthesis pathways were identified for Subtype 5 and Subtype 6. In addition, Subtype 6 was enriched in several pathways involved in water, mineral and vitamin absorption. These different biological pathways may reflect different characteristics of the identified subtypes. We also observed some common pathways among different subtypes. As shown in **Figure 3-7** for example, calcium signaling pathway is significantly enriched in Subtype 1 and Subtype 5, CAMs enriched in both Subtype 2 and Subtype 4, cytokine-cytokine receptor interaction enriched in both Subtype 2 and Subtype 6, thyroid hormone synthesis enriched in Subtype 3 and Subtype 6, phenylalanine metabolism enriched in Subtype 3, Subtype 4 and Subtype 6, and vitamin digestion and absorption enriched in Subtype 4 and Subtype 6.

Since we only used the patient-directional factorized matrix for the subtyping, we did not fully utilize the latent information in the other two factorized matrices, such as the data type-directional matrix with data types by the latent ranks, and the gene-directional matrix with genes by the latent ranks. In the future, we could explore new approaches that can effectively take advantage of all the latent information to identify the significant genes or biological pathways for each of the identified BC subtypes.

This study was focused on developing the innovative unsupervised BTF-CNMF method for subtyping BCs into prognostically different subgroups. Many other studies have been

undertaken to identify BC or its subtypes related genes using individual gene expression, CNV, DNA methylation data or their combinations. In the future it will be interesting to explore how to apply some of the innovative algorithms to identify subtype-specific genes based on the prognostic groups identified using our BTF-CNMF method.

3.5 Conclusion

One of the major innovations in our proposed method is that it uses a model-driven approach to integrate three high-dimensional omics data sets for tumor stratification. Comparing with other network-based approaches, such as SNF, the model-based method can handle high-dimensional data directly. Our results have demonstrated that the proposed innovative approach BTF-CNMF is a promising strategy to efficiently use publicly available multi-omics data to identify BC subtypes, which is better than other state-of-the-art genomic-based BC subtyping methods. In addition, our method could be easily transferred to identify subtypes of other complex diseases or be used in conjunction with other novel downstream subtyping algorithms.

4 Chapter 4: GNN for BC proteomic heterogeneity study

4.1 Introduction

Genome-wide survival models using multiple types of omics data (gene expression, CNV, DNA methylation) have identified the prognostic associations of each gene in the human genome (Smith and Sheltzer, 2022). Cross-omics meta-analysis could cumulate each gene's prognostic effects at different biological levels. However, these meta-prognostic significances may still be small in individual genes. Likewise, although some key driver genes are frequently mutated in the BC, most low-frequency mutations are likely to be of minor significance (Zheng *et al.*, 2021). In addition, they could not reflect a global view of the system-level biology of BC (Yang *et al.*, 2014).

Organizing the gene-level survival risks or mutations into gene set level has been very useful in identifying higher-order systems of genes under high survival or mutation burdens, many of which would otherwise be missed (Zheng *et al.*, 2021). To do so, an accurate and comprehensive knowledge map of biological systems needs to be determined first, so that the survival/mutation burdens of these biological systems can be evaluated later. Biological networks are valuable for understanding the heterogeneity of BC in a systematic view (Furlong, 2013). Protein-protein interaction (PPI) networks are a kind of commonly used biological network in BC research (Zheng *et al.*, 2021). They could be modeled using an affinity graph, where nodes are proteins with their features, and edges are undirected and weighted connections between the interacted proteins. To better incorporate the existing knowledge of proteins into the PPI network, their amino acid sequences can be fed into the graph as node features for creating the comprehensive knowledge map of biological systems. The biological systems of BC can be defined at a range of biophysical scales from individual proteins, small groups of proteins to large groups of proteins. Analyzing the

survival/mutation burdens using only one of these scales could miss survival/mutation burdens at all others. Zheng et al. integrated large amounts of existing protein-protein association data with their own generated PPI data of BC and constructed a hierarchical protein system map. They then developed a group-lasso-based statistical model and successfully identified 395 significant biological systems under mutation burdens on all scales of the hierarchy which can be used as clinical biomarkers (Zheng *et al.*, 2021).

A limitation of Zheng *et al.*'s work is that the mutations may not have direct clinical associations. Instead, the gene-level survival risks estimated by Smith and Sheltzer in their genome-wide survival analyses (Smith and Sheltzer, 2022) are more closely related to patients' prognosis, thus using this survival information to select biological systems in the hierarchical map may identify better biomarkers. Another limitation of Zheng *et al*.'s work is that they used a tradition topological method called CliXO to construct the hierarchy. CliXO takes the weighted graph as input, calculates the maximal cliques, and adds the identified maximal cliques to the ontology as the threshold of edge weight of this graph is lowered (Kramer *et al.*, 2014). **Figure 4-1** explains how CliXO works. Although the CliXO algorithm is widely used in detecting hierarchical communities of a network (Zheng *et al.*, 2021; Qin *et al.*, 2021), it only considers the edge weights, and it has several hyperparameters that control the threshold decreasing strategies which need to be pre-specified and tuned manually to achieve a propitiate decision. More efficient and automated algorithms are needed to address these drawbacks. Moreover, to better incorporate the existing knowledge of gene products and their functions into PPI network, the terms in Gene Ontology (GO) (The Gene Ontology Consortium *et al.*, 2004; Carbon *et al.*, 2021) could be involved. GO has three branches: Biological Process (BP), Cell Component (CC), and the Molecular Function (MF). Hence, it is useful to incorporate the extra information of GO into the

PPI graph for discovering hierarchical functional modules or hierarchical protein community of BC systematically (Zheng *et al.*, 2021) (Fionda, 2018).

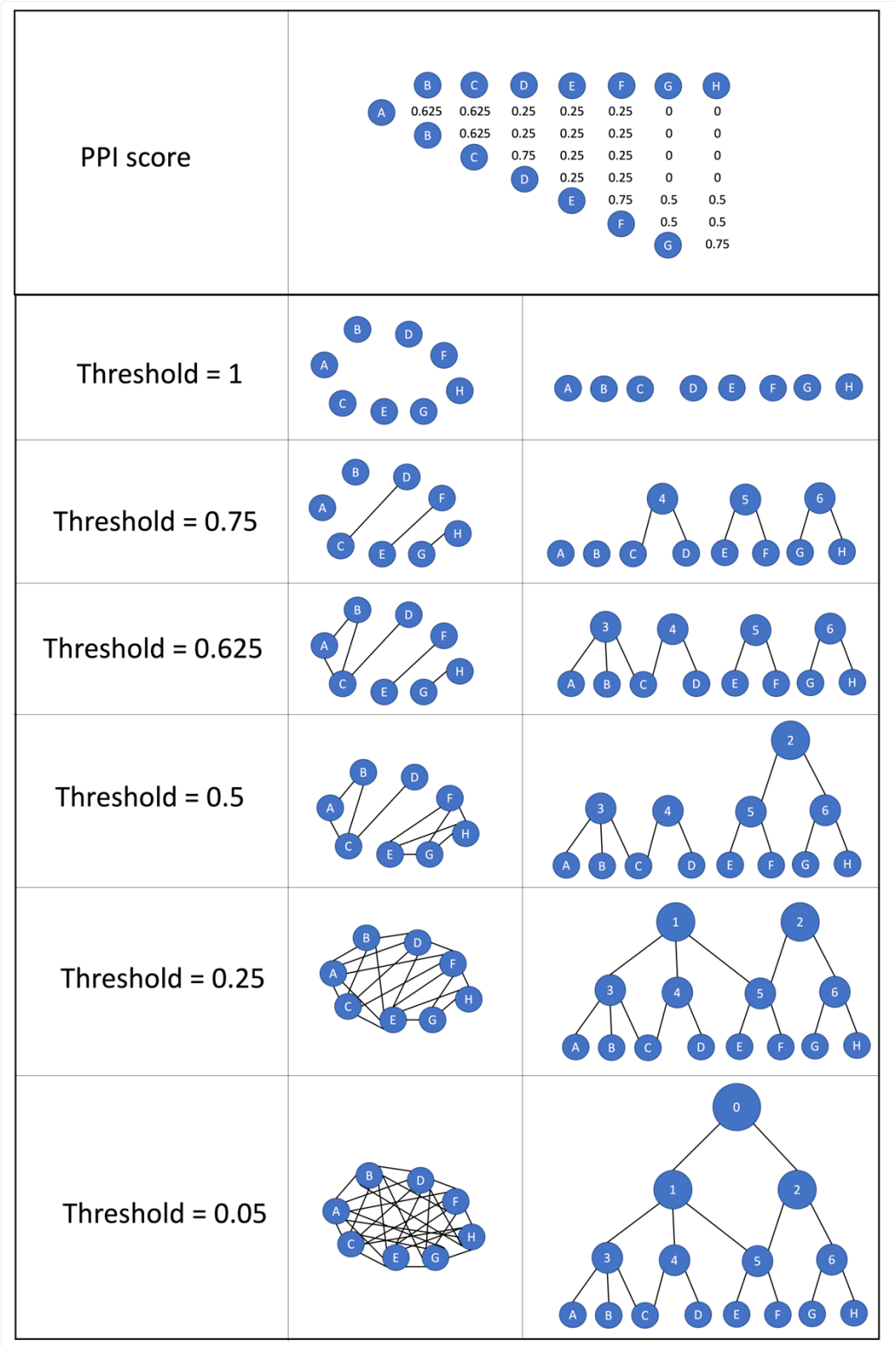


Figure 4-1. A simple explanation of the CliXO algorithm. The bottom right hierarchy is the final output. It has 8 proteins (A-H) and 7 biological systems (0-6) with a hierarchical structure. The PPI scores between each protein pairs are in the first row. As the threshold is decreasing, edges with the scores that is equal to or larger than the threshold are added to the graph. At each new threshold, maximal cliques identified in the graph are added to the predicted hierarchy. Eventually, a hierarchical tree is built. It should be noted that a clique is a subset of vertices of an undirected graph such that every two distinct vertices in the clique are adjacent and a maximal clique is a clique that cannot be extended by including one more adjacent vertex, meaning it is not a subset of a larger clique.

Recently, GNN has shown great capability in automatically analyzing biological networks and it is also powerful in combining other auxiliary data sources like the protein sequence to achieve better model performance (Long *et al.*, 2020, 2022, 2021). GNN could take both the node features, edge features, and the graph structure into consideration for representation learning. The node features can be initialed using the protein sequence information, which will be aggregated by the features of the neighbor nodes in each layer of the GNN and eventually propagated to the final target. Yang *et al.* proposed a GNN autoencoder to learn node representations for PPI prediction and achieved very high accuracy (F. Yang *et al.*, 2020). Likewise, Long *et al.* used the same strategy of GNN autoencoder to learn node embeddings and applied the learned embeddings to drug-target interaction (DTI) and synthetic lethality (SL) predictions (Long *et al.*, 2022). These successful studies suggest that GNN can be a potentially useful framework for building clinically relevant biological systems for BC.

4.2 Materials and methods

4.2.1 Framework of the meta-survival/mutation enriched hierarchical biological system detection

We build a GNN-based framework to identify meta-survival/mutation enriched hierarchical biological systems (**Figure 4-2)**. There are four major steps in the framework: 1) Use pretrained transformer models to extract features from protein amino acid sequence data; 2) Combine Zheng *et al.*'s data (Zheng *et al.*, 2021) and the features generated in the first step to generate an affinity graph; 3) Use GO terms as label to construct a hierarchical tree of protein communities/biological systems for BC in a GNN way; 4) Use the group-lasso-based statistical method called HiSig (Zheng *et al.*, 2021) to pinpoint which BC protein communities/biological systems are under strong prognostic pressure/mutation burden based on the meta-survival risk scores and mutation counts. This study could decipher the BC prognostic risk/mutation pressure that is raised from tumor genomes and gradually converged to their higher-order entities (such as transcriptome and proteome). The identified protein communities under high prognostic risk/mutation pressures could be potential biomarkers for BC.

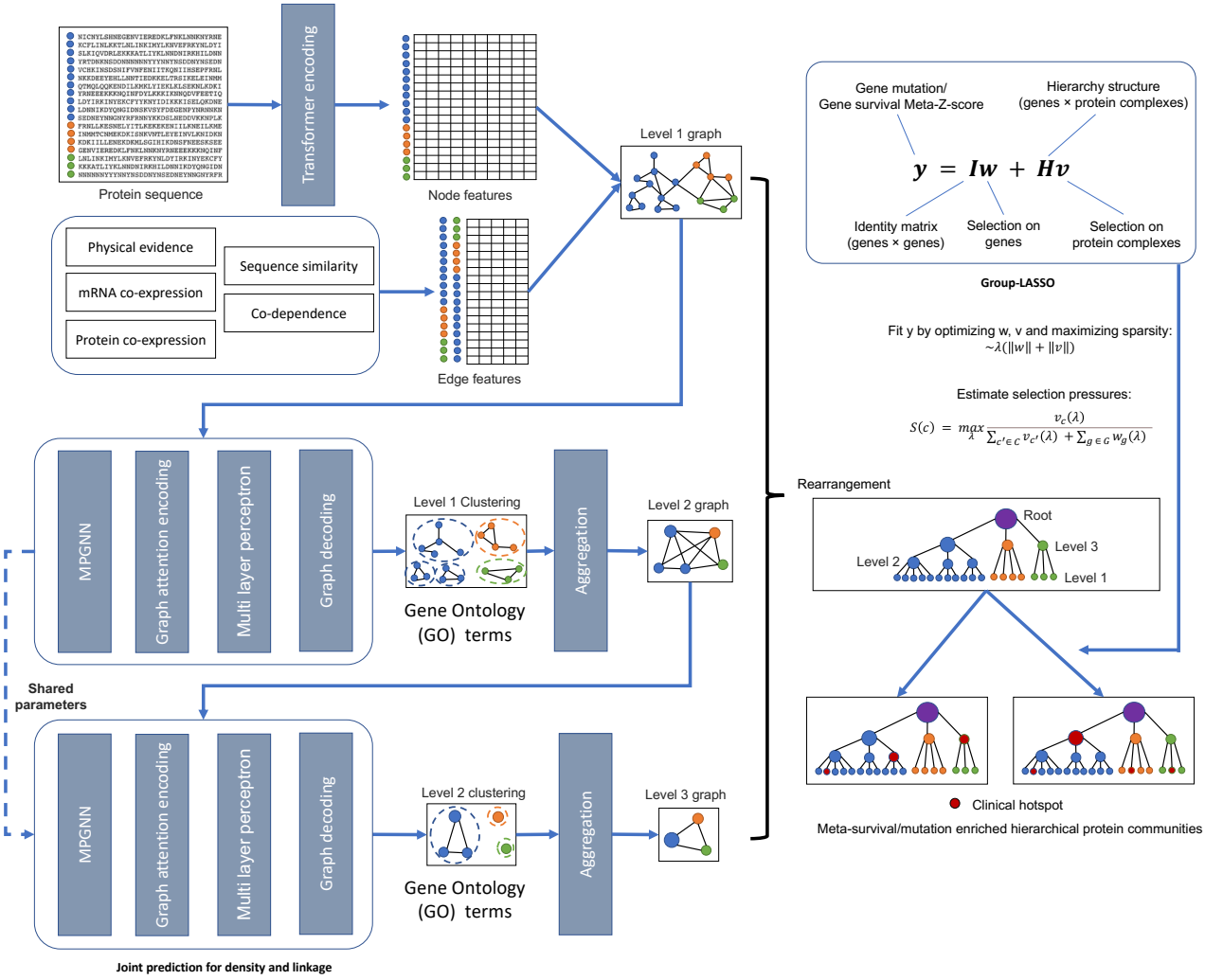


Figure 4-2. The meta-survival/mutation enriched hierarchical biological system detection framework. A pre-trained deep learning model was used to extract 1,280 numeric features from the amino sequence of each protein. These numeric features were stored in the nodes of the graph (a node in the graph is a protein). The pairwise protein association/interaction information was collected from previously published evidence (physical interaction evidence, mRNA co-expression, protein co-expression, sequence similarity, and co-dependence). There pair-wised features were stored in the edges that link two proteins. The graph was then input into a message passing GNN **(**MPGNN) block to embed edge features and node features for new node representations. The learned new node features were then input into a supervised GNN module to predict the node GO annotations (labels). There are several GNN modules, and each of

them configured with an aggregation module to aggregate a cluster of nodes identified in current level of the graph to a new node in the next level of the graph. The new graph was input into another GNN module, and the procedure was repeated again. Eventually a hierarchical tree of these predicted clusters was constructed by rearranging the nodes in each level of the graph.

4.2.2 Data sources

There were three data sources used in this study, including protein amino sequence data, PPI data, GO terms, gene-level meta-survival effects, and gene-level mutation counts. The protein amino sequence data of 19,035 proteins were downloaded from the Uniport database, which is the world's leading high-quality, comprehensive, and freely accessible resource of protein sequences (Bateman *et al.*, 2021). The PPI data were downloaded from Zheng *et al.*'s paper. They first identified 1722 PPIs from BC cell lines (MDA-MB-231, MCF-7, MCF-10A), which were quantified by mass spectrometry using PPI confidence-scoring algorithms. Then, these experimental results were integrated with a broad collection of previously identified human PPI data and other 4 types of protein-protein association evidence, namely mRNA co-expression, protein co-expression, sequence similarity, and co-dependence. Finally, the combined PPI data consist of 1.8×10^8 protein pairs. Each protein pair has 5 scores to indicate their associations. The human GO terms were downloaded from GO database (The Gene Ontology Consortium *et al.*, 2004). There are 5,764 unique genes with annotated GO terms. The gene-level meta-survival effects were downloaded from Smith and Sheltzer's paper (Smith and Sheltzer, 2022). They constructed Cox regression models for all gene by omics data types and cancer types in TCGA and found more than 112,000 gene features with significant survival associations. We downloaded the BC-specific z-scores of these Cox models. Each gene can have 4 different z-scores corresponding to 4 omics data used in the Cox regressions: CNV z-score, gene expression z-score,

DNA methylation z-score, and mutation z-score. After matching these data sources, we got the final graph with 4,968 nodes/genes/proteins, and 710,751 edges/protein pairs.

4.2.3 Protein feature extraction

We utilized a recently published pre-trained deep learning model, Evolutionary Scale Modeling (ESM)-1b Transformer (Rives *et al.*, 2021), to extract 1,280 numeric features for each protein in our dataset. ESM-1b is a 33 layers' Transformer-based deep contextual language model with approximate 650 million parameters, which was trained on 250 million protein amino sequences. It has been proved to be able to learn deep representations from proteins with very good quality (Rives et al., 2021; Mahmud et al., 2021). The protein features generated in the first step was then used to construct an affinity graph together with the PPI scores from Zheng *et al*.'s study.

4.2.4 Hierarchical biological system tree construction

GNN-based hierarchical clustering model, such as Hi-Lander, has been proposed to solve supervised image clustering problem and achieved state-of-the-art performance (Xing *et al.*, 2022). However, the Hi-Lander model uses k nearest neighbors (k -NN) to initial the links/edges of the graph in different scales in the hierarchy. It cannot take the protein-protein interactions/associations into consideration, which is the key information for studying biological systems. To address this problem, we modified the initialization procedure of the first-level graph of the Hi-Lander and added a message passing GNN (MPGNN) (Gilmer *et al.*, 2017) head for learning node representations from both node features and edge features. We named the proposed method as MPGNN-HiLander.

The graph $G_0 = \{V, E, X_V, X_E\}$ generated in section 4.2.2 consisted of both node features X_V (1,280 ESM-1b extracted protein features) for 4,968 nodes (V) and edge features X_E (5 types of protein-protein association scores) for 710,751 edges (E). It was used as the first level of the graph in our hierarchical clustering framework (**Figure 4-1**). A clustering function ϕ can take G as input and output an edge subset $E' \subset E$ i.e., $E' = \phi(G)$. The resulting graph $G' = \{V, E', X'_V\}$ (edge features X_E were embedded into node features X_V, thus were not involved in second level and above) was then split into connected components, with each corresponding to a cluster of nodes. Nodes in the same cluster were aggregated using an aggregation function ψ into a new node which constructs the next level graph. This single-level clustering and aggregation operation can be generalized to multiple levels and form the hierarchy eventually. Given a $G = \{V, E, X_V\}$, a sequence of graphs $G_l = \{V_l, E_l, X_{v_l}\}$ can be generated iteratively, where $i = 1 \ldots |V_i|$ and $l = 1 \ldots$, were iteratively generated using a base cluster function ϕ and an aggregation function ψ. We constructed and realized the clustering function and aggregation function of second level and beyond using the same loss functions as the Hi-Lander method (Xing *et al.*, 2022).

The data were randomly split into training/validation/test sets in a ratio of 70%:10%:20%. The model's performance on the test set was evaluated using three metrics: Pairwise F-score (Fp), BCubed F-score (Fb), and Normalized Mutual Information (NMI). CliXO (Zheng *et al.*, 2021), GCN-V (L. Yang *et al.*, 2020), GCN-V+E (L. Yang *et al.*, 2020), and Hi-LANDER (Xing *et al.*, 2022) were used as baselines. After the model was well-trained, we applied it to the whole dataset to generate the biological system hierarchy for downstream analysis.

4.2.5 Clinical hotspot detection

After the hierarchical tree was built, we applied HiSig algorithm (**Equation 4-1**) (Xing *et al.*, 2022) to detect which biological systems are under mutation/survival pressure.

Equation 4-1

$$\boldsymbol{y} = \boldsymbol{Iw} + \boldsymbol{Hv}$$

Here, *y* is the corrected mutation counts or meta-z-score for all genes, *I* is an identity matrix with the size of genes by genes. *H* is a gene by systems matrix. If a gene is a member of a biological system, the corresponding element in *H* will be 1, otherwise it will be 0. While *w* and *v* are weight vectors to be estimated.

The corrected mutation counts were calculated as the log transformed observed counts minus the expected counts. The expected counts were computed using MutSigCV 1.4 with default settings. It accounts for covariates of mutation tendency, including gene length, mRNA expression level, replication timing, and trinucleotide context of the mutation, which are integral parts of the MutSigCV statistical model (Lawrence *et al.*, 2013) . Meta-z-score was calculated using **Equation 4-2**.

Equation 4-2

$$meta - z = \frac{\sum_{i=1}^{K} Z_i}{\sqrt{K}}$$

Where K is the number of platforms, in this case, K = 4 (CNV, gene expression, DNA methylation, mutation). The Z_i were calculated by dividing the coefficient β in the Cox regression model (**Equation 4-3**) by its standard error. The univariate Cox regression model used to generate the z-scores could be formulated as follow:

Equation 4-3

$$h(t, X) = h_0(t)e^{\beta X}$$

Where t is the survival time, $h(t, X)$ is the hazard function, h_0(t) is the baseline time-dependent risk of death, X is a time-independent gene feature been tested.

By solving **Equation 4-1** in a LASSO regulated way (fit y by optimizing w, v and maximizing sparsity: $\sim\lambda(\|w\| + \|v\|)$, the weight vectors w and v could be estimated under different λ. They model the positive mutation/survival selection pressures on genes and systems, respectively. We tested 500 different λs, which resulted 500 optimal solutions of w and v. Each biological system c was then assigned a selective pressure *S(c)* over all λs. The selective pressure *S(c)* is defined as the maximum fraction of the weight of a biological system c among all weights:

Equation 4-4

$$S(c) = \max_{\lambda} \frac{v_c(\lambda)}{\sum_{c' \in C} v_{c'}(\lambda) + \sum_{g \in G} w_g(\lambda)}$$

The statistical significance of the *S(c)* was tested in a permutation way. First, an empirical *P* value was calculated by comparing *S*(c) of the actual hierarchy against 10,000 random hierarchies in which the hierarchy structure H is permuted with respect to gene labels (i.e., permuting the rows in *H*). Then the false discovery rate (FDR) was calculated using the Benjamini-Hochberg procedure, and a pre-defined cut-off of FDR< 0.25 was used to select the significant biological systems (clinical hotspots).

4.2.6 Evaluation of the identified clinical hotspots

Since the biological systems were identified from BC cell line data and publicly available cancer PPI information. To further evaluate their significance in BC cohorts. We checked the genetic alterations of these biological systems in TCGA-BRCA and METABRIC cohorts using the CBioPortal OncoPrint function (Gao *et al.*, 2013). The survival difference between the alternated patients and the non-alternated patients was also evaluated for each biological system.

4.3 Results

The performance of the proposed MPGNN-HiLander and baselines could be found in **Table 4-1**. The proposed MPGNN-HiLander achieved comparable performance with HiLander but had better performance than other baseline models including CliXO.

Table 4-1. GO annotation prediction performance.

Method	BP			CC			MF			GO		
	Fp	Fb	NMI	Fp	Fb	NMI	Fp	Fb	NMI	Fp	Fb	NMI
CliXO	0.08	0.15	0.51	0.05	0.09	0.49	0.06	0.13	0.59	0.11	0.13	0.57
GCN-V	0.24	0.31	0.66	0.27	0.29	0.69	0.24	0.35	0.60	0.23	0.26	0.59
GCN-V+E	0.24	0.32	0.67	0.31	0.29	0.68	0.28	0.34	0.62	0.24	0.31	0.60
Hi-Lander	**0.54**	0.49	**0.72**	0.56	**0.59**	**0.77**	**0.62**	**0.55**	**0.70**	0.53	**0.56**	0.66
MPGNN-HiLander	0.52	**0.50**	0.71	**0.57**	**0.59**	**0.77**	0.60	**0.55**	0.69	**0.54**	**0.56**	**0.67**

There were 878 biological systems identified by the well-trained MPGNN-HiLander model. They formed a 6-layer hierarchy. The detailed information of the hierarchy could be found in **Appendix B**. Among the 878 systems, 131 are under significant survival pressure (**Appendix**

C), and 60 are under significant mutation pressure (**Appendix D)** and 16 are under both survival and mutation pressures. These 16 biological systems are shown in **Table 4-2**.

Table 4-2. The identified biological systems with both survival and mutation pressures.

System	Genes	Survival			mutation		
		Pressure	p-value	FDR	Pressure	p-value	FDR
C48	MAML1,PYGO2,CBLL1,NCOA6,SS18L1,EP400,BCL9,SAP130,BCL9L,ARID1A,CITED2,DAZAP2,SS18,ARID1B	0.004706	0.02	0.1416	0.007887	0.02	0.1418
C58	ATAD5,RBM44,BRCA2,ICE1,SGO2,BRCA1,CRYBG3,TET1,CEP192	0.014937	0.01	0.0896	0.024231	0.02	0.1418
C59	BCL3,FBXL19,SET,SMARCA5,MPHOSPH8,NXF1,PAN3,BRF1,BRD9,KAT6B,BRPF1,JADE1,KDM5A,BRD1,WWOX,SRCAP,HNRNPU,KAT8,HMGB1,DPF3,DNMT3B,KMT2E,KMT2C,PIAS4,KAT2B,ERN2,DPF2,ASH2L,DGKI,ERN1,MSH6,KAT6A,SMARCA1,PHF12,ING4,ING5,HMGB2,PHF20L1,TBC1D5,ING3,INO80,TIMM50,BOD1L2,CHD4,RIOX2,MCRS1,PHF10,SMARCB1,ING2,L3MBTL1,HNRNPUL1,JADE3,PARP1,JADE2,CXXC1,DPF1,KDM2B,SCML2,DCLRE1C,CBX5,LIN28A,BOD1,DBI,CBX3,ECT2,JMJD6,MBTD1,SMARCA2,BRPF3,CHD5,KAT7,GLYR1,RFC1,BRD7,NT5DC3,PAXIP1,DNMT3A,KAT5,BRD8,CSNK2B,SMARCA4,LIG3,PRICKLE4,SETD5,ORC1,CHD3,PHF20,PBRM1,PHF14,ANP32E,KAT2A,EXOSC10	0.038839	0.01	0.0896	0.153701	0.02	0.1418
C175	HNRNPH1,GRPEL2,HNRNPH3,TFAM,ESRP1,GRPEL1,ESRP2,GRSF1,METTL18,HNRNPR,A1CF,SYNCRIP,CYB5B,HNRNPF,COQ3,HNRNPH2,TREX1	0.003392	0.02	0.1416	0.000985	0.03	0.2477
C198	KMT2D,FLYWCH1	0.001861	0.02	0.1416	0.000993	0.03	0.2477
C243	HSP90AA1,HSP90AB1,PMS1,HSP90B1,MLH1	0.017458	0.01	0.0896	0.000998	0.03	0.2477
C356	MYO1G,MYO6,MYO5A,MYO1B,MYO3B,MYO7B,MYO18A,MYO19,MYO1D,MYO3A,MYO5B,MYH15,MYO1A,MYO10,MYO7A,MYH7B,MYO1E,MYO9A,MYO1F,MYO16,MYO5C,MYO9B,MYO1C	0.002267	0.01	0.0896	0.001996	0.03	0.2321
C375	ATF6B,ATF6,NFIL3,BACH2,NFE2L2,MYC,BACH1,CREB3,NFE2L3,NFE2,ATF4	0.009705	0.03	0.2011	0.002925	0.03	0.2321
C392	POM121C,NUP153,NUP58,NUP50,NUP54,NUP62,POM121,NUP98	0.0163	0.01	0.0896	0.001197	0.03	0.2477
C395	ODF2,CEP83,EEA1,SCLT1,CEP290,CEP128,AKAP9,TPR,RABEP1	0.007122	0.01	0.0896	0.020845	0.02	0.1418
C416	PHKA1,PHKA2,PHKB	0.007611	0.01	0.0896	0.004959	0.02	0.1418
C560	IQGAP1,DMD,SYNE1,SPTA1,SPTBN1,SPTAN1,UTRN,FLNA,ACTN4,SYNE2,SPTB,SPTBN5,SPTBN4,SPTBN2	0.047516	0.01	0.0896	0.008456	0.02	0.1418
C619	XAB2,TTC30A,MAU2,SART3,PEX5,ANAPC7,EMC2,NAA16,CRNKL1,PEX5L,TTC37,SF3B1,UTP6,TTC21B,PRPF6,PRPF39,FNTA,ANAPC5,COPE,NAA25,IFT88,GTF3C3,CTR9,TTC30B,TTC8,NAA15,RABGGTA,TTC26,SRP72,CNOT10,TTC21A	0.002781	0.01	0.0896	0.003944	0.03	0.2477
C639	WASF2,CEP170,RAD21,APBB1IP,PCM1,LEO1,WASF3,INO80D,EPC2,DROSHA,EIF4G3,KANSL2,EPC1,TAF1,KANSL1L,GTF2E1,SUPT20H,WNK1,CHTOP,SYNPO2,WBP11,TAF1L,ATXN7L3,KANSL1,WASF1,EIF4G1,LMO7,SF3B2,AKAP6,TOB1,HIPK2,MSL2,WASHC1	0.025465	0.01	0.0896	0.003139	0.03	0.2477
C721	SPDL1,UGT3A2,ODF2,CEP83,CUBN,RGPD3,FBXL12,EEA1,CNTRL,ITGBL1,RGPD1,TCF25,COMMD1,LRP6,AMHR2,TGS1,RGPD4,EIF2AK4,RIPK1,DFFB,CD14,ZZZ3,RGPD5,CHMP7,HPS3,FBXO33,VRK2,SWI5,MEPCE,TLR4,HSD17B8,ASCC1,RPA4,RHBDD2,SCLT1,FGL1,NBAS,SPECC1L,CCAR2,GTF3C4,JARID2,OTULIN,IKBKE,ANAPC4,CCDC68,RAG1,LENG8,SNAP47,UBL4A,SSX2IP,CEP290,INCENP,ALS2,FBX	0.023813	0.01	0.0896	0.001951	0.03	0.2477

	L15,LRP1,INTS8,FGG,LDLR,MTERF4,LRP2,RGPD8,LRP5,FANCG,RGPD2,CEP128,IKBKB,VIPAS39,RNF8,CHAF1A,TAF1A,LAS1L,SPATS2L,ING1,UGT3A1,MCM3AP,AKAP9,PDE3B,FBXL8,FBXO24,AZI2,CLIP2,CFAP206,TEFM,MAP7,LUM,RANBP2,TLR2,GRHPR,LRRK2,HSD17B10,INTS10,KNTC1,FGB,TRAPPC12,GLIPR1L1,ITLN1,VLDLR,DDX20,FBXO38,TLR7,TLR1,CTU2,TPR,WDR93,CBR4,TRAF3IP1,FBXO48,FBXL3,TBK1,STN1,NME8,AIMP2,SMG7,RABEP1,FGL2,CGNL1,LRP1B,FBXO39,RIPK3,TLR6,TLR10,CHUK,WTAP,MAP3K5,SPECC1,SMG6,FBXL22,NEK10,LRP8,DTX3L,SHTN1,FBXL18						
C829	RBM12B,RNASEH2B,LARP7,MTA1,EIF3D,EAF2,POLE3,MTA3,CC2D2A,EIF2S2,TAF13,SRP9,HYPK,RANBP3,BAZ1A,NPLOC4,PARP11,INTS12,POLR3G,RBM41,RNPC3,MEAF6,NECAP1,THRAP3,TAF12,MIER1,DR1,BCL10,DMAP1,CDC5L,MKRN3,ARID4B,PIK3C2G,SUPT6H,EAF1,MTA2,PIK3C2B,CARD8,RANBP3L,NCOR2,GEMIN6,TAF8,POLR2D,MLLT1,SNUPN,CCNT2,CCDC47,RRP7A,PNN,CDC73,ZMAT5,NEMF,CPSF4,CCNT1,TEP1,POLR3D,SOCS5,SOCS6,POLR2F,CREBZF,EHBP1L1,PRIM1,POLR1D,ZC3H13,CLIC6,LIN9,NPM1,CENPS,ATG3,LIG4,PEA15,RERE,PRPF18,UBAP1,ZCCHC8,L1TD1,KRIT1,SUB1,TAF7,POLR3GL,MPHOSPH6,TRDN,TAF9B,CUL7,DNAJC28,PAF1,KAT14,TAF11,NDUFA12,ZFC3H1,TADA2B,BPTF,HTATSF1,ANAPC15,TAF7L,FBXO21,FBXO8,MED18,APOBR,CCNK,BAG6,BAK1,CLNK,SNAPC3,PIK3R6,SEM1,PRKAB1,ZNHIT6,PEX13,LARP6,PEX3,RCOR3,EXD1,KMT2A,BAZ1B,TADA3,MAP4,PKD1L3,RCOR1,BAX,KDM6B,PCP4,LAGE3,RSPH4A,NACAD,DPY30,CPSF4L,CHRAC1,NCBP3,PPM1E,SLF1,HDGF,CSTA,RBM28,KMT2B,PPP2R3A,TIPIN,MIER2,NXF3,NFYB,NXF5,CARD16,CCDC8,NEDD4,MKS1,MLLT3,NECAP2,BCLAF1,CAT,PIK3C2A,TRIP4,URI1,SMARCC2,PPP1R15A,FRG1,TAF9,PPP1R8,TNKS1BP1,ERCC6,BAHD1,SOCS4,ZC3H8,RCOR2,ARID4A,ANKRA2,PRKAB2,MAPK6,SMARCC1,CEP76,NCOR1,PIK3R5,TADA2A,CHTF8,MIER3,BCL2L2,PPP4R2,PSMG1,DRAP1,THG1L,EZH2,IFT46,LCP2,RSF1,CLNS1A,MCL1,BTF3,BCL2L1,CRCP,EZH1,MAVS,GTF3C6,SUPT3H,SLF2,BCL2,CUL9,SMTNL1	0.004067	0.02	0.1416	0.004204	0.03	0.2321

BC patients with genetic alterations in 9 of these biological systems suffered from poor prognosis than patients without genetic alterations. The 5-year survival differences of METABRIC patients with and without genetic alterations in the 9 biological systems can be found in **Figure 4-3**, and the OncoPrint visualizations of the genetic alterations of some significant biological systems in METABRIC cohort are shown in **Figure 4-4**.

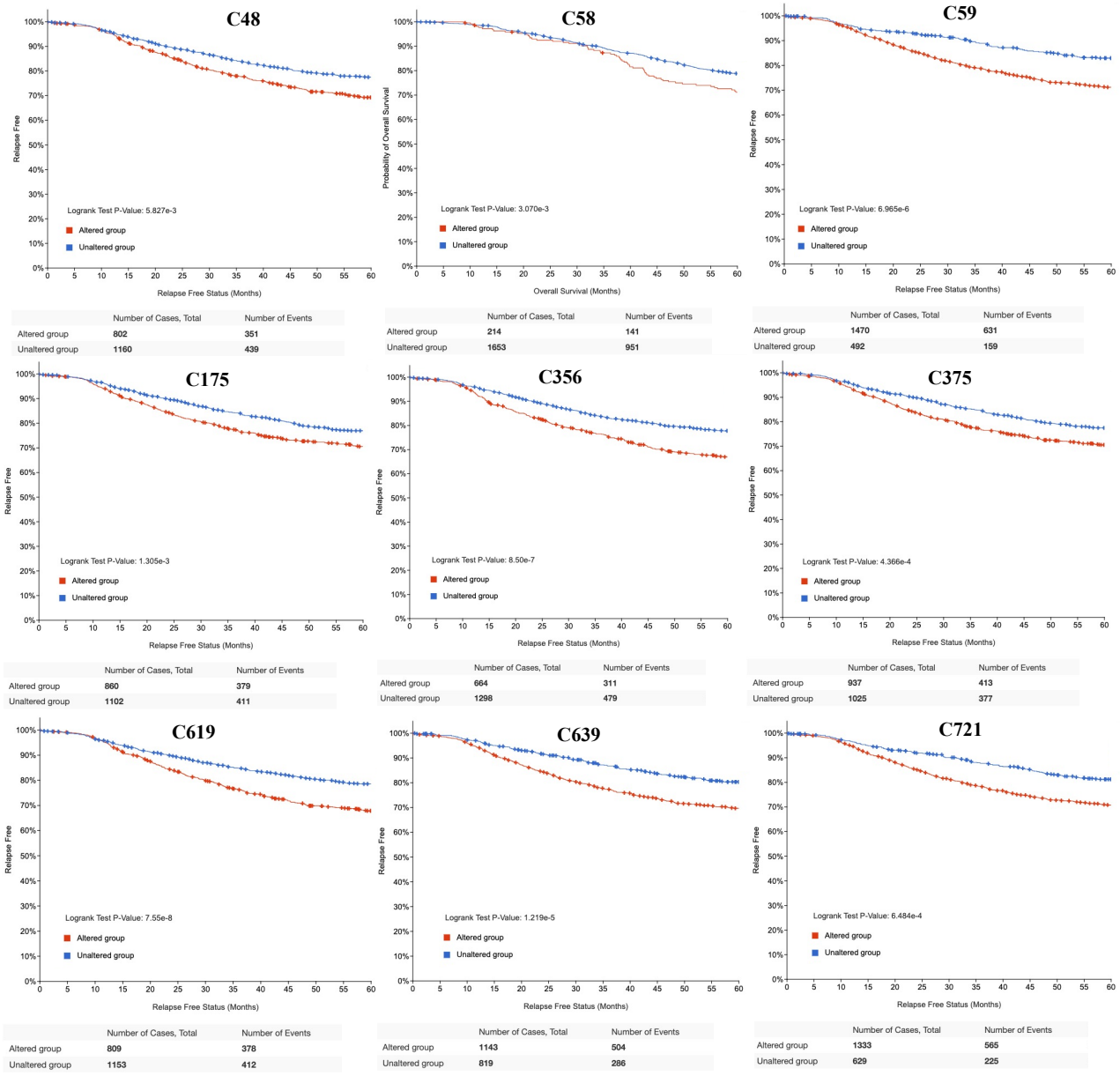


Figure 4-3. Survival differences of METABRIC patients with and without genetic alterations in the 9 biological systems.

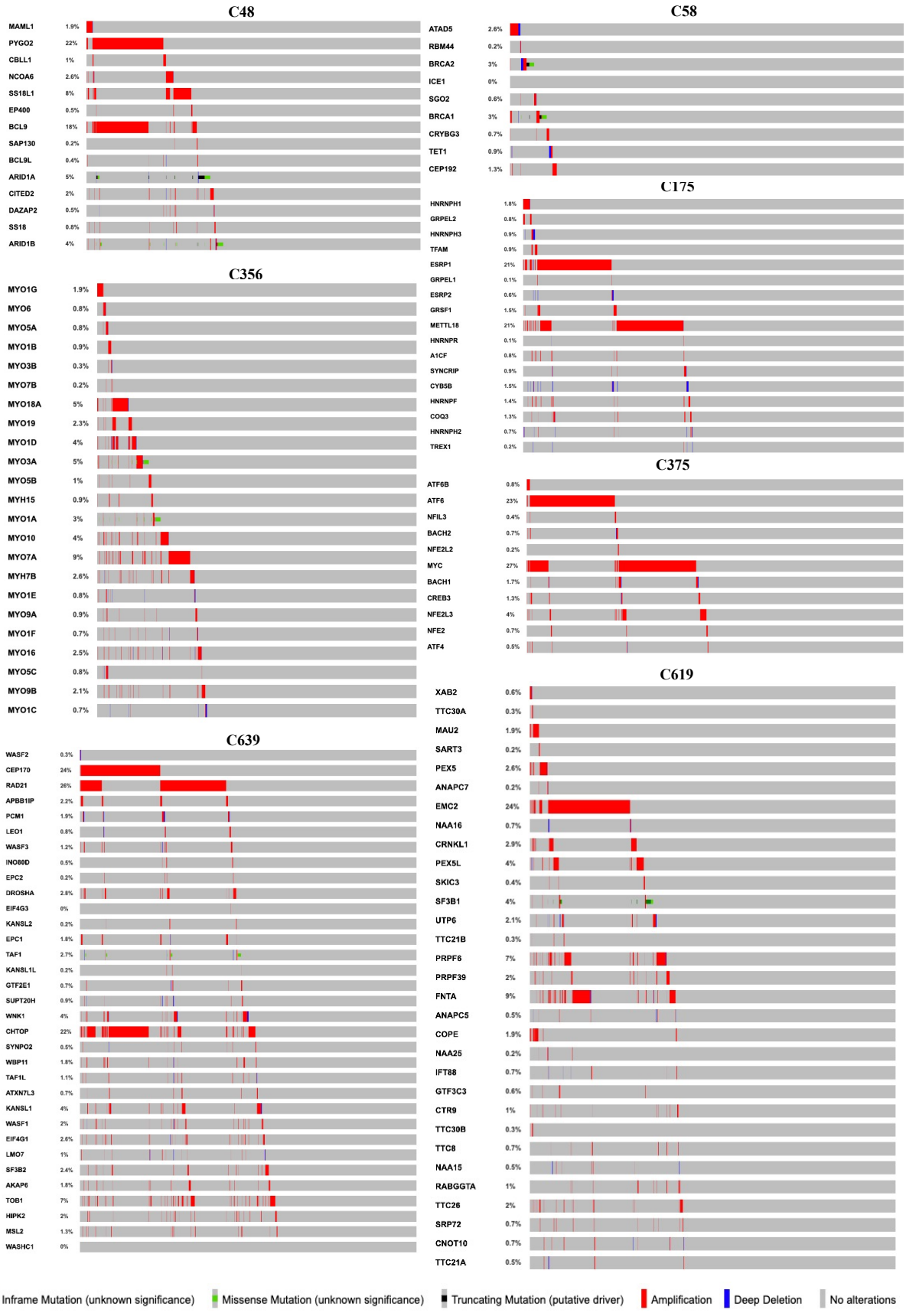

C48
MAML1 1.9%
PYGO2 22%
CBLL1 1%
NCOA6 2.6%
SS18L1 8%
EP400 0.5%
BCL9 18%
SAP130 0.2%
BCL9L 0.4%
ARID1A 5%
CITED2 2%
DAZAP2 0.5%
SS18 0.8%
ARID1B 4%
C58
ATAD5 2.6%
RBM44 0.2%
BRCA2 3%
ICE1 0%
SGO2 0.6%
BRCA1 3%
CRYBG3 0.7%
TET1 0.9%
CEP192 1.3%
C356
MYO1G 1.9%
MYO6 0.8%
MYO5A 0.8%
MYO1B 0.9%
MYO3B 0.3%
MYO7B 0.2%
MYO18A 5%
MYO19 2.3%
MYO1D 4%
MYO3A 5%
MYO5B 1%
MYH15 0.9%
MYO1A 3%
MYO10 4%
MYO7A 9%
MYH7B 2.6%
MYO1E 0.8%
MYO9A 0.9%
MYO1F 0.7%
MYO16 2.5%
MYO5C 0.8%
MYO9B 2.1%
MYO1C 0.7%
C175
HNRNPH1 1.8%
GRPEL2 0.8%
HNRNPH3 0.9%
TFAM 0.9%
ESRP1 21%
GRPEL1 0.1%
ESRP2 0.6%
GRSF1 1.5%
METTL18 21%
HNRNPR 0.1%
A1CF 0.8%
SYNCRIP 0.9%
CYB5B 1.5%
HNRNPF 1.4%
COQ3 1.3%
HNRNPH2 0.7%
TREX1 0.2%
C375
ATF6B 0.8%
ATF6 23%
NFIL3 0.4%
BACH2 0.7%
NFE2L2 0.2%
MYC 27%
BACH1 1.7%
CREB3 1.3%
NFE2L3 4%
NFE2 0.7%
ATF4 0.5%
C639
WASF2 0.3%
CEP170 24%
RAD21 26%
APBB1IP 2.2%
PCM1 1.9%
LEO1 0.8%
WASF3 1.2%
ING80D 0.5%
EPC2 0.2%
DROSHA 2.8%
EIF4G3 0%
KANSL2 0.2%
EPC1 1.8%
TAF1 2.7%
KANSL1L 0.2%
GTF2E1 0.7%
SUPT20H 0.9%
WNK1 4%
CHTOP 22%
SYNPO2 0.5%
WBP11 1.8%
TAF1L 1.1%
ATXN7L3 0.7%
KANSL1 4%
WASF1 2%
EIF4G1 2.6%
LMO7 1%
SF3B2 2.4%
AKAP6 1.8%
TOB1 7%
HIPK2 2%
MSL2 1.3%
WASHC1 0%
C619
XAB2 0.6%
TTC30A 0.3%
MAU2 1.9%
SART3 0.2%
PEX5 2.6%
ANAPC7 0.2%
EMC2 24%
NAA16 0.7%
CRNKL1 2.9%
PEX5L 4%
SKIC3 0.4%
SF3B1 4%
UTP6 2.1%
TTC21B 0.3%
PRPF6 7%
PRPF39 2%
FNTA 9%
ANAPC5 0.5%
COPE 1.9%
NAA25 0.2%
IFT88 0.7%
GTF3C3 0.6%
CTR9 1%
TTC30B 0.3%
TTC8 0.7%
NAA15 0.5%
RABGGTA 1%
TTC26 2%
SRP72 0.7%
CNOT10 0.7%
TTC21A 0.5%
Inframe Mutation (unknown significance)
Missense Mutation (unknown significance)
Truncating Mutation (putative driver)
Amplification
Deep Deletion
No alterations

Figure 4-4. Genetic alterations of some significant biological systems in METABRIC cohort.

4.4 Discussion

The proposed GNN-based hierarchical clustering model performed very well on BC graph structured PPI data. It successfully identified hierarchically structured biologically meaningful protein systems from protein sequence data, pair-wise protein interaction/association data, and GO annotation data, which could help us understand the heterogeneity of BC systematically. In addition, the survival and mutation pressure tests successfully pinpointed several statistically significant hotspots on the created hierarchy map. Among them, C58 system (ATAD5, RBM44, BRCA2, ICE1, SGO2, BRCA1, CRYBG3, TET1, CEP192) contains BRCA1 and BRCA2, which are known BC driver genes. They are both tumor suppressor genes, and they code proteins that are involved in transcription, DNA repair of double-stranded breaks, and recombination. Mutations in BRCA1 and BRCA2 are responsible for more than 40% of inherited BCs (Petrucelli *et al.*, 2022). Although the mutations in BRCA1 and BRCA2 have long been noticed to play important roles in BC, they are often missed in previous genome wide analyses because of their rare mutation frequency in population. We can identify them due to the aggregated effects from other members in the C58 system, which is exactly what the proposed hierarchical map wants to achieve. According to the evaluation results on the data from METABRIC, several clinical hotspots have the potential to be developed as BC prognostic biomarkers (C48, C58, C59, C175, C356, C375, C619, C639, C721). The 5-year survival differences showed that BC patients with genetic alterations in METABRIC in these 9 biological systems suffered from poor prognosis than patients without genetic alterations.

There are some limitations in this study. First, to match the GO annotations, we significantly decreased the number of proteins from 19,035 to 4,968 and decreased the number of pair-wised protein interactions/associations from 1.8×10^8 to 710,751 in the graph. Although this made the analysis more concise and efficient, some information may be missed. Second, some visualizations and manual annotations need to be added to the created biological system hierarchy. More biological evaluations (such as gene set enrichment analysis) may be performed after visualizations are added. A potential future direction is that more clinically relevant information such as DepMap dependency score (Tsherniak *et al.*, 2017) and drug response information could be used to select the hotspots from the hierarchy tree.

4.5 Conclusion

In conclusion, we showed that the GNN model can organize the structured proteomics data into clinically valuable hierarchies. This kind of biological hierarchy map provides a new perspective to explore BC heterogeneity. Layers of screening will increase the chance of identifying the rare mutations and minor survival hazards. The study has promising potential to map more clinically relevant information to the identified hierarchy to provide different types of biomarkers for BC.

5 Chapter 5: BC radiogenomic biomarker discovery

This chapter is modified from the published paper below. I performed all data analyses, paper writing and revision.

***Q Liu**, P Hu. 2022. A novel integrative computational framework for breast cancer radiogenomic biomarker discovery. Computational and Structural Biotechnology Journal. 66:102111, https://doi.org/10.1016/j.csbj.2022.05.031. (IF: 6.155). I did the data analysis, manuscript writing and revision with the guidance from Dr. Pingzhao Hu.*

5.1 Introduction

BC is a solid tumor with very high heterogeneity that comes from a variety of cellular function gain and loss during the development of the tumor. The widely accepted cancer theory indicates that there might be ten large biological capabilities acquired during the course of human tumors (Hanahan and Weinberg, 2011). These abnormal biological capabilities include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis, energy metabolism, evading immune destruction, and tumor microenvironment (Hanahan and Weinberg, 2011). They influence intermediate phenotypes such as tumor morphology and then eventually change the clinical outcomes such as overall survival (OS). Therefore, it is critical to characterize the cellular heterogeneity comprehensively. Meanwhile, the intermediate tumor morphology is also worthy of consideration in estimating the clinical outcome of patients. The cellular heterogeneity could be detected in different biological levels using a variety of modern molecular biological techniques which could generate high-throughput measurements (gene expression, CNV, and DNA methylation). These measurements contain rich and valuable information about the molecular

heterogeneity but are hard to be understood directly by human. A lot of computational tools have been developed to help human experts summarize the heterogeneity of cancer cells from these high-dimensional molecular biology measurements (Rutman and Kuo, 2009). For example, in **Chapter 3**, we have already showed that the BTF features that extracted from multi-omics data performed very well in BC subtyping.

However, although multi-omic measurements could provide us with rich information about the cellular tumor heterogeneity, the genomic examination is invasive and sometimes expensive. In addition, it may not be able to capture the dynamic and macroscopic information of the whole tumor as the biopsy is often taking at a certain time point on a small bulk of tumor tissue. Radiomics is a research field where high-throughput medical image features are used to describe disease phenotypes (Lambina *et al.*, 2012). It could be used as an auxiliary or surrogate of the multi-genomic analysis. Medical imaging is non-invasive, so it is often used as a disease monitoring method and thus performed at multiple time points during the course of the BC. And the imaging region of interest (ROI) usually covers the entire tumor and even the tissues around the tumor. Current radiomic studies are facing feature subjectivity and interpretability trade-off issues. Traditional computational engineering methods usually involve human experts' pre-processing which introduces the subjectivity. Although human understandable image features such as tumor morphological features and the first-order, second-order statistic features of the image pixel distribution (Van Griethuysen *et al.*, 2017) could be generated using the traditional feature engineering methods, they are pre-defined and limited by human knowledge therefore may not be able to fully represent the image heterogeneity. Combining radiomics with genomics leads to the field of radiogenomics, which has a goal of noninvasively uncovering the radiogenomic biomarkers that could indicate the clinical outcomes of the patients (Rutman and Kuo, 2009).

Besides the challenges in radiomics and genomics, radiogenomics also faces the unpaired data problem. Currently, the publicly available BC datasets are usually incomplete to do a biomarker-oriented radiogenomics study. For example, a dataset may contain medical images and genomics data for the same patients, which provides us with enough information for feature extraction and radiogenomic mapping, but the patients' clinical outcomes might be hard to obtain as this may need long-term observation. Hence the prognostic significance of the image features could not be evaluated (i.e. the image features cannot be identified as prognostic biomarkers) (Clark *et al.*, 2013; Tomczak *et al.*, 2015). Effective utilization of the unpaired imaging, genomic, and clinical data should be considered wisely. Recently, with the fast development of deep learning techniques, deep learning-based feature extraction approaches have been widely used in radiomics (Visvikis *et al.*, 2019). Deep learning is highly flexible and accurate in analyzing multi-modal volumetric and dynamic medical images in a fully automatic and non-linear manner (Nie *et al.*, 2019). But image features extracted by sophisticated deep learning models are considered not human understandable. Therefore, it is critical to explore potential tools to increase their explainability (Barredo Arrieta *et al.*, 2020).

In this study, we propose a deep learning-based radiogenomic framework for prognostic biomarker identification. Our framework includes the following five modules: a deep learning-based multi-modal image feature extraction module with build-in saliency maps for the explanation of the deep learning model; a BTF multi-genomic feature extraction module with GSEA (Subramanian *et al.*, 2005) to explore the biological meaning of the extracted features; a radiogenomic leverage module consists of a series of predictive models to impute the unpaired imaging, genomic, and survival data; a prognostic biomarker identification module which uses survival analysis to evaluate the prognostic significance of each radiogenomic feature; and a

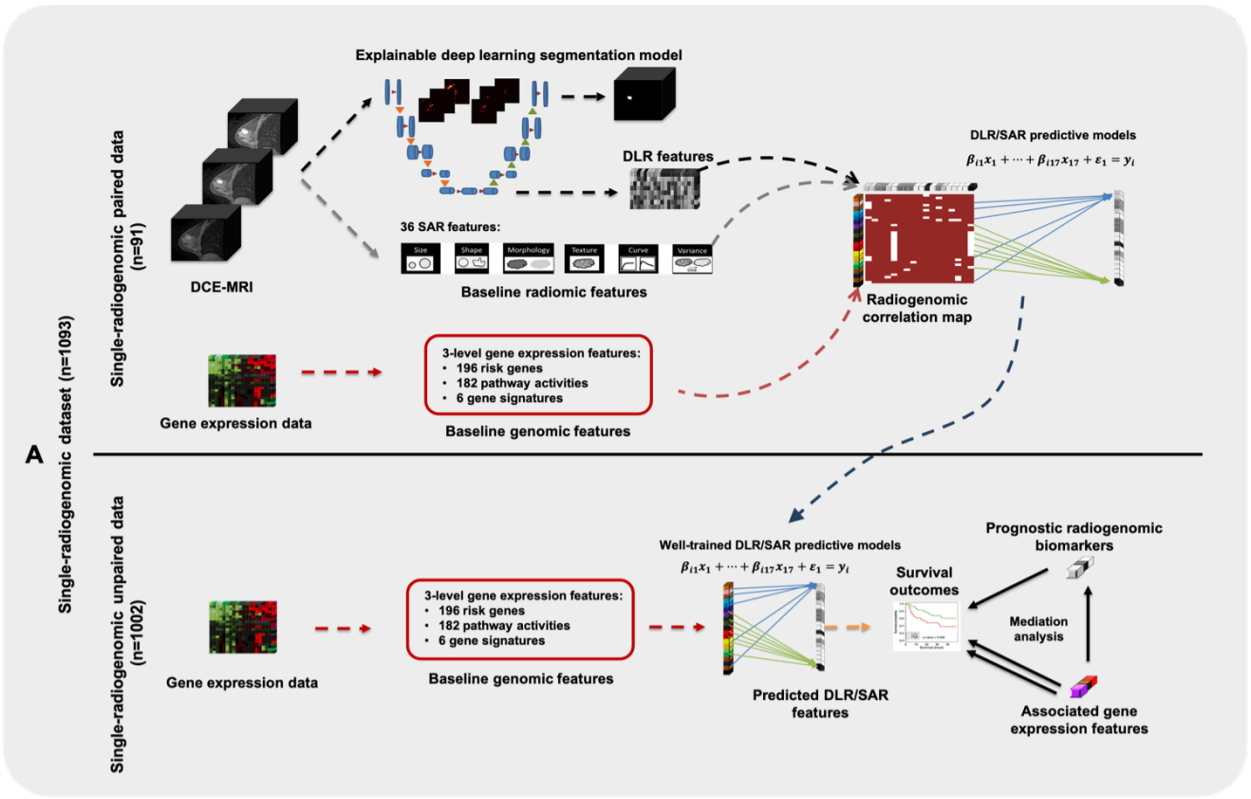
statistic mediation analysis module to provide potential biological causal inference of the identified prognostic biomarkers. It is expected that the identified radiogenomic biomarkers will have better prognostic significance than the traditional ones.

5.2 Material and methods

The overall design of this study is shown in **Figure 5-1.** Single-radiogenomic stage **(Figure 5-1A)** is a baseline workflow with only gene expression as the genomic data source. This is to test whether multi-genomic features have better radiogenomic associations than the single-genomic features. Multi-radiogenomic stage **(Figure 5-1B)** is the proposed workflow.

5.2.1 Formation of the datasets

Multi-source genomic data (gene expression, CNV, and DNA methylation) are provided by TCGA-BRCA project (Tomczak *et al.*, 2015). Medical image data, specifically, the 3D DCE-MRI volumes of a sub-cohort of the BRCA, are collected from The Cancer Imaging Archive (TCIA) (Clark *et al.*, 2013) platform. Part of these medical images has segmentation labels and 36 traditional semi-auto radiomic (SAR) features provided by the TCIA Breast Phenotype Research Group (Burnside *et al.*, 2016b). The data of BRCA cohort from both TCGA and TCIA form two datasets for this study: BRCA single-radiogenomic dataset, and BRCA multi-radiogenomic dataset. The exact data matching and formation workflow can be found in **Figure 5-2**. The demographic information of the sub-cohorts is listed in **Table 5-1**.


A
Single-radiogenomic dataset (n=1093)
Single-radiogenomic paired data (n=91)
Explainable deep learning segmentation model
DCE-MRI
DLR features
36 SAR features:
Size
Shape
Morphology
Texture
Curve
Variance
Baseline radiomic features
DLR/SAR predictive models
$\beta_{i1}x_1 + \cdots + \beta_{i17}x_{17} + \varepsilon_1 = y_i$
Radiogenomic correlation map
Gene expression data
3-level gene expression features:
• 196 risk genes
• 182 pathway activities
• 6 gene signatures
Baseline genomic features
Single-radiogenomic unpaired data (n=1002)
Gene expression data
3-level gene expression features:
• 196 risk genes
• 182 pathway activities
• 6 gene signatures
Baseline genomic features
Well-trained DLR/SAR predictive models
$\beta_{i1}x_1 + \cdots + \beta_{i17}x_{17} + \varepsilon_1 = y_i$
Predicted DLR/SAR features
Survival outcomes
Prognostic radiogenomic biomarkers
Mediation analysis
Associated gene expression features

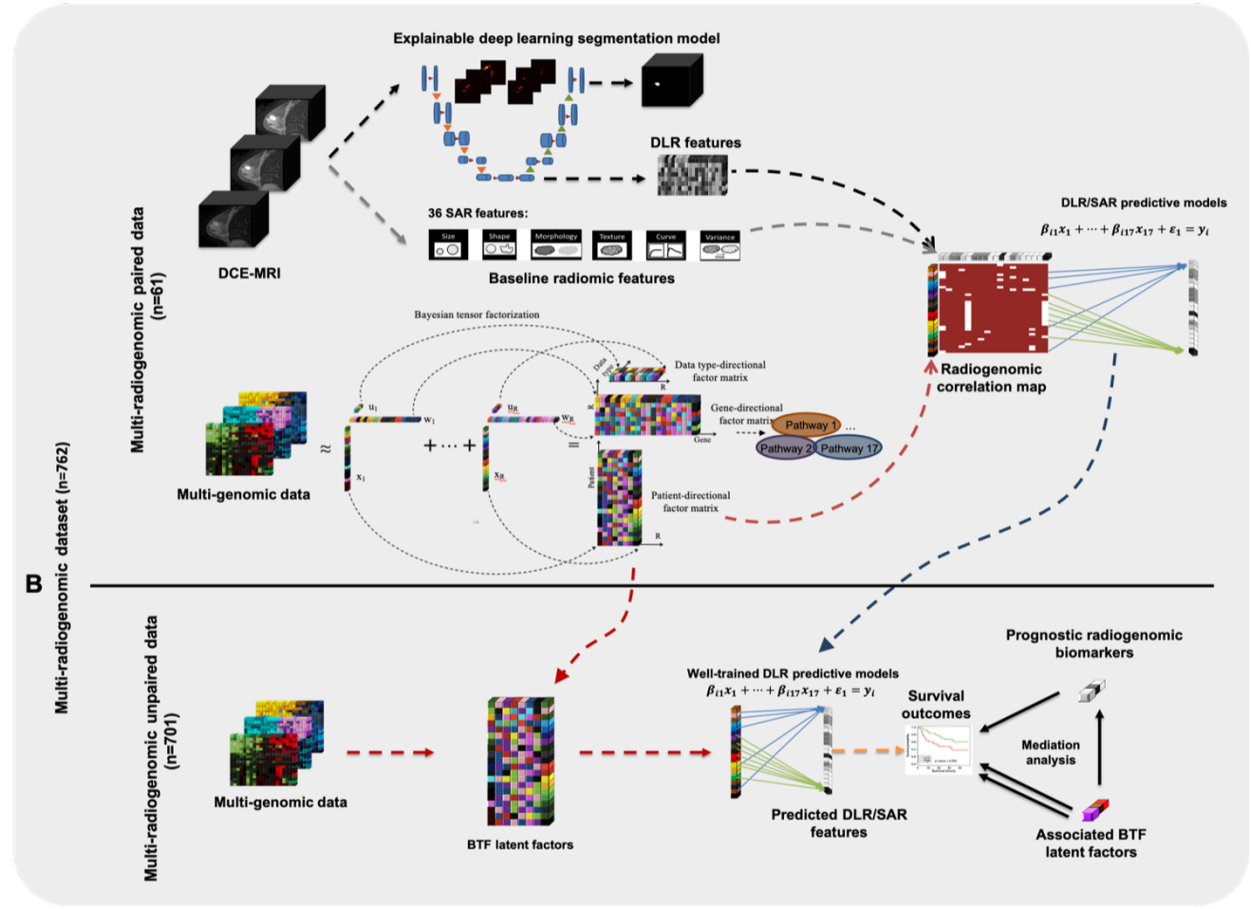

B
Multi-radiogenomic dataset (n=762)
Multi-radiogenomic paired data (n=61)
Explainable deep learning segmentation model
DCE-MRI
DLR features
36 SAR features:
Size
Shape
Morphology
Texture
Curve
Variance
Baseline radiomic features
DLR/SAR predictive models
$\beta_{i1}x_1 + \cdots + \beta_{i17}x_{17} + \varepsilon_1 = y_i$
Radiogenomic correlation map
Bayesian tensor factorization
Data type-directional factor matrix
Gene-directional factor matrix
Patient-directional factor matrix
Pathway 1
Pathway 2
Pathway 17
Multi-genomic data
Multi-radiogenomic unpaired data (n=701)
Multi-genomic data
BTF latent factors
Well-trained DLR predictive models
$\beta_{i1}x_1 + \cdots + \beta_{i17}x_{17} + \varepsilon_1 = y_i$
Predicted DLR/SAR features
Survival outcomes
Prognostic radiogenomic biomarkers
Mediation analysis
Associated BTF latent factors

Figure 5-1. The overall workflow of this study. A deep learning model (3DU-net) was built, trained, and validated to segment the tumor region from the raw three-dimensional DCE-MRI image. After the 3DU-net was well-trained, deep learning-based radiomic (DLR) features were extracted from the last hidden layer in the encoding phase of the model. Gradient-based saliency maps were generated to show the importance of each input pixel to the 3DU-net in making its segmentation decision. **A: Single-radiogenomic stage.** In this stage, we first focus on the paired data (top panel of A). Three-level gene expression features (197 BC risk gene expressions, 182 KEGG pathway activities, and 6 well-established BC gene signatures) are generated. Then, lasso models are built to predict each DLR feature and semi-auto radiomic (SAR) feature using these three-level gene expression features. After the predictive lasso models are well-trained and validated, we turn to the unpaired data (bottom panel of A). We generate the same three-level gene expression features using the unpaired data, then we apply the well-trained lasso models to get the predicted DLR and SAR features. In this way, we could generate the DLR and SAR features for the 1002 patients without medical images. Then, we performed survival analysis on the predicted DLR and SAR features. The significant ones are the identified prognostic radiogenomic biomarkers. Mediation analysis is then performed on these identified radiogenomic biomarkers to gain insight into the potential biological mechanisms underlying them. **B: Multi-radiogenomic stage.** In this stage, similar procedures of the single-radiogenomic stage are performed. We first focus on the paired data (top panel of B). We perform BTF on the multi-genomic data tensor to extract 17 BTF features. We also run GSEA to identify the key biological pathway of each BTF feature. These key pathways could explain the key functions of the identified multi-genomic BTF features. Then we train lasso models to utilize these 17 BTF features for predicting the DLR and SAR features. After the lasso models are well-trained and well-validated, we turn to the unpaired data (bottom panel of B). We obtain the BTF features using the multi-genomic data, then we apply the well-trained lasso models in the previous step to get the predicted DLR and SAR features. In this way, we could get the DLR and SAR features for the 701 patients without medical images. Then, we perform survival analysis on the predicted DLR and SAR features. The significant ones are the identified

radiogenomic biomarkers. Mediation analysis is then performed on each of these identified radiogenomic biomarkers to gain insights into the potential biological mechanisms of them.

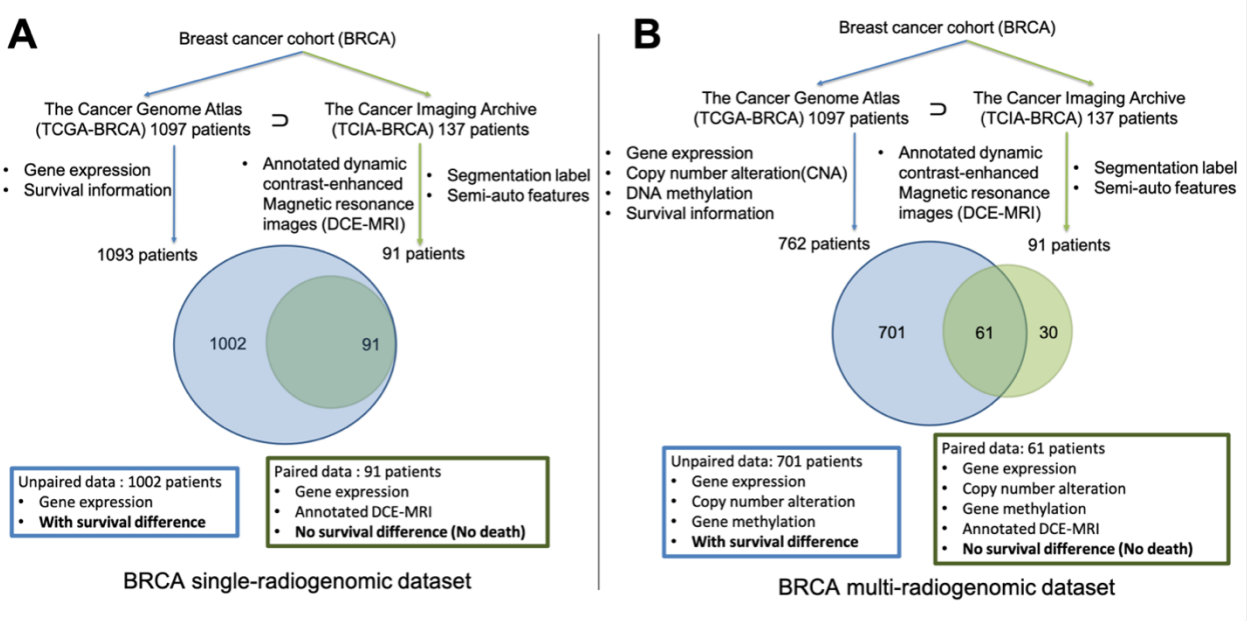


Figure 5-2. BRCA single- and multi- radiogenomic datasets organization flowcharts. TCGA provides genomic data and clinical data of a cohort consisting of 1097 BRCA patients. TCIA provides medical images of a subset of the TCGA-BRCA cohort (137 patients). Four of the 1097 BRCA patients have no meaningful survival information (death days or last contact days are negative numbers) thus were excluded. Ninety-one of 137 TCIA patients have annotated dynamic contrast-enhanced magnetic resonance images (DCE-MRI). **A: BRCA single-radiogenomic dataset.** One-thousand and ninety-three (1097 - 4) TCGA-BRCA patients all have gene expression data. Thus, gene expression data were used as baseline single-genomic information in this study to compare with the multi-genomic information. Those 91 patients with annotated DCE-MRI are all included in the 1093 TCGA-BRCA patients. This means, 91 patients have paired gene expression data and annotated image data. However, no survival difference is observed among these 91 patients, because they were all alive according to the last follow-up. Therefore, we cannot perform survival analysis using the paired data. The rest of 1002 (1093 - 91) patients only have gene expression data

(no DCE-MRI data), but there exists survival differences among them. **B: BRCA multi-radiogenomic dataset.** Only 762 of the 1093 TCGA-BRCA patients have matched gene expression, CNV, and DNA methylation data. Sixty-one of those 91 TCIA-BRCA patients with annotated DCE-MRI are included in the 762 TCGA-BRCA patients with multi-genomic data. This means, 61 patients have paired multi-genomic data and annotated image data. However, no survival difference is observed among these 61 patients. The rest of 701 (762 - 61) patients only have multi-genomic data (no medical image data), but there exists survival difference among them.

Table 5-1. Demographics of BRCA sub-cohorts.

		Single-radiogenomic dataset		Multi-radiogenomic dataset	
		Unpaired data	Paired data	Unpaired data	Paired data
Number of patients		1002	91	701	61
Age at diagnosis	≥65	329	14	215	8
	<65	673	77	486	53
	Mean	58.9	53.6	58.3	54.0
	Min	26	29	26	29
	Max	90	82	90	82
	Standard deviation	13.3	11.5	13.2	11.5
Stage	I	160	22	107	14
	II	561	58	386	39
	III	237	11	187	8
	X or IV	33	0	15	0
	Other	11	0	6	0
ER Status	Positive	729	77	499	53
	Negative	224	14	159	8
	Not Evaluated	49	0	43	0
PR Status	Positive	625	72	434	48
	Negative	325	19	221	13
	Indeterminate	4	0	2	0
	Not Evaluated	48	0	44	0
HER2 Status	Positive	150	14	81	7
	Negative	512	49	353	34
	Indeterminate	12	0	12	0
	Equivocal	157	22	118	18
	Not Evaluated/Available	171	6	137	2
Pam50 subtype	LumA	498	63	366	43
	LumB	197	11	132	9
	Her2	78	4	44	1
	Basal	178	12	126	7
	Normal	39	1	33	1
	NA	12	0	0	0

5.2.2 Explainable deep learning-based image feature extraction

DCE-MRI volumes of the same patients were acquired at different time points with an interval of dozens of seconds (Gordon *et al.*, 2014). That is where the "dynamic" comes from and it is a very strong advantage of DCE-MRI. Besides, the number of DCE-MRI volumes (i.e., time points) varies among patients, depending on the exam pipeline of the imaging institute and the patient's individual conditions (such as blood flow velocities). To handle this problem, a multi-modal three-dimensional deep learning model (Tulder and Bruijne, 2018; Zhang *et al.*, 2018; Vukotić *et al.*, 2016; Srivastava and Salakhutdinov, 2014) called 3DU-Net (Çiçek *et al.*, 2016) was applied to incorporate DCE-MRI volumes acquired at different time points to extract fused and dynamic DLR features. The modality of the input was set to the maximum number of volumes a given patient can have, which is 8. If the given patient has less than 8 DCE-MRI volumes, the position with absent volume was set to empty. The output of the 3DU-Net is the tumor segments provided by TCIA Breast Phenotype Research Group (Burnside *et al.*, 2016b). Two gradient-based saliency maps (Gradient map (Simonyan *et al.*, 2014) and Gradient*image map (Shrikumar *et al.*, 2016)) were embedded in the structure of the 3DU-Net to support explaining the segmentation decision. The detailed model structure of 3DU-Net is shown in **Figure 5-3**.

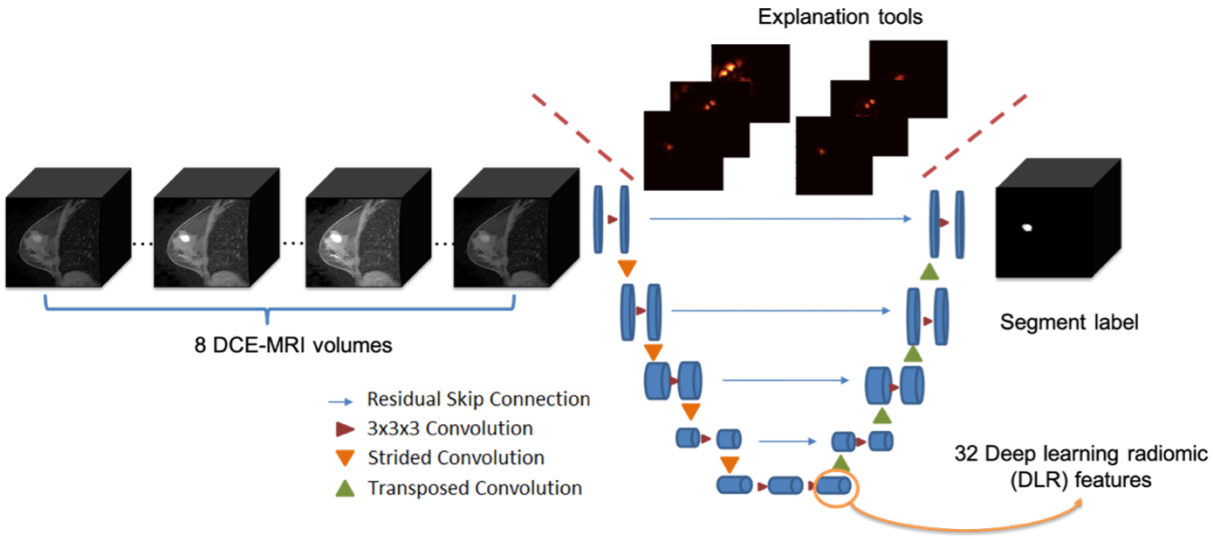


Figure 5-3. The structure of explainable 3DU-Net. The modality of the input is set to 8, which is the maximum number of volumes a patient can have. If a patient has fewer than 8 DCE-MRI volumes, the positions with absent volumes are set to empty. The output is the tumor segment annotation. The last hidden layer of the encoder phase is the DLR features. Two explanation tools (Gradient map and Gradient*image map) are used to increase the explainability of the 3DU-net.

The number of DLR features was set as 32, which is comparable with the 36 SAR features provided with the data. The 91 patients with annotated DCE-MRI data were involved in training and validating the 3DU-net. Patients were randomly split into training set (71 patients), validation set (10 patients), and test set (10 patients). We did hyperparameter tuning for the 3DU-net on stride, learning rate, and dropout ratio. The best hyperparameter combination was then used in the final model. After 1000 epochs, the performance of the segmentation task was measured by Dice similarity coefficient (DSC). Given a reference segmentation S_{lab}, the DSC of a predicted segmentation S_{pred} is defined as:

Equation 5-1

$$\mathbf{DSC} = \frac{2|S_{pred} \cap S_{lab}|}{|S_{pred}| + |S_{lab}|}$$

Then the well-trained 3DU-net was applied to the whole 91 patients for DLR feature extraction. We then performed pair-wise correlation analysis among all DLR features to see if they are correlated with each other.

5.2.3 BTF for multi-genomic feature extraction

Using the R package "tensorBF" (Khan and Ammad-ud-din, 2016), we implemented the BTF algorithm to extract latent factors (patient-directional projection matrix) from the gene expression, CNV, and DNA methylation data for the 762 patients from TCGA-BRCA. More details of the BTF algorithm can be found in Chapter 3. We did GSEA using the gene-directional projection matrix to further explore the potential key biological functions of each latent factor. Three-level gene expression features were generated as the baseline, including 196 BC risk genes identified by previous studies (Baxter *et al.*, 2018; Wu *et al.*, 2018), 182 pathway activities calculated using the Single Sample GSEA (ssGSEA) function (Barbie *et al.*, 2009) which was implemented in the GenePattern toolkit (Reich *et al.*, 2006), and 6 commercialized BC gene signatures calculated using R package "genefu" (Gendoo *et al.*, 2016). We then performed pair-wise correlation analysis among all BTF features and all three-level gene expression features to see if they are correlated with each other.

5.2.4 Leveraging strategy for radiogenomic feature imputation

To perform a biomarker-orientated radiogenomic research, ideally, we need to have matched medical images, genomic profiles, and clinical outcomes measured on the same set of patients, in which we can first identify radiomic biomarkers associated with clinical outcomes (e.g., prognosis), then we can associate the radiomic biomarkers with patients' genomic profiles to gain insight into their biological mechanisms. However, in the majority of cases, we only have one or two sets of data sources measured on the same patients. To solve the challenge, we used a leverage strategy (O. Gevaert *et al.*, 2012). We first focused on the paired part of the radiogenomic dataset, where we have the paired genomic data and medical image data for 61 patients. we trained lasso models (Friedman *et al.*, 2009) to predict each radiomic feature $\boldsymbol{y_i}$ using the genomic features *x* (**Equation 5-2, 5-3**).

Equation 5-2

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_g x_{ig} + \varepsilon}$$

Equation 5-3

$$\boldsymbol{\widehat{\beta^{lasso}} = arg \min_{\beta} \sum_{n=1}^{N} \frac{1}{2} (y_n - \beta x_n)^2 + \lambda \sum_{j=1}^{g} |\beta_j|}$$

The 61 samples were randomly split into training set (43 samples) and test set (18 samples). Prediction performances were evaluated using Root Mean Square Error (RMSE). Then we turned to the unpaired part of the radiogenomic dataset, where we only have the genomic data and patients' clinical outcomes without medical images for 762 patients. We applied the predictive models which were well-trained in the previous step to get the predicted radiomic features from the genomic features. In this way, we could get a completed paired dataset for further analysis. We also generated radiogenomic correlation maps between the radiomic features and the genomic features to explore the potential biological explanations for the relationship between them.

5.2.5 Survival analysis for radiogenomic biomarker identification

We further applied the function "surv_cutpoint" and "surv_categorize" in the R package "survminer" (Kassambara *et al.*, 2017) to select the optimized cut-off of each radiogenomic feature to categorize the patients into high and low-risk groups. For each radiogenomic feature, "survminer" looks for the cut-off where the log-rank test for survival analysis can produce the maximum statistic (lowest *p-value*). We classified the patients into the high-risk group and the low-risk group based on the cut-off for each radiogenomic feature. Then, we utilized the KM plot to show the survival difference between the high-risk group and the low-risk group.

5.2.6 Mediation analysis

The complexity of deep learning models and their low reproducibility have weakened their applications in clinical practice (Yan *et al.*, 2020). Hence, to enhance the biological interpretation of the DLR biomarkers, mediation analysis (MacKinnon *et al.*, 2007) between the genomic features and the BC prognosis through the identified radiogenomic biomarkers are implemented to reason on these radiogenomic biomarkers both biologically and statistically. By testing and estimating the mediation effects of the identified radiogenomic biomarkers on the relationship between genomic features and patient survival, biological interpretation of these radiogenomic biomarkers could be well made. We first regressed the survival outcome variable y_s against each genomic feature x_g (**Equation 5-4**). The effect β_{te} is the total (direct and indirect) effect of the genomic feature on the survival outcome. Then, we regressed the mediator (identified radiogenomic biomarker) m against x_g (**Equation 5-5**). The effect β_g is the effect of a genomic feature on a mediator. Lastly, we regressed the survival outcome y_s against both m and x_g

(**Equation 5-6**). β_{cme} is the indirect effect of the genomic feature on the survival outcome that goes through the radiogenomic biomarker. β_{de} is the direct effect of the genomic feature on the survival outcome.

Equation 5-4

$$\boldsymbol{y_s = \beta_{te} x_g + \varepsilon_t}$$

Equation 5-5

$$\boldsymbol{m = \beta_g x_g + \varepsilon_g}$$

Equation 5-6

$$\boldsymbol{y_s = \beta_{de} x_g + \beta_{cme} m + \varepsilon_m}$$

These mediation analyses are done using R package "mediation" (Tingley *et al.*, 2014). The significance of the estimated effects was tested and corrected using Benjamini-Hochberg multiple testing method (Benjamini and Hochberg, 1995).

5.3 Results

5.3.1 Radiomic and genomic features

The hyperparameter tuning for the 3DU-net could be found in **Table 5-2**. The best hyperparameter combination was stride = 1, learning rate = 0.001, dropout ratio = 0.2. The segmentation performance DSC of the well-trained 3DU-net on the test set is 0.44. The explanation saliency maps are shown in **Figure 5-4** and 36 DLR features were extracted from the well-trained 3DU-net. According to the saliency maps, the important pixels fall into and around the tumor regions, which means our DL model made a certain segmentation decision mainly based on the

tumor as well as surround-tumor regions. Using BTF algorithm, 17 multi-genomic factors were acquired. Their key biological functions identified by GSEA are shown in **Table 5-3**. These key functions range from cell division, blood vessel formation, immune response, intercellular signal transmitting, and so on, which quite fit the well-accepted cancer hallmark hypothesis (Hanahan and Weinberg, 2011). This means that the multi-omics BTF method captures cancer heterogeneity very well. The pair-wise correlation analysis among radiomic features and genomic features could be found in **Figure 5-5.** DLR features and BTF features are less redundant and could capture more information than SAR features and traditional gene expression features.

Table 5-2. 3DU-net hyperparameters tunning. Bold ones were used in the final model.

Stride	Learning rate	Dropout ratio	DSC (Validation set)	DSC (Test set)
1	0.0001	0.2	0.32	Not Applicable (N/A)*
2	0.0001	0.2	0.16	N/A
1	**0.001**	**0.2**	**0.42**	**0.44**
2	0.001	0.2	0.19	N/A
1	0.005	0.2	0.18	N/A
2	0.005	0.2	0.20	N/A
1	0.0001	0.4	0.28	N/A
2	0.0001	0.4	0.12	N/A
1	0.001	0.4	0.37	N/A
2	0.001	0.4	0.19	N/A
1	0.005	0.4	0.29	N/A
2	0.005	0.4	0.10	N/A
1	0.0001	0.6	0.37	N/A
2	0.0001	0.6	0.18	N/A
1	0.001	0.6	0.20	N/A
2	0.001	0.6	0.11	N/A
1	0.005	0.6	0.31	N/A
2	0.005	0.6	0.17	N/A

*Test set was applied to only the best trained model.

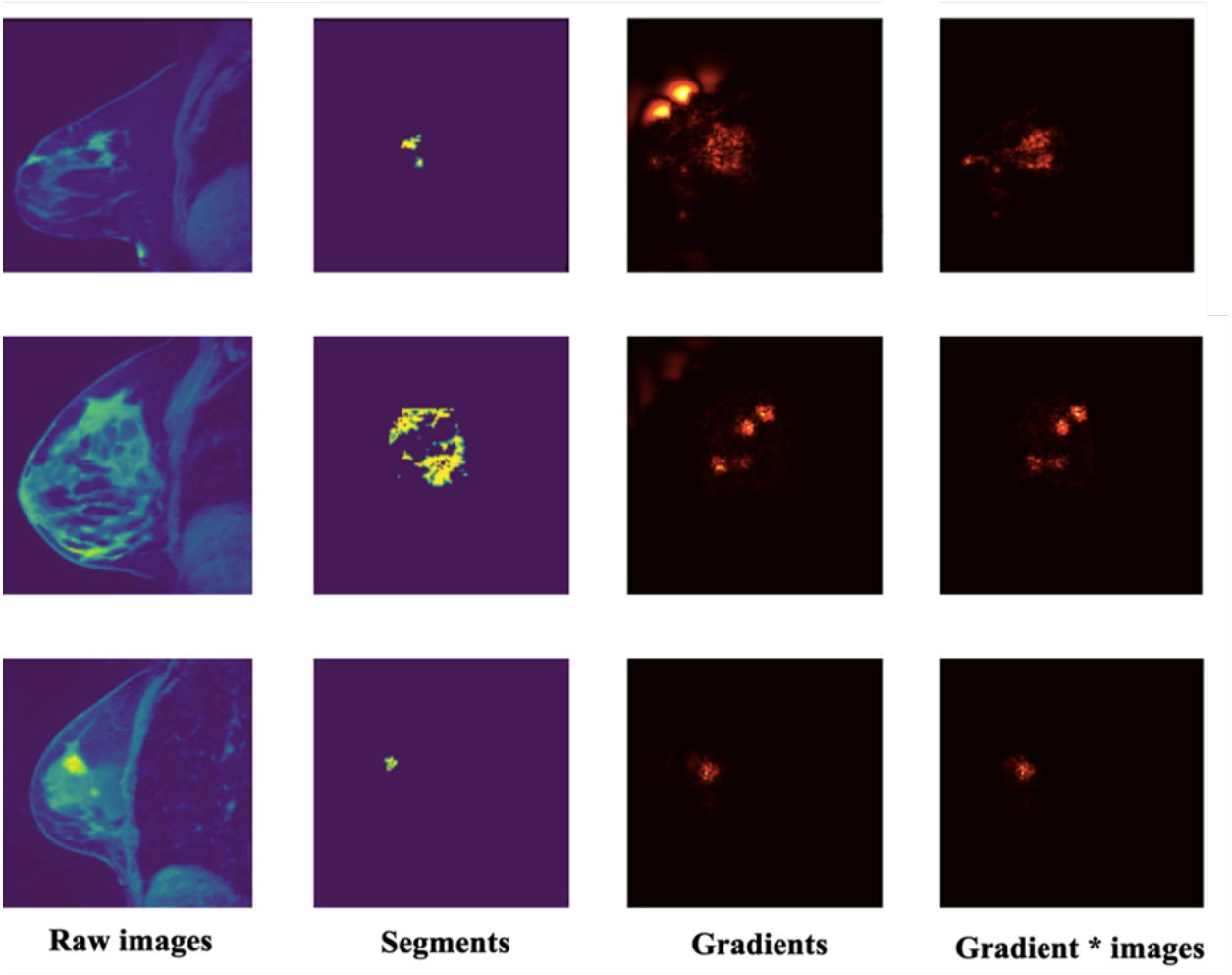


Figure 5-4. Image data and explanation saliency visualization. The first column is the raw images. The second column is the predicted tumor segments. The third and fourth columns are two kinds of saliency map generated using gradient method and gradient*input method separately.

Table 5-3. Key enriched pathways for the multi-genomic Bayesian tensor factors

BTF	Key pathways	NES	p-value	FDR	Key pathway genes
1	Chemokine signaling pathway	1.57	0.0059	0.25	CXCL5|CXCL1|CCL8|CXCL3|CCL13|PRKX|CXCL6|CCL8|CXCL2|CCL5|CCL2|CCL4|CCL25|ADCY3|GNB4|CCL3|RAC2|RELA|PPBP|CCL23|CCL24|ROCK2|ITK
2	Cytokine receptor interaction	1.96	<0.001	<0.001	CCL21|CCL19|CCL14|CXCL14|IL17B|CXCL2|CD40LG|CXCL12|TSLP|TNFSF11|CNTFR|CXCL1|CCL5|TPO|CCL23|IL21R|CCL13|ACVRL1|IL12B|CCL11|PDGFRB|CCL2|CD70|CCL16|CCL18|CCL4|IL7|CSF1R|TNFSF4|IL11RA|CXCL6|CTF1|CXCL3|EPOR|EDA2R|CCL3

3	Huntington's disease	-2	<0.001	0.0029	DNALI1|DNAI1|COX7B|CREB3|CLTA|NDUFS3|UQCRC1|COX6A1|AP2S1|NDUFA2|COX5B|NDUFA7|BBC3
4	Natural killer cell mediated cytotoxcity	-2	<0.001	0.0024	IFNA7|KIR2DL1|NCR1|ULBP1|ULBP2|RAC2|ZAP70|CD244|VAV1|KIR2DL4|CD48|GZMB|
5	Hematopoietic cell lineage	2.13	<0.001	<0.001	MS4A1|CR2|FCER2|CD5|CD2|CR1|CD1E|CD1B|CD1C|CD1D|CSF1R|CD33|IL7|TPO|CD1A|IL5RA
6	Starch and sucrose metabolism	-1.93	0.0029	0.03	PYGL|UGT2B10|UGT2B11
7	Cell cycle	-1.83	0.0047	0.08	CCNA1|ANAPC10|CCND2|CDKN1B|ANAPC7|CDC23|PTTG1|CDK1|CDC25C|ESPL1
8	Retinol metabolism	1.76	0.0063	0.16	ADH1C|UGT2B11|UGT2B10|UGT1A6|UGT1A7|UGT1A9|RPE65|CYP1A2|CYP2C19|ADH4|UGT1A8|UGT2A1|ALDH1A1
9	Steroid hormone biosynthesis	-1.65	0.0011	0.05	UGT2B11|CYP7B1|HSD17B7|CYP11B2|CYP11B1
10	Leukocyte trans endothelial migration	1.92	<0.001	0.04	MSN|CXCL12|CYBA|CYBB|JAM2|MYL2|CTNND1|SIPA1|CLDN22
11	VEGF signaling pathway	1.63	0.02	0.22	PTGS2|PLA2G10|CHP2|PLA2G4E|PLA2G2F
12	Olfactory transduction	-1.76	<0.001	0.10	OR51V1|OR2W1|OR12D2|OR2T5|OR7D4|OR10G7|OR2A2|PRKX|OR10G8|OR1I1|GUCA1C|OR14J1|OR10H5|OR11A1|OR13J1|OR10C1|
13	Oocyte meiosis	1.46	0.05	0.39	ADCY8|MOS|YWHAQ|SMC1A|CHP2
14	Drug metabolism cytochrome	2.32	<0.001	<0.001	GSTM1|UGT2B11|UGT2B10|GSTO2|MAOA|GSTM3|UGT1A7|ADH1C|UGT1A6|ALDH1A3|UGT1A9|UGT1A10
15	Endocytosis	1.81	0.0048	0.09	DNAJC6|CBLC|PSD2|CHMP4A|RAB11FIP4|EHD1|HSPA1A|RET|ARF6|CSF1R|VPS37C|AP2S1|RAB11B|ARAP3|RAB5A
16	Antigen processing and presentation	1.47	<0.001	0.0097	CD74|HSPA1A|PSME1|HSPA1B|PSME2|NFYB|HSPA2
17	Metabolism of Xenobiotics by cytochrome	2.16	<0.001	0.0010	UGT2A1|CYP2C9|UGT1A7|UGT1A3|UGT1A5|UGT1A8|GSTM1|UGT1A10|GSTA3|UGT1A4|CYP1A2|UGT1A1|UGT1A9|UGT1A6|CYP2F1|AKR1C4|CYP2C18|GSTA5|ADH4|CYP2C19|UGT2B10|ALDH1A3

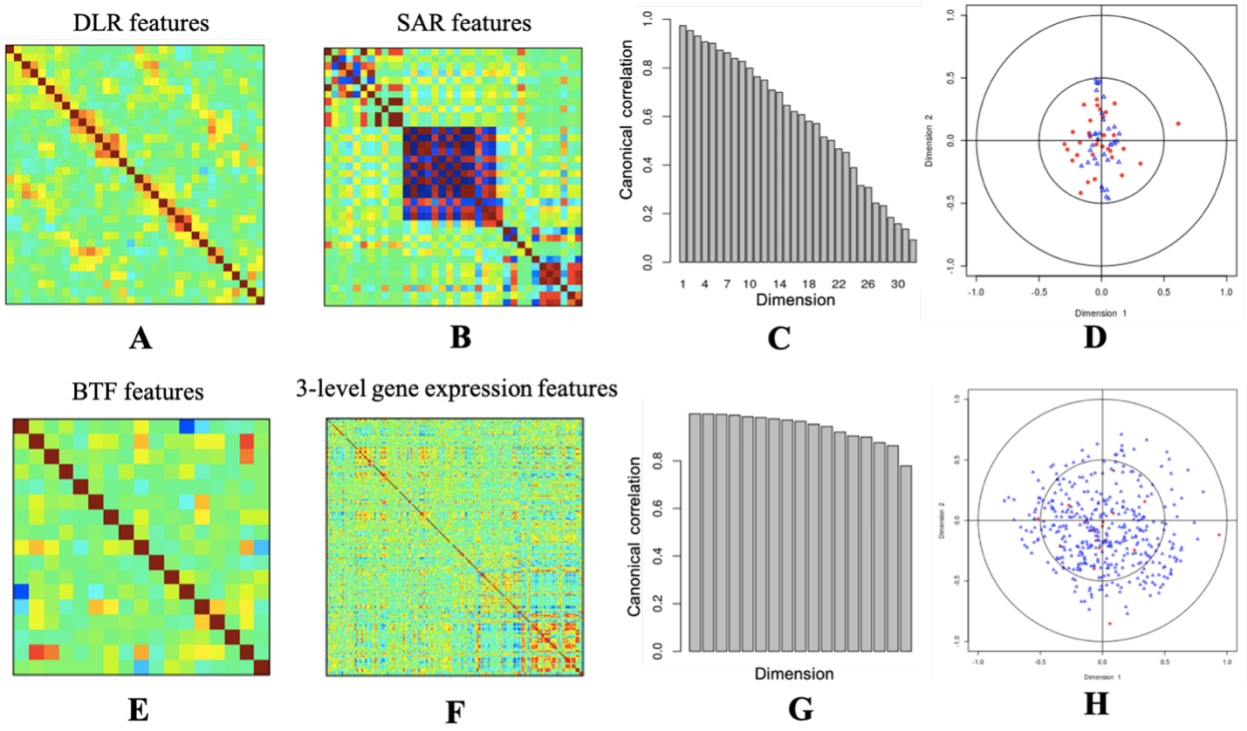


Figure 5-5. The radiomic feature correlation analysis and genomic feature correlation analysis. A: The pairwise DLR feature correlations. Columns and rows are 32 DLR features. The darker colors represent the higher correlations. **B: The pairwise SAR feature correlations.** Columns and rows are 36 SAR features. The darker colors represent the higher correlations. As we can see, some of the SAR features are correlated with each other. **C: The canonical correlations of the two radiomic feature matrices (DLR and SAR).** The x-axis is the canonical dimensions, while the y-axis is the correlation of the correlations between the DLR features and SAR features in each dimension. It is telling us, these two feature matrices are highly correlated with each other, which also means, the DLR features are able to capture the majority of information that the SAR features captured. **D: The scalar plot of the first two dimensions of DLR features and SAR features.** Blue ones are the SAR features, while red ones are the DLR features. DLR features may capture more information than the SAR features because the red dots are more widely spread. **E: The pairwise BTF multi-genomics feature correlations.** Columns and rows are 17 BTF features. The darker colors represent the higher correlations. **F: The pairwise three-level gene expression feature correlations.** Columns and rows are 197 (risk gene expressions) + 182 (pathway activities) + 6 (gene signatures) = 385 gene expression features. The darker colors represent the higher correlations. According to the results, we could see that BTF features are more independent than the baseline three-level gene expression features. **G: The canonical correlations of the two genomic feature matrices (BTF and three-level gene expression features).** The x-axis is the canonical dimensions, while the y-axis is the correlation of the correlations between the BTF features and three-level gene expression features in each dimension. **H: The scalar plot of the first two dimensions of BTF features and three-level gene expression features.** Blue ones are the three-level gene expression features, while red ones are the BTF features. The BTF feature matrix and the three-level gene expression feature matrix are highly correlated with each other, which also means, the BTF features are able to capture the majority of information that the three-level gene expression features captured.

5.3.2 Leveraging strategy for prognostic radiogenomic biomarker identification

The RMSE of the radiogenomic predictive lasso models could be found in **Table 5-4** (for DLR feature prediction) and **Table 5-5** (for SAR feature prediction). A lower RMSE means a better performance. As we can see, the multi-genomic BTF features perform overall better in predicting the DLR features than the baseline gene expression features. The radiogenomic correlation maps could be found in **Figure 5-6**.

Table 5-4. Performance of predictive LASSO models for each DLR feature.

Radiomic feature	Gene expression feature			BTF feature		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
DLR_1	83.08	68.11	0.4	57.22	35.87	0.75
DLR_2	71.6	47.31	0.26	49.81	27.36	0.76
DLR_3	65.16	53.33	0.82	21.08	17.65	0.1
DLR_4	60.55	49.51	0.71	22.72	16.18	0.09
DLR_5	61.45	46.26	0.3	16.64	11.69	0.07
DLR_6	60.13	44.88	0.25	28.14	24.48	0.15
DLR_7	53.99	40.35	0.24	50.74	35.39	0.43
DLR_8	54.66	44.39	0.25	53.93	40.58	1
DLR_9	107.26	87.58	4.36	80.61	60.65	3.52
DLR_10	98.83	79.18	2.85	25.76	25.51	0.13
DLR_11	45.66	34.04	0.23	12.01	12.01	0.06
DLR_12	24.76	17.26	0.12	30.44	13.04	0.12

DLR_13	27.85	18.32	0.13	30.63	13.29	0.13
DLR_14	27.52	18.57	0.13	33.67	20.68	0.16
DLR_15	104.52	81.16	4.08	69.06	50.32	2.26
DLR_16	89.73	65.58	2.42	53.86	33.54	1.74
DLR_17	102.34	85.94	0.51	44.35	36.26	0.3
DLR_18	77.37	60.19	0.49	47.01	31.37	0.59
DLR_19	58.2	46.47	0.77	27.88	26.51	0.15
DLR_20	62.56	42.71	0.22	22.39	17.25	0.11
DLR_21	37.11	27.53	0.15	32.35	26.22	0.19
DLR_22	64.53	51.65	0.3	25.29	18.7	0.09
DLR_23	50.77	37.48	0.24	29	23.1	0.15
DLR_24	57.25	45.85	0.27	52.08	33.93	1.14
DLR_25	63.32	49.08	1.14	70.53	38.77	2.57
DLR_26	22.65	17.78	0.1	12.98	9.12	0.05
DLR_27	21.42	12.33	0.09	4.84	4.84	0.02
DLR_28	32.88	16.05	0.17	3.03	3.03	0.02
DLR_29	7.35	5.67	0.03	0.01	0.01	0
DLR_30	17.17	13.57	0.07	2.79	2.79	0.01
DLR_31	60.98	44.31	0.95	17.77	16.69	0.09
DLR_32	94.82	76.64	2.78	53.41	41.07	1.59

Table 5-5. Performance of predictive LASSO models for each SAR feature.

	Gene expression features			BTF features		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE
SAR_1	1.45	1.16	0.79	0.8	0.59	0.27
SAR_2	123.96	105.41	1	131.96	96.43	0.64
SAR_3	0.02	0.02	1.32	0.02	0.02	1.35
SAR_4	0	0	2.1	0	0	0
SAR_5	0.26	0.21	2.79	0.2	0.16	1.43
SAR_6	1.43	1.12	0.81	0.81	0.6	0.31
SAR_7	0.6	0.47	0.45	0.3	0.2	0.17
SAR_8	93.36	72.46	353.62	10.28	4.51	2.17
SAR_9	147.92	128.1	0.93	131.97	131.31	1.47
SAR_10	0.67	0.52	521.79	0.15	0.05	2.98
SAR_11	0.17	0.13	0	0.02	0.01	0
SAR_12	12.71	11.27	0.33	10.46	8.39	0.38
SAR_13	0.08	0.07	0.09	0.07	0.06	0.07
SAR_14	0.17	0.15	0.06	0.18	0.13	0.06
SAR_15	6.13	5.31	0.34	4.59	3.75	0.36
SAR_16	0	0	0.17	0	0	0.09
SAR_17	0.14	0.13	0.02	0.15	0.1	0.02
SAR_18	0.04	0.03	0.09	0.04	0.03	0.07
SAR_19	0.04	0.04	0.27	0.05	0.03	0.17
SAR_20	0.07	0.06	0.08	0.06	0.05	0.06
SAR_21	0.08	0.07	0.09	0.07	0.05	0.06
SAR_22	0.81	0.7	0.02	0.83	0.58	0.02
SAR_23	0.02	0.02	0	0.01	0.01	0

SAR_24	14.96	12.98	0.05	16.5	11.09	0.04
SAR_25	2.23	1.75	0.02	2.11	1.35	0.02
SAR_26	0.2	0.16	0.23	0.13	0.12	0.18
SAR_27	0.19	0.15	0.28	0.1	0.08	0.14
SAR_28	0.06	0.05	0.24	0.08	0.06	0.24
SAR_29	0	0	1.84	0	0	1.07
SAR_30	0.01	0.01	0.79	0	0	0.44
SAR_31	12321.17	9011.66	6.58	5030.68	3546.42	1.79
SAR_32	11.14	7.73	0.47	6	4.7	0.28
SAR_33	9478.19	6306.81	4.36	4554.12	3011.8	1.26
SAR_34	0.35	0.32	0.36	0.37	0.24	0.28
SAR_35	163.74	92.75	0.93	55.17	43.45	0.59
SAR_36	28.85	20.23	0.87	14.35	11.19	0.43

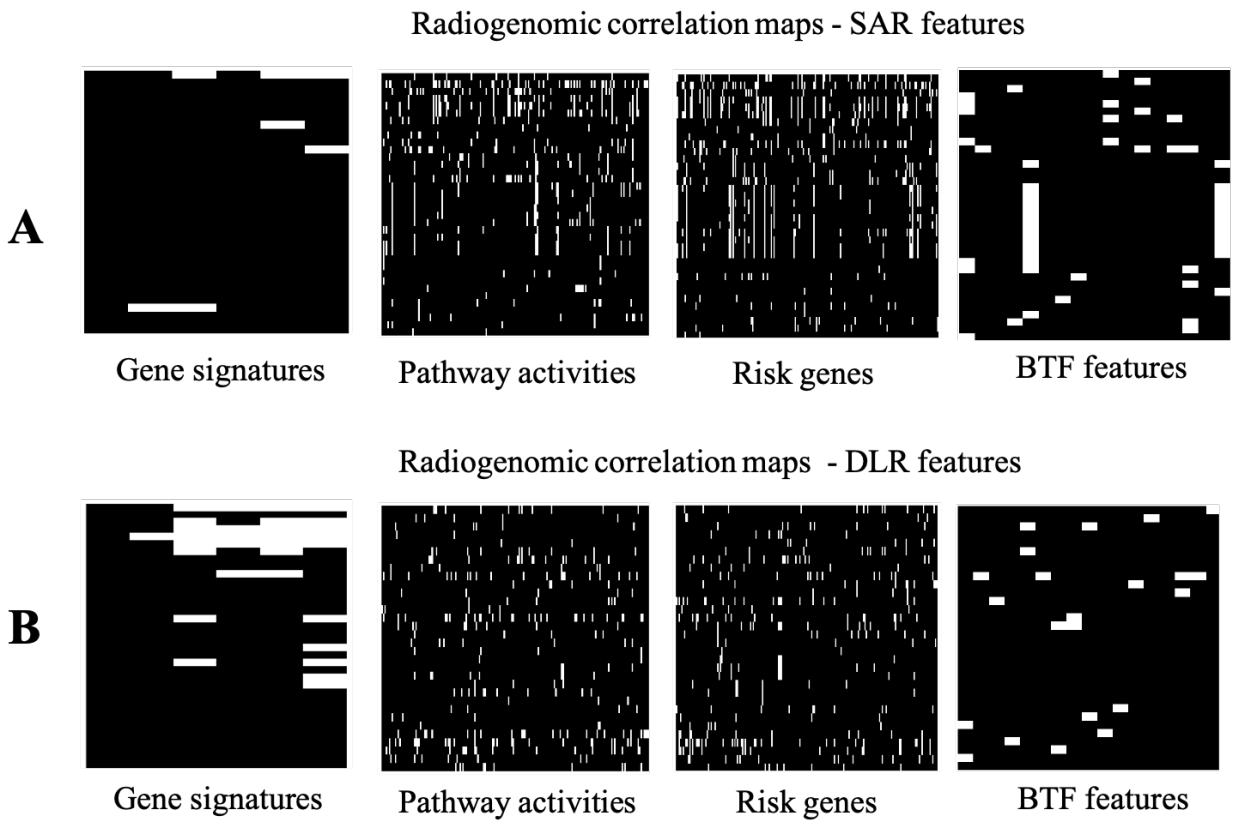


Figure 5-6. Radiogenomic correlation maps between the radiomic features and the genomic features. A: SAR feature-based radiogenomic correlation maps. Rows are 36 SAR features. Columns are different genomic features (6 gene signature scores, 182 pathway activities, 197 risk gene expressions). **B: DLR feature-based radiogenomic correlation maps.** Rows are 32 SAR features. Columns are different genomic features.

Twenty-three DLR features are significant in the survival analyses, which means we have identified 23 significant prognostic biomarkers using the proposed method and they have overall lower log-rank p-values than the SAR features (**Figure 5-7A**). The KM plots of the most prognostically significant DLR biomarker (DLR_8) and SAR biomarker (Maximum enhancement) are shown in **Figure 5-7B, C**. **Table 5-6** is showing the significant results of the mediation analyses.

The most significant DLR biomarker (DLR_8) is a significant mediator of the BTF_4 (Natural killer cell mediated cytotoxicity)'s effect on patient survival.

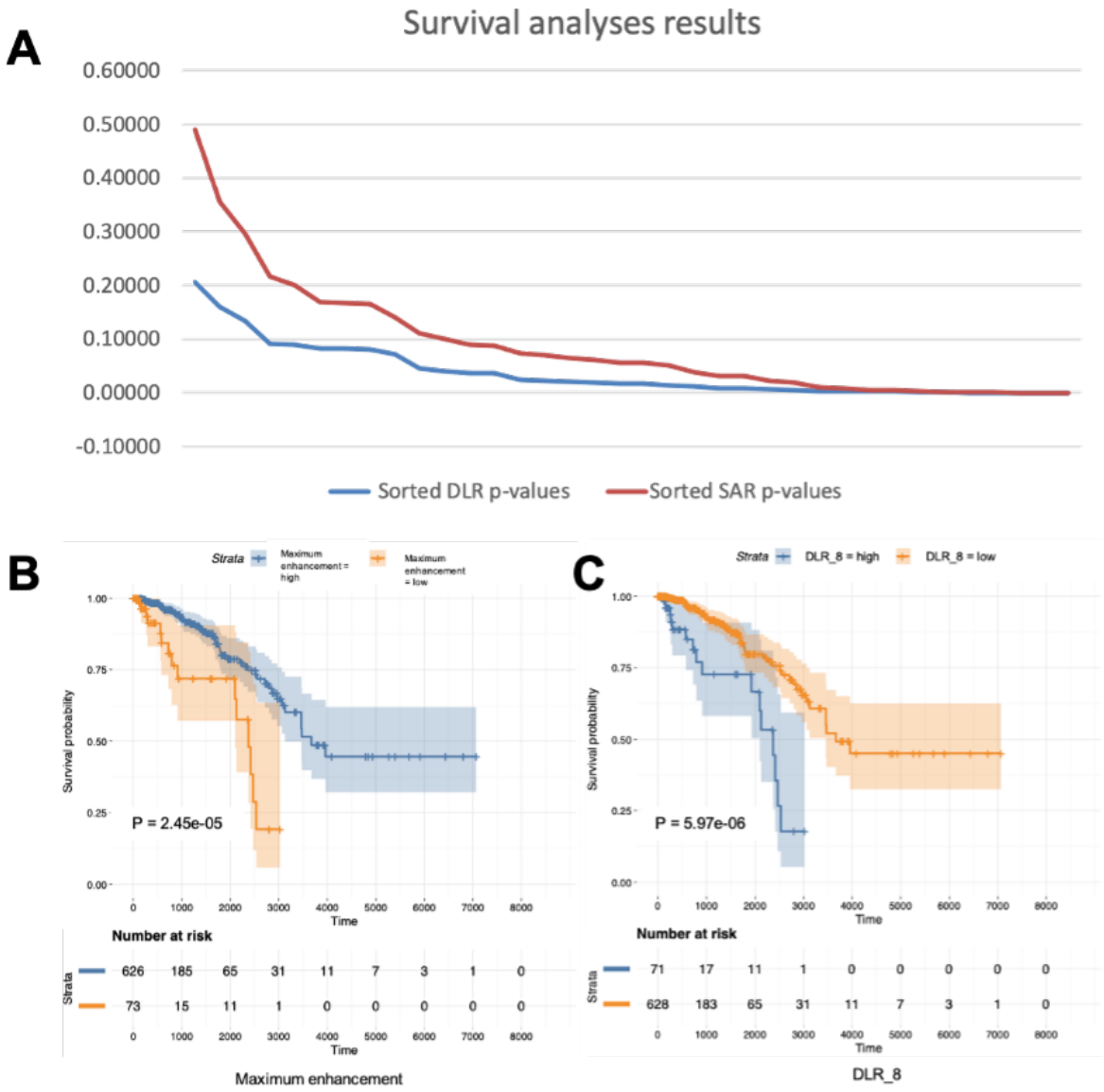


Figure 5-7. Prognostically significant DLR features and SAR features. A: The sorted *p-values* of survival analyses. DLR features showed overall lower p-values than SAR features. **B: The most prognostically significant SAR feature.** Maximum enhancement has the lowest p-value (2.45e-05) among

all SAR features in the survival analyses. **C: The most prognostically significant DLR feature.** DLR-8 has the lowest *p-value* (5.97e-06) among all DLR features in the survival analyses**.**

Table 5-6. The significant results of mediation analyses of the identified biomarkers

Independent variable	Mediator	ACME*	ACME_pvalue	ADE*	ADE_pvalue	TE*	TE_pvalue
Metabolism of xenobiotics by cytochrome	DLR_7	-0.03	0.044	0.21	<2e-16	0.18	0.004
BTF_4	DLR_8	30.38	0.05	-118.63	<2e-16	-88.25	0.026
BTF_7	DLR_2	9.00	0.046	-87.52	0.036	-78.53	0.05

*ACME (average causal mediation effects): indirect effect of the IV on the DV that goes through the mediator.
*ADE (average direct effects): direct effect of the IV on the DV.
*TE (total effect): direct and indirect effect of the IV on the DV.

5.4 Discussion

Two advanced mathematical methods, BTF and deep learning, were used to estimate the multi-level genomic and morphological heterogeneity of BC. BTF plus GSEA successfully provided us with biologically meaningful multi-genomic features. And their key biological functions are highly related to the known hallmarks of cancer (Hanahan and Weinberg, 2011), including signaling, cell cycle, metabolism, and immune related pathways. BTF features are more advanced than the single-source genomic features because they not only consider multiple genomic sources, but also consider the interaction between them. Deep learning could extract image features automatically and objectively but its explainability needs to be increased. The proposed workflow increased the explainability of the deep learning-based image feature extraction in two ways, one is by adding two explanation tools into the model structure, the other is by introducing domain knowledge to support the extracted image features. According to our experiment, the DLR features performed better than the traditional SAR features, thus, we believe

once the explainability issue is addressed, deep learning will have a bright future in healthcare data analyzing.

Leveraging strategy is often seen in the biomedical field (Lonsdale *et al.*, 2013; Olivier Gevaert *et al.*, 2012) because healthcare data are often not easy to get and thus will lead to the unpaired data problem. This is the first time that the leveraging strategy is being introduced into DL-based radiogenomics. It successfully solved the unpaired data problem in this case. Taking advantage of the estimated radiogenomic features which representing the multi-level tumor heterogeneity, we successfully identified several prognostic biomarkers using the proposed workflow. The most prognostically significant radiogenomic biomarker has a potential intermediate effect on the causal relationship between the function of nature killer cells and patient's survival time. The identified BC prognostic radiogenomic biomarkers are non-invasive and effectively representing both medical imaging and multi-genomic information. They are clinically more feasible because they could be obtained from medical images only, no need to perform the invasive biopsy.

5.5 Conclusion

In conclusion, we provided a comprehensive radiogenomic workflow which could overcome major difficulties of current radiogenomic studies. Our experiments showed that the proposed workflow could identify non-invasive, objective, automatic, integrated, and explainable BC radiogenomic biomarkers with great prognostic significance compared with the baselines. Our results also uncover genetic mechanisms regulating clinical phenotypes. Such mechanisms could promote medical imaging as a non-invasive examination of probing BC molecular status, then support clinical decisions and ultimately improve patient care.

6 Chapter 6: Significance, Limitations and Future Directions

Part of this chapter (mainly Section 6.2) is from a published review paper of mine: ***Q Liu****, P Hu. 2022. Extendable and explainable deep learning for pan-cancer radiogenomics research. Current Opinion in Chemical Biology. 66:102111, https://doi.org/10.1016/j.cbpa.2021.102111. (IF: 8.972). I did the literature reviewing, manuscript writing, and manuscript revision with the guidance from Dr. Pingzhao Hu.*

6.1 Significance

This thesis aims to explore the BC heterogeneity using machine learning algorithms at different data resolutions ranging from single genomics, multi-genomics, to proteomics, radiogenomics. Four independent and coherent studies were conducted and reported in the Chapters 2-5. The objective of Chapter 2 is to develop a diagnostic and prognostic gene expression signature for HER2+/ER+ BC, which is a subgroup of BC and has not been well-studied. Chapter 2 is the first to explore the heterogeneity associated with HER2+/ER+ BCs. We identified and validated two subgroups of HER2+/ER+ breast tumors with reproducible prognostic and other properties. We also provided a well-trained HER2+/ER+ subgroup classifier to assign new patients into these two subgroups. Potential biological explanations of the identified subgroups were discussed to link the subtype information we learned in this study with existing knowledge of BC. Several machine learning models such as CoC, XGboost, and DeepDep were used in this study, which confirmed the powerful ability of machine learning to analyze single genomic data. Most importantly, our findings may provide guidance for future target therapy of the HER2+/ER+ BC patients.

The objective of the Chapter 3 is to use integrated multi-genomics data to stratify BC. We achieved this goal by developing a strategy that combines BTF with CoC. This chapter explored the possibility of machine learning for multi-genomics data integration and proposed a novel subtyping scheme for BC. BTF took a tensor that combines different types of omics data as input, and output learned cross-omics representations directly. The proposed method could be easily transferred to identify subtypes of other complex diseases or be used in conjunction with other novel downstream subtyping algorithms.

In Chapter 4, we built a hierarchical tree to represent BC heterogeneity in a systematical view. GNN was proved to be capable of organizing the structured proteomic data into clinically valuable hierarchy. The built hierarchy map provided a new perspective to explore BC heterogeneity. Group-lasso-based layers of screening increased the chance of identifying the biologically and clinically relevant biological systems with rare mutations and/or minor survival hazards. Promising potential is foreseen as more clinically relevant information can be mapped to the identified hierarchy, thereby providing a new type of biomarker for BC.

Chapter 5 extended the multi-genomics by incorporating radiomics into biomarker discovery. We proposed a BTF and deep learning-based framework with lasso-based leverage strategy for BC prognostic radiogenomic biomarker identification. The proposed framework identified non-invasive, objective, automatic, integrated, and explainable BC radiogenomic biomarkers with significant prognostic significance. The relationship among the heterogeneity of the genetic mechanisms, the intermediate image phenotypes, and the clinical outcomes were also explored for BCs. This study could promote medical imaging as a non-invasive examination of probing BC molecular status, then support clinical decisions and ultimately improve patient care.

Overall, we explored different machine learning algorithms, different data types, different framework designs, etc. to deepen machine learning's applications in deciphering BC heterogeneity at different biological levels.

6.2 Limitations and future works

The first limitation of the study is that only publicly available data were used, which poses the unresolvable sample size problem, especially in the Chapter 5 where image data for only 61 patients was available. Similar issues can be seen in the Chapter 2, only 123 samples in TCGA-BRCA cohort were HER2+/ER+, making the multi-omics-based exploration for HER2+/ER+ BCs difficult because more samples are needed for capturing the cross-omics relationships. In the future, further evaluations should be conducted once more samples with high quality omics data are available. Another limitation is that due to the computational complexity to analyze both the MRI data and the graph structured proteomic data, we did not use all the available information for the current analysis. To control the time and cost, we had to sacrifice some information by compressing the MRI images to a lower resolution. More advanced parallel computing techniques or super-computing resources are needed to fully utilize the high-dimensional data.

The more challenging part is the biological interpretation of the results from machine learning/deep learning models, which requires continuous efforts. We partially addressed this issue by using interpretation tools to link the results to exist knowledge, but more standard explanation pipelines and quantitative metrics for evaluating the explainability are needed to regulate the use of complex models in the human healthcare domain.

Chapter 5 solved the unpaired data problem using the leverage strategy proposed by Gevaert *et al*. (O. Gevaert *et al.*, 2012). Other strategies such as deep generative model-based image synthesis can be explored to solve this problem so that the framework could be end-to-end, and the generated synthetic images could be shared to the public for further analysis without privacy issues.

Preclinical tumor models could be considered as a potential tool to validate the identified deep learning radiogenomic biomarkers. Such biological validation is well applied in some drug discovery and biomarker identification (Binder *et al.*, 2018). It could provide solid experimental evidence to gain the trust of clinicians and patients (Bodalal *et al.*, 2019). Shaikh et al. suggested to incorporate radiogenomics into existing and future clinical trials, to co-develop radiogenomic biomarkers for immunotherapy response and prognosis (Shaikh *et al.*, 2018).

The developed subtyping schemes/machine learning tools and the identified biomarkers could be archived into a publicly available BC multimodal biomarker bank (BioBank) to support both the research community and clinic use in the future.

Reference

van ’t Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–6.

Abel,H.J. and Duncavage,E.J. (2013) Detection of structural DNA variation from next generation sequencing data: A review of informatic approaches. *Cancer Genet.*, **206**, 432–440.

Alanko,J. and Ivaska,J. (2016) Endosomes: Emerging Platforms for Integrin-Mediated FAK Signalling. *Trends Cell Biol.*, **26**, 391–398.

Almahairi,A. *et al.* (2018) Augmented CycleGAN: Learning many-to-many mappings from unpaired data. *arXiv*.

Anders,C.K. *et al.* (2008) Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *J. Clin. Oncol.*, **26**, 3324–3330.

Barbie,D.A. *et al.* (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*.

Barredo Arrieta,A. *et al.* (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion*, **58**, 82–115.

Baselga,J. *et al.* (2012) Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): A randomised, open-label, multicentre, phase 3 trial. *Lancet*, **379**, 633–640.

Basse,C. and Arock,M. (2015) The increasing roles of epigenetics in breast cancer: Implications for pathogenicity, biomarkers, prevention and treatment. *Int. J. cancer*, **137**, 2785–2794.

Bateman,A. *et al.* (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids*

Res., **49**, D480–D489.

Baxter,J.S. *et al.* (2018) Capture Hi-C identifies putative target genes at 33 breast cancer risk loci. *Nat. Commun.*

Bender,L.M. and Nahta,R. (2008) Her2 cross talk and therapeutic resistance in breast cancer. *Front. Biosci.*, **13**, 3906–3912.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Van den Berge,K. *et al.* (2019) RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. *Annu. Rev. Biomed. Data Sci.*, **2**, 139–173.

Binder,Z.A. *et al.* (2018) Epidermal Growth Factor Receptor Extracellular Domain Mutations in Glioblastoma Present Opportunities for Clinical Imaging and Therapeutic Development. *Cancer Cell*, **34**, 163-177.e7.

Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev.*, **16**, 6–21.

Bismeijer,T. *et al.* (2018) Molecular characterization of breast and lung tumors by integration of multiple data types with functional sparse-factor analysis. *PLoS Comput. Biol.*, **14**, 1–28.

Blattler,A. and Farnham,P.J. (2013) Cross-talk between site-specific transcription factors and DNA methylation states. *J. Biol. Chem.*, **288**, 34287–34294.

Bodalal,Z. *et al.* (2019) Radiogenomics: bridging imaging and genomics. *Abdom. Radiol.*, **44**, 1960–1984.

Bontempi,G. *et al.* (2010) A fuzzy gene expression-based computational approach improves breast cancer prognostication. *Genome Biol.*, **11**, R18.

Brandão,M. *et al.* (2020) What Is the Real Impact of Estrogen Receptor Status on the Prognosis and Treatment of HER2-Positive Early Breast Cancer? *Clin. Cancer Res.*, **26**, 2783–2788.

Brown,R.W. *et al.* (2014) Magnetic Resonance Imaging : Second Edition John Wiley & Sons.

Brunet,J.P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 4164–4169.

Burnside,E.S. *et al.* (2016a) Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage. *Cancer*, **122**, 748–757.

Burnside,E.S. *et al.* (2016b) Using computer-extracted image phenotypes from tumors on breast magnetic resonance imaging to predict breast cancer pathologic stage. *Cancer*, **122**, 748–757.

Burstein,M.D. *et al.* (2015) Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-negative Breast Cancer. *Clin. Cancer Res.*, **21**, 1688.

Cameron,D. *et al.* (2017) 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. *Lancet*, **389**, 1195–1205.

Canadian Cancer Statistics Advisory Committee (2022) Canadian Cancer Statistics 2022.

Carbon,S. *et al.* (2021) The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.

Carey,L.A. *et al.* (2016) Molecular heterogeneity and response to neoadjuvant human epidermal growth factor receptor 2 targeting in CALGB 40601, a randomized phase III trial of paclitaxel plus trastuzumab with or without lapatinib. *J. Clin. Oncol.*, **34**, 542–549.

Carroll,J.D. and Chang,J.J. (1970) Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, **35**, 283–319.

Cecco,L. De (2021) Gene expression profiling of primary HER2-positive breast cancers treated with neoadjuvant trastuzumab.

Chalise,P. *et al.* (2016) InterSIM: Simulation tool for multiple integrative 'omic datasets'. *Comput. Methods Programs Biomed.*, **128**, 69–74.

Chatterjee,R. and Vinson,C. (2012) CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. *Biochim. Biophys. Acta (BBA)-Gene Regul. Mech.*, **1819**, 763–770.

Chen,E.Y. *et al.* (2013) Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**.

Chiu,Y.C. *et al.* (2021) Predicting and characterizing a cancer dependency map of tumors with deep learning. *Sci. Adv.*, **7**.

Çiçek,Ö. *et al.* (2016) 3D U-net: Learning dense volumetric segmentation from sparse annotation. In, *International conference on medical image computing and computer-assisted intervention*., pp. 424–432.

Clark,K. *et al.* (2013) The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *J. Digit. Imaging*, **26**, 1045–1057.

Colaprico,A. *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.*, **44**, e71–e71.

Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Curtis,C. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

Dai,X. *et al.* (2015) Breast cancer intrinsic subtype classification, clinical use and future trends. *Am. J. Cancer Res.*, **5**, 2929–2943.

Dai,X. *et al.* (2016) Cancer hallmarks, biomarkers and breast cancer molecular subtypes. *J. Cancer*, **7**, 1281–1294.

Dean-Colomb,W. and Esteva,F.J. (2008) Her2-positive breast cancer: herceptin and beyond. *Eur. J. Cancer*, **44**, 2806–2812.

Deng,J. *et al.* (2010) ImageNet: A large-scale hierarchical image database. Institute of Electrical and Electronics Engineers (IEEE), pp. 248–255.

Domcke,S. *et al.* (2015) Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature*, **528**, 575.

Duan,R. *et al.* (2021) Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLOS Comput. Biol.*, **17**, e1009224.

Elston,C.W. and Ellis,I.O. (1991) Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, **19**, 403–410.

Fang,J. (2019) Tightly integrated genomic and epigenomic data mining using tensor decomposition. *Bioinformatics*, **35**, 112–118.

Fang,L. and Wang,K. (2018) Identification of Copy Number Variants from SNP Arrays Using PennCNV. In, *Copy Number Variants*. Springer, pp. 1–28.

Fasoulis,R. (2021) Graph representation learning for structural proteomics. *Emerg. Top. Life Sci.*, **5**, 789–802.

Fass,L. (2008) Imaging and cancer: A review. *Mol. Oncol.*, **2**, 115–152.

Feuk,L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85.

Filipits,M. *et al.* (2011) A New Molecular Predictor of Distant Recurrence in ER-Positive, HER2-Negative Breast Cancer Adds Independent Information to Conventional Clinical

Risk Factors. *Clin. Cancer Res.*, **17**, 6012–6020.

Fionda,V. (2018) Networks in biology. *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, **1**–**3**, 915–921.

Friedman,J. *et al.* (2009) glmnet: Lasso and elastic-net regularized generalized linear models. *R Packag. version*, **1**.

Furlong,L.I. (2013) Human diseases through the lens of network biology. *Trends Genet.*, **29**, 150–159.

Gao,J. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**.

Gendoo,D.M.A. *et al.* (2016) Genefu: An R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics*.

Gevaert,Olivier *et al.* (2012) Identifying Prognostic Imaging Biomarkers by Leveraging Public Gene Expression Microarray Data. *Radiology*, **264**, 387–396.

Gevaert,O. *et al.* (2012) Non – Small Cell Lung Cancer : Identifying Prognostic Imaging Biomarkers by Leveraging Public. *Radiology*, **264**, 387–396.

Gianni,L. *et al.* (2012) Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): A randomised multicentre, open-label, phase 2 trial. *Lancet Oncol.*, **13**, 25–32.

Gilmer,J. *et al.* (2017) Neural Message Passing for Quantum Chemistry. In, *34th International Conference on Machine Learning*., pp. 1263–1272.

Gingras,I. *et al.* (2017) HER2-positive breast cancer is lost in translation: Time for patient-centered research. *Nat. Rev. Clin. Oncol.*, **14**, 669–681.

Van Goethem,M. *et al.* (2006) Magnetic resonance imaging in breast cancer. *Eur. J. Surg. Oncol.*, **32**, 901–910.

Goldman,M.J. *et al.* (2020) Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.*, **38**, 675–678.

Gonçalves,E. *et al.* (2017) Widespread Post-transcriptional Attenuation of Genomic Copy-Number Variation in Cancer. *Cell Syst.*, **5**, 386-398.e4.

Goodfellow,I.J. *et al.* (2014) Generative Adversarial Networks. 1–9.

Gordon,Y. *et al.* (2014) Dynamic contrast-enhanced magnetic resonance imaging: fundamentals and application to the evaluation of the peripheral perfusion. *Cardiovasc. Diagn. Ther.*, **4**, 147–14764.

Gori,M. *et al.* (2005) A new model for earning in raph domains. *Proc. Int. Jt. Conf. Neural Networks*, **2**, 729–734.

Van Griethuysen,J.J.M. *et al.* (2017) Computational radiomics system to decode the radiographic phenotype. *Cancer Res.*, **77**, e104–e107.

Grossman,R.L. *et al.* (2016) Toward a shared vision for cancer genomic data. *N. Engl. J. Med.*, **375**, 1109–1112.

Guedj,M. *et al.* (2012) A refined molecular taxonomy of breast cancer. *Oncogene*, **31**, 1196–1206.

Gupta,R. *et al.* (2010) Advances in genome-wide DNA methylation analysis. *Biotechniques*, **49**, iii–xi.

Ha,R., Chang,P., *et al.* (2019) Convolutional Neural Network Using a Breast MRI Tumor Dataset Can Predict Oncotype Dx Recurrence Score. *J. Magn. Reson. Imaging*, **49**, 518–524.

Ha,R., Mutasa,S., *et al.* (2019) Predicting Breast Cancer Molecular Subtype with MRI Dataset Utilizing Convolutional Neural Network Algorithm. *J. Digit. Imaging*, **32**, 276–282.

Haibe-Kains,B. *et al.* (2012) A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes. *JNCI J. Natl. Cancer Inst.*, **104**, 311.

Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: The next generation. *Cell*, **144**, 646–674.

Harbeck,N. *et al.* (2019) Breast cancer.

Harshman,R. a (1970) Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. *UCLA Work. Pap. Phonetics*, **16**, 1–84.

Håstad,J. (1990) Tensor rank is NP-complete. *J. Algorithms*, **11**, 644–654.

Herschkowitz,J.I. *et al.* (2007) Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.*, **8**, 1–17.

Hillar,C.J. and Lim,L.H. (2013) Most tensor problems are NP-Hard. *J. ACM*, **60**, 1–39.

Hitchcock,F.L. (1927) The Expression of a Tensor or a Polyadic as a Sum of Products. *J. Math. Phys.*, **6**, 164–189.

Hu,D.J. *et al.* (2019) High WDR34 mRNA expression as a potential prognostic biomarker in patients with breast cancer as determined by integrated bioinformatics analysis. *Oncol. Lett.*, **18**, 3177–3187.

Hu,Z. *et al.* (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, **7**, 96.

Huang,S. *et al.* (2017) More is better: Recent progress in multi-omics data integration methods. *Front. Genet.*, **8**, 1–12.

Huang,T. *et al.* (2012) Dysfunctions associated with methylation, microRNA expression and

gene expression in lung cancer. *PLoS One*, **7**, e43441.

Huang,T. *et al.* (2015) Novel candidate key drivers in the integrative network of genes, microRNAs, methylations, and copy number variations in squamous cell lung carcinoma. *Biomed Res. Int.*, **2015**, 358125.

Huang,T. and Cai,B.-Q.L. and Y.-D. (2016) The Integrative Network of Gene Expression, MicroRNA, Methylation and Copy Number Variation in Colon and Rectal Cancer. *Curr. Bioinform.*, **11**, 59–65.

Hussain,Z. *et al.* (2017) Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, **2017**, 979–984.

Hwang,K.T. *et al.* (2019) Impact of breast cancer subtypes on prognosis of women with operable invasive breast cancer: A Population-based Study Using SEER Database. *Clin. Cancer Res.*, **25**, 1970–1979.

Jha,K. *et al.* (2022) Prediction of protein–protein interaction using graph neural networks. *Sci. Rep.*, **12**, 1–12.

Jin,C. Bin *et al.* (2019) Deep CT to MR synthesis using paired and unpaired data. *Sensors (Switzerland)*, **19**, 1–19.

Jin,Y.W. and Hu,P. (2020) Tumor-infiltrating cd8 t cells predict clinical breast cancer outcomes in young women. *Cancers (Basel).*, **12**.

Joel S. Parker *et al.* (2009) Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

Johnson,R.H. *et al.* (2015) Gene expression in 'young adult type' breast cancer: A retrospective analysis. *Oncotarget*, **6**, 13688–13702.

Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond.

Nat. Rev. Genet., **13**, 484.

Jönsson,G. *et al.* (2010) Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. *Breast Cancer Res.*, **12**.

Jovanovic,J. *et al.* (2010) The epigenetics of breast cancer. *Mol. Oncol.*, **4**, 242–254.

Jozefowicz,R. *et al.* (2015) An empirical exploration of Recurrent Network architectures. *32nd Int. Conf. Mach. Learn. ICML 2015*, **3**, 2332–2340.

Justin Donaldson,M. (2016) T-Distributed Stochastic Neighbor Embedding for R (t-SNE). 2–5.

Kalli,S. *et al.* (2018) American joint committee on cancer's staging system for breast cancer, eighth edition: What the radiologist needs to know. *Radiographics*, **38**, 1921–1933.

Kaplan,E.L. and Meier,P. (1958) Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.*, **53**, 457–481.

Kassambara,A. *et al.* (2017) survminer: Drawing Survival Curves using'ggplot2'. *R Packag. version 0.3*, **1**.

Khan,S. and Ammad-ud-din,M. (2016) tensorBF: an R package for Bayesian tensor factorization. *bioRxiv*, 097048.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Kiers,H.A.L. (2000) Towards a standardized notation and terminology in multiway analysis. *J. Chemom.*, **14**, 105–122.

Koboldt,D.C. *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Kolda,T.G. and Bader,B.W. (2009) Tensor decompositions and applications. *SIAM Rev.*, **51**,

455–500.

Koren,S. and Bentires-Alj,M. (2015) Breast Tumor Heterogeneity: Source of Fitness, Hurdle for Therapy. *Mol. Cell*, **60**, 537–546.

Kotsiantis,S.B. *et al.* (2006) Machine learning: A review of classification and combining techniques. *Artif. Intell. Rev.*, **26**, 159–190.

Kramer,M. *et al.* (2014) Inferring gene ontologies from pairwise similarity data. *Bioinformatics*, **30**, 34–42.

Krizhevsky,A. *et al.* (2017) ImageNet Classification with Deep Convolutional Neural Networks Alex. *Commun. ACM*, **60**, 84–90.

Kuhn,M. (2008) Building predictive models in R using the caret package. *J. Stat. Softw.*, **28**, 1–26.

Lambina,P. *et al.* (2012) Radiomics: Extracting more information from medical images using advanced feature analysis. **48**, 441–446.

Lawrence,M.S. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.

Lecun,Y. *et al.* (2015) Deep learning. *Nature*, **521**, 436–444.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey JD,T.L. (2022) sva: Surrogate Variable Analysis.

Lehmann,B.D. *et al.* (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, **121**, 2750.

Lehmann,B.D. *et al.* (2016) Refinement of Triple-Negative Breast Cancer Molecular Subtypes:

Implications for Neoadjuvant Chemotherapy Selection. *PLoS One*, **11**.

Li,B. *et al.* (2016) Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol.*, **17**.

Li,W. *et al.* (2015) Identifying human genome-wide CNV, LOH and UPD by targeted sequencing of selected regions. *PLoS One*, **10**, 1–18.

Liu,J. *et al.* (2018) An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*, **173**, 400-416.e11.

Liu,Y. *et al.* (2021) Retraction: Arginine methylation of SHANK2 by PRMT7 promotes human breast cancer metastasis through activating endosomal FAK signalling. *Elife*, **10**, 1–27.

Lloyd,S.P. (1982) Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory*, **28**, 129–137.

Loi,S. *et al.* (2010) PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor – positive breast cancer. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 10208–10213.

Long,Y. *et al.* (2020) Ensembling graph attention networks for human microbe-drug association prediction. *Bioinformatics*, **36**, I779–I786.

Long,Y. *et al.* (2021) Graph contextualized attention network for predicting synthetic lethality in human cancers. *Bioinformatics*, **37**, 2432–2440.

Long,Y. *et al.* (2022) Pre-training graph neural networks for link prediction in biomedical networks. *Bioinformatics*, **38**, 2254–2262.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Lundberg,S.M. and Lee,S.I. (2017) A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, **2017-Decem**, 4766–4775.

MacKinnon,D.P. *et al.* (2007) Mediation Analysis. *Annu. Rev. Psychol.*, **58**, 593–614.

Mahmud,S.M.H. *et al.* (2021) PreDTIs: Prediction of drug-Target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques. *Brief. Bioinform.*, **22**, 1–20.

Malhotra,G.K. *et al.* (2010) Histological, molecular and functional subtypes of breast cancers. *Cancer Biol. Ther.*, **10**, 955.

Martelotto,L.G. *et al.* (2014) Breast cancer intra-tumor heterogeneity. *Breast Cancer Res.*, **16**, 1–11.

Martens,J.H.A. *et al.* (2010) PML-RARα/RXR Alters the Epigenetic Landscape in Acute Promyelocytic Leukemia. *Cancer Cell*, **17**, 173–185.

Martínez,M.T. *et al.* (2019) Breast Cancer in Very Young Patients in a Spanish Cohort: Age as an Independent Bad Prognostic Indicator. *Breast Cancer Basic Clin. Res.*, **13**.

Mermel,C.H. *et al.* (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.*, **12**.

Mikołajczyk,A. and Grochowski,M. (2018) Data augmentation for improving deep learning in image classification problem. *2018 Int. Interdiscip. PhD Work. IIPhDW 2018*, 117–122.

Mnih,V. *et al.* (2015) Human-level control through deep reinforcement learning. *Nature*, **518**, 529–533.

Mo,Q. *et al.* (2018) A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, **19**, 71–86.

MÖcks,J. (1988) Topographic components model for event-related potentials and some biophysical considerations. *IEEE Trans. Biomed. Eng.*, **35**, 482–484.

Monti,S. *et al.* (2003) Consensus Clustering A Resampling-Based Method for Class Discovery

and Vi - Monti - Mach Learn. *Mach. Learn.*, **52**, 91–118.

Moor,J. (2006) The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. *AI Mag.*, **27**, 87.

Neagu,A. *et al.* (2021) Proteomics and its applications in breast cancer. *Am. J. Cancer Res.*, **11**, 4006–4049.

Nguyen,Q.-H. *et al.* (2020) Multi-Omics Analysis Detects Novel Prognostic Subgroups of Breast Cancer . *Front. Genet.* , **11**, 1265.

Nie,D. *et al.* (2019) Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages. *Sci. Rep.*, **9**, 1–14.

Ogata,H. *et al.* (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.

Paik,S. *et al.* (2004) A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *N. Engl. J. Med.*, **351**, 2817–2826.

Paleri,V. *et al.* (2010) TNM classification of malignant tumours 7th edition: what's new for head and neck? *Clin. Otolaryngol.*, **35**, 270–272.

Parker,J.S. *et al.* (2009) Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

Pasculli,B. *et al.* (2018) Epigenetics of breast cancer: biology and clinical implication in the era of precision medicine. In, *Seminars in cancer biology*. Elsevier, pp. 22–35.

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747.

Petrucelli,N. *et al.* (2022) BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer. *GeneReviews®*.

Polyak,K. (2011) Heterogeneity in breast cancer. *J. Clin. Invest.*, **121**, 3786–3788.

Prat,A. *et al.* (2010) Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.*, **12**.

Prat,A. and Perou,C.M. (2011) Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.*, **5**, 5–23.

Qian,Y. *et al.* (2021) Prognostic cancer gene expression signatures: Current status and challenges. *Cells*, **10**, 1–17.

Qin,Y. *et al.* (2021) A multi-scale map of cell structure fusing protein images and interactions. *Nature*, **600**, 536–542.

Rao,V.S. *et al.* (2014) Protein-Protein Interaction Detection : Methods and Analysis. *Int. J. Proteomics*, **2014**.

Rappoport,N. *et al.* (2019) NEMO: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, **35**, 3348–3356.

Reich,M. *et al.* (2006) GenePattern 2.0. *Nat. Genet.*, **38**, 500.

Ritchie,M.D. *et al.* (2015) Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.*, **16**, 85–97.

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

Rives,A. *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **118**.

Robertson,K.D. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol. 2010 113*, **11**, 1–9.

Rousseeuw,P.J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster

analysis. *J. Comput. Appl. Math.*, **20**, 53–65.

Runge,V.M. (2015) Contrast-enhanced clinical magnetic resonance imaging University Press of Kentucky.

Russnes,H.G. *et al.* (2017) Breast Cancer Molecular Stratification: From Intrinsic Subtypes to Integrative Clusters. *Am. J. Pathol.*, **187**, 2152–2162.

Rutman,A.M. and Kuo,M.D. (2009) Radiogenomics: Creating a link between molecular diagnostics and diagnostic imaging. *Eur. J. Radiol.*, **70**, 232–241.

Rye,I.H. *et al.* (2018) Intratumor heterogeneity defines treatment-resistant HER2+ breast tumors. *Mol. Oncol.*, **12**, 1838–1855.

Shaikh,F. *et al.* (2018) Translational Radiomics: Defining the Strategy Pipeline and Considerations for Application—Part 1: From Methodology to Clinical Implementation. *J. Am. Coll. Radiol.*, **15**, 538–542.

Shen,W. *et al.* (2019) Genome-wide copy number variation detection using NGS: data analysis and interpretation. In, *Tumor Profiling*. Springer, pp. 113–124.

Shlien,A. and Malkin,D. (2009) Copy number variations and cancer. *Genome Med.*, **1**, 1–9.

Shrikumar,A. *et al.* (2016) Not Just a Black Box: Interpretable Deep Learning by Propagating Activation Differences. *arXiv*, **1**, 0–5.

Simonyan,K. *et al.* (2014) Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd Int. Conf. Learn. Represent. ICLR 2014 - Work. Track Proc.*, 1–8.

Smith,J.C. and Sheltzer,J.M. (2022) Genome-wide identification and analysis of prognostic features in human cancers. *Cell Rep.*, **38**, 110569.

Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nat.*

Rev. Genet., **14**, 204.

Song,M. *et al.* (2020) Targeting ubiquitin protein ligase E3 component N-recognin 5 in cancer cells induces a CD8+ T cell mediated immune response. *Oncoimmunology*, **9**.

Sørlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.*, **98**, 10869–10874.

Sørlie,T. *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.*, **100**, 8418–8423.

Sotiriou,C. *et al.* (2003) Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 10393–10398.

Sotiriou,C. *et al.* (2006) Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.*, **98**, 262–272.

Srivastava,N. and Salakhutdinov,R. (2014) Multimodal learning with Deep Boltzmann Machines. *J. Mach. Learn. Res.*, **15**, 2949–2980.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.

Tang,J. *et al.* (2009) Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. *IEEE Trans. Inf. Technol. Biomed.*, **13**, 236–251.

Tang,Y. *et al.* (2018) Bayesian tensor factorization for multi-way analysis of multi-dimensional EEG. *Neurocomputing*, **318**, 162–174.

The Gene Ontology Consortium *et al.* (2004) Gene Ontology: tool for the unification of biology. *Hydrol. und Wasserbewirtschaftung*, **48**, 2–11.

Therneau,T.M. and Grambsch,P.M. (2000) Modeling Survival Data: Extending the Cox Model

Springer New York, New York, NY.

Thingholm,L.B. *et al.* (2016) Strategies for integrated analysis of genetic, epigenetic, and gene expression variation in cancer: Addressing the challenges. *Front. Genet.*, **7**, 1–13.

Tingley,D. *et al.* (2014) Mediation: R package for causal mediation analysis.

Tomczak,K. *et al.* (2015) The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkol.*, **1A**, A68–A77.

Tsherniak,A. *et al.* (2017) Defining a Cancer Dependency Map. *Cell*, **170**, 564–576.

Tulder,G. Van and Bruijne,M. De (2018) Learning Cross-Modality Representations from Multi-Modal Images. 1–11.

Turashvili,G. and Brogi,E. (2017) Tumor heterogeneity in breast cancer. *Front. Med.*, **4**.

Visvikis,D. *et al.* (2019) Artificial intelligence, machine (deep) learning and radio(geno)mics: definitions and nuclear medicine imaging applications. *Eur. J. Nucl. Med. Mol. Imaging*, **46**, 2630–2637.

Vogelstein,B. *et al.* (2013) Cancer genome landscapes. *Science (80-.).*, **339**, 1546–1558.

Vukotić,V. *et al.* (2016) Multimodal and crossmodal representation learning from textual and visual features with bidirectional deep neural networks for video hyperlinking. *Iv L-MM 2016 - Proc. 2016 ACM Work. Vis. Lang. Integr. Meets Multimed. Fusion, co-located with ACM Multimed. 2016*, 37–44.

Wang,B. *et al.* (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

Wang,H. *et al.* (2016) DCAF4L2 promotes colorectal cancer invasion and metastasis via mediating degradation of NFκb negative regulator PPM1B. *Am. J. Transl. Res.*, **8**, 405.

Wang,Z. *et al.* (2018) A framework for analyzing DNA methylation data from Illumina Infinium

HumanMethylation450 BeadChip. *BMC Bioinformatics*, **19**, 115.

Weigel,M.T. and Dowsett,M. (2010) Current and emerging biomarkers in breast cancer: Prognosis and prediction. *Endocr. Relat. Cancer*, **17**, 245–262.

Wickham,H. (2016) ggplot2: Elegant Graphics for Data Analysis Springer-Verlag New York, New York, New York, USA.

Wilkerson,M.D. and Hayes,D.N. (2010) ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics*, **26**, 1572–1573.

Wolff,A.C. *et al.* (2018) Human epidermal growth factor receptor 2 testing in breast cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Arch. Pathol. Lab. Med.*, **142**, 1364–1382.

Wu,D. *et al.* (2015) Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics 2015 161*, **16**, 1–10.

Wu,L. *et al.* (2018) A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat. Genet.*

Xiang,G. *et al.* (2022) UBR5 targets tumor suppressor CDC73 proteolytically to promote aggressive breast cancer. *Cell Death Dis. 2022 135*, **13**, 1–14.

Xing,Y. *et al.* (2022) Learning Hierarchical Graph Neural Networks for Image Clustering. 3447–3457.

Xiong,L. *et al.* (2010) Temporal collaborative filtering with Bayesian probabilistic tensor factorization. *Proc. 10th SIAM Int. Conf. Data Mining, SDM 2010*, 211–222.

Xu,T. *et al.* (2017) CancerSubtypes: An R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics*, **33**, 3131–3133.

Yan,L. *et al.* (2020) An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.*, **2**, 283–288.

Yang,F. *et al.* (2020) Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. *BMC Bioinformatics*, **21**, 1–16.

Yang,L. *et al.* (2020) Learning to Cluster Faces via Confidence and Connectivity Estimation. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 13366–13375.

Yang,Y. *et al.* (2014) Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.*, **5**, 1–9.

Yoon,H.-J. *et al.* (2018) Deep radiogenomics for predicting clinical phenotypes in invasive breast cancer. In, Krupinski,E.A. (ed), *14th International Workshop on Breast Imaging (IWBI 2018)*. SPIE, p. 75.

Zagami,P. and Carey,L.A. (2022) Triple negative breast cancer: Pitfalls and progress. *npj Breast Cancer*, **8**.

Zhang,A. (2009) Protein interaction networks: computational analysis Cambridge University Press.

Zhang,J. *et al.* (2019) Hierarchical Convolutional Neural Networks for Segmentation of Breast Tumors in MRI With Application to Radiogenomics. *IEEE Trans. Med. Imaging*, **38**, 435–447.

Zhang,Z. *et al.* (2018) Translating and Segmenting Multimodal Medical Volumes with Cycle- and Shape-Consistency Generative Adversarial Network. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 9242–9251.

Zhao,Q. *et al.* (2015) Bayesian CP factorization of incomplete tensors with automatic rank determination. *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**, 1751–1763.

Zheng,F. *et al.* (2021) Interpretation of cancer mutations using a multiscale map of protein systems. *Science (80-.).*, **374**.

Zhu,Z. *et al.* (2017) Deep learning for identifying radiogenomic associations in breast cancer. *arXiv*.

Appendices

Appendix A: The gene differential expression results for the 6 subtypes and their overlaps.

A1. C1 vs Normal

	logFC	AveExpr	t	P.Value	adj.P.Val	B
CDCA8	0.60697882	6.37108046	35.6188351	2.04E-172	4.13E-168	383.669651
NUF2	0.85974301	6.26557126	34.9097954	6.28E-168	6.35E-164	373.367215
KIF4A	0.84648467	6.35962962	34.3164446	3.67E-164	2.47E-160	364.718545
PLK1	0.71476188	6.39482509	34.2079478	1.80E-163	9.09E-160	363.134588
NEK2	0.88762653	6.34404171	34.0853055	1.08E-162	4.38E-159	361.343224
FOXM1	0.68053858	6.49206266	33.9342968	9.91E-162	3.34E-158	359.136248
TPX2	0.60312014	6.56774487	33.9166431	1.28E-161	3.71E-158	358.878151
UBE2C	0.84537734	6.39652708	33.8470165	3.56E-161	9.01E-158	357.860027
HJURP	0.79153782	6.2629067	33.8151153	5.69E-161	1.28E-157	357.393451
CDC20	0.74853714	6.42480429	33.6406326	7.36E-160	1.49E-156	354.840469
IQGAP3	0.77177979	6.42188467	33.5288764	3.80E-159	6.98E-156	353.204365
CEP55	0.7153502	6.34298029	33.2713468	1.67E-157	2.61E-154	349.431508
KIF2C	0.73162428	6.37430417	33.2711112	1.68E-157	2.61E-154	349.428054
NDC80	0.7668924	6.26641156	33.175746	6.81E-157	9.84E-154	348.03003
CDCA3	0.69301871	6.24280639	32.940094	2.18E-155	2.94E-152	344.573436
EZH2	0.48246034	6.41056425	32.6349192	1.95E-153	2.46E-150	340.093021
MELK	0.84754567	6.28831274	32.5026081	1.37E-152	1.63E-149	338.149157
UBE2T	0.6070985	6.36588439	32.391289	7.06E-152	7.94E-149	336.513102
BUB1	0.72636475	6.36316398	32.3264109	1.84E-151	1.96E-148	335.559344
KIFC1	0.62720915	6.4348331	32.0412786	1.23E-149	1.25E-146	331.365642
KIF18B	0.90054458	6.17840068	32.008235	2.01E-149	1.93E-146	330.879433
CDCA5	0.58683962	6.4036834	31.9868525	2.75E-149	2.53E-146	330.564786
PKMYT1	0.93501727	6.29174483	31.7246674	1.32E-147	1.16E-144	326.705329
AURKB	0.88878417	6.18397469	31.6155019	6.61E-147	5.57E-144	325.097673
AURKA	0.55409741	6.38487415	31.5645185	1.40E-146	1.14E-143	324.346717
PTTG1	0.63653735	6.358503	31.5406506	2.00E-146	1.55E-143	323.995129
CCNB2	0.68513742	6.35844758	31.3876426	1.92E-145	1.43E-142	321.740812
TROAP	0.84597022	6.2219041	31.2346294	1.84E-144	1.33E-141	319.485733
RAD54L	0.77217395	6.15592248	30.9501108	1.23E-142	8.60E-140	315.290899
GTSE1	0.69015905	6.22876743	30.887742	3.10E-142	2.09E-139	314.371096

CENPA	0.83825506	6.09325606	30.8497149	5.44E-142	3.55E-139	313.810236
WDR62	0.68213595	6.2419295	30.8262761	7.70E-142	4.82E-139	313.464524
CKS2	0.46460694	6.47078181	30.8249089	7.86E-142	4.82E-139	313.444359
FAM54A	0.75119632	5.97500164	30.8075357	1.02E-141	6.04E-139	313.188103
SKA1	0.77802019	6.12435005	30.7933179	1.25E-141	7.24E-139	312.978385
CDK1	0.53270759	6.49757011	30.6520469	1.01E-140	5.69E-138	310.894363
SPC25	0.72789285	6.1274963	30.5722042	3.30E-140	1.80E-137	309.716362
SKA3	0.70872698	6.18963205	30.4523629	1.94E-139	1.04E-136	307.948013
ANLN	0.6470611	6.46011454	30.3212996	1.35E-138	7.01E-136	306.01382
EXO1	0.82164296	6.21994198	30.2627908	3.22E-138	1.63E-135	305.150289
FAM64A	0.85460121	6.07720101	30.2375282	4.67E-138	2.31E-135	304.777423
PDE2A	-0.5883774	6.27116985	-30.145645	1.82E-137	8.77E-135	303.421213
KIF20A	0.74483558	6.38061838	30.061297	6.34E-137	2.98E-134	302.176129
BIRC5	0.84936607	6.30609349	30.055385	6.92E-137	3.18E-134	302.088858
DLGAP5	0.80208881	6.27899041	29.8379962	1.73E-135	7.77E-133	298.879647
SDPR	-0.8214422	6.26241248	-29.725608	9.13E-135	3.93E-132	297.220394
FAM83D	0.72013222	6.29439873	29.7255309	9.14E-135	3.93E-132	297.219255
CDC25C	0.9220935	6.03053886	29.6856871	1.65E-134	6.95E-132	296.631006
CDCA2	0.75088362	6.15123726	29.6758393	1.91E-134	7.87E-132	296.485612
NCAPG	0.59958129	6.36860084	29.6086249	5.16E-134	2.09E-131	295.493251
HDGF	0.20302698	6.90256714	29.5999583	5.87E-134	2.33E-131	295.365296
KIF23	0.5289883	6.41583227	29.4613458	4.57E-133	1.78E-130	293.318786
ASF1B	0.53488768	6.4165402	29.4087732	9.94E-133	3.79E-130	292.54259
CENPF	0.51382499	6.63953848	29.3368457	2.88E-132	1.08E-129	291.480642
CCNB1	0.40150026	6.54486023	29.265988	8.23E-132	3.00E-129	290.434501
UHRF1	0.70130081	6.33767246	29.265347	8.31E-132	3.00E-129	290.425037
TTK	0.69373467	6.26441373	29.1573635	4.11E-131	1.46E-128	288.830814
CCNA2	0.55802991	6.38551035	29.1194028	7.21E-131	2.51E-128	288.270392
TRIP13	0.51129896	6.35813633	28.9885933	5.00E-130	1.71E-127	286.339306
NCAPH	0.60063888	6.29985916	28.9182866	1.42E-129	4.77E-127	285.301458
CDKN3	0.63154974	6.18240124	28.8541361	3.66E-129	1.21E-126	284.354529
HMMR	0.68376584	6.27443208	28.8335094	4.96E-129	1.62E-126	284.050066
TUBA1C	0.25388621	6.79109944	28.7283898	2.35E-128	7.55E-126	282.49852
PBK	0.7777686	6.25257777	28.684216	4.52E-128	1.43E-125	281.846566
RAD51	0.56539169	6.20782871	28.6212718	1.15E-127	3.57E-125	280.917633
KIF14	0.85258669	6.17148577	28.5579103	2.93E-127	8.98E-125	279.982602
NUSAP1	0.52727534	6.49254783	28.4812951	9.11E-127	2.75E-124	278.852078
ECE2	0.43596429	6.2976477	28.4745202	1.01E-126	2.99E-124	278.752112
DEPDC1	0.72611506	6.23369815	28.4647291	1.16E-126	3.41E-124	278.607645
CDC45	0.84595027	6.15982582	28.4510564	1.42E-126	4.12E-124	278.405908
ASPM	0.74271399	6.38765214	28.4157458	2.40E-126	6.84E-124	277.884924

GINS1	0.46133706	6.35418956	28.3953454	3.25E-126	9.12E-124	277.583941
TK1	0.49518311	6.54806901	28.3554533	5.86E-126	1.62E-123	276.995405
OIP5	0.67005317	6.02364443	28.3052	1.23E-125	3.37E-123	276.254056
BUB1B	0.63436256	6.3426436	28.0957372	2.73E-124	7.36E-122	273.164624
RAD51AP1	0.50354118	6.26408958	27.9534605	2.24E-123	5.95E-121	271.066772
MYBL2	0.7445989	6.4498735	27.9279792	3.26E-123	8.56E-121	270.691112
RECQL4	0.56050218	6.38602512	27.9121819	4.12E-123	1.07E-120	270.458228
UBE2S	0.47778078	6.39618186	27.8696062	7.72E-123	1.98E-120	269.830615
TMEM206	0.29100989	6.35227982	27.8654015	8.22E-123	2.08E-120	269.768636
TACC3	0.40848522	6.52529071	27.8614255	8.72E-123	2.18E-120	269.710029
NEIL3	0.93951822	5.88302004	27.8436761	1.13E-122	2.79E-120	269.448404
CTPS	0.29588473	6.51811799	27.7660312	3.57E-122	8.70E-120	268.304037
TNS1	-0.2972854	6.78984699	-27.763645	3.70E-122	8.90E-120	268.268878
FAM72B	0.62229027	6.12246268	27.684681	1.19E-121	2.83E-119	267.105267
LDB2	-0.390196	6.37885702	-27.560132	7.48E-121	1.76E-118	265.270361
KIAA0101	0.64342023	6.31068887	27.5040304	1.71E-120	3.98E-118	264.444031
ORC6L	0.63655984	6.13354798	27.4969953	1.90E-120	4.37E-118	264.340419
RRM2	0.60843669	6.50072185	27.4826619	2.35E-120	5.33E-118	264.129324
PRC1	0.42202612	6.5552189	27.4361731	4.66E-120	1.05E-117	263.44471
ZWINT	0.40892776	6.49771018	27.4052115	7.37E-120	1.64E-117	262.988803
CPA1	-1.1468728	4.93190324	-27.342222	1.87E-119	4.10E-117	262.0614
MKI67	0.53023198	6.61606625	27.2999177	3.49E-119	7.58E-117	261.438649
TNXB	-0.6587976	6.47346201	-27.200096	1.52E-118	3.27E-116	259.969486
CKAP2L	0.74594684	6.15039967	27.184446	1.92E-118	4.08E-116	259.739188
MMP11	0.63260559	6.7487667	27.1698495	2.38E-118	5.01E-116	259.524405
LMNB1	0.39179125	6.57488951	27.1404543	3.67E-118	7.64E-116	259.091891
SHCBP1	0.56979166	6.24370337	27.1220971	4.81E-118	9.92E-116	258.821808
C16orf59	0.62389644	6.14327509	27.0903783	7.67E-118	1.57E-115	258.355177
DEPDC1B	0.73909022	6.07918754	27.0819006	8.70E-118	1.76E-115	258.230465
KIF11	0.43801771	6.4902621	27.0521545	1.35E-117	2.70E-115	257.792908
TENC1	-0.2709077	6.67923919	-27.014064	2.37E-117	4.69E-115	257.232669
ATP1A2	-1.3667498	5.88595134	-26.911207	1.08E-116	2.12E-114	255.720185
CCNE1	0.67661631	6.07665305	26.8974197	1.32E-116	2.57E-114	255.517491
PSRC1	0.43623474	6.2709282	26.8389211	3.13E-116	6.03E-114	254.657564
TOP2A	0.55659416	6.66122445	26.745671	1.24E-115	2.36E-113	253.287161
STMN1	0.30008442	6.7673766	26.6981032	2.49E-115	4.71E-113	252.588282
RCC2	0.17682445	6.8235672	26.6811757	3.20E-115	5.99E-113	252.33961
DONSON	0.30549938	6.36702971	26.6782923	3.34E-115	6.19E-113	252.297253
E2F1	0.51036415	6.36777872	26.6646519	4.08E-115	7.50E-113	252.096881
POC1A	0.38619205	6.29818866	26.6279039	7.01E-115	1.28E-112	251.557125
MCM10	0.72798424	6.20364161	26.6265848	7.14E-115	1.29E-112	251.537752

C6orf129	0.36055007	6.37154955	26.5580988	1.96E-114	3.50E-112	250.532033
MCM2	0.32692152	6.60137228	26.5543871	2.07E-114	3.67E-112	250.477535
SGOL1	0.83220629	5.986741	26.4929768	5.11E-114	8.98E-112	249.575968
TYMS	0.42456125	6.47602767	26.4814469	6.05E-114	1.06E-111	249.406723
RNASEH2A	0.30848706	6.49430864	26.4478955	9.92E-114	1.71E-111	248.914272
CENPL	0.30993058	6.30507886	26.4433194	1.06E-113	1.82E-111	248.847112
SPAG5	0.4520607	6.47385191	26.4055886	1.85E-113	3.14E-111	248.293408
HMGA1	0.3118396	6.6816958	26.4035634	1.90E-113	3.21E-111	248.26369
C1orf135	0.55939432	6.0337158	26.3611652	3.55E-113	5.94E-111	247.641603
RAPGEF3	-0.4559099	6.3909226	-26.325326	6.02E-113	9.98E-111	247.115848
PYCR1	0.35636299	6.69412871	26.3103838	7.50E-113	1.23E-110	246.896663
SEMA3G	-0.5051074	6.46067543	-26.302276	8.45E-113	1.38E-110	246.777738
FZD4	-0.3619879	6.62180745	-26.27885	1.19E-112	1.93E-110	246.434158
HSD17B13	-1.1817678	5.27888574	-26.211129	3.23E-112	5.18E-110	245.441109
KIF15	0.63628929	6.22176186	26.1995742	3.82E-112	6.09E-110	245.271708
AOC3	-0.4943992	6.556616	-26.119101	1.25E-111	1.97E-109	244.092133
MMRN2	-0.3165011	6.55425596	-26.100328	1.64E-111	2.58E-109	243.81702
MAD2L1	0.40755648	6.39085974	26.0726632	2.47E-111	3.84E-109	243.411638
STIL	0.44055476	6.34512442	26.0331316	4.41E-111	6.81E-109	242.83247
GPRIN1	0.52667621	6.22642399	26.0122299	6.00E-111	9.19E-109	242.526286
FANCA	0.52274704	6.27992197	26.0103397	6.17E-111	9.38E-109	242.4986
EBF1	-0.4816471	6.32928608	-25.990705	8.23E-111	1.24E-108	242.211013
ACSM5	-0.9382351	5.78556586	-25.988921	8.45E-111	1.27E-108	242.184879
CENPM	0.67100355	6.16264049	25.9432707	1.65E-110	2.46E-108	241.516347
HN1	0.32251207	6.67601384	25.8279662	8.97E-110	1.32E-107	239.828431
GPR172A	0.29477216	6.63841502	25.8033788	1.29E-109	1.89E-107	239.468629
VPS72	0.20160997	6.62099057	25.7428847	3.13E-109	4.55E-107	238.583581
RFC4	0.27843706	6.437178	25.7369093	3.41E-109	4.93E-107	238.496174
LOC572558	-0.9937113	4.8688959	-25.718994	4.44E-109	6.37E-107	238.234128
NOSTRIN	-0.5645576	6.4029238	-25.716599	4.60E-109	6.55E-107	238.199105
ORC1L	0.6416463	6.10607762	25.6424957	1.36E-108	1.93E-106	237.115481
KPNA2	0.30879349	6.7287381	25.6309811	1.61E-108	2.27E-106	236.947141
ERCC6L	0.59097108	6.13705386	25.624699	1.77E-108	2.47E-106	236.855304
CNIH2	0.82005234	5.94439043	25.6104827	2.18E-108	3.02E-106	236.647488
CDT1	0.60849105	6.29595265	25.60864	2.24E-108	3.08E-106	236.620552
DDX39	0.26964866	6.59999891	25.6009426	2.51E-108	3.42E-106	236.508038
GPAM	-0.4230588	6.5186204	-25.591026	2.90E-108	3.93E-106	236.363087
ESPL1	0.60112764	6.31110809	25.5764773	3.59E-108	4.84E-106	236.150457
E2F2	0.59752947	6.19979158	25.5663266	4.16E-108	5.58E-106	236.002111
C16orf75	0.47344361	6.32883909	25.5478235	5.46E-108	7.26E-106	235.731721
PEAR1	-0.4511954	6.24711804	-25.544824	5.70E-108	7.54E-106	235.687898

LMNB2	0.26581233	6.63829791	25.488974	1.29E-107	1.70E-105	234.871925
MYOC	-1.4374562	5.09969529	-25.471117	1.68E-107	2.19E-105	234.61109
THSD1	-0.3566395	6.20983908	-25.457806	2.04E-107	2.65E-105	234.416676
SPARCL1	-0.2949675	6.85298812	-25.451535	2.24E-107	2.88E-105	234.32509
BYSL	0.24229617	6.48816397	25.4413158	2.60E-107	3.33E-105	234.175848
CA4	-1.6394567	5.19276764	-25.375984	6.76E-107	8.60E-105	233.221944
LMOD1	-0.4644862	6.4816809	-25.330626	1.31E-106	1.66E-104	232.559904
CCT3	0.18071817	6.90816075	25.329741	1.33E-106	1.67E-104	232.546984
LRRC70	-0.5071504	5.9750323	-25.298409	2.10E-106	2.63E-104	232.08977
DTL	0.52667673	6.41790358	25.2665558	3.35E-106	4.16E-104	231.625032
CXorf36	-0.3261839	6.46536425	-25.240404	4.92E-106	6.06E-104	231.243548
ILF2	0.17623605	6.83967444	25.2308302	5.66E-106	6.93E-104	231.103909
CD34	-0.2728433	6.65777938	-25.216625	6.96E-106	8.48E-104	230.896724
BTNL9	-0.7931567	6.17418505	-25.178131	1.22E-105	1.48E-103	230.335409
ADRB2	-0.6961599	6.00473997	-25.110584	3.28E-105	3.95E-103	229.350743
LHFP	-0.3210609	6.54996981	-25.107822	3.42E-105	4.09E-103	229.310489
FLAD1	0.21971698	6.62432136	25.0243112	1.16E-104	1.38E-102	228.09373
MTHFD2	0.23790772	6.67896551	25.0217082	1.20E-104	1.42E-102	228.055816
NFKBIL2	0.44087917	6.37964704	25.0168834	1.29E-104	1.52E-102	227.985539
ZNF695	0.88068756	5.85716221	25.0036863	1.56E-104	1.83E-102	227.793327
SPC24	0.97872536	5.6453362	24.9348144	4.28E-104	4.97E-102	226.790493
PPP1R12B	-0.2524633	6.61356813	-24.900979	7.01E-104	8.10E-102	226.297996
NKAPL	-0.7017022	5.48110736	-24.873536	1.05E-103	1.20E-101	225.898615
CAV1	-0.3500515	6.691822	-24.750171	6.32E-103	7.22E-101	224.104228
H2AFZ	0.21180123	6.78826193	24.7431719	7.00E-103	7.95E-101	224.002477

A2. C2 vs Normal

	logFC	AveExpr	t	P.Value	adj.P.Val	B
MMP11	0.6733733	6.7487667	31.2253932	2.11E-144	4.26E-140	319.130542
COL10A1	0.99582633	6.51546296	29.2150533	1.75E-131	1.77E-127	289.525858
NEK2	0.67743641	6.34404171	28.0868651	3.11E-124	2.10E-120	272.910342
KIF4A	0.63546893	6.35962962	27.8147605	1.74E-122	8.78E-119	268.906577
PKMYT1	0.72943729	6.29174483	26.7216471	1.76E-115	7.12E-112	252.84918
UBE2C	0.59459978	6.39652708	25.7035051	5.57E-109	1.62E-105	237.95003
NUF2	0.58628733	6.26557126	25.7031869	5.60E-109	1.62E-105	237.945385
IQGAP3	0.54061525	6.42188467	25.3578077	8.82E-107	2.23E-103	232.908133
HJURP	0.54836709	6.2629067	25.2934689	2.26E-106	5.08E-103	231.97087

A3. C3 vs Normal

	logFC	AveExpr	t	P.Value	adj.P.Val	B
MMP11	0.77261037	6.7487667	33.2695166	1.72E-157	3.47E-153	349.318131
NEK2	0.81821548	6.34404171	31.5018717	3.54E-146	3.58E-142	323.358993
KIF4A	0.77389688	6.35962962	31.4556046	7.02E-146	4.73E-142	322.677897
COL10A1	1.11791051	6.51546296	30.4553824	1.86E-139	9.40E-136	307.939993
PKMYT1	0.88261617	6.29174483	30.024866	1.09E-136	4.40E-133	301.590701
UBE2C	0.73712014	6.39652708	29.5896495	6.83E-134	2.30E-130	295.170371
UBE2T	0.54600898	6.36588439	29.2079188	1.94E-131	5.62E-128	289.538714
NUF2	0.71638619	6.26557126	29.1647065	3.69E-131	9.32E-128	288.901229
HJURP	0.67985363	6.2629067	29.1196648	7.18E-131	1.61E-127	288.236766
TPX2	0.51127654	6.56774487	28.8268143	5.48E-129	1.11E-125	283.916966
CDKN3	0.62539442	6.18240124	28.6474701	7.79E-128	1.43E-124	281.271938
AURKA	0.50020715	6.38487415	28.568977	2.49E-127	4.19E-124	280.114439
KIAA0101	0.66541094	6.31068887	28.5182791	5.27E-127	8.20E-124	279.366872
PLK1	0.59417033	6.39482509	28.5107309	5.89E-127	8.51E-124	279.255574
CDK1	0.49385455	6.49757011	28.4905878	7.94E-127	1.07E-123	278.958569
CEP55	0.60912348	6.34298029	28.4046061	2.83E-126	3.58E-123	277.690873
MELK	0.72961538	6.28831274	28.0530984	5.13E-124	6.10E-121	272.509929
UHRF1	0.66819569	6.33767246	27.9566261	2.13E-123	2.40E-120	271.088533
IQGAP3	0.64030534	6.42188467	27.8897405	5.74E-123	6.11E-120	270.103209
ASF1B	0.504523	6.4165402	27.8116681	1.82E-122	1.84E-119	268.953255
SPC25	0.65758991	6.1274963	27.6914818	1.07E-121	1.03E-118	267.183363
RRM2	0.61083023	6.50072185	27.6627707	1.64E-121	1.51E-118	266.760626
TROAP	0.74434932	6.2219041	27.5543285	8.15E-121	7.16E-118	265.1642
ZWINT	0.40994012	6.49771018	27.544744	9.38E-121	7.91E-118	265.023122
TK1	0.47890787	6.54806901	27.4950488	1.95E-120	1.58E-117	264.291695
NUSAP1	0.50627032	6.49254783	27.4180455	6.09E-120	4.74E-117	263.158524
CPA1	-1.1457498	4.93190324	-27.386726	9.68E-120	7.16E-117	262.697692
CDCA8	0.46545379	6.37108046	27.3851105	9.91E-120	7.16E-117	262.673925
BUB1	0.61322566	6.36316398	27.3624385	1.39E-119	9.66E-117	262.340359
CCNB2	0.59559128	6.35844758	27.3565354	1.51E-119	9.93E-117	262.253513
SKA3	0.63500868	6.18963205	27.3560537	1.52E-119	9.93E-117	262.246426
HMMR	0.64673913	6.27443208	27.3433046	1.84E-119	1.16E-116	262.058865
HSD17B6	0.65249401	6.06502729	27.2514245	7.13E-119	4.37E-116	260.707351

CDC25C	0.84226493	6.03053886	27.1864571	1.86E-118	1.11E-115	259.751924
KIF2C	0.59580971	6.37430417	27.1655516	2.53E-118	1.46E-115	259.444523
ANLN	0.57520913	6.46011454	27.0246478	2.02E-117	1.14E-114	257.373137
FOXM1	0.53887252	6.49206266	26.9403908	7.01E-117	3.83E-114	256.134944
DLGAP5	0.72000615	6.27899041	26.8543817	2.49E-116	1.33E-113	254.871364
PYCR1	0.36259197	6.69412871	26.8401244	3.07E-116	1.59E-113	254.661944
NDC80	0.61799806	6.26641156	26.8043407	5.21E-116	2.63E-113	254.136373
CDCA5	0.49025716	6.4036834	26.792163	6.23E-116	3.07E-113	253.95753
EXO1	0.72286492	6.21994198	26.6940669	2.64E-115	1.27E-112	252.51716
CCNB1	0.36505932	6.54486023	26.679185	3.29E-115	1.55E-112	252.298691
FAM54A	0.64605696	5.97500164	26.5647707	1.77E-114	8.16E-112	250.61948
KIF20A	0.65562543	6.38061838	26.5298531	2.97E-114	1.33E-111	250.107157
PTTG1	0.5310708	6.358503	26.3834103	2.56E-113	1.13E-110	247.959282
CKS2	0.39587698	6.47078181	26.3334709	5.34E-113	2.30E-110	247.227121
CDC20	0.5831161	6.42480429	26.2746892	1.27E-112	5.34E-110	246.365524
OIP5	0.61969332	6.02364443	26.2461453	1.93E-112	7.96E-110	245.947219
NCAPG	0.52879753	6.36860084	26.1813079	5.00E-112	2.02E-109	244.997237
EZH2	0.38596824	6.41056425	26.1760577	5.40E-112	2.14E-109	244.920325
GTSE1	0.58177034	6.22876743	26.104796	1.54E-111	5.99E-109	243.876563
WDR62	0.57590869	6.2419295	26.093691	1.81E-111	6.92E-109	243.713941
NEIL3	0.87778496	5.88302004	26.0820218	2.15E-111	8.06E-109	243.543065
LOC572558	-1.0022923	4.8688959	-26.008773	6.31E-111	2.30E-108	242.470673
PPAPDC1A	1.24547997	5.96498458	26.0082303	6.36E-111	2.30E-108	242.462731
RAD51	0.5110846	6.20782871	25.9396466	1.74E-110	6.18E-108	241.458985
KIFC1	0.50457912	6.4348331	25.8439254	7.10E-110	2.48E-107	240.05864
SHCBP1	0.54011693	6.24370337	25.7766646	1.91E-109	6.53E-107	239.075061
SPAG5	0.43822876	6.47385191	25.664436	9.88E-109	3.33E-106	237.43467
BUB1B	0.57780742	6.3426436	25.6577021	1.09E-108	3.62E-106	237.336275
GINS1	0.41512631	6.35418956	25.6177404	1.96E-108	6.39E-106	236.752433
LMNB1	0.36687554	6.57488951	25.4807899	1.46E-107	4.68E-105	234.752559
CDCA3	0.53418898	6.24280639	25.4569606	2.07E-107	6.49E-105	234.40474
NCAPH	0.52735754	6.29985916	25.4563439	2.08E-107	6.49E-105	234.395739
BIRC5	0.71475794	6.30609349	25.3581828	8.77E-107	2.69E-104	232.963462
POC1A	0.36631187	6.29818866	25.3230724	1.47E-106	4.43E-104	232.451364
FAM83D	0.61170386	6.29439873	25.3157239	1.63E-106	4.86E-104	232.344197
KIF18B	0.70861531	6.17840068	25.2521753	4.14E-106	1.21E-103	231.417633
C16orf59	0.57874907	6.14327509	25.195595	9.47E-106	2.74E-103	230.592973
CDC45	0.74225347	6.15982582	25.0286572	1.09E-104	3.09E-102	228.16155
PAQR4	0.37964057	6.56718813	25.0022257	1.60E-104	4.49E-102	227.776819
SLC35A2	0.19853559	6.60419465	24.9644219	2.77E-104	7.69E-102	227.22667
SKA1	0.62808127	6.12435005	24.9237402	5.03E-104	1.37E-101	226.634793

CENPF	0.43432926	6.63953848	24.862743	1.22E-103	3.30E-101	225.747652
AURKB	0.69623516	6.18397469	24.8308436	1.95E-103	5.19E-101	225.283853

A4. C4 vs Normal

	logFC	AveExpr	t	P.Value	adj.P.Val	B
MMP11	0.73449949	6.7487667	35.4388219	2.81E-171	5.68E-167	380.827504
COL10A1	1.06113677	6.51546296	32.3914358	7.05E-152	7.13E-148	336.357097
NEK2	0.73439203	6.34404171	31.6809678	2.51E-147	1.70E-143	325.9195
PKMYT1	0.74770832	6.29174483	28.4998931	6.92E-127	3.50E-123	279.050816
PAFAH1B3	0.36863652	6.58546139	28.2405352	3.21E-125	1.30E-121	275.229629
NUF2	0.60911908	6.26557126	27.7852574	2.69E-122	9.06E-119	268.52571
IQGAP3	0.5688466	6.42188467	27.7622291	3.78E-122	1.09E-118	268.186776
KIF4A	0.60913342	6.35962962	27.7414519	5.13E-122	1.30E-118	267.880988
GABRD	0.76114717	5.94238325	27.658282	1.75E-121	3.94E-118	266.657079
RAG1AP1	0.25301933	6.6685569	26.8094924	4.83E-116	9.76E-113	254.181913
PPAPDC1A	1.13978237	5.96498458	26.6684519	3.86E-115	7.09E-112	252.112211
KIAA0101	0.55335481	6.31068887	26.5728957	1.57E-114	2.65E-111	250.710585
HSD17B6	0.56702533	6.06502729	26.5348698	2.76E-114	4.29E-111	250.152962
PAQR4	0.35714613	6.56718813	26.3544419	3.92E-113	5.67E-110	247.508276
UBE2T	0.4351228	6.36588439	26.0804145	2.20E-111	2.97E-108	243.495529
SPC25	0.55108181	6.1274963	26.0021315	6.96E-111	8.80E-108	242.350106
NUAK2	0.38110107	6.28795794	25.9446773	1.62E-110	1.93E-107	241.50972
CPA1	-0.9641995	4.93190324	-25.823734	9.55E-110	1.07E-106	239.741448
UHRF1	0.54724454	6.33767246	25.6545466	1.14E-108	1.22E-105	237.269665
CDC25C	0.70433676	6.03053886	25.4733504	1.63E-107	1.64E-104	234.624933
LOC572558	-0.8740794	4.8688959	-25.414302	3.86E-107	3.72E-104	233.76364
RYR3	-0.7150914	5.66792286	-25.384179	6.00E-107	5.51E-104	233.324381
HAGHL	0.76237692	6.14657488	25.3023061	1.99E-106	1.75E-103	232.130878
C16orf59	0.51414105	6.14327509	25.0794732	5.17E-105	4.36E-102	228.885531
MRAS	-0.2933438	6.54432769	-25.065472	6.34E-105	5.13E-102	228.681771
ASF1B	0.40430849	6.4165402	24.9724266	2.47E-104	1.92E-101	227.328127
PYCR1	0.30020605	6.69412871	24.8993263	7.18E-104	5.38E-101	226.265233

A5. C5 vs Normal

	logFC	AveExpr	t	P.Value	adj.P.Val	B
MMP11	0.64281301	6.7487667	29.3487172	2.42E-132	4.89E-128	291.357232
COL10A1	0.97373768	6.51546296	28.126617	1.73E-124	1.75E-120	273.377356
NEK2	0.642177	6.34404171	26.214521	3.07E-112	2.07E-108	245.331097

A6. C6 vs Normal

	logFC	AveExpr	t	P.Value	adj.P.Val	B
MMP11	0.7487638	6.7487667	34.2479321	1.00E-163	2.02E-159	363.639627
NEK2	0.78546225	6.34404171	32.1216274	3.77E-150	3.81E-146	332.490788
KIF4A	0.72699165	6.35962962	31.3868632	1.94E-145	1.31E-141	321.679606
PKMYT1	0.86296163	6.29174483	31.1820195	4.00E-144	2.02E-140	318.662492
UBE2C	0.70724758	6.39652708	30.1561967	1.56E-137	6.30E-134	303.538962
COL10A1	1.01787147	6.51546296	29.4546259	5.04E-133	1.70E-129	293.188054
CDC25C	0.85667676	6.03053886	29.3713854	1.73E-132	5.00E-129	291.959832
CPA1	-1.1537892	4.93190324	-29.294112	5.43E-132	1.37E-128	290.819671
TPX2	0.48900849	6.56774487	29.2860453	6.12E-132	1.37E-128	290.700642
NUF2	0.67665076	6.26557126	29.2602834	8.96E-132	1.81E-128	290.32053
KIAA0101	0.63771252	6.31068887	29.0309921	2.67E-130	4.91E-127	286.937529
UBE2T	0.50924388	6.36588439	28.9354475	1.10E-129	1.85E-126	285.527957
IQGAP3	0.62195186	6.42188467	28.7751555	1.18E-128	1.83E-125	283.163368
AURKA	0.47050197	6.38487415	28.543669	3.62E-127	5.23E-124	279.749119
UHRF1	0.63989399	6.33767246	28.4375825	1.74E-126	2.34E-123	278.18471
TROAP	0.72293385	6.2219041	28.4259605	2.06E-126	2.61E-123	278.013338
WDR62	0.58984453	6.2419295	28.3872284	3.66E-126	4.36E-123	277.442232
PLK1	0.55678826	6.39482509	28.3786032	4.16E-126	4.67E-123	277.315058
SPC25	0.62907828	6.1274963	28.1383935	1.45E-124	1.55E-121	273.773915
ASF1B	0.47982884	6.4165402	28.0954484	2.74E-124	2.77E-121	273.140965
ZWINT	0.39287645	6.49771018	28.0399905	6.22E-124	5.99E-121	272.323664
H2AFY	0.16970158	6.82948803	28.0356749	6.63E-124	6.10E-121	272.260067
HJURP	0.61573708	6.2629067	28.0136575	9.18E-124	8.08E-121	271.935613
KIF18B	0.73772643	6.17840068	27.9246126	3.43E-123	2.89E-120	270.623564
KLHL29	-0.5566033	6.12874383	-27.688655	1.12E-121	9.06E-119	267.147945
CDK1	0.45093495	6.49757011	27.632476	2.57E-121	2.00E-118	266.320694
SPRY2	-0.4260585	6.35917977	-27.528134	1.20E-120	8.98E-118	264.784537

PBK	0.69952658	6.25257777	27.4746462	2.64E-120	1.91E-117	263.997219
CCNB1	0.35372319	6.54486023	27.4584617	3.36E-120	2.34E-117	263.759013
RYR3	-0.8148531	5.66792286	-27.42096	5.84E-120	3.94E-117	263.207097
CEP55	0.55164549	6.34298029	27.3241751	2.44E-119	1.59E-116	261.782953
DMD	-0.5881811	6.34933294	-27.184529	1.91E-118	1.21E-115	259.728809
HMMR	0.60509492	6.27443208	27.173709	2.24E-118	1.38E-115	259.569689
NUSAP1	0.47217928	6.49254783	27.1621726	2.66E-118	1.58E-115	259.400035
CDCA8	0.43451727	6.37108046	27.1549197	2.96E-118	1.71E-115	259.293378
RECQL4	0.5111195	6.38602512	27.1065494	6.05E-118	3.40E-115	258.582129
TOP2A	0.52943269	6.66122445	27.0932379	7.36E-118	4.02E-115	258.386414
C16orf59	0.58458493	6.14327509	27.0324547	1.80E-117	9.60E-115	257.492832
MELK	0.66078787	6.28831274	26.9868669	3.53E-117	1.83E-114	256.822755
CKS2	0.38157976	6.47078181	26.9610461	5.17E-117	2.61E-114	256.44327
PAFAH1B3	0.37086187	6.58546139	26.9332327	7.79E-117	3.84E-114	256.034536
BUB1	0.56734767	6.36316398	26.8897839	1.48E-116	7.12E-114	255.396107
KIF2C	0.55459651	6.37430417	26.859111	2.32E-116	1.09E-113	254.945463
ATP6AP1	0.18875755	6.85206833	26.7828702	7.14E-116	3.28E-113	253.825545
PTTG1	0.50496059	6.358503	26.6464617	5.33E-115	2.40E-112	251.822583
CDKN3	0.54744984	6.18240124	26.6366839	6.16E-115	2.71E-112	251.67905
NEIL3	0.8420171	5.88302004	26.5752638	1.52E-114	6.54E-112	250.777554
CDCA5	0.45756995	6.4036834	26.5610279	1.88E-114	7.90E-112	250.568636
FOXM1	0.4999091	6.49206266	26.5468164	2.31E-114	9.54E-112	250.360089
CAV1	-0.3522274	6.691822	-26.521897	3.34E-114	1.35E-111	249.994433
POC1A	0.36109086	6.29818866	26.5146227	3.71E-114	1.47E-111	249.887704
CAV2	-0.3760988	6.53056101	-26.391034	2.29E-113	8.90E-111	248.074832
SPAG5	0.42412829	6.47385191	26.3834535	2.56E-113	9.77E-111	247.963666
NDC80	0.57229591	6.26641156	26.3658762	3.31E-113	1.24E-110	247.705915
HSD17B13	-1.1159058	5.27888574	-26.358239	3.71E-113	1.36E-110	247.593929
KIFC1	0.48380318	6.4348331	26.320945	6.42E-113	2.32E-110	247.04714
LOC572558	-0.9545994	4.8688959	-26.311778	7.35E-113	2.61E-110	246.912749
POLR3K	0.26413621	6.47594938	26.2615854	1.54E-112	5.36E-110	246.177011
PYCR1	0.333798	6.69412871	26.245426	1.95E-112	6.68E-110	245.940177
TNS1	-0.2638661	6.78984699	-26.243498	2.01E-112	6.76E-110	245.911924
CCNB2	0.53681759	6.35844758	26.1904556	4.37E-112	1.45E-109	245.134649
NCAPG	0.49685035	6.36860084	26.1294984	1.07E-111	3.49E-109	244.241626
RAD51	0.4826421	6.20782871	26.0195642	5.39E-111	1.73E-108	242.631729
KIF20A	0.60528211	6.38061838	26.0159953	5.68E-111	1.79E-108	242.579478
TK1	0.42589055	6.54806901	25.9719214	1.08E-110	3.37E-108	241.9343
MYOC	-1.3749805	5.09969529	-25.946885	1.57E-110	4.80E-108	241.567865
DLGAP5	0.65488731	6.27899041	25.9447251	1.62E-110	4.88E-108	241.536253
CCDC64	0.54979364	6.37917025	25.8775436	4.34E-110	1.29E-107	240.553212

FAM13A	-0.3325536	6.46929617	-25.835734	8.01E-110	2.35E-107	239.941601
HSD17B6	0.5816795	6.06502729	25.80477	1.26E-109	3.64E-107	239.488718
GINS1	0.39322711	6.35418956	25.7755278	1.94E-109	5.52E-107	239.061091
CDCA3	0.50896745	6.24280639	25.7635191	2.31E-109	6.49E-107	238.885499
BIRC5	0.68006631	6.30609349	25.6279551	1.69E-108	4.67E-106	236.904045
LYVE1	-0.7965458	6.01919277	-25.61967	1.90E-108	5.21E-106	236.783001
LOC286367	-0.5477514	5.92333826	-25.552824	5.07E-108	1.37E-105	235.806531
BUB1B	0.54153052	6.3426436	25.5423668	5.91E-108	1.57E-105	235.653806
SKA1	0.60414267	6.12435005	25.4648099	1.84E-107	4.84E-105	234.521403
SQLE	0.34534635	6.64626784	25.4515955	2.24E-107	5.80E-105	234.328511
MRAS	-0.3140393	6.54432769	-25.438108	2.72E-107	6.97E-105	234.13165
ADAMTS5	-0.5278858	6.37270753	-25.4327	2.95E-107	7.45E-105	234.052709
PPP1R12B	-0.2414478	6.61356813	-25.361606	8.35E-107	2.08E-104	233.015307
TRPM3	-0.8585407	4.87826079	-25.357373	8.88E-107	2.19E-104	232.95356
ARHGAP39	0.33308967	6.45196482	25.3291117	1.34E-106	3.27E-104	232.541301
NKAPL	-0.6707547	5.48110736	-25.321165	1.51E-106	3.63E-104	232.425397
OIP5	0.56242926	6.02364443	25.30231	1.99E-106	4.73E-104	232.1504
GABRD	0.73220208	5.94238325	25.2225653	6.38E-106	1.50E-103	230.987702
LRRC3B	-1.0633462	5.03805967	-25.08153	5.02E-105	1.17E-102	228.93278
ANXA1	-0.2875062	6.76618226	-25.055217	7.37E-105	1.69E-102	228.549596
CA4	-1.517469	5.19276764	-25.013707	1.35E-104	3.07E-102	227.945243
CENPM	0.60528151	6.16264049	24.9225587	5.11E-104	1.14E-101	226.618756
GTSE1	0.52289348	6.22876743	24.9221503	5.14E-104	1.14E-101	226.612814
DTL	0.48725494	6.41790358	24.8939305	7.76E-104	1.71E-101	226.202296
MAZ	0.1926078	6.83363415	24.8652543	1.18E-103	2.57E-101	225.785219
KIF23	0.41787394	6.41583227	24.7848964	3.81E-103	8.20E-101	224.616901

A7. Overlapped genes among the identified subtypes

Subtypes	No. of overlapping genes	Genes
Subtype 1, Subtype 2, Subtype 3, Subtype 4, Subtype 5, Subtype 6	2	NEK2, MMP11
Subtype 1, Subtype 2, Subtype 3, Subtype 4, Subtype 6	4	IQGAP3, KIF4A, NUF2, PKMYT1
Subtype 2, Subtype 3, Subtype 4, Subtype 5, Subtype 6	1	COL10A1
Subtype 1, Subtype 2, Subtype 3, Subtype 6	2	UBE2C, HJURP
Subtype 1, Subtype 3, Subtype 4, Subtype 6	10	CPA1, UHRF1, ASF1B, KIAA0101, CDC25C, SPC25, LOC572558, C16orf59, UBE2T, PYCR1
Subtype 1, Subtype 3, Subtype 6	39	CDK1, TK1, MELK, NEIL3, CEP55, TPX2, POC1A, KIF18B, ZWINT, CCNB1, NDC80, RAD51, SPAG5, OIP5, HMMR, GTSE1, CDKN3, BUB1, NCAPG, CDCA3, GINS1, WDR62, KIFC1, SKA1, PTTG1, TROAP, KIF2C, BIRC5, KIF20A, CDCA5, PLK1, AURKA, NUSAP1, CKS2, BUB1B, FOXM1, DLGAP5, CCNB2, CDCA8
Subtype 3, Subtype 4, Subtype 6	1	HSD17B6
Subtype 1, Subtype 3	14	AURKB, SKA3, NCAPH, CDC20, FAM83D, RRM2, SHCBP1, LMNB1, CDC45, EXO1, ANLN, CENPF, FAM54A, EZH2

Subtype 1, Subtype 6	13	NKAPL, TNS1, DTL, TOP2A, RECQL4, PPP1R12B, MYOC, CAV1, PBK, CA4, CENPM, HSD17B13, KIF23
Subtype 3, Subtype 4	2	PPAPDC1A, PAQR4
Subtype 4, Subtype 6	4	RYR3, PAFAH1B3,GABRD, MRAS
Subtype 1	94/178	HDGF, ESPL1, DONSON, CENPA, ZNF695, CD34, TYMS, CDT1, BYSL, GPRIN1, RNASEH2A, CKAP2L, KIF11, MKI67, CTPS, FAM72B, AOC3, MTHFD2, C16orf75, TNXB, ADRB2, LRRC70, DEPDC1B, HN1, FANCA, SGOL1, FZD4, SEMA3G, ORC6L, LDB2, CDCA2, DDX39, HMGA1, CENPL, MMRN2, PDE2A, NOSTRIN, ASPM, UBE2S, H2AFZ, TUBA1C, RAPGEF3, E2F1, PEAR1, ORC1L, RFC4, CCNA2, LMOD1, TTK, RAD54L, SPARCL1, BTNL9, ATP1A2, C6orf129, ECE2, ILF2, CCT3, PSRC1, SDPR, LMNB2, CNIH2, VPS72, MCM2, C1orf135, KPNA2, TRIP13, MYBL2, LHFP, TACC3, KIF15, ERCC6L, RCC2, GPR172A, STMN1, TMEM206, MCM10, THSD1, FAM64A, SPC24, FLAD1, STIL, KIF14, TENC1, GPAM, CCNE1, MAD2L1, EBF1, NFKBIL2, ACSM5, DEPDC1, CXorf36, RAD51AP1, E2F2, PRC1
Subtype 3	1/76	SLC35A2
Subtype 4	3/27	RAG1AP1, NUAK2, HAGHL
Subtype 6	18/94	TRPM3, ANXA1, ADAMTS5, POLR3K, CCDC64, ATP6AP1, DMD, LOC286367, MAZ, LRRC3B, SPRY2, SQLE, ARHGAP39, CAV2, FAM13A, KLHL29, LYVE1, H2AFY

Appendix B: The data of the 7-layer BC biological system hierarchy.

The first column is the gene name. Second column is the biological systems at the second layer, third column is the biological system at the third layer, etc.

A1CF	175	59	9	11	0
AAAS	103	133	50	6	1
AAR2	320	56	26	11	0
AATF	376	79	26	11	0
ABAT	561	59	9	11	0
ABCA2	517	59	9	11	0
ABCB6	581	59	9	11	0
ABCB8	581	59	9	11	0
ABCB9	581	59	9	11	0
ABCC8	38	69	9	11	0
ABCC9	38	69	9	11	0
ABCD4	581	59	9	11	0
ABCG5	168	59	9	11	0
ABCG8	168	59	9	11	0
ABHD12	299	69	9	11	0
ABHD6	299	69	9	11	0
ABI1	0	15	9	11	0
ABI2	0	15	9	11	0
ABI3	473	15	9	11	0
ABL1	292	15	9	11	0
ABT1	605	79	26	11	0
ABTB1	168	59	9	11	0
ACD	1	46	34	11	0
ACIN1	30	150	52	11	0
ACR	181	17	9	11	0
ACTA1	2	0	4	10	1
ACTA2	2	0	4	10	1
ACTB	2	0	4	10	1
ACTC1	2	0	4	10	1
ACTG1	2	0	4	10	1
ACTG2	2	0	4	10	1
ACTL10	3	59	9	11	0
ACTL6A	5	59	9	11	0
ACTL6B	5	59	9	11	0
ACTL7A	3	59	9	11	0
ACTL7B	3	59	9	11	0
ACTL8	3	59	9	11	0
ACTN4	560	59	9	11	0
ACTR10	245	59	9	11	0
ACTR1A	4	59	9	11	0
ACTR1B	4	59	9	11	0
ACTR2	5	59	9	11	0
ACTR3	5	59	9	11	0
ACTR5	245	59	9	11	0
ACTR6	245	59	9	11	0
ACTR8	245	59	9	11	0
ACVR1	6	1	9	11	0
ACVR1B	6	1	9	11	0
ACVR1C	6	1	9	11	0
ACVR2A	6	1	9	11	0
ACVR2B	6	1	9	11	0
ACVRL1	6	1	9	11	0
ADAM10	268	69	9	11	0
ADAM8	268	69	9	11	0
ADAR	448	59	9	11	0
ADCYAP1R1	7	59	9	11	0
ADD1	8	15	9	11	0
ADD2	8	15	9	11	0
ADGRA2	593	69	9	11	0
ADIPOQ	62	17	9	11	0
ADNP	617	150	52	11	0
ADRA2A	271	59	9	11	0
ADRB2	246	51	9	11	0
ADRB3	221	51	9	11	0
ADRM1	527	122	33	7	1
ADSL	248	59	9	11	0
AEBP2	524	13	3	11	0
AFDN	49	15	9	11	0
AFF1	9	2	0	11	0
AFF3	9	2	0	11	0
AFF4	9	2	0	11	0
AFG3L2	453	59	9	11	0
AFTPH	654	150	52	11	0
AGFG1	171	15	9	11	0
AGK	299	69	9	11	0
AGL	10	3	27	7	1
AGO1	170	59	9	11	0
AGO2	170	59	9	11	0
AGO3	170	59	9	11	0
AGO4	170	59	9	11	0

CRLF2	255	37	9	11	0
CRNKL1	619	99	9	11	0
CRTAP	400	15	9	11	0
CRX	11	4	1	11	0
CRYAA	130	59	9	11	0
CRYAB	130	59	9	11	0
CRYBG3	58	150	52	11	0
CSDE1	199	125	49	7	1
CSF1	233	75	34	11	0
CSF1R	322	69	9	11	0
CSF2	155	150	52	11	0
CSF2RA	255	37	9	11	0
CSF2RB	208	37	9	11	0
CSF3R	158	37	9	11	0
CSNK1A1	96	59	9	11	0
CSNK1E	96	59	9	11	0
CSNK2A1	96	59	9	11	0
CSNK2A2	96	59	9	11	0
CSNK2A3	96	59	9	11	0
CSNK2B	59	15	9	11	0
CSTA	157	36	20	11	0
CSTF1	582	133	50	6	1
CSTF2	129	55	28	5	1
CSTF2T	129	55	28	5	1
CT45A10	131	150	52	11	0
CT45A2	131	150	52	11	0
CT45A3	131	150	52	11	0
CT45A6	131	150	52	11	0
CTBP1	70	22	6	10	1
CTBP2	70	22	6	10	1
CTC1	1	46	34	11	0
CTDNEP1	541	59	9	11	0
CTDP1	49	15	9	11	0
CTHRC1	79	17	9	11	0
CTLA4	515	37	9	11	0
CTNNA1	177	141	26	11	0
CTNNA2	177	141	26	11	0
CTNNB1	443	147	49	7	1
CTNNBIP1	331	75	34	11	0
CTNNBL1	443	147	49	7	1
CTNND1	421	15	9	11	0
CTR9	619	99	9	11	0
CTSB	181	17	9	11	0
CTU1	283	59	9	11	0
CTU2	253	64	9	11	0
CUBN	298	64	9	11	0
CUL1	132	59	9	11	0
CUL2	132	59	9	11	0
CUL3	132	59	9	11	0
CUL4A	132	59	9	11	0
CUL4B	132	59	9	11	0
CUL5	132	59	9	11	0
CUL7	418	67	20	11	0
CUL9	418	67	20	11	0
CWC15	534	79	26	11	0
CWC22	566	132	0	11	0
CWC25	477	144	9	11	0
CWC27	437	95	15	5	1
CWF19L1	133	30	15	5	1
CWF19L2	605	79	26	11	0
CXADR	322	69	9	11	0
CXCR4	246	51	9	11	0
CXXC1	59	15	9	11	0
CYB5B	175	59	9	11	0
CYB5R3	378	69	9	11	0
CYBA	368	73	47	11	0
CYBB	378	69	9	11	0
CYC1	268	69	9	11	0
CYCS	636	59	9	11	0
CYFIP1	622	141	26	11	0
CYFIP2	622	141	26	11	0
DAB2IP	49	15	9	11	0
DACH1	11	4	1	11	0
DAD1	114	7	9	11	0
DAG1	134	16	26	11	0
DAP3	466	69	9	11	0
DAPK1	13	59	9	11	0
DAZAP1	135	55	28	5	1
DAZAP2	48	16	26	11	0
DAZL	99	22	6	10	1

HYAL2	268	69	9	11	0
HYOU1	245	59	9	11	0
HYPK	247	36	20	11	0
ICE1	58	150	52	11	0
ICE2	319	150	52	11	0
ID2	373	12	9	11	0
IDH3A	248	59	9	11	0
IDH3B	248	59	9	11	0
IDH3G	248	59	9	11	0
IFITM1	68	59	9	11	0
IFITM2	68	59	9	11	0
IFITM3	68	59	9	11	0
IFNL1	382	80	24	11	0
IFNLR1	208	37	9	11	0
IFT122	249	99	9	11	0
IFT140	642	79	26	11	0
IFT172	249	99	9	11	0
IFT20	233	75	34	11	0
IFT22	509	9	9	11	0
IFT27	250	25	9	11	0
IFT43	155	150	52	11	0
IFT46	157	36	20	11	0
IFT52	143	32	9	11	0
IFT57	251	16	26	11	0
IFT74	251	16	26	11	0
IFT80	249	99	9	11	0
IFT81	251	16	26	11	0
IFT88	619	99	9	11	0
IGF1	407	106	9	11	0
IGF1R	170	59	9	11	0
IGF2BP1	404	22	6	10	1
IGF2R	64	17	9	11	0
IGFALS	221	51	9	11	0
IGFBP3	252	17	9	11	0
IGFBP5	252	17	9	11	0
IGFBP6	252	17	9	11	0
IGHMBP2	170	59	9	11	0
IGLL1	238	59	9	11	0
IGLL5	238	59	9	11	0
IK	534	79	26	11	0
IKBKB	253	64	9	11	0
IKBKE	253	64	9	11	0
IKBKG	403	15	9	11	0
IKZF1	254	59	9	11	0
IKZF4	254	59	9	11	0
IKZF5	617	150	52	11	0
IL10RB	255	37	9	11	0
IL11RA	158	37	9	11	0
IL12A	155	150	52	11	0
IL12B	255	37	9	11	0
IL12RB1	158	37	9	11	0
IL12RB2	158	37	9	11	0
IL13RA1	255	37	9	11	0
IL13RA2	255	37	9	11	0
IL18R1	322	69	9	11	0
IL18RAP	322	69	9	11	0
IL23A	382	80	24	11	0
IL23R	208	37	9	11	0
IL27RA	158	37	9	11	0
IL2RA	90	37	9	11	0
IL2RB	208	37	9	11	0
IL2RG	255	37	9	11	0
IL3RA	255	37	9	11	0
IL4R	208	37	9	11	0
IL5RA	255	37	9	11	0
IL6	155	150	52	11	0
IL6R	158	37	9	11	0
IL6ST	255	37	9	11	0
ILDR1	524	13	3	11	0
ILF2	256	55	28	5	1
ILF3	257	55	28	5	1
ILVBL	561	59	9	11	0
IMMP1L	258	118	9	11	0
IMMP2L	258	118	9	11	0
IMMT	321	16	26	11	0
IMP3	645	79	26	11	0
IMP4	333	71	21	11	0
IMPACT	49	15	9	11	0
IMPG2	259	150	52	11	0
INCENP	558	64	9	11	0

NLRP9	358	59	9	11	0
NME8	481	64	9	11	0
NOB1	194	48	13	5	1
NOC2L	620	59	9	11	0
NOC4L	360	59	9	11	0
NOD2	358	59	9	11	0
NOL10	645	79	26	11	0
NOL11	388	83	26	11	0
NOL6	473	15	9	11	0
NOLC1	241	135	26	11	0
NOMO2	593	69	9	11	0
NONO	129	55	28	5	1
NOP10	340	75	34	11	0
NOP14	635	79	26	11	0
NOP56	377	59	9	11	0
NOP58	377	59	9	11	0
NOP9	620	59	9	11	0
NOS1	49	15	9	11	0
NOTCH1	593	69	9	11	0
NOTCH2	593	69	9	11	0
NOTCH3	593	69	9	11	0
NOX1	378	69	9	11	0
NOX4	378	69	9	11	0
NOXA1	473	15	9	11	0
NOXO1	473	15	9	11	0
NPAS2	379	10	9	11	0
NPAS4	380	13	3	11	0
NPHS2	381	59	9	11	0
NPIPA1	331	75	34	11	0
NPLOC4	157	36	20	11	0
NPM1	157	36	20	11	0
NPPA	382	80	24	11	0
NPPB	382	80	24	11	0
NPPC	382	80	24	11	0
NPR1	170	59	9	11	0
NPR3	491	69	9	11	0
NPRL2	383	83	26	11	0
NPRL3	383	83	26	11	0
NR0B2	385	22	6	10	1
NR1H2	385	22	6	10	1
NR1H3	385	22	6	10	1
NR1H4	385	22	6	10	1
NR1I2	385	22	6	10	1
NR2E3	472	22	6	10	1
NR3C1	384	12	9	11	0
NR3C2	384	12	9	11	0
NR4A1	385	22	6	10	1
NR4A2	385	22	6	10	1
NR4A3	385	22	6	10	1
NR5A1	385	22	6	10	1
NR5A2	385	22	6	10	1
NR6A1	385	22	6	10	1
NRIP1	386	150	52	11	0
NRP1	593	69	9	11	0
NRP2	593	69	9	11	0
NRXN2	206	106	9	11	0
NSA2	496	71	21	11	0
NSFL1C	629	103	33	7	1
NSL1	155	150	52	11	0
NSMCE1	124	99	9	11	0
NSMCE2	209	99	9	11	0
NSMCE3	306	13	3	11	0
NSMCE4A	577	16	26	11	0
NSRP1	477	144	9	11	0
NSUN3	324	69	9	11	0
NSUN4	324	69	9	11	0
NT5DC3	59	15	9	11	0
NTN4	289	69	9	11	0
NTNG1	289	69	9	11	0
NTNG2	289	69	9	11	0
NTRK1	322	69	9	11	0
NTRK2	491	69	9	11	0
NTRK3	491	69	9	11	0
NUDCD3	321	16	26	11	0
NUDT21	333	71	21	11	0
NUF2	251	16	26	11	0
NUFIP1	524	13	3	11	0
NUMA1	101	24	26	11	0
NUP107	387	81	25	7	1
NUP133	388	83	26	11	0

SDHC	125	73	47	11	0
SDHD	125	73	47	11	0
SDR16C5	299	69	9	11	0
SEC11A	519	118	9	11	0
SEC11C	519	118	9	11	0
SEC13	25	133	50	6	1
SEC22B	572	16	26	11	0
SEC23A	520	55	28	5	1
SEC23B	520	55	28	5	1
SEC24A	520	55	28	5	1
SEC24B	520	55	28	5	1
SEC24C	520	55	28	5	1
SEC24D	520	55	28	5	1
SEC31A	328	79	26	11	0
SEC31B	328	79	26	11	0
SEC61A1	13	59	9	11	0
SEC61A2	13	59	9	11	0
SEC61B	68	59	9	11	0
SEC61G	532	59	9	11	0
SEC63	605	79	26	11	0
SECISBP2	521	150	52	11	0
SECISBP2L	521	150	52	11	0
SEH1L	474	108	36	5	1
SEL1L	93	59	9	11	0
SELENOS	605	79	26	11	0
SEM1	434	93	20	11	0
SEMG1	522	120	34	11	0
SENP2	280	59	9	11	0
SENP3	280	59	9	11	0
SERF1A	523	121	44	11	0
SERPINA5	158	37	9	11	0
SERPINB6	170	59	9	11	0
SERPINE1	158	37	9	11	0
SERPINF2	158	37	9	11	0
SESN2	34	16	26	11	0
SESN3	34	16	26	11	0
SET	59	15	9	11	0
SETD1A	524	13	3	11	0
SETD1B	524	13	3	11	0
SETD5	59	15	9	11	0
SF1	573	15	9	11	0
SF3A1	573	15	9	11	0
SF3A2	129	55	28	5	1
SF3A3	525	79	26	11	0
SF3B1	619	99	9	11	0
SF3B2	639	15	9	11	0
SF3B3	249	99	9	11	0
SF3B4	547	22	6	10	1
SF3B5	331	75	34	11	0
SF3B6	547	22	6	10	1
SFPQ	129	55	28	5	1
SFR1	530	150	52	11	0
SFTPA1	526	17	9	11	0
SFTPA2	526	17	9	11	0
SFTPD	526	17	9	11	0
SGCA	134	16	26	11	0
SGCB	134	16	26	11	0
SGCD	134	16	26	11	0
SGCE	134	16	26	11	0
SGCG	134	16	26	11	0
SGF29	60	141	26	11	0
SGIP1	49	15	9	11	0
SGO2	58	150	52	11	0
SGTA	527	122	33	7	1
SGTB	527	122	33	7	1
SH3GLB1	420	59	9	11	0
SHARPIN	473	15	9	11	0
SHC1	374	13	3	11	0
SHISA7	374	13	3	11	0
SHISA9	374	13	3	11	0
SHMT2	561	59	9	11	0
SHOC2	299	69	9	11	0
SHPRH	280	59	9	11	0
SHTN1	558	64	9	11	0
SIAH1	49	15	9	11	0
SIN3A	537	125	49	7	1
SIN3B	537	125	49	7	1
SIPA1	473	15	9	11	0
SIRT1	49	15	9	11	0
SIRT2	49	15	9	11	0

AHCTF1	388	83	26	11	0	DBF4	580	150	52	11	0	ING1	191	64	9	11	0	NUP153	392	83	26	11	0	SIRT3	49	15	9	11	0
AHI1	148	83	26	11	0	DBF4B	136	150	52	11	0	ING2	59	15	9	11	0	NUP155	388	83	26	11	0	SIX1	528	22	6	10	1
AHNAK	129	55	28	5	1	DBI	59	15	9	11	0	ING3	59	15	9	11	0	NUP160	388	83	26	11	0	SIX2	528	22	6	10	1
AHR	379	10	9	11	0	DBP	137	13	3	11	0	ING4	59	15	9	11	0	NUP188	389	82	26	11	0	SIX3	528	22	6	10	1
AHRR	379	10	9	11	0	DBT	146	34	9	11	0	ING5	59	15	9	11	0	NUP205	389	82	26	11	0	SIX4	49	15	9	11	0
AHSP	331	75	34	11	0	DCAF1	649	147	49	7	1	INHA	295	106	9	11	0	NUP210	593	69	9	11	0	SIX5	529	123	3	11	0
AICDA	358	59	9	11	0	DCAF10	138	31	50	6	1	INHBA	407	106	9	11	0	NUP210L	593	69	9	11	0	SIX6	528	22	6	10	1
AIM2	586	150	52	11	0	DCAF11	138	31	50	6	1	INIP	260	75	34	11	0	NUP214	390	83	26	11	0	SKA1	457	16	26	11	0
AIMP1	179	43	19	5	1	DCAF12	140	79	26	11	0	INO80	59	15	9	11	0	NUP35	391	84	35	11	0	SKA2	233	75	34	11	0
AIMP2	481	64	9	11	0	DCAF13	139	146	50	6	1	INO80B	428	79	26	11	0	NUP37	140	79	26	11	0	SKA3	530	150	52	11	0
AIP	362	15	9	11	0	DCAF15	148	83	26	11	0	INO80C	501	71	21	11	0	NUP43	140	79	26	11	0	SKAP1	49	15	9	11	0
AJUBA	270	15	9	11	0	DCAF16	331	75	34	11	0	INO80D	639	15	9	11	0	NUP50	392	83	26	11	0	SKI	11	4	1	11	0
AKAP10	49	15	9	11	0	DCAF17	185	150	52	11	0	INO80E	524	13	3	11	0	NUP54	392	83	26	11	0	SKIL	49	15	9	11	0
AKAP17A	630	144	9	11	0	DCAF4	140	79	26	11	0	INSC	299	69	9	11	0	NUP58	392	83	26	11	0	SKIV2L	170	59	9	11	0
AKAP4	155	150	52	11	0	DCAF4L1	140	79	26	11	0	INSIG1	261	7	9	11	0	NUP62	392	83	26	11	0	SKOR1	11	4	1	11	0
AKAP5	117	150	52	11	0	DCAF5	138	31	50	6	1	INSIG2	261	7	9	11	0	NUP62CL	321	16	26	11	0	SKP1	168	59	9	11	0
AKAP6	639	15	9	11	0	DCAF6	138	31	50	6	1	INSM1	11	4	1	11	0	NUP85	387	81	25	7	1	SKP2	299	69	9	11	0
AKAP7	131	150	52	11	0	DCAF7	474	108	36	5	1	INSR	170	59	9	11	0	NUP88	45	83	26	11	0	SLA2	552	129	9	11	0
AKAP8L	129	55	28	5	1	DCAF8	138	31	50	6	1	INSRR	170	59	9	11	0	NUP93	387	81	25	7	1	SLBP	531	15	9	11	0
AKAP9	395	64	9	11	0	DCHS1	95	59	9	11	0	INTS1	263	56	26	11	0	NUP98	392	83	26	11	0	SLC11A2	13	59	9	11	0
AKIRIN2	11	4	1	11	0	DCLRE1C	59	15	9	11	0	INTS10	481	64	9	11	0	NUPR1	218	75	34	11	0	SLC12A7	111	59	9	11	0
AKT1	446	59	9	11	0	DCP2	483	109	37	5	1	INTS11	127	30	15	5	1	NUS1	299	69	9	11	0	SLC17A6	271	59	9	11	0
AKT2	446	59	9	11	0	DCTN1	240	16	26	11	0	INTS12	262	36	20	11	0	NUTF2	394	55	28	5	1	SLC17A7	271	59	9	11	0
AKTIP	321	16	26	11	0	DCTN2	318	16	26	11	0	INTS13	265	83	26	11	0	NVL	453	59	9	11	0	SLC18A3	299	69	9	11	0
ALB	143	32	9	11	0	DCTN3	318	16	26	11	0	INTS14	265	83	26	11	0	NXF1	59	15	9	11	0	SLC1A2	532	59	9	11	0
ALG14	299	69	9	11	0	DCTN4	45	83	26	11	0	INTS2	263	56	26	11	0	NXF3	393	85	20	11	0	SLC1A3	532	59	9	11	0
ALK	64	17	9	11	0	DCTN5	141	69	9	11	0	INTS3	567	83	26	11	0	NXF5	393	85	20	11	0	SLC1A6	532	59	9	11	0
ALS2	253	64	9	11	0	DCTN6	141	69	9	11	0	INTS4	360	59	9	11	0	NXT1	394	55	28	5	1	SLC25A4	70	22	6	10	1
ALX4	11	4	1	11	0	DCUN1D1	142	57	9	11	0	INTS5	264	56	26	11	0	NXT2	394	55	28	5	1	SLC25A5	70	22	6	10	1
ALYREF	129	55	28	5	1	DCUN1D2	142	57	9	11	0	INTS6	265	83	26	11	0	OBSL1	593	69	9	11	0	SLC25A6	70	22	6	10	1
AMBRA1	138	31	50	6	1	DCUN1D3	142	57	9	11	0	INTS6L	265	83	26	11	0	OCLN	212	58	48	5	1	SLC26A6	111	59	9	11	0
AMFR	23	9	9	11	0	DCUN1D4	142	57	9	11	0	INTS7	123	147	49	7	1	ODF2	395	64	9	11	0	SLC27A5	170	59	9	11	0
AMHR2	485	64	9	11	0	DCUN1D5	142	57	9	11	0	INTS8	481	64	9	11	0	OGDH	408	59	9	11	0	SLC3A2	642	79	26	11	0
AMN	134	16	26	11	0	DCX	403	15	9	11	0	INTS9	127	30	15	5	1	OGDHL	408	59	9	11	0	SLC41A1	532	59	9	11	0
AMN1	299	69	9	11	0	DDA1	331	75	34	11	0	IPO4	266	147	49	7	1	OGG1	166	15	9	11	0	SLC51A	600	7	9	11	0
ANAPC1	319	150	52	11	0	DDB1	249	99	9	11	0	IPO5	266	147	49	7	1	OGT	93	59	9	11	0	SLC51B	233	75	34	11	0
ANAPC10	299	69	9	11	0	DDB2	140	79	26	11	0	IPO7	266	147	49	7	1	OLFM2	79	17	9	11	0	SLC6A3	111	59	9	11	0
ANAPC11	70	22	6	10	1	DDIT3	345	75	34	11	0	IQGAP1	560	59	9	11	0	OLFM3	79	17	9	11	0	SLC6A6	111	59	9	11	0
ANAPC13	331	75	34	11	0	DDOST	143	32	9	11	0	IRAK1	473	15	9	11	0	OLFM4	79	17	9	11	0	SLC7A1	111	59	9	11	0
ANAPC15	434	93	20	11	0	DDR1	170	59	9	11	0	IRF4	164	13	3	11	0	OLR1	90	37	9	11	0	SLC7A5	111	59	9	11	0
ANAPC16	218	75	34	11	0	DDR2	170	59	9	11	0	IRF9	164	13	3	11	0	OOEP	134	16	26	11	0	SLC9A1	13	59	9	11	0
ANAPC2	176	15	9	11	0	DDX1	144	59	9	11	0	IRS1	473	15	9	11	0	ORAI1	68	59	9	11	0	SLC9A3R1	403	15	9	11	0
ANAPC4	481	64	9	11	0	DDX17	161	22	6	10	1	IRS2	473	15	9	11	0	ORC1	59	15	9	11	0	SLF1	311	67	20	11	0
ANAPC5	619	99	9	11	0	DDX19B	161	22	6	10	1	ISCU	371	52	16	6	1	ORC2	396	135	26	11	0	SLF2	425	36	20	11	0
ANAPC7	619	99	9	11	0	DDX20	253	64	9	11	0	ISG15	419	59	9	11	0	ORC3	397	86	9	11	0	SLIRP	533	124	9	11	0
ANG	304	80	24	11	0	DDX21	144	59	9	11	0	ISL1	528	22	6	10	1	ORC4	397	86	9	11	0	SLITRK5	491	69	9	11	0
ANKIB1	23	9	9	11	0	DDX23	144	59	9	11	0	ISY1	605	79	26	11	0	ORC5	397	86	9	11	0	SLU7	534	79	26	11	0
ANKRA2	311	67	20	11	0	DDX39B	161	22	6	10	1	ITCH	541	59	9	11	0	ORC6	82	13	3	11	0	SLX1A	473	15	9	11	0
ANKRD1	27	69	9	11	0	DDX3X	161	22	6	10	1	ITFG2	45	83	26	11	0	ORMDL1	398	62	18	11	0	SLX4	30	150	52	11	0
ANKRD39	27	69	9	11	0	DDX41	144	59	9	11	0	ITGA1	267	69	9	11	0	ORMDL2	398	62	18	11	0	SMAD1	535	125	49	7	1
ANKRD9	12	56	26	11	0	DDX46	144	59	9	11	0	ITGA10	267	69	9	11	0	ORMDL3	398	62	18	11	0	SMAD2	535	125	49	7	1
ANO1	13	59	9	11	0	DDX5	161	22	6	10	1	ITGA11	267	69	9	11	0	OS9	252	17	9	11	0	SMAD3	535	125	49	7	1
ANO2	13	59	9	11	0	DDX58	49	15	9	11	0	ITGA2	267	69	9	11	0	OSGEP	413	59	9	11	0	SMAD4	535	125	49	7	1
ANO6	13	59	9	11	0	DDX6	144	59	9	11	0	ITGA2B	267	69	9	11	0	OSMR	255	37	9	11	0	SMAD5	535	125	49	7	1
ANP32E	59	15	9	11	0	DECR1	299	69	9	11	0	ITGA3	267	69	9	11	0	OST4	368	73	47	11	0	SMAD6	536	15	9	11	0
ANXA2	426	89	28	5	1	DEK	241	135	26	11	0	ITGA4	267	69	9	11	0	OSTC	399	62	18	11	0	SMAD7	536	15	9	11	0
AP1B1	123	147	49	7	1	DENND4C	265	83	26	11	0	ITGA5	267	69	9	11	0	OSTM1	470	106	9	11	0	SMAD9	535	125	49	7	1
AP1G1	123	147	49	7	1	DEPDC1	49	15	9	11	0	ITGA6	267	69	9	11	0	OTULIN	253	64	9	11	0	SMARCA1	59	15	9	11	0
AP1G2	123	147	49	7	1	DEPDC5	265	83	26	11	0	ITGA7	267	69	9	11	0	OTX2	11	4	1	11	0	SMARCA2	59	15	9	11	0
AP1M1	14	5	2	10	1	DERL1	111	59	9	11	0	ITGA8	267	69	9	11	0	OXA1L	316	79	26	11	0	SMARCA4	59	15	9	11	0
AP1M2	14	5	2	10	1	DERL2	111	59	9	11	0	ITGA9	267	69	9	11	0	P2RX1	518	69	9	11	0	SMARCA5	59	15	9	11	0
AP1S1	15	6	2	10	1	DERL3	111	59	9	11	0	ITGAD	267	69	9	11	0	P2RX6	518	69	9	11	0	SMARCB1	59	15	9	11	0
AP1S2	15	6	2	10	1	DESI1	541	59	9	11	0	ITGAE	267	69	9	11	0	P3H1	400	15	9	11	0	SMARCC1	344	70	20	11	0
AP1S3	15	6	2	10	1	DET1	148	83	26	11	0	ITGAL	267	69	9	11	0	P3H3	400	15	9	11	0	SMARCC2	344	70	20	11	0
AP2A1	123	147	49	7	1	DFFA	233	75	34	11	0	ITGAM	267	69	9	11	0	P3H4	400	15	9	11	0	SMARCD1	537	125	49	7	1
AP2A2	123	147	49	7	1	DFFB	253	64	9	11	0	ITGAV	267	69	9	11	0	P4HA1	13	59	9	11	0	SMARCD2	537	125	49	7	1
AP2B1	123	147	49	7	1	DGCR8	448	59	9	11	0	ITGAX	267	69	9	11	0	P4HB	213	52	16	6	1	SMARCD3	537	125	49	7	1
AP2M1	14	5	2	10	1	DGKI	59	15	9	11	0	ITGB1	268	69	9	11	0	PA2G4	124	99	9	11	0	SMARCE1	573	15	9	11	0
AP2S1	15	6	2	10	1	DHDDS	13	59	9	11	0	ITGB2	268	69	9	11	0	PAAF1	35	31	50	6	1	SMC1A	283	59	9	11	0
AP3B1	16	147	49	7	1	DHRS7B	299	69	9	11	0	ITGB3	268	69	9	11	0	PABPC1	163	22	6	10	1	SMC1B	283	59	9	11	0
AP3B2	16	147	49	7	1	DHX15	280	59	9	11	0	ITGB4	268	69	9	11	0	PABPC1L	163	22	6	10	1	SMC2	283	59	9	11	0
AP3D1	16	147	49	7	1	DHX16	144	59	9	11	0	ITGB5	268	69	9	11	0	PABPC3	163	22	6	10	1	SMC3	283	59	9	11	0
AP3M1	14	5	2	10	1	DHX29	280	59	9	11	0	ITGB6	268	69	9	11	0	PABPC4	163	22	6	10	1	SMC4	283	59	9	11	0
AP3M2	14	5	2	10	1	DHX32	284	59	9	11	0	ITGB7	268	69	9	11	0	PABPC4L	163	22	6	10	1	SMC5	538	59	9	11	0
AP3S1	15	6	2	10	1	DHX33	280	59	9	11	0	ITGB8	268	69	9	11	0	PABPC5	163	22	6	10	1	SMC6	538	59	9	11	0
AP3S2	15	6	2	10	1	DHX35	280	59	9	11	0	ITGBL1	298	64	9	11	0	PABPN1	570	31	50	6	1	SMCR8	65	35	26	11	0
AP4B1	123	147	49	7	1	DHX38	144	59	9	11	0	ITLN1	197	64	9	11	0	PADI4	401	141	26	11	0	SMDT1	369	73	47	11	0
AP4E1	123	147	49	7	1	DHX8	144	59	9	11	0	ITPR1	269	83	26	11	0	PAF1	434	93	20	11	0	SMG1	263	56	26	11	0
AP4M1	14	5	2	10	1	DHX9	280	59	9	11	0	ITPR2	269	83	26	11	0	PAFAH1B1	582	133	50	6	1	SMG5	328	79	26	11	0
AP4S1	15	6	2	10	1	DIABLO	338	73	47	11	0	ITPR3	269	83	26	11	0	PAFAH1B2	324	69	9	11	0	SMG6	539	64	9	11	0

AP5B1	264	56	26	11	0	DICER1	49	15	9	11	0	IVNS1ABP	168	59	9	11	0	PAFAH1B3	324	69	9	11	0	SMG7	539	64	9	11	0
AP5M1	17	35	26	11	0	DIS3	23	9	9	11	0	IZUMO1	90	37	9	11	0	PAGR1	402	56	26	11	0	SMN1	49	15	9	11	0
AP5S1	349	75	34	11	0	DIS3L	23	9	9	11	0	JADE1	59	15	9	11	0	PAIP1	115	56	26	11	0	SMNDC1	605	79	26	11	0
AP5Z1	620	59	9	11	0	DIS3L2	23	9	9	11	0	JADE2	59	15	9	11	0	PAK1	70	22	6	10	1	SMTN	49	15	9	11	0
APAF1	18	69	9	11	0	DISC1	136	150	52	11	0	JADE3	59	15	9	11	0	PALB2	117	150	52	11	0	SMTNL1	540	36	20	11	0
APBB1IP	639	15	9	11	0	DIXDC1	403	15	9	11	0	JAK2	491	69	9	11	0	PAM16	154	118	9	11	0	SMTNL2	49	15	9	11	0
APC	421	15	9	11	0	DKC1	145	33	15	5	1	JAKMIP1	240	16	26	11	0	PAN2	328	79	26	11	0	SMU1	582	133	50	6	1
APC2	421	15	9	11	0	DLAT	146	34	9	11	0	JAM2	322	69	9	11	0	PAN3	59	15	9	11	0	SMURF2	541	59	9	11	0
APEX1	123	147	49	7	1	DLD	147	59	9	11	0	JAM3	322	69	9	11	0	PANX1	299	69	9	11	0	SNAI3	270	15	9	11	0
APH1A	19	7	9	11	0	DLG1	170	59	9	11	0	JARID2	539	64	9	11	0	PARD3	170	59	9	11	0	SNAP23	572	16	26	11	0
APH1B	19	7	9	11	0	DLG3	170	59	9	11	0	JAZF1	254	59	9	11	0	PARD3B	403	15	9	11	0	SNAP25	572	16	26	11	0
API5	123	147	49	7	1	DLG4	170	59	9	11	0	JCHAIN	331	75	34	11	0	PARD6A	403	15	9	11	0	SNAP29	572	16	26	11	0
APOA1	20	8	24	11	0	DLST	146	34	9	11	0	JDP2	31	12	9	11	0	PARD6B	403	15	9	11	0	SNAP47	253	64	9	11	0
APOA2	20	8	24	11	0	DMAP1	344	70	20	11	0	JMJD1C	319	150	52	11	0	PARD6G	403	15	9	11	0	SNAPC1	320	56	26	11	0
APOA4	20	8	24	11	0	DMD	560	59	9	11	0	JMJD6	59	15	9	11	0	PARP1	59	15	9	11	0	SNAPC3	418	67	20	11	0
APOA5	20	8	24	11	0	DMXL1	148	83	26	11	0	JPH1	270	15	9	11	0	PARP11	418	67	20	11	0	SNAPC4	586	150	52	11	0
APOB	143	32	9	11	0	DMXL2	148	83	26	11	0	JPH2	270	15	9	11	0	PARP4	123	147	49	7	1	SNAPIN	542	16	26	11	0
APOBEC1	358	59	9	11	0	DNAAF2	586	150	52	11	0	JPH3	270	15	9	11	0	PARP9	358	59	9	11	0	SNCA	543	126	46	13	1
APOBEC3F	358	59	9	11	0	DNAH1	149	35	26	11	0	JPH4	270	15	9	11	0	PAX2	99	22	6	10	1	SND1	199	125	49	7	1
APOBEC3G	358	59	9	11	0	DNAH10	150	35	26	11	0	JRK	586	150	52	11	0	PAXIP1	59	15	9	11	0	SNF8	124	99	9	11	0
APOBR	540	36	20	11	0	DNAH11	150	35	26	11	0	JUN	200	12	9	11	0	PBRM1	59	15	9	11	0	SNIP1	129	55	28	5	1
APOC1	20	8	24	11	0	DNAH12	149	35	26	11	0	JUNB	200	12	9	11	0	PBX1	528	22	6	10	1	SNRNP200	144	59	9	11	0
APOC2	20	8	24	11	0	DNAH14	149	35	26	11	0	JUND	200	12	9	11	0	PBX2	528	22	6	10	1	SNRNP25	405	59	9	11	0
APOC3	20	8	24	11	0	DNAH17	150	35	26	11	0	JUP	443	147	49	7	1	PBX3	528	22	6	10	1	SNRNP27	566	132	0	11	0
APOC4	20	8	24	11	0	DNAH2	150	35	26	11	0	KANSL1	639	15	9	11	0	PBXIP1	233	75	34	11	0	SNRNP35	544	132	0	11	0
APOE	20	8	24	11	0	DNAH3	149	35	26	11	0	KANSL1L	639	15	9	11	0	PCBP1	404	22	6	10	1	SNRNP40	646	146	50	6	1
APOF	155	150	52	11	0	DNAH5	150	35	26	11	0	KANSL2	639	15	9	11	0	PCBP2	404	22	6	10	1	SNRNP48	545	127	26	11	0
APOH	102	69	9	11	0	DNAH6	149	35	26	11	0	KANSL3	11	4	1	11	0	PCBP3	404	22	6	10	1	SNRNP70	544	132	0	11	0
APOL1	586	150	52	11	0	DNAH7	149	35	26	11	0	KAT14	418	67	20	11	0	PCBP4	404	22	6	10	1	SNRPA	547	22	6	10	1
APOM	134	16	26	11	0	DNAH8	150	35	26	11	0	KAT2A	59	15	9	11	0	PCCA	636	59	9	11	0	SNRPA1	299	69	9	11	0
APOO	125	73	47	11	0	DNAH9	150	35	26	11	0	KAT2B	59	15	9	11	0	PCCB	413	59	9	11	0	SNRPB	546	55	28	5	1
APOOL	125	73	47	11	0	DNAI1	328	79	26	11	0	KAT5	59	15	9	11	0	PCF11	129	55	28	5	1	SNRPB2	547	22	6	10	1
APP	206	106	9	11	0	DNAI2	328	79	26	11	0	KAT6A	59	15	9	11	0	PCGF1	405	59	9	11	0	SNRPC	548	55	28	5	1
APPBP2	196	69	9	11	0	DNAJB11	362	15	9	11	0	KAT6B	59	15	9	11	0	PCGF2	405	59	9	11	0	SNRPD1	302	65	19	5	1
AQP4	274	58	48	5	1	DNAJC10	151	32	9	11	0	KAT7	59	15	9	11	0	PCGF3	405	59	9	11	0	SNRPD2	302	65	19	5	1
AQR	626	82	26	11	0	DNAJC11	152	118	9	11	0	KAT8	59	15	9	11	0	PCGF5	405	59	9	11	0	SNRPD3	302	65	19	5	1
AR	385	22	6	10	1	DNAJC15	154	118	9	11	0	KATNA1	453	59	9	11	0	PCGF6	405	59	9	11	0	SNRPE	302	65	19	5	1
ARCN1	14	5	2	10	1	DNAJC17	153	118	9	11	0	KATNAL1	453	59	9	11	0	PCID2	454	99	9	11	0	SNRPF	302	65	19	5	1
ARF1	462	49	14	5	1	DNAJC19	154	118	9	11	0	KATNAL2	453	59	9	11	0	PCM1	639	15	9	11	0	SNRPG	302	65	19	5	1
ARFGEF1	567	83	26	11	0	DNAJC28	157	36	20	11	0	KATNB1	422	146	50	6	1	PCNA	463	102	26	11	0	SNRPN	546	55	28	5	1
ARHGAP29	49	15	9	11	0	DNAJC9	153	118	9	11	0	KATNBL1	131	150	52	11	0	PCP4	406	87	20	11	0	SNTA1	549	15	9	11	0
ARHGAP33	473	15	9	11	0	DNAL1	299	69	9	11	0	KBTBD6	168	59	9	11	0	PCSK9	268	69	9	11	0	SNTB1	549	15	9	11	0
ARHGEF2	49	15	9	11	0	DNAL4	612	16	26	11	0	KBTBD7	168	59	9	11	0	PCYOX1	105	26	26	11	0	SNTB2	549	15	9	11	0
ARHGEF7	49	15	9	11	0	DNALI1	321	16	26	11	0	KBTBD8	168	59	9	11	0	PDCD10	457	16	26	11	0	SNTG1	549	15	9	11	0
ARID1A	48	16	26	11	0	DNHD1	149	35	26	11	0	KCNA1	271	59	9	11	0	PDCD11	652	144	9	11	0	SNTG2	549	15	9	11	0
ARID1B	48	16	26	11	0	DNM1	315	59	9	11	0	KCNA2	271	59	9	11	0	PDCD6	142	57	9	11	0	SNU13	501	71	21	11	0
ARID2	21	83	26	11	0	DNM1L	315	59	9	11	0	KCNA3	271	59	9	11	0	PDCD7	339	73	47	11	0	SNUPN	157	36	20	11	0
ARID4A	22	36	20	11	0	DNM2	315	59	9	11	0	KCNA4	271	59	9	11	0	PDCL3	308	32	9	11	0	SNW1	534	79	26	11	0
ARID4B	22	36	20	11	0	DNM3	315	59	9	11	0	KCNA5	271	59	9	11	0	PDE3B	539	64	9	11	0	SNX1	551	59	9	11	0
ARID5A	11	4	1	11	0	DNMT3A	59	15	9	11	0	KCNA6	271	59	9	11	0	PDE4B	271	59	9	11	0	SNX12	550	128	9	11	0
ARIH1	23	9	9	11	0	DNMT3B	59	15	9	11	0	KCNA7	271	59	9	11	0	PDE4D	271	59	9	11	0	SNX17	552	129	9	11	0
ARIH2	23	9	9	11	0	DNTTIP1	313	15	9	11	0	KCNAB1	272	59	9	11	0	PDGFA	407	106	9	11	0	SNX2	551	59	9	11	0
ARL6	513	119	9	11	0	DOCK1	567	83	26	11	0	KCNAB2	272	59	9	11	0	PDGFB	407	106	9	11	0	SNX3	550	128	9	11	0
ARL6IP1	261	7	9	11	0	DONSON	155	150	52	11	0	KCNAB3	272	59	9	11	0	PDGFRA	322	69	9	11	0	SNX31	552	129	9	11	0
ARMC8	299	69	9	11	0	DOT1L	11	4	1	11	0	KCNB1	271	59	9	11	0	PDGFRB	322	69	9	11	0	SNX4	551	59	9	11	0
ARNT	24	10	9	11	0	DPF1	59	15	9	11	0	KCNC1	271	59	9	11	0	PDHA1	408	59	9	11	0	SNX5	551	59	9	11	0
ARNT2	24	10	9	11	0	DPF2	59	15	9	11	0	KCNC2	271	59	9	11	0	PDHA2	408	59	9	11	0	SNX6	551	59	9	11	0
ARNTL	24	10	9	11	0	DPF3	59	15	9	11	0	KCNC3	271	59	9	11	0	PDHB	248	59	9	11	0	SOCS1	552	129	9	11	0
ARNTL2	379	10	9	11	0	DPH1	248	59	9	11	0	KCNC4	271	59	9	11	0	PDHX	146	34	9	11	0	SOCS2	552	129	9	11	0
ARPC1A	25	133	50	6	1	DPH2	502	79	26	11	0	KCND1	271	59	9	11	0	PDIA3	151	32	9	11	0	SOCS3	552	129	9	11	0
ARPC1B	25	133	50	6	1	DPM1	636	59	9	11	0	KCND2	271	59	9	11	0	PDIA6	151	32	9	11	0	SOCS4	553	130	20	11	0
ARPC2	331	75	34	11	0	DPM2	398	62	18	11	0	KCND3	271	59	9	11	0	PDK1	123	147	49	7	1	SOCS5	553	130	20	11	0
ARPC3	331	75	34	11	0	DPM3	398	62	18	11	0	KCNE1	233	75	34	11	0	PDK2	123	147	49	7	1	SOCS6	553	130	20	11	0
ARPC4	331	75	34	11	0	DPP10	268	69	9	11	0	KCNE2	233	75	34	11	0	PDLIM1	292	15	9	11	0	SOCS7	374	13	3	11	0
ARPC5	26	79	26	11	0	DPP6	268	69	9	11	0	KCNF1	276	59	9	11	0	PDLIM2	292	15	9	11	0	SOD1	554	131	45	12	1
ARPC5L	26	79	26	11	0	DPY30	157	36	20	11	0	KCNG1	276	59	9	11	0	PDLIM3	292	15	9	11	0	SORBS1	271	59	9	11	0
ARVCF	421	15	9	11	0	DPYSL3	413	59	9	11	0	KCNG2	276	59	9	11	0	PDLIM4	292	15	9	11	0	SOS1	49	15	9	11	0
ASB1	27	69	9	11	0	DPYSL5	413	59	9	11	0	KCNG3	276	59	9	11	0	PDLIM5	292	15	9	11	0	SOST	407	106	9	11	0
ASB12	27	69	9	11	0	DQX1	284	59	9	11	0	KCNH1	271	59	9	11	0	PDLIM7	292	15	9	11	0	SOX12	555	4	1	11	0
ASB2	27	69	9	11	0	DR1	579	36	20	11	0	KCNH2	271	59	9	11	0	PDP1	438	59	9	11	0	SOX15	555	4	1	11	0
ASB4	27	69	9	11	0	DRAP1	579	36	20	11	0	KCNH4	271	59	9	11	0	PDPR	248	59	9	11	0	SOX17	555	4	1	11	0
ASB6	27	69	9	11	0	DRC1	60	141	26	11	0	KCNIP1	273	57	9	11	0	PDRG1	412	16	26	11	0	SOX18	555	4	1	11	0
ASCC1	253	64	9	11	0	DRG1	228	59	9	11	0	KCNIP2	273	57	9	11	0	PDSS1	419	59	9	11	0	SOX2	555	4	1	11	0
ASCC2	627	82	26	11	0	DROSHA	639	15	9	11	0	KCNIP3	273	57	9	11	0	PDSS2	419	59	9	11	0	SOX4	555	4	1	11	0
ASCC3	28	59	9	11	0	DSCC1	229	16	26	11	0	KCNIP4	273	57	9	11	0	PDZD11	297	59	9	11	0	SOX8	555	4	1	11	0
ASCL1	29	11	1	11	0	DSN1	155	150	52	11	0	KCNJ11	274	58	48	5	1	PDZD7	403	15	9	11	0	SOX9	555	4	1	11	0
ASCL2	583	11	1	11	0	DTL	138	31	50	6	1	KCNJ14	274	58	48	5	1	PEA15	418	67	20	11	0	SP1	556	10	9	11	0
ASCL4	583	11	1	11	0	DTNA	23	9	9	11	0	KCNJ16	274	58	48	5	1	PECAM1	409	37	9	11	0	SP3	556	10	9	11	0
ASCL5	583	11	1	11	0	DTNBP1	457	16	26	11	0	KCNJ2	274	58	48	5	1	PEF1	326	57	9	11	0	SPAG1	362	15	9	11	0
ASF1A	439	15	9	11	0	DTX3L	481	64	9	11	0	KCNJ3	274	58	48	5	1	PELP1	620	59	9	11	0	SPAG4	526	17	9	11	0

ASF1B	439	15	9	11	0	DUOX1	156	69	9	11	0	KCNJ4	274	58	48	5	1	PES1	605	79	26	11	0	SPATS2L	539	64	9	11	0
ASH2L	59	15	9	11	0	DUOX2	156	69	9	11	0	KCNJ5	274	58	48	5	1	PEX11A	410	16	26	11	0	SPC24	233	75	34	11	0
ASXL1	30	150	52	11	0	DUSP22	420	59	9	11	0	KCNJ6	274	58	48	5	1	PEX11B	410	16	26	11	0	SPC25	557	16	26	11	0
ASXL2	30	150	52	11	0	DVL1	403	15	9	11	0	KCNJ8	274	58	48	5	1	PEX11G	410	16	26	11	0	SPCS1	261	7	9	11	0
ASXL3	33	150	52	11	0	DYDC2	131	150	52	11	0	KCNK1	271	59	9	11	0	PEX12	627	82	26	11	0	SPCS2	457	16	26	11	0
ATAD5	58	150	52	11	0	DYNC1H1	150	35	26	11	0	KCNK2	271	59	9	11	0	PEX13	46	14	20	11	0	SPCS3	143	32	9	11	0
ATF1	556	10	9	11	0	DYNC1I1	328	79	26	11	0	KCNK4	271	59	9	11	0	PEX14	411	83	26	11	0	SPDL1	118	64	9	11	0
ATF2	200	12	9	11	0	DYNC1I2	328	79	26	11	0	KCNK6	271	59	9	11	0	PEX19	605	79	26	11	0	SPECC1	558	64	9	11	0
ATF3	31	12	9	11	0	DYNC1LI1	509	9	9	11	0	KCNMA1	271	59	9	11	0	PEX2	627	82	26	11	0	SPECC1L	558	64	9	11	0
ATF4	375	12	9	11	0	DYNC1LI2	509	9	9	11	0	KCNMB1	275	106	9	11	0	PEX3	46	14	20	11	0	SPEN	129	55	28	5	1
ATF5	32	13	3	11	0	DYNC2H1	150	35	26	11	0	KCNMB2	275	106	9	11	0	PEX5	619	99	9	11	0	SPG7	453	59	9	11	0
ATF6	375	12	9	11	0	DYNC2LI1	509	9	9	11	0	KCNMB3	275	106	9	11	0	PEX5L	619	99	9	11	0	SPHK2	473	15	9	11	0
ATF6B	375	12	9	11	0	DYNLL1	612	16	26	11	0	KCNMB4	275	106	9	11	0	PFDN1	412	16	26	11	0	SPI1	373	12	9	11	0
ATF7	200	12	9	11	0	DYNLL2	612	16	26	11	0	KCNN1	271	59	9	11	0	PFDN2	412	16	26	11	0	SPOP	168	59	9	11	0
ATF7IP	33	150	52	11	0	DYNLRB1	15	6	2	10	1	KCNN4	276	59	9	11	0	PFDN4	412	16	26	11	0	SPOPL	168	59	9	11	0
ATF7IP2	33	150	52	11	0	DYNLRB2	15	6	2	10	1	KCNQ1	271	59	9	11	0	PFDN5	412	16	26	11	0	SPSB1	559	31	50	6	1
ATG101	307	75	34	11	0	DYNLT1	612	16	26	11	0	KCNQ2	271	59	9	11	0	PFDN6	412	16	26	11	0	SPSB2	559	31	50	6	1
ATG12	405	59	9	11	0	DYNLT3	612	16	26	11	0	KCNQ3	271	59	9	11	0	PFKFB1	170	59	9	11	0	SPSB3	559	31	50	6	1
ATG13	319	150	52	11	0	DYRK1A	96	59	9	11	0	KCNQ4	271	59	9	11	0	PFKL	413	59	9	11	0	SPSB4	559	31	50	6	1
ATG14	34	16	26	11	0	DYRK2	96	59	9	11	0	KCNQ5	271	59	9	11	0	PFKM	413	59	9	11	0	SPTA1	560	59	9	11	0
ATG16L1	450	146	50	6	1	DZIP1	49	15	9	11	0	KCNS1	276	59	9	11	0	PFKP	413	59	9	11	0	SPTAN1	560	59	9	11	0
ATG16L2	35	31	50	6	1	E2F1	588	15	9	11	0	KCNS2	276	59	9	11	0	PGGT1B	123	147	49	7	1	SPTB	560	59	9	11	0
ATG3	157	36	20	11	0	E2F2	588	15	9	11	0	KCNS3	276	59	9	11	0	PGM5	413	59	9	11	0	SPTBN1	560	59	9	11	0
ATG5	36	20	26	11	0	E2F3	588	15	9	11	0	KCNV2	276	59	9	11	0	PHAX	414	150	52	11	0	SPTBN2	560	59	9	11	0
ATM	263	56	26	11	0	E2F4	588	15	9	11	0	KCTD10	277	15	9	11	0	PHB2	42	99	9	11	0	SPTBN4	560	59	9	11	0
ATP10A	37	69	9	11	0	E2F5	588	15	9	11	0	KCTD13	277	15	9	11	0	PHC1	415	10	9	11	0	SPTBN5	560	59	9	11	0
ATP10B	37	69	9	11	0	E2F6	588	15	9	11	0	KCTD16	277	15	9	11	0	PHC2	415	10	9	11	0	SPTLC1	561	59	9	11	0
ATP10D	37	69	9	11	0	E2F7	49	15	9	11	0	KCTD17	277	15	9	11	0	PHC3	415	10	9	11	0	SPTLC2	561	59	9	11	0
ATP11A	38	69	9	11	0	E2F8	49	15	9	11	0	KCTD2	277	15	9	11	0	PHF10	59	15	9	11	0	SPTLC3	561	59	9	11	0
ATP11B	38	69	9	11	0	EAF1	157	36	20	11	0	KCTD5	277	15	9	11	0	PHF12	59	15	9	11	0	SPTSSA	562	32	9	11	0
ATP11C	38	69	9	11	0	EAF2	157	36	20	11	0	KDM1A	23	9	9	11	0	PHF14	59	15	9	11	0	SPTSSB	562	32	9	11	0
ATP12A	39	22	6	10	1	EBI3	158	37	9	11	0	KDM1B	105	26	26	11	0	PHF20	59	15	9	11	0	SRA1	321	16	26	11	0
ATP1A1	39	22	6	10	1	EBNA1BP2	634	79	26	11	0	KDM2B	59	15	9	11	0	PHF20L1	59	15	9	11	0	SRCAP	59	15	9	11	0
ATP1A2	39	22	6	10	1	ECT2	59	15	9	11	0	KDM3A	49	15	9	11	0	PHF21A	200	12	9	11	0	SREBF1	563	15	9	11	0
ATP1A3	39	22	6	10	1	EDA	79	17	9	11	0	KDM3B	319	150	52	11	0	PHF5A	307	75	34	11	0	SREBF2	563	15	9	11	0
ATP1A4	39	22	6	10	1	EEA1	395	64	9	11	0	KDM5A	59	15	9	11	0	PHKA1	416	88	27	7	1	SREK1	566	132	0	11	0
ATP1B1	40	69	9	11	0	EED	642	79	26	11	0	KDM6A	93	59	9	11	0	PHKA2	416	88	27	7	1	SRFBP1	564	150	52	11	0
ATP1B2	40	69	9	11	0	EEF1A1	353	49	14	5	1	KDM6B	540	36	20	11	0	PHKB	416	88	27	7	1	SRP14	501	71	21	11	0
ATP1B3	40	69	9	11	0	EEF1A2	353	49	14	5	1	KDR	322	69	9	11	0	PHKG1	417	69	9	11	0	SRP19	501	71	21	11	0
ATP1B4	40	69	9	11	0	EEF1AKMT3	324	69	9	11	0	KEAP1	168	59	9	11	0	PHKG2	417	69	9	11	0	SRP54	315	59	9	11	0
ATP23	593	69	9	11	0	EEF1B2	107	27	8	2	1	KHDRBS1	278	55	28	5	1	PIAS4	59	15	9	11	0	SRP68	454	99	9	11	0
ATP2A1	39	22	6	10	1	EEF1D	107	27	8	2	1	KHDRBS3	278	55	28	5	1	PICALM	171	15	9	11	0	SRP72	619	99	9	11	0
ATP2B4	39	22	6	10	1	EEF1E1	112	32	9	11	0	KIDINS220	13	59	9	11	0	PIDD1	473	15	9	11	0	SRP9	418	67	20	11	0
ATP4B	40	69	9	11	0	EEF2	228	59	9	11	0	KIF11	280	59	9	11	0	PIGA	299	69	9	11	0	SRPRA	565	59	9	11	0
ATP6AP1	470	106	9	11	0	EEF2KMT	324	69	9	11	0	KIF12	284	59	9	11	0	PIGC	261	7	9	11	0	SRPRB	513	119	9	11	0
ATP6AP1L	470	106	9	11	0	EEFSEC	228	59	9	11	0	KIF13A	279	59	9	11	0	PIGH	299	69	9	11	0	SRRM1	566	132	0	11	0
ATP6AP2	470	106	9	11	0	EFCAB7	326	57	9	11	0	KIF13B	279	59	9	11	0	PIGK	226	59	9	11	0	SRRM2	566	132	0	11	0
ATP6V0A1	41	59	9	11	0	EFL1	228	59	9	11	0	KIF14	280	59	9	11	0	PIGM	261	7	9	11	0	SRRT	129	55	28	5	1
ATP6V0A2	41	59	9	11	0	EFTUD2	228	59	9	11	0	KIF15	280	59	9	11	0	PIGP	398	62	18	11	0	SRSF1	609	132	0	11	0
ATP6V0A4	41	59	9	11	0	EGFR	170	59	9	11	0	KIF16B	279	59	9	11	0	PIGQ	275	106	9	11	0	SRSF2	609	132	0	11	0
ATP6V0B	532	59	9	11	0	EHBP1	49	15	9	11	0	KIF17	281	59	9	11	0	PIGR	515	37	9	11	0	SS18	48	16	26	11	0
ATP6V0C	532	59	9	11	0	EHBP1L1	540	36	20	11	0	KIF18A	280	59	9	11	0	PIGS	143	32	9	11	0	SS18L1	48	16	26	11	0
ATP6V0D1	454	99	9	11	0	EID3	577	16	26	11	0	KIF18B	284	59	9	11	0	PIGT	143	32	9	11	0	SSB	544	132	0	11	0
ATP6V0D2	454	99	9	11	0	EIF1	432	115	41	10	1	KIF19	284	59	9	11	0	PIGU	261	7	9	11	0	SSBP3	99	22	6	10	1
ATP6V0E1	398	62	18	11	0	EIF1B	432	115	41	10	1	KIF1A	279	59	9	11	0	PIGV	299	69	9	11	0	SSPN	600	7	9	11	0
ATP6V0E2	398	62	18	11	0	EIF2A	25	133	50	6	1	KIF1B	279	59	9	11	0	PIH1D1	321	16	26	11	0	SSR4	134	16	26	11	0
ATP6V1A	315	59	9	11	0	EIF2AK4	253	64	9	11	0	KIF1C	279	59	9	11	0	PIH1D2	321	16	26	11	0	SSRP1	124	99	9	11	0
ATP6V1B1	315	59	9	11	0	EIF2B1	159	38	10	8	1	KIF20A	282	59	9	11	0	PIK3C2A	418	67	20	11	0	SSU72	541	59	9	11	0
ATP6V1B2	315	59	9	11	0	EIF2B2	159	38	10	8	1	KIF20B	282	59	9	11	0	PIK3C2B	418	67	20	11	0	SSX2IP	118	64	9	11	0
ATP6V1C1	119	141	26	11	0	EIF2B3	642	79	26	11	0	KIF21A	283	59	9	11	0	PIK3C2G	418	67	20	11	0	ST18	49	15	9	11	0
ATP6V1C2	119	141	26	11	0	EIF2B4	159	38	10	8	1	KIF21B	283	59	9	11	0	PIK3C3	419	59	9	11	0	STAC3	362	15	9	11	0
ATP6V1D	605	79	26	11	0	EIF2B5	642	79	26	11	0	KIF22	284	59	9	11	0	PIK3CA	419	59	9	11	0	STAG1	567	83	26	11	0
ATP6V1E1	42	99	9	11	0	EIF2S1	179	43	19	5	1	KIF23	280	59	9	11	0	PIK3CB	419	59	9	11	0	STAG2	567	83	26	11	0
ATP6V1E2	42	99	9	11	0	EIF2S2	157	36	20	11	0	KIF24	280	59	9	11	0	PIK3CD	419	59	9	11	0	STAG3	567	83	26	11	0
ATP6V1F	501	71	21	11	0	EIF2S3	228	59	9	11	0	KIF25	284	59	9	11	0	PIK3CG	419	59	9	11	0	STAM	568	59	9	11	0
ATP6V1G1	42	99	9	11	0	EIF3A	454	99	9	11	0	KIF26A	284	59	9	11	0	PIK3R1	420	59	9	11	0	STAM2	568	59	9	11	0
ATP6V1G2	42	99	9	11	0	EIF3B	25	133	50	6	1	KIF26B	284	59	9	11	0	PIK3R2	420	59	9	11	0	STAP1	552	129	9	11	0
ATP6V1G3	42	99	9	11	0	EIF3C	454	99	9	11	0	KIF27	280	59	9	11	0	PIK3R3	420	59	9	11	0	STAT1	569	83	26	11	0
ATP6V1H	649	147	49	7	1	EIF3CL	454	99	9	11	0	KIF2A	279	59	9	11	0	PIK3R4	360	59	9	11	0	STAT2	569	83	26	11	0
ATP8A1	38	69	9	11	0	EIF3D	157	36	20	11	0	KIF2C	281	59	9	11	0	PIK3R5	418	67	20	11	0	STAT3	569	83	26	11	0
ATP8A2	38	69	9	11	0	EIF3E	454	99	9	11	0	KIF3A	283	59	9	11	0	PIK3R6	418	67	20	11	0	STAT4	569	83	26	11	0
ATP8B1	38	69	9	11	0	EIF3F	165	39	9	11	0	KIF3B	283	59	9	11	0	PITX1	11	4	1	11	0	STAT5A	569	83	26	11	0
ATP8B2	38	69	9	11	0	EIF3G	544	132	0	11	0	KIF3C	283	59	9	11	0	PITX2	11	4	1	11	0	STAT5B	569	83	26	11	0
ATP8B3	37	69	9	11	0	EIF3H	165	39	9	11	0	KIF4A	280	59	9	11	0	PIWIL2	419	59	9	11	0	STAT6	569	83	26	11	0
ATP8B4	38	69	9	11	0	EIF3I	641	146	50	6	1	KIF4B	280	59	9	11	0	PKD1	90	37	9	11	0	STAU1	448	59	9	11	0
ATR	263	56	26	11	0	EIF3J	605	79	26	11	0	KIF5A	283	59	9	11	0	PKD1L1	233	75	34	11	0	STAU2	448	59	9	11	0
ATRIP	389	82	26	11	0	EIF3K	454	99	9	11	0	KIF5B	283	59	9	11	0	PKD1L3	418	67	20	11	0	STIL	319	150	52	11	0
ATXN2	524	13	3	11	0	EIF3L	160	99	9	11	0	KIF5C	283	59	9	11	0	PKD2	517	59	9	11	0	STIP1	527	122	33	7	1
ATXN7	30	150	52	11	0	EIF3M	454	99	9	11	0	KIF6	280	59	9	11	0	PKD2L1	517	59	9	11	0	STK11	104	25	9	11	0
ATXN7L3	639	15	9	11	0	EIF4A1	161	22	6	10	1	KIF7	284	59	9	11	0	PKN1	170	59	9	11	0	STK3	283	59	9	11	0

AUP1	642	79	26	11	0	EIF4A2	161	22	6	10	1	KIF9	280	59	9	11	0	PKN2	170	59	9	11	0	STK4	283	59	9	11	0
AURKA	43	59	9	11	0	EIF4A3	161	22	6	10	1	KIFAP3	443	147	49	7	1	PKNOX1	528	22	6	10	1	STN1	253	64	9	11	0
AURKAIP1	339	73	47	11	0	EIF4B	129	55	28	5	1	KIFC1	284	59	9	11	0	PKP1	421	15	9	11	0	STRA6	275	106	9	11	0
AURKB	43	59	9	11	0	EIF4E	162	79	26	11	0	KIFC2	284	59	9	11	0	PKP3	421	15	9	11	0	STRADA	491	69	9	11	0
AURKC	43	59	9	11	0	EIF4E2	162	79	26	11	0	KIFC3	284	59	9	11	0	PLA2G7	299	69	9	11	0	STRADB	491	69	9	11	0
AXIN1	49	15	9	11	0	EIF4E3	162	79	26	11	0	KIN	605	79	26	11	0	PLA2R1	181	17	9	11	0	STRAP	641	146	50	6	1
AXIN2	49	15	9	11	0	EIF4EBP3	313	15	9	11	0	KIT	322	69	9	11	0	PLAT	181	17	9	11	0	STRN	570	31	50	6	1
AXL	170	59	9	11	0	EIF4G1	639	15	9	11	0	KLC1	283	59	9	11	0	PLAU	181	17	9	11	0	STRN3	570	31	50	6	1
AZI2	118	64	9	11	0	EIF4G2	567	83	26	11	0	KLC2	283	59	9	11	0	PLAUR	304	80	24	11	0	STRN4	570	31	50	6	1
B2M	238	59	9	11	0	EIF4G3	639	15	9	11	0	KLC3	283	59	9	11	0	PLCB1	170	59	9	11	0	STT3A	13	59	9	11	0
B3GALT4	221	51	9	11	0	EIF4H	129	55	28	5	1	KLC4	283	59	9	11	0	PLCB2	170	59	9	11	0	STT3B	13	59	9	11	0
B4GALT1	299	69	9	11	0	EIF5A	504	71	21	11	0	KLF4	374	13	3	11	0	PLCB3	170	59	9	11	0	STUB1	527	122	33	7	1
B9D1	419	59	9	11	0	EIF5A2	504	71	21	11	0	KLF5	374	13	3	11	0	PLCG1	170	59	9	11	0	STX10	572	16	26	11	0
B9D2	419	59	9	11	0	EIF5AL1	504	71	21	11	0	KLHDC1	285	69	9	11	0	PLEKHA2	49	15	9	11	0	STX11	571	16	26	11	0
BABAM1	265	83	26	11	0	EIF6	501	71	21	11	0	KLHDC10	285	69	9	11	0	PLEKHA7	49	15	9	11	0	STX12	571	16	26	11	0
BACH1	375	12	9	11	0	ELANE	181	17	9	11	0	KLHDC2	285	69	9	11	0	PLK1	417	69	9	11	0	STX16	571	16	26	11	0
BACH2	375	12	9	11	0	ELAVL1	163	22	6	10	1	KLHDC3	285	69	9	11	0	PLK4	49	15	9	11	0	STX17	572	16	26	11	0
BAD	44	68	24	11	0	ELAVL2	163	22	6	10	1	KLHL12	168	59	9	11	0	PLN	368	73	47	11	0	STX18	571	16	26	11	0
BAG2	233	75	34	11	0	ELAVL3	163	22	6	10	1	KLHL13	168	59	9	11	0	PLOD1	491	69	9	11	0	STX19	571	16	26	11	0
BAG3	321	16	26	11	0	ELL	164	13	3	11	0	KLHL2	168	59	9	11	0	PLP1	600	7	9	11	0	STX1A	571	16	26	11	0
BAG5	457	16	26	11	0	ELL2	164	13	3	11	0	KLHL20	168	59	9	11	0	PLRG1	422	146	50	6	1	STX1B	571	16	26	11	0
BAG6	46	14	20	11	0	ELL3	117	150	52	11	0	KLHL21	168	59	9	11	0	PLTP	268	69	9	11	0	STX2	571	16	26	11	0
BAHD1	540	36	20	11	0	ELMO1	269	83	26	11	0	KLHL22	168	59	9	11	0	PLXDC1	423	106	9	11	0	STX3	571	16	26	11	0
BAIAP2L2	49	15	9	11	0	ELOA	524	13	3	11	0	KLHL24	168	59	9	11	0	PLXNA1	424	69	9	11	0	STX4	571	16	26	11	0
BAK1	46	14	20	11	0	ELOB	405	59	9	11	0	KLHL25	168	59	9	11	0	PLXNA2	424	69	9	11	0	STX5	571	16	26	11	0
BAP1	49	15	9	11	0	ELOC	168	59	9	11	0	KLHL3	168	59	9	11	0	PLXNA3	424	69	9	11	0	STX6	572	16	26	11	0
BARD1	49	15	9	11	0	ELOF1	507	117	43	10	1	KLHL40	168	59	9	11	0	PLXNA4	424	69	9	11	0	STX7	571	16	26	11	0
BARX2	11	4	1	11	0	ELOVL6	378	69	9	11	0	KLHL41	168	59	9	11	0	PLXNB1	424	69	9	11	0	STX8	572	16	26	11	0
BATF	31	12	9	11	0	ELP2	25	133	50	6	1	KLHL42	168	59	9	11	0	PLXNB2	424	69	9	11	0	STXBP1	454	99	9	11	0
BATF2	201	123	3	11	0	ELP3	636	59	9	11	0	KLHL7	168	59	9	11	0	PLXNB3	424	69	9	11	0	STXBP5	249	99	9	11	0
BATF3	31	12	9	11	0	ELP4	466	69	9	11	0	KLHL8	168	59	9	11	0	PLXNC1	424	69	9	11	0	STXBP6	460	69	9	11	0
BAX	46	14	20	11	0	ELP5	466	69	9	11	0	KLHL9	168	59	9	11	0	PLXND1	424	69	9	11	0	SUB1	157	36	20	11	0
BAZ1A	157	36	20	11	0	ELP6	466	69	9	11	0	KLK3	181	17	9	11	0	PMAIP1	331	75	34	11	0	SUCLA2	248	59	9	11	0
BAZ1B	157	36	20	11	0	EMC1	424	69	9	11	0	KLK8	181	17	9	11	0	PMF1	457	16	26	11	0	SUCLG1	248	59	9	11	0
BAZ2A	30	150	52	11	0	EMC10	470	106	9	11	0	KLRC1	286	60	9	11	0	PMPCA	632	118	9	11	0	SUCLG2	248	59	9	11	0
BBIP1	331	75	34	11	0	EMC2	619	99	9	11	0	KLRC2	286	60	9	11	0	PMPCB	632	118	9	11	0	SUDS3	60	141	26	11	0
BBS1	45	83	26	11	0	EMC3	398	62	18	11	0	KLRC3	286	60	9	11	0	PMS1	243	59	9	11	0	SUFU	307	75	34	11	0
BBS2	45	83	26	11	0	EMC4	398	62	18	11	0	KLRD1	286	60	9	11	0	PMS2	70	22	6	10	1	SUGP1	573	15	9	11	0
BBS4	93	59	9	11	0	EMC6	114	7	9	11	0	KMT2A	540	36	20	11	0	PNN	425	36	20	11	0	SUGT1	124	99	9	11	0
BBS5	124	99	9	11	0	EMC7	143	32	9	11	0	KMT2B	540	36	20	11	0	PNPT1	180	44	12	4	1	SUMO1	405	59	9	11	0
BBS7	45	83	26	11	0	EMC8	165	39	9	11	0	KMT2C	59	15	9	11	0	POLA1	429	90	9	11	0	SUMO4	405	59	9	11	0
BBS9	45	83	26	11	0	EMC9	165	39	9	11	0	KMT2D	198	13	3	11	0	POLA2	133	30	15	5	1	SUN1	526	17	9	11	0
BCAS2	457	16	26	11	0	EME1	586	150	52	11	0	KMT2E	59	15	9	11	0	POLB	426	89	28	5	1	SUN2	526	17	9	11	0
BCAS4	54	16	26	11	0	EME2	166	15	9	11	0	KNDC1	136	150	52	11	0	POLD1	429	90	9	11	0	SUPT16H	124	99	9	11	0
BCCIP	320	56	26	11	0	EMG1	167	40	11	3	1	KNTC1	481	64	9	11	0	POLD2	133	30	15	5	1	SUPT20H	639	15	9	11	0
BCKDHA	408	59	9	11	0	EMID1	79	17	9	11	0	KPNA1	287	59	9	11	0	POLD3	427	135	26	11	0	SUPT20HL1	574	150	52	11	0
BCKDHB	248	59	9	11	0	EMILIN1	526	17	9	11	0	KPNA2	287	59	9	11	0	POLD4	428	79	26	11	0	SUPT3H	579	36	20	11	0
BCKDK	123	147	49	7	1	EMILIN2	526	17	9	11	0	KPNA3	287	59	9	11	0	POLE	429	90	9	11	0	SUPT4H1	501	71	21	11	0
BCL10	72	19	20	11	0	EML1	642	79	26	11	0	KPNA4	287	59	9	11	0	POLE2	133	30	15	5	1	SUPT5H	199	125	49	7	1
BCL11A	11	4	1	11	0	EML2	642	79	26	11	0	KPNA5	287	59	9	11	0	POLE3	579	36	20	11	0	SUPT6H	157	36	20	11	0
BCL11B	11	4	1	11	0	ENC1	168	59	9	11	0	KPNA6	287	59	9	11	0	POLE4	100	23	7	1	1	SUPT7L	577	16	26	11	0
BCL2	46	14	20	11	0	ENG	252	17	9	11	0	KPNA7	287	59	9	11	0	POLG	429	90	9	11	0	SUPV3L1	28	59	9	11	0
BCL2L1	46	14	20	11	0	ENO1	169	97	41	10	1	KPNB1	266	147	49	7	1	POLG2	341	71	21	11	0	SUV39H1	362	15	9	11	0
BCL2L11	44	68	24	11	0	ENO2	169	97	41	10	1	KPTN	45	83	26	11	0	POLR1A	430	91	29	8	1	SUZ12	265	83	26	11	0
BCL2L2	46	14	20	11	0	ENO3	169	97	41	10	1	KRI1	635	79	26	11	0	POLR1B	430	91	29	8	1	SWI5	191	64	9	11	0
BCL3	59	15	9	11	0	ENO4	123	147	49	7	1	KRIT1	418	67	20	11	0	POLR1C	179	43	19	5	1	SWSAP1	466	69	9	11	0
BCL7A	47	150	52	11	0	ENTHD1	117	150	52	11	0	KRR1	199	125	49	7	1	POLR1D	431	36	20	11	0	SYF2	534	79	26	11	0
BCL7B	47	150	52	11	0	ENY2	607	79	26	11	0	KRTCAP2	398	62	18	11	0	POLR1E	320	56	26	11	0	SYK	491	69	9	11	0
BCL7C	47	150	52	11	0	EP300	415	10	9	11	0	KSR1	49	15	9	11	0	POLR2A	430	91	29	8	1	SYMPK	567	83	26	11	0
BCL9	48	16	26	11	0	EP400	48	16	26	11	0	KXD1	542	16	26	11	0	POLR2B	430	91	29	8	1	SYN2	568	59	9	11	0
BCL9L	48	16	26	11	0	EPAS1	379	10	9	11	0	L1TD1	288	61	20	11	0	POLR2C	179	43	19	5	1	SYNCRIP	175	59	9	11	0
BCLAF1	592	36	20	11	0	EPB41	271	59	9	11	0	L3MBTL1	59	15	9	11	0	POLR2D	157	36	20	11	0	SYNE1	560	59	9	11	0
BCOR	386	150	52	11	0	EPB41L2	271	59	9	11	0	LAGE3	46	14	20	11	0	POLR2E	179	43	19	5	1	SYNE2	560	59	9	11	0
BCR	49	15	9	11	0	EPC1	639	15	9	11	0	LAMA1	289	69	9	11	0	POLR2F	157	36	20	11	0	SYNE3	284	59	9	11	0
BCS1L	453	59	9	11	0	EPC2	639	15	9	11	0	LAMA2	289	69	9	11	0	POLR2G	179	43	19	5	1	SYNE4	233	75	34	11	0
BDP1	33	150	52	11	0	EPHA1	170	59	9	11	0	LAMA3	289	69	9	11	0	POLR2H	307	75	34	11	0	SYNJ1	575	59	9	11	0
BECN1	34	16	26	11	0	EPHA10	170	59	9	11	0	LAMA5	289	69	9	11	0	POLR2I	23	9	9	11	0	SYNPO2	639	15	9	11	0
BEST1	212	58	48	5	1	EPHA2	170	59	9	11	0	LAMB1	289	69	9	11	0	POLR2J	432	115	41	10	1	SYNRG	171	15	9	11	0
BEST2	212	58	48	5	1	EPHA3	170	59	9	11	0	LAMB2	289	69	9	11	0	POLR2J2	432	115	41	10	1	SYVN1	23	9	9	11	0
BEST3	212	58	48	5	1	EPHA4	170	59	9	11	0	LAMB3	268	69	9	11	0	POLR2J3	432	115	41	10	1	SZT2	576	56	26	11	0
BEST4	212	58	48	5	1	EPHA5	170	59	9	11	0	LAMC1	289	69	9	11	0	POLR2K	507	117	43	10	1	TAB1	438	59	9	11	0
BET1	572	16	26	11	0	EPHA6	170	59	9	11	0	LAMC2	268	69	9	11	0	POLR2L	433	92	30	9	1	TADA1	577	16	26	11	0
BET1L	572	16	26	11	0	EPHA7	170	59	9	11	0	LAMTOR1	349	75	34	11	0	POLR2M	319	150	52	11	0	TADA2A	344	70	20	11	0
BEX1	331	75	34	11	0	EPHA8	170	59	9	11	0	LAMTOR2	290	16	26	11	0	POLR3A	430	91	29	8	1	TADA2B	344	70	20	11	0
BICD2	240	16	26	11	0	EPHB1	170	59	9	11	0	LAMTOR3	290	16	26	11	0	POLR3B	430	91	29	8	1	TADA3	434	93	20	11	0
BIK	50	75	34	11	0	EPHB2	170	59	9	11	0	LAMTOR4	290	16	26	11	0	POLR3C	124	99	9	11	0	TAF1	639	15	9	11	0
BIN1	123	147	49	7	1	EPHB3	170	59	9	11	0	LAMTOR5	290	16	26	11	0	POLR3D	157	36	20	11	0	TAF10	512	15	9	11	0
BIRC2	51	59	9	11	0	EPHB4	170	59	9	11	0	LAPTM5	291	62	18	11	0	POLR3E	229	16	26	11	0	TAF11	579	36	20	11	0
BIRC3	51	59	9	11	0	EPHB6	170	59	9	11	0	LARP6	418	67	20	11	0	POLR3F	124	99	9	11	0	TAF12	579	36	20	11	0
BIRC5	51	59	9	11	0	EPN1	171	15	9	11	0	LARP7	418	67	20	11	0	POLR3G	434	93	20	11	0	TAF13	579	36	20	11	0

BLM	280	59	9	11	0	EPN2	171	15	9	11	0	LAS1L	481	64	9	11	0	POLR3GL	434	93	20	11	0	TAF1A	481	64	9	11	0
BLOC1S1	542	16	26	11	0	EPN3	171	15	9	11	0	LAT	296	123	3	11	0	POLR3H	179	43	19	5	1	TAF1B	82	13	3	11	0
BLOC1S2	52	16	26	11	0	EPPIN	64	17	9	11	0	LBH	331	75	34	11	0	POLR3K	23	9	9	11	0	TAF1C	578	141	26	11	0
BLOC1S3	53	56	26	11	0	EPS15	49	15	9	11	0	LCAT	268	69	9	11	0	POLRMT	429	90	9	11	0	TAF1D	331	75	34	11	0
BLOC1S4	54	16	26	11	0	EPS15L1	49	15	9	11	0	LCP2	553	130	20	11	0	POM121	392	83	26	11	0	TAF1L	639	15	9	11	0
BLOC1S5	457	16	26	11	0	EPS8	473	15	9	11	0	LDB1	265	83	26	11	0	POM121C	392	83	26	11	0	TAF2	360	59	9	11	0
BLOC1S6	542	16	26	11	0	EPS8L1	473	15	9	11	0	LDB2	265	83	26	11	0	PON1	143	32	9	11	0	TAF3	524	13	3	11	0
BMF	55	46	34	11	0	EPS8L2	473	15	9	11	0	LDB3	292	15	9	11	0	POP1	435	141	26	11	0	TAF4	415	10	9	11	0
BMI1	405	59	9	11	0	ERBB2	170	59	9	11	0	LDHA	293	59	9	11	0	POP4	501	71	21	11	0	TAF4B	415	10	9	11	0
BMP2	407	106	9	11	0	ERBB3	170	59	9	11	0	LDHAL6B	293	59	9	11	0	POP5	162	79	26	11	0	TAF5	422	146	50	6	1
BMPR1A	6	1	9	11	0	ERBB4	170	59	9	11	0	LDHB	293	59	9	11	0	POP7	502	79	26	11	0	TAF5L	422	146	50	6	1
BMPR1B	6	1	9	11	0	ERC1	240	16	26	11	0	LDHC	293	59	9	11	0	PORCN	299	69	9	11	0	TAF6	11	4	1	11	0
BMPR2	49	15	9	11	0	ERCC1	586	150	52	11	0	LDLR	298	64	9	11	0	POT1	586	150	52	11	0	TAF6L	11	4	1	11	0
BMS1	645	79	26	11	0	ERCC2	172	59	9	11	0	LEF1	555	4	1	11	0	POU2AF1	319	150	52	11	0	TAF7	157	36	20	11	0
BNIP1	572	16	26	11	0	ERCC3	173	59	9	11	0	LENG8	539	64	9	11	0	POU2F1	556	10	9	11	0	TAF7L	157	36	20	11	0
BOD1	59	15	9	11	0	ERCC4	174	41	9	11	0	LEO1	639	15	9	11	0	POU3F1	11	4	1	11	0	TAF8	579	36	20	11	0
BOD1L1	33	150	52	11	0	ERCC5	49	15	9	11	0	LEPR	208	37	9	11	0	POU4F1	11	4	1	11	0	TAF9	579	36	20	11	0
BOD1L2	59	15	9	11	0	ERCC6	418	67	20	11	0	LGALS1	294	63	26	11	0	PPAN	645	79	26	11	0	TAF9B	579	36	20	11	0
BOLA1	56	71	21	11	0	ERCC6L2	280	59	9	11	0	LGALS2	294	63	26	11	0	PPARG	385	22	6	10	1	TAL1	373	12	9	11	0
BOLA2	56	71	21	11	0	ERCC8	140	79	26	11	0	LGALS3	294	63	26	11	0	PPARGC1B	30	150	52	11	0	TANK	580	150	52	11	0
BOLA3	56	71	21	11	0	ERG	373	12	9	11	0	LHB	295	106	9	11	0	PPCDC	248	59	9	11	0	TAOK2	421	15	9	11	0
BOP1	570	31	50	6	1	ERGIC2	518	69	9	11	0	LHCGR	246	51	9	11	0	PPFIA1	236	59	9	11	0	TAP1	581	59	9	11	0
BORCS5	542	16	26	11	0	ERGIC3	518	69	9	11	0	LHX3	528	22	6	10	1	PPIA	39	22	6	10	1	TAP2	581	59	9	11	0
BORCS6	440	56	26	11	0	ERH	331	75	34	11	0	LIFR	255	37	9	11	0	PPIB	436	94	31	12	1	TAPBP	90	37	9	11	0
BORCS7	233	75	34	11	0	ERI1	642	79	26	11	0	LIG3	59	15	9	11	0	PPIE	436	94	31	12	1	TARBP2	448	59	9	11	0
BORCS8	233	75	34	11	0	ERLIN1	42	99	9	11	0	LIG4	157	36	20	11	0	PPIF	436	94	31	12	1	TAS1R3	322	69	9	11	0
BPTF	157	36	20	11	0	ERLIN2	42	99	9	11	0	LILRA4	237	59	9	11	0	PPIH	436	94	31	12	1	TBC1D5	59	15	9	11	0
BRAP	57	59	9	11	0	ERN1	59	15	9	11	0	LIMD1	270	15	9	11	0	PPIL1	436	94	31	12	1	TBC1D7	299	69	9	11	0
BRCA1	58	150	52	11	0	ERN2	59	15	9	11	0	LIME1	296	123	3	11	0	PPIL2	437	95	15	5	1	TBK1	253	64	9	11	0
BRCA2	58	150	52	11	0	ESAM	515	37	9	11	0	LIN28A	59	15	9	11	0	PPIL3	436	94	31	12	1	TBKBP1	403	15	9	11	0
BRCC3	165	39	9	11	0	ESPN	292	15	9	11	0	LIN37	319	150	52	11	0	PPM1E	418	67	20	11	0	TBL1X	582	133	50	6	1
BRD1	59	15	9	11	0	ESR1	385	22	6	10	1	LIN52	331	75	34	11	0	PPM1F	438	59	9	11	0	TBL1XR1	582	133	50	6	1
BRD7	59	15	9	11	0	ESRP1	175	59	9	11	0	LIN54	415	10	9	11	0	PPP1CA	2	0	4	10	1	TBL3	642	79	26	11	0
BRD8	59	15	9	11	0	ESRP2	175	59	9	11	0	LIN7A	297	59	9	11	0	PPP1CB	2	0	4	10	1	TBP	573	15	9	11	0
BRD9	59	15	9	11	0	ETF1	124	99	9	11	0	LIN7B	297	59	9	11	0	PPP1CC	2	0	4	10	1	TBPL1	341	71	21	11	0
BRF1	59	15	9	11	0	ETFA	413	59	9	11	0	LIN7C	297	59	9	11	0	PPP1R10	129	55	28	5	1	TBX15	373	12	9	11	0
BRF2	82	13	3	11	0	ETFB	413	59	9	11	0	LIN9	157	36	20	11	0	PPP1R12A	49	15	9	11	0	TBX18	373	12	9	11	0
BRIP1	280	59	9	11	0	ETV3	373	12	9	11	0	LIPC	593	69	9	11	0	PPP1R15A	540	36	20	11	0	TBX2	11	4	1	11	0
BRK1	542	16	26	11	0	EVC	233	75	34	11	0	LMO2	292	15	9	11	0	PPP1R15B	117	150	52	11	0	TBX5	373	12	9	11	0
BRMS1	60	141	26	11	0	EVC2	233	75	34	11	0	LMO4	292	15	9	11	0	PPP1R3B	439	15	9	11	0	TCAP	345	75	34	11	0
BRMS1L	60	141	26	11	0	EXD1	311	67	20	11	0	LMO7	639	15	9	11	0	PPP1R3C	439	15	9	11	0	TCEA1	123	147	49	7	1
BRPF1	59	15	9	11	0	EXOC1	119	141	26	11	0	LMTK2	49	15	9	11	0	PPP1R3D	439	15	9	11	0	TCEA2	123	147	49	7	1
BRPF3	59	15	9	11	0	EXOC2	177	141	26	11	0	LOX	252	17	9	11	0	PPP1R3E	439	15	9	11	0	TCF12	555	4	1	11	0
BTBD1	168	59	9	11	0	EXOC3	119	141	26	11	0	LOXL4	181	17	9	11	0	PPP1R3F	440	56	26	11	0	TCF15	583	11	1	11	0
BTF3	157	36	20	11	0	EXOC3L1	176	15	9	11	0	LPA	604	101	9	11	0	PPP1R3G	439	15	9	11	0	TCF21	584	12	9	11	0
BTRC	193	31	50	6	1	EXOC3L2	176	15	9	11	0	LPL	593	69	9	11	0	PPP1R7	299	69	9	11	0	TCF25	539	64	9	11	0
BUB1B	49	15	9	11	0	EXOC3L4	176	15	9	11	0	LRP1	298	64	9	11	0	PPP1R8	157	36	20	11	0	TCF3	555	4	1	11	0
BUB3	469	146	50	6	1	EXOC4	177	141	26	11	0	LRP1B	298	64	9	11	0	PPP1R9B	403	15	9	11	0	TCF4	555	4	1	11	0
BUD13	477	144	9	11	0	EXOC5	119	141	26	11	0	LRP2	298	64	9	11	0	PPP2CA	2	0	4	10	1	TCF7	555	4	1	11	0
BUD31	307	75	34	11	0	EXOC6	177	141	26	11	0	LRP5	298	64	9	11	0	PPP2CB	2	0	4	10	1	TCF7L1	555	4	1	11	0
BYSL	645	79	26	11	0	EXOC6B	177	141	26	11	0	LRP6	298	64	9	11	0	PPP2R1A	123	147	49	7	1	TCF7L2	555	4	1	11	0
C1D	318	16	26	11	0	EXOC7	177	141	26	11	0	LRP8	298	64	9	11	0	PPP2R1B	123	147	49	7	1	TCIRG1	41	59	9	11	0
C1QA	62	17	9	11	0	EXOC8	119	141	26	11	0	LRPPRC	338	73	47	11	0	PPP2R2A	441	108	36	5	1	TCL1A	331	75	34	11	0
C1QB	62	17	9	11	0	EXOG	178	42	48	5	1	LRRC43	49	15	9	11	0	PPP2R2B	441	108	36	5	1	TCP1	85	59	9	11	0
C1QC	62	17	9	11	0	EXOSC1	179	43	19	5	1	LRRC46	586	150	52	11	0	PPP2R2C	441	108	36	5	1	TCTN1	470	106	9	11	0
C1QL1	61	17	9	11	0	EXOSC10	59	15	9	11	0	LRRC55	221	51	9	11	0	PPP2R2D	441	108	36	5	1	TCTN2	470	106	9	11	0
C1QL2	61	17	9	11	0	EXOSC2	179	43	19	5	1	LRRC8A	299	69	9	11	0	PPP2R3A	553	130	20	11	0	TEAD1	164	13	3	11	0
C1QL3	61	17	9	11	0	EXOSC3	179	43	19	5	1	LRRC8B	299	69	9	11	0	PPP2R3B	326	57	9	11	0	TEAD2	164	13	3	11	0
C1QTNF1	62	17	9	11	0	EXOSC4	180	44	12	4	1	LRRC8C	299	69	9	11	0	PPP2R5A	442	59	9	11	0	TEAD3	164	13	3	11	0
C1QTNF2	62	17	9	11	0	EXOSC5	180	44	12	4	1	LRRC8D	299	69	9	11	0	PPP2R5B	442	59	9	11	0	TEAD4	164	13	3	11	0
C1QTNF3	62	17	9	11	0	EXOSC6	180	44	12	4	1	LRRC8E	299	69	9	11	0	PPP2R5C	442	59	9	11	0	TECPR1	585	134	9	11	0
C1QTNF5	62	17	9	11	0	EXOSC7	180	44	12	4	1	LRRK2	300	64	9	11	0	PPP2R5D	442	59	9	11	0	TEFM	253	64	9	11	0
C1QTNF6	62	17	9	11	0	EXOSC8	180	44	12	4	1	LRWD1	578	141	26	11	0	PPP2R5E	442	59	9	11	0	TEK	593	69	9	11	0
C1QTNF7	62	17	9	11	0	EXOSC9	180	44	12	4	1	LSM1	302	65	19	5	1	PPP3CA	271	59	9	11	0	TELO2	620	59	9	11	0
C1QTNF9	62	17	9	11	0	EXT2	491	69	9	11	0	LSM10	302	65	19	5	1	PPP3CB	271	59	9	11	0	TEN1	492	83	26	11	0
C1QTNF9B	62	17	9	11	0	EYA1	520	55	28	5	1	LSM11	524	13	3	11	0	PPP3CC	271	59	9	11	0	TEP1	418	67	20	11	0
C2	181	17	9	11	0	EYA3	520	55	28	5	1	LSM14A	301	55	28	5	1	PPP3R1	273	57	9	11	0	TERF1	586	150	52	11	0
C2CD6	63	150	52	11	0	EZH1	344	70	20	11	0	LSM14B	301	55	28	5	1	PPP4C	271	59	9	11	0	TERF2	586	150	52	11	0
C2orf49	131	150	52	11	0	EZH2	344	70	20	11	0	LSM2	302	65	19	5	1	PPP4R1	16	147	49	7	1	TERF2IP	586	150	52	11	0
C3	267	69	9	11	0	EZR	283	59	9	11	0	LSM3	302	65	19	5	1	PPP4R2	157	36	20	11	0	TERT	473	15	9	11	0
C5	267	69	9	11	0	F10	181	17	9	11	0	LSM4	302	65	19	5	1	PPP4R3A	443	147	49	7	1	TES	292	15	9	11	0
C6	64	17	9	11	0	F11R	322	69	9	11	0	LSM5	302	65	19	5	1	PPP4R3B	443	147	49	7	1	TESPA1	117	150	52	11	0
C7	64	17	9	11	0	F13A1	268	69	9	11	0	LSM6	302	65	19	5	1	PPP4R4	360	59	9	11	0	TET1	58	150	52	11	0
C8A	64	17	9	11	0	F13B	102	69	9	11	0	LSM7	302	65	19	5	1	PPP5C	271	59	9	11	0	TEX10	263	56	26	11	0
C8B	64	17	9	11	0	F2	181	17	9	11	0	LSM8	302	65	19	5	1	PPWD1	645	79	26	11	0	TF	268	69	9	11	0
C8G	590	59	9	11	0	F2RL2	246	51	9	11	0	LSR	524	13	3	11	0	PRAM1	444	13	3	11	0	TFAM	175	59	9	11	0
C9	64	17	9	11	0	F3	255	37	9	11	0	LTBP1	79	17	9	11	0	PRAME	299	69	9	11	0	TFAP4	31	12	9	11	0
C9orf72	65	35	26	11	0	F7	181	17	9	11	0	LTF	268	69	9	11	0	PRDM10	49	15	9	11	0	TFCP2	587	125	49	7	1
C9orf78	534	79	26	11	0	F9	181	17	9	11	0	LTK	170	59	9	11	0	PRDM16	11	4	1	11	0	TFDP1	588	15	9	11	0
CAB39	649	147	49	7	1	FAAP100	578	141	26	11	0	LTN1	263	56	26	11	0	PRDM4	617	150	52	11	0	TFDP2	588	15	9	11	0

CAB39L	649	147	49	7	1	FAAP20	136	150	52	11	0	LTV1	376	79	26	11	0	PRDX3	224	52	16	6	1	TFIP11	589	135	26	11	0
CACHD1	67	69	9	11	0	FAAP24	586	150	52	11	0	LUC7L	303	79	26	11	0	PRELID1	299	69	9	11	0	TFPT	440	56	26	11	0
CACNA1A	66	59	9	11	0	FABP1	182	45	26	11	0	LUC7L2	303	79	26	11	0	PRICKLE4	59	15	9	11	0	TFR2	590	59	9	11	0
CACNA1B	66	59	9	11	0	FADD	350	69	9	11	0	LUC7L3	303	79	26	11	0	PRIM1	157	36	20	11	0	TFRC	590	59	9	11	0
CACNA1C	66	59	9	11	0	FAF1	308	32	9	11	0	LUM	596	64	9	11	0	PRIM2	626	82	26	11	0	TGFBR1	6	1	9	11	0
CACNA1D	66	59	9	11	0	FAF2	308	32	9	11	0	LY6G5B	304	80	24	11	0	PRIMPOL	49	15	9	11	0	TGFBR2	6	1	9	11	0
CACNA1E	66	59	9	11	0	FAM186B	155	150	52	11	0	LY6G5C	304	80	24	11	0	PRKAA1	96	59	9	11	0	TGFBR3	252	17	9	11	0
CACNA1F	66	59	9	11	0	FAM86B2	324	69	9	11	0	LY6G6C	304	80	24	11	0	PRKAA2	96	59	9	11	0	TGFBRAP1	638	99	9	11	0
CACNA1G	66	59	9	11	0	FAM8A1	212	58	48	5	1	LY6G6D	304	80	24	11	0	PRKAB1	445	36	20	11	0	TGM3	268	69	9	11	0
CACNA1H	66	59	9	11	0	FAM98A	183	16	26	11	0	LY96	134	16	26	11	0	PRKAB2	445	36	20	11	0	TGS1	558	64	9	11	0
CACNA1I	66	59	9	11	0	FAM98B	183	16	26	11	0	LYN	170	59	9	11	0	PRKACA	446	59	9	11	0	THBD	268	69	9	11	0
CACNA1S	66	59	9	11	0	FAM98C	184	56	26	11	0	LYRM4	339	73	47	11	0	PRKACB	446	59	9	11	0	THBS1	289	69	9	11	0
CACNA2D1	67	69	9	11	0	FANCA	264	56	26	11	0	LZTR1	285	69	9	11	0	PRKAG1	170	59	9	11	0	THEMIS	349	75	34	11	0
CACNA2D2	67	69	9	11	0	FANCB	185	150	52	11	0	MAD1L1	403	15	9	11	0	PRKAG2	170	59	9	11	0	THG1L	157	36	20	11	0
CACNA2D3	67	69	9	11	0	FANCC	263	56	26	11	0	MAD2L1	320	56	26	11	0	PRKAG3	111	59	9	11	0	THOC1	263	56	26	11	0
CACNA2D4	67	69	9	11	0	FANCD2	263	56	26	11	0	MAD2L2	320	56	26	11	0	PRKAR1A	271	59	9	11	0	THOC2	567	83	26	11	0
CACNB1	270	15	9	11	0	FANCE	264	56	26	11	0	MAEA	209	99	9	11	0	PRKAR1B	271	59	9	11	0	THOC3	591	133	50	6	1
CACNB2	270	15	9	11	0	FANCF	186	46	34	11	0	MAF	305	11	1	11	0	PRKAR2A	271	59	9	11	0	THOC5	320	56	26	11	0
CACNB3	270	15	9	11	0	FANCG	481	64	9	11	0	MAFB	305	11	1	11	0	PRKAR2B	271	59	9	11	0	THOC6	35	31	50	6	1
CACNB4	270	15	9	11	0	FANCI	263	56	26	11	0	MAFF	31	12	9	11	0	PRKCA	447	59	9	11	0	THOC7	318	16	26	11	0
CACNG1	68	59	9	11	0	FANCL	23	9	9	11	0	MAFG	31	12	9	11	0	PRKCB	447	59	9	11	0	THRAP3	592	36	20	11	0
CACNG2	68	59	9	11	0	FANCM	49	15	9	11	0	MAFK	31	12	9	11	0	PRKCI	447	59	9	11	0	THRB	472	22	6	10	1
CACNG3	68	59	9	11	0	FAP	268	69	9	11	0	MAGED1	306	13	3	11	0	PRKCSH	268	69	9	11	0	TICAM1	242	123	3	11	0
CACNG4	68	59	9	11	0	FARSA	248	59	9	11	0	MAGI2	170	59	9	11	0	PRKCZ	447	59	9	11	0	TIE1	593	69	9	11	0
CACNG6	68	59	9	11	0	FARSB	430	91	29	8	1	MAGOH	307	75	34	11	0	PRKDC	620	59	9	11	0	TIMELESS	627	82	26	11	0
CACNG7	68	59	9	11	0	FAS	603	37	9	11	0	MAGOHB	307	75	34	11	0	PRKRA	448	59	9	11	0	TIMM10	595	16	26	11	0
CACNG8	68	59	9	11	0	FAU	419	59	9	11	0	MAGT1	308	32	9	11	0	PRKRIP1	634	79	26	11	0	TIMM10B	595	16	26	11	0
CACTIN	573	15	9	11	0	FBF1	187	144	9	11	0	MAK16	309	79	26	11	0	PRKX	446	59	9	11	0	TIMM13	595	16	26	11	0
CACYBP	124	99	9	11	0	FBL	324	69	9	11	0	MALT1	74	59	9	11	0	PRLR	208	37	9	11	0	TIMM17A	594	136	46	13	1
CALCR	246	51	9	11	0	FBXL12	191	64	9	11	0	MAML1	48	16	26	11	0	PRM1	449	96	24	11	0	TIMM17B	594	136	46	13	1
CALCRL	246	51	9	11	0	FBXL13	188	31	50	6	1	MAP1A	310	66	9	11	0	PRM2	449	96	24	11	0	TIMM21	369	73	47	11	0
CALM1	69	18	4	10	1	FBXL14	188	31	50	6	1	MAP1B	310	66	9	11	0	PRMT1	636	59	9	11	0	TIMM22	594	136	46	13	1
CALM2	69	18	4	10	1	FBXL15	191	64	9	11	0	MAP1S	310	66	9	11	0	PRMT5	419	59	9	11	0	TIMM23	594	136	46	13	1
CALM3	69	18	4	10	1	FBXL16	188	31	50	6	1	MAP2	319	150	52	11	0	PROC	181	17	9	11	0	TIMM23B	594	136	46	13	1
CALR	268	69	9	11	0	FBXL17	188	31	50	6	1	MAP3K5	253	64	9	11	0	PROP1	11	4	1	11	0	TIMM29	338	73	47	11	0
CAMK1G	417	69	9	11	0	FBXL18	191	64	9	11	0	MAP3K7	484	69	9	11	0	PRPF18	157	36	20	11	0	TIMM44	316	79	26	11	0
CAMK2A	70	22	6	10	1	FBXL19	59	15	9	11	0	MAP4	540	36	20	11	0	PRPF19	450	146	50	6	1	TIMM50	59	15	9	11	0
CAMK2B	70	22	6	10	1	FBXL2	188	31	50	6	1	MAP7	558	64	9	11	0	PRPF3	605	79	26	11	0	TIMM8A	595	16	26	11	0
CAMK2D	70	22	6	10	1	FBXL20	188	31	50	6	1	MAPK6	311	67	20	11	0	PRPF31	377	59	9	11	0	TIMM8B	595	16	26	11	0
CAMK2G	70	22	6	10	1	FBXL22	191	64	9	11	0	MAPK8IP2	374	13	3	11	0	PRPF38A	566	132	0	11	0	TIMM9	595	16	26	11	0
CAMKMT	324	69	9	11	0	FBXL3	191	64	9	11	0	MAPKAP1	552	129	9	11	0	PRPF38B	566	132	0	11	0	TINF2	117	150	52	11	0
CAMLG	71	150	52	11	0	FBXL4	299	69	9	11	0	MAPKAPK5	417	69	9	11	0	PRPF39	619	99	9	11	0	TIPIN	157	36	20	11	0
CAND1	266	147	49	7	1	FBXL5	299	69	9	11	0	MAPT	310	66	9	11	0	PRPF4	422	146	50	6	1	TJP1	270	15	9	11	0
CANX	268	69	9	11	0	FBXL6	188	31	50	6	1	MARCO	526	17	9	11	0	PRPF40A	573	15	9	11	0	TLE1	99	22	6	10	1
CAPG	420	59	9	11	0	FBXL7	299	69	9	11	0	MARK4	271	59	9	11	0	PRPF40B	573	15	9	11	0	TLE2	99	22	6	10	1
CAPN1	170	59	9	11	0	FBXL8	191	64	9	11	0	MASP1	181	17	9	11	0	PRPF4B	270	15	9	11	0	TLE3	99	22	6	10	1
CAPN2	170	59	9	11	0	FBXO10	299	69	9	11	0	MAT1A	413	59	9	11	0	PRPF6	619	99	9	11	0	TLE4	99	22	6	10	1
CAPN3	170	59	9	11	0	FBXO11	299	69	9	11	0	MAT2A	413	59	9	11	0	PRPF8	149	35	26	11	0	TLE6	140	79	26	11	0
CAPNS1	326	57	9	11	0	FBXO15	189	80	24	11	0	MAT2B	299	69	9	11	0	PRPS1	451	97	41	10	1	TLR1	596	64	9	11	0
CAPNS2	326	57	9	11	0	FBXO17	190	47	9	11	0	MATN1	312	69	9	11	0	PRPS1L1	451	97	41	10	1	TLR10	596	64	9	11	0
CAPRIN2	524	13	3	11	0	FBXO2	190	47	9	11	0	MATN2	312	69	9	11	0	PRPS2	451	97	41	10	1	TLR2	596	64	9	11	0
CAPZA1	457	16	26	11	0	FBXO21	418	67	20	11	0	MATN3	312	69	9	11	0	PRPSAP1	451	97	41	10	1	TLR4	596	64	9	11	0
CAPZA2	457	16	26	11	0	FBXO24	191	64	9	11	0	MATN4	312	69	9	11	0	PRPSAP2	451	97	41	10	1	TLR6	596	64	9	11	0
CAPZB	307	75	34	11	0	FBXO25	189	80	24	11	0	MAU2	619	99	9	11	0	PRR5	49	15	9	11	0	TLR7	596	64	9	11	0
CARD10	473	15	9	11	0	FBXO27	190	47	9	11	0	MAVS	540	36	20	11	0	PRR5L	49	15	9	11	0	TM7SF2	581	59	9	11	0
CARD11	403	15	9	11	0	FBXO31	189	80	24	11	0	MAX	31	12	9	11	0	PSEN1	13	59	9	11	0	TMED10	597	17	9	11	0
CARD16	72	19	20	11	0	FBXO32	189	80	24	11	0	MBD2	313	15	9	11	0	PSEN2	13	59	9	11	0	TMED3	597	17	9	11	0
CARD8	72	19	20	11	0	FBXO33	191	64	9	11	0	MBD3	313	15	9	11	0	PSENEN	602	138	47	11	0	TMED7	597	17	9	11	0
CARD9	403	15	9	11	0	FBXO38	191	64	9	11	0	MBIP	319	150	52	11	0	PSG9	239	59	9	11	0	TMEM102	598	56	26	11	0
CARMIL1	299	69	9	11	0	FBXO39	191	64	9	11	0	MBL2	526	17	9	11	0	PSMA1	452	98	32	8	1	TMEM107	600	7	9	11	0
CASC3	524	13	3	11	0	FBXO4	192	80	24	11	0	MBP	314	68	24	11	0	PSMA2	452	98	32	8	1	TMEM147	599	137	47	11	0
CASP1	74	59	9	11	0	FBXO42	285	69	9	11	0	MBTD1	59	15	9	11	0	PSMA3	452	98	32	8	1	TMEM17	600	7	9	11	0
CASP10	74	59	9	11	0	FBXO44	190	47	9	11	0	MCCC1	636	59	9	11	0	PSMA4	452	98	32	8	1	TMEM183A	189	80	24	11	0
CASP12	74	59	9	11	0	FBXO45	559	31	50	6	1	MCCC2	413	59	9	11	0	PSMA5	452	98	32	8	1	TMEM199	318	16	26	11	0
CASP2	74	59	9	11	0	FBXO48	191	64	9	11	0	MCL1	46	14	20	11	0	PSMA6	452	98	32	8	1	TMEM216	600	7	9	11	0
CASP3	73	59	9	11	0	FBXO6	190	47	9	11	0	MCM2	315	59	9	11	0	PSMA7	452	98	32	8	1	TMEM231	275	106	9	11	0
CASP4	74	59	9	11	0	FBXO7	321	16	26	11	0	MCM3	315	59	9	11	0	PSMA8	452	98	32	8	1	TMEM249	601	75	34	11	0
CASP5	74	59	9	11	0	FBXO8	418	67	20	11	0	MCM3AP	539	64	9	11	0	PSMB1	452	98	32	8	1	TMEM258	602	138	47	11	0
CASP8	74	59	9	11	0	FBXO9	189	80	24	11	0	MCM4	315	59	9	11	0	PSMB10	452	98	32	8	1	TMEM30A	40	69	9	11	0
CASP9	74	59	9	11	0	FBXW11	193	31	50	6	1	MCM5	315	59	9	11	0	PSMB2	452	98	32	8	1	TMEM30B	40	69	9	11	0
CASQ2	151	32	9	11	0	FBXW4	140	79	26	11	0	MCM6	315	59	9	11	0	PSMB3	452	98	32	8	1	TMEM67	299	69	9	11	0
CAT	157	36	20	11	0	FBXW5	138	31	50	6	1	MCM7	315	59	9	11	0	PSMB4	452	98	32	8	1	TNC	593	69	9	11	0
CATSPER1	75	37	9	11	0	FBXW7	193	31	50	6	1	MCM8	315	59	9	11	0	PSMB5	452	98	32	8	1	TNFAIP1	277	15	9	11	0
CATSPER2	517	59	9	11	0	FBXW8	140	79	26	11	0	MCM9	315	59	9	11	0	PSMB6	452	98	32	8	1	TNFAIP2	176	15	9	11	0
CATSPER3	517	59	9	11	0	FCER1G	88	15	9	11	0	MCOLN1	517	59	9	11	0	PSMB7	452	98	32	8	1	TNFRSF11B	603	37	9	11	0
CATSPERB	134	16	26	11	0	FCF1	194	48	13	5	1	MCRS1	59	15	9	11	0	PSMB8	452	98	32	8	1	TNFRSF1A	603	37	9	11	0
CATSPERG	134	16	26	11	0	FCGR3A	409	37	9	11	0	MCU	316	79	26	11	0	PSMB9	452	98	32	8	1	TNFRSF1B	603	37	9	11	0
CAV1	76	83	26	11	0	FCN1	195	17	9	11	0	MCUB	316	79	26	11	0	PSMC1	453	59	9	11	0	TNK2	292	15	9	11	0
CAV2	76	83	26	11	0	FCN2	195	17	9	11	0	MDM2	317	150	52	11	0	PSMC2	453	59	9	11	0	TNKS	292	15	9	11	0
CAV3	76	83	26	11	0	FCN3	195	17	9	11	0	MDM4	317	150	52	11	0	PSMC3	453	59	9	11	0	TNKS1BP1	540	36	20	11	0

CBFB	331	75	34	11	0
CBL	49	15	9	11	0
CBLL1	48	16	26	11	0
CBR4	77	64	9	11	0
CBX2	78	13	3	11	0
CBX3	59	15	9	11	0
CBX4	78	13	3	11	0
CBX5	59	15	9	11	0
CBX6	78	13	3	11	0
CBX7	78	13	3	11	0
CBX8	78	13	3	11	0
CC2D2A	418	67	20	11	0
CCAR2	539	64	9	11	0
CCBE1	79	17	9	11	0
CCDC102A	240	16	26	11	0
CCDC103	233	75	34	11	0
CCDC113	251	16	26	11	0
CCDC115	457	16	26	11	0
CCDC12	605	79	26	11	0
CCDC120	80	15	9	11	0
CCDC47	157	36	20	11	0
CCDC51	605	79	26	11	0
CCDC61	403	15	9	11	0
CCDC65	60	141	26	11	0
CCDC68	118	64	9	11	0
CCDC8	540	36	20	11	0
CCDC9	81	144	9	11	0
CCM2	49	15	9	11	0
CCNA1	82	13	3	11	0
CCNA2	82	13	3	11	0
CCNB1	82	13	3	11	0
CCNB2	82	13	3	11	0
CCNB3	522	120	34	11	0
CCNC	82	13	3	11	0
CCND1	82	13	3	11	0
CCND2	82	13	3	11	0
CCND3	82	13	3	11	0
CCNE1	82	13	3	11	0
CCNE2	82	13	3	11	0
CCNF	82	13	3	11	0
CCNG1	82	13	3	11	0
CCNG2	82	13	3	11	0
CCNH	82	13	3	11	0
CCNI	82	13	3	11	0
CCNI2	82	13	3	11	0
CCNJ	82	13	3	11	0
CCNJL	82	13	3	11	0
CCNK	83	36	20	11	0
CCNL1	98	13	3	11	0
CCNL2	98	13	3	11	0
CCNO	82	13	3	11	0
CCNT1	83	36	20	11	0
CCNT2	83	36	20	11	0
CCNY	98	13	3	11	0
CCP110	84	150	52	11	0
CCT2	85	59	9	11	0
CCT3	85	59	9	11	0
CCT4	85	59	9	11	0
CCT5	85	59	9	11	0
CCT6A	85	59	9	11	0
CCT6B	85	59	9	11	0
CCT7	85	59	9	11	0
CCT8	85	59	9	11	0
CCZ1	86	20	26	11	0
CCZ1B	87	20	26	11	0
CD14	191	64	9	11	0
CD19	90	37	9	11	0
CD2	90	37	9	11	0
CD200R1	409	37	9	11	0
CD247	88	15	9	11	0
CD28	515	37	9	11	0
CD2AP	362	15	9	11	0
CD2BP2	89	79	26	11	0
CD34	233	75	34	11	0
CD36	40	69	9	11	0
CD3D	90	37	9	11	0
CD3E	90	37	9	11	0
CD3G	90	37	9	11	0
CD4	90	37	9	11	0
CD40	603	37	9	11	0

FCRL5	239	59	9	11	0
FDX2	56	71	21	11	0
FEM1A	196	69	9	11	0
FEM1B	196	69	9	11	0
FEM1C	196	69	9	11	0
FEN1	377	59	9	11	0
FGA	206	106	9	11	0
FGB	197	64	9	11	0
FGFR1	170	59	9	11	0
FGFR2	170	59	9	11	0
FGFR3	170	59	9	11	0
FGFR4	170	59	9	11	0
FGG	197	64	9	11	0
FGL1	197	64	9	11	0
FGL2	197	64	9	11	0
FH	248	59	9	11	0
FIP1L1	524	13	3	11	0
FIS1	607	79	26	11	0
FKBP1A	362	15	9	11	0
FKBP1B	362	15	9	11	0
FKBP4	362	15	9	11	0
FKBP8	362	15	9	11	0
FLNA	560	59	9	11	0
FLOT1	42	99	9	11	0
FLOT2	42	99	9	11	0
FLT1	593	69	9	11	0
FLT3	322	69	9	11	0
FLT4	593	69	9	11	0
FLYWCH1	198	13	3	11	0
FMR1	199	125	49	7	1
FN1	593	69	9	11	0
FNTA	619	99	9	11	0
FNTB	123	147	49	7	1
FOS	200	12	9	11	0
FOSL1	200	12	9	11	0
FOSL2	200	12	9	11	0
FOXE3	555	4	1	11	0
FOXF1	555	4	1	11	0
FOXF2	555	4	1	11	0
FOXH1	201	123	3	11	0
FOXO3	374	13	3	11	0
FOXP3	11	4	1	11	0
FOXRED1	248	59	9	11	0
FRG1	157	36	20	11	0
FRMPD1	49	15	9	11	0
FRMPD4	49	15	9	11	0
FSHB	295	106	9	11	0
FSHR	246	51	9	11	0
FTH1	202	59	9	11	0
FTL	202	59	9	11	0
FTSJ3	645	79	26	11	0
FUBP3	404	22	6	10	1
FXN	341	71	21	11	0
FXR1	199	125	49	7	1
FXR2	199	125	49	7	1
FXYD1	68	59	9	11	0
FXYD2	68	59	9	11	0
FZD1	7	59	9	11	0
FZD8	7	59	9	11	0
FZR1	203	31	50	6	1
G3BP1	204	55	28	5	1
G3BP2	204	55	28	5	1
GABARAP	405	59	9	11	0
GABBR1	491	69	9	11	0
GABBR2	491	69	9	11	0
GABRA2	205	59	9	11	0
GABRA4	205	59	9	11	0
GABRB2	205	59	9	11	0
GABRB3	205	59	9	11	0
GABRD	205	59	9	11	0
GABRE	205	59	9	11	0
GABRG1	205	59	9	11	0
GABRP	205	59	9	11	0
GABRR1	205	59	9	11	0
GABRR2	205	59	9	11	0
GADD45GIP1	339	73	47	11	0
GAN	168	59	9	11	0
GANAB	268	69	9	11	0
GAPDH	293	59	9	11	0
GAR1	302	65	19	5	1

MDN1	299	69	9	11	0
MEAF6	157	36	20	11	0
MED1	319	150	52	11	0
MED10	318	16	26	11	0
MED11	318	16	26	11	0
MED12	115	56	26	11	0
MED12L	115	56	26	11	0
MED13	319	150	52	11	0
MED13L	319	150	52	11	0
MED14	320	56	26	11	0
MED15	415	10	9	11	0
MED16	388	83	26	11	0
MED17	233	75	34	11	0
MED18	418	67	20	11	0
MED19	319	150	52	11	0
MED20	331	75	34	11	0
MED21	318	16	26	11	0
MED22	457	16	26	11	0
MED23	263	56	26	11	0
MED24	263	56	26	11	0
MED25	265	83	26	11	0
MED26	30	150	52	11	0
MED27	457	16	26	11	0
MED28	321	16	26	11	0
MED29	321	16	26	11	0
MED30	321	16	26	11	0
MED31	589	135	26	11	0
MED4	457	16	26	11	0
MED6	320	56	26	11	0
MED7	457	16	26	11	0
MED8	321	16	26	11	0
MED9	321	16	26	11	0
MEF2A	555	4	1	11	0
MEF2B	137	13	3	11	0
MEF2C	555	4	1	11	0
MEFV	614	59	9	11	0
MEIS1	528	22	6	10	1
MEN1	586	150	52	11	0
MEP1A	593	69	9	11	0
MEP1B	593	69	9	11	0
MEPCE	539	64	9	11	0
MERTK	322	69	9	11	0
MET	424	69	9	11	0
METTL1	323	15	9	11	0
METTL14	323	15	9	11	0
METTL18	175	59	9	11	0
METTL21A	324	69	9	11	0
METTL22	324	69	9	11	0
METTL23	324	69	9	11	0
METTL3	323	15	9	11	0
MFAP1	534	79	26	11	0
MFF	325	16	26	11	0
MFN1	299	69	9	11	0
MGA	319	150	52	11	0
MICU1	326	57	9	11	0
MICU2	326	57	9	11	0
MICU3	326	57	9	11	0
MID1	49	15	9	11	0
MIER1	327	70	20	11	0
MIER2	327	70	20	11	0
MIER3	327	70	20	11	0
MIF4GD	568	59	9	11	0
MIOS	328	79	26	11	0
MIS12	84	150	52	11	0
MITF	200	12	9	11	0
MKLN1	285	69	9	11	0
MKRN3	157	36	20	11	0
MKS1	418	67	20	11	0
MLEC	79	17	9	11	0
MLH1	243	59	9	11	0
MLH3	580	150	52	11	0
MLLT1	329	36	20	11	0
MLLT10	200	12	9	11	0
MLLT3	329	36	20	11	0
MLST8	641	146	50	6	1
MLXIPL	374	13	3	11	0
MMGT1	331	75	34	11	0
MMRN1	526	17	9	11	0
MMRN2	526	17	9	11	0
MMS19	620	59	9	11	0

PSMC3IP	124	99	9	11	0
PSMC4	453	59	9	11	0
PSMC5	453	59	9	11	0
PSMC6	453	59	9	11	0
PSMD1	123	147	49	7	1
PSMD10	27	69	9	11	0
PSMD11	454	99	9	11	0
PSMD12	454	99	9	11	0
PSMD13	454	99	9	11	0
PSMD14	455	59	9	11	0
PSMD2	123	147	49	7	1
PSMD3	454	99	9	11	0
PSMD4	265	83	26	11	0
PSMD5	123	147	49	7	1
PSMD6	454	99	9	11	0
PSMD7	165	39	9	11	0
PSMD8	456	99	9	11	0
PSMD9	292	15	9	11	0
PSME1	457	16	26	11	0
PSME2	457	16	26	11	0
PSME3	457	16	26	11	0
PSME4	649	147	49	7	1
PSMF1	321	16	26	11	0
PSMG1	418	67	20	11	0
PSMG2	299	69	9	11	0
PSMG3	458	100	47	11	0
PSMG4	331	75	34	11	0
PTBP2	99	22	6	10	1
PTCD3	338	73	47	11	0
PTGES3	124	99	9	11	0
PTGS2	156	69	9	11	0
PTH1R	7	59	9	11	0
PTK2B	491	69	9	11	0
PTN	64	17	9	11	0
PTOV1	457	16	26	11	0
PTPA	626	82	26	11	0
PTPN1	420	59	9	11	0
PTPN11	420	59	9	11	0
PTPN6	420	59	9	11	0
PTPRA	420	59	9	11	0
PTPRB	459	101	9	11	0
PTPRN2	271	59	9	11	0
PTPRQ	459	101	9	11	0
PUF60	99	22	6	10	1
PWP2	642	79	26	11	0
PXK	460	69	9	11	0
PXMP2	410	16	26	11	0
PXN	292	15	9	11	0
PYCARD	350	69	9	11	0
PYDC1	350	69	9	11	0
PYGO2	48	16	26	11	0
PYHIN1	586	150	52	11	0
PYM1	605	79	26	11	0
QPRT	362	15	9	11	0
QRSL1	413	59	9	11	0
QTRT1	461	59	9	11	0
QTRT2	461	59	9	11	0
RAB11A	462	49	14	5	1
RAB3A	462	49	14	5	1
RAB3GAP1	320	56	26	11	0
RAB3GAP2	388	83	26	11	0
RABEP1	395	64	9	11	0
RABGAP1	49	15	9	11	0
RABGGTA	619	99	9	11	0
RABGGTB	123	147	49	7	1
RAC1	462	49	14	5	1
RAC2	462	49	14	5	1
RAC3	462	49	14	5	1
RACGAP1	420	59	9	11	0
RACK1	216	108	36	5	1
RAD1	463	102	26	11	0
RAD17	397	86	9	11	0
RAD18	49	15	9	11	0
RAD21	639	15	9	11	0
RAD23A	464	103	33	7	1
RAD23B	464	103	33	7	1
RAD50	538	59	9	11	0
RAD51	453	59	9	11	0
RAD51AP1	564	150	52	11	0
RAD51AP2	465	104	34	11	0

TNN	604	101	9	11	0
TNNC1	354	57	9	11	0
TNNC2	354	57	9	11	0
TNNI1	605	79	26	11	0
TNNI2	605	79	26	11	0
TNNI3	605	79	26	11	0
TNNT1	605	79	26	11	0
TNNT2	605	79	26	11	0
TNNT3	605	79	26	11	0
TNP1	606	75	34	11	0
TNP2	345	75	34	11	0
TNR	593	69	9	11	0
TNRC6A	129	55	28	5	1
TNXB	593	69	9	11	0
TOB1	639	15	9	11	0
TOLLIP	541	59	9	11	0
TOMM20	607	79	26	11	0
TOMM22	125	73	47	11	0
TOMM40	637	118	9	11	0
TOMM40L	637	118	9	11	0
TOMM5	368	73	47	11	0
TOMM6	368	73	47	11	0
TOMM7	368	73	47	11	0
TOMM70	93	59	9	11	0
TOP1	608	79	26	11	0
TOP2A	315	59	9	11	0
TOP2B	315	59	9	11	0
TOP3A	315	59	9	11	0
TOP3B	315	59	9	11	0
TOPBP1	49	15	9	11	0
TOPORS	653	150	52	11	0
TOX4	200	12	9	11	0
TP53	536	15	9	11	0
TP53BP1	319	150	52	11	0
TP53RK	483	109	37	5	1
TP63	536	15	9	11	0
TPM1	283	59	9	11	0
TPM2	283	59	9	11	0
TPM3	283	59	9	11	0
TPM4	283	59	9	11	0
TPR	395	64	9	11	0
TPRKB	331	75	34	11	0
TRA	90	37	9	11	0
TRA2A	609	132	0	11	0
TRA2B	609	132	0	11	0
TRADD	184	56	26	11	0
TRAF1	49	15	9	11	0
TRAF2	49	15	9	11	0
TRAF3	49	15	9	11	0
TRAF3IP1	539	64	9	11	0
TRAF4	49	15	9	11	0
TRAF5	49	15	9	11	0
TRAF6	49	15	9	11	0
TRAF7	193	31	50	6	1
TRAPPC1	611	20	26	11	0
TRAPPC10	45	83	26	11	0
TRAPPC11	45	83	26	11	0
TRAPPC12	539	64	9	11	0
TRAPPC13	610	139	26	11	0
TRAPPC2	611	20	26	11	0
TRAPPC2L	611	20	26	11	0
TRAPPC3	612	16	26	11	0
TRAPPC3L	612	16	26	11	0
TRAPPC4	611	20	26	11	0
TRAPPC5	612	16	26	11	0
TRAPPC6A	612	16	26	11	0
TRAPPC6B	612	16	26	11	0
TRAPPC8	45	83	26	11	0
TRAPPC9	45	83	26	11	0
TRAT1	233	75	34	11	0
TRDN	613	36	20	11	0
TRERF1	524	13	3	11	0
TREX1	175	59	9	11	0
TRIAP1	370	73	47	11	0
TRIM21	614	59	9	11	0
TRIM28	49	15	9	11	0
TRIM40	614	59	9	11	0
TRIM54	49	15	9	11	0
TRIP4	157	36	20	11	0
TRMT10B	333	71	21	11	0

CD44	91	106	9	11	0	GATA1	555	4	1	11	0	MNAT1	542	16	26	11	0	RAD51B	466	69	9	11	0	TRMT10C	333	71	21	11	0
CD69	90	37	9	11	0	GATA2	555	4	1	11	0	MOCS1	636	59	9	11	0	RAD51C	466	69	9	11	0	TRMT112	615	140	26	11	0
CD74	252	17	9	11	0	GATA4	555	4	1	11	0	MOCS2	57	59	9	11	0	RAD51D	466	69	9	11	0	TRMT6	333	71	21	11	0
CD79A	90	37	9	11	0	GATA6	555	4	1	11	0	MON1A	473	15	9	11	0	RAD52	467	105	35	11	0	TRMT61A	324	69	9	11	0
CD79B	90	37	9	11	0	GATAD2A	200	12	9	11	0	MON1B	473	15	9	11	0	RAD54L	468	59	9	11	0	TRMT61B	333	71	21	11	0
CD80	515	37	9	11	0	GATAD2B	200	12	9	11	0	MORF4L1	330	125	49	7	1	RAD9A	463	102	26	11	0	TRPC1	616	59	9	11	0
CD8A	515	37	9	11	0	GATB	248	59	9	11	0	MORF4L2	330	125	49	7	1	RAD9B	463	102	26	11	0	TRPC3	13	59	9	11	0
CD8B	515	37	9	11	0	GATC	334	71	21	11	0	MPC1	125	73	47	11	0	RADIL	473	15	9	11	0	TRPC4	13	59	9	11	0
CD9	631	51	9	11	0	GBA3	268	69	9	11	0	MPC2	125	73	47	11	0	RAE1	469	146	50	6	1	TRPC4AP	389	82	26	11	0
CDC16	93	59	9	11	0	GCFC2	60	141	26	11	0	MPHOSPH10	376	79	26	11	0	RAG1	253	64	9	11	0	TRPC5	13	59	9	11	0
CDC20	203	31	50	6	1	GCH1	413	59	9	11	0	MPHOSPH6	434	93	20	11	0	RAG2	285	69	9	11	0	TRPC6	13	59	9	11	0
CDC23	93	59	9	11	0	GCHFR	307	75	34	11	0	MPHOSPH8	59	15	9	11	0	RALY	129	55	28	5	1	TRPM4	13	59	9	11	0
CDC26	92	75	34	11	0	GCLC	413	59	9	11	0	MPP4	473	15	9	11	0	RAMP1	470	106	9	11	0	TRPS1	617	150	52	11	0
CDC27	93	59	9	11	0	GCLM	272	59	9	11	0	MPP7	103	133	50	6	1	RAMP2	470	106	9	11	0	TRPV3	13	59	9	11	0
CDC37	605	79	26	11	0	GCM1	319	150	52	11	0	MPV17L2	410	16	26	11	0	RAMP3	470	106	9	11	0	TRPV6	13	59	9	11	0
CDC40	422	146	50	6	1	GCN1	620	59	9	11	0	MR1	238	59	9	11	0	RAN	462	49	14	5	1	TRRAP	649	147	49	7	1
CDC42	462	49	14	5	1	GCSH	146	34	9	11	0	MRE11	133	30	15	5	1	RANBP1	107	27	8	2	1	TSC1	389	82	26	11	0
CDC45	94	21	5	0	1	GDF11	407	106	9	11	0	MREG	331	75	34	11	0	RANBP10	559	31	50	6	1	TSC2	649	147	49	7	1
CDC5L	344	70	20	11	0	GDI1	105	26	26	11	0	MRGBP	49	15	9	11	0	RANBP17	649	147	49	7	1	TSEN2	49	15	9	11	0
CDC73	157	36	20	11	0	GEMIN2	115	56	26	11	0	MRPL1	333	71	21	11	0	RANBP2	481	64	9	11	0	TSEN34	586	150	52	11	0
CDCA8	131	150	52	11	0	GEMIN4	263	56	26	11	0	MRPL10	334	71	21	11	0	RANBP3	471	107	20	11	0	TSEN54	623	141	26	11	0
CDH1	95	59	9	11	0	GEMIN5	140	79	26	11	0	MRPL11	334	71	21	11	0	RANBP3L	471	107	20	11	0	TSG101	618	143	48	5	1
CDH10	95	59	9	11	0	GEMIN6	157	36	20	11	0	MRPL12	334	71	21	11	0	RANBP9	266	147	49	7	1	TSHR	246	51	9	11	0
CDH11	95	59	9	11	0	GEMIN7	349	75	34	11	0	MRPL13	334	71	21	11	0	RANGAP1	299	69	9	11	0	TSN	119	141	26	11	0
CDH12	95	59	9	11	0	GEMIN8	457	16	26	11	0	MRPL14	334	71	21	11	0	RAP1A	462	49	14	5	1	TSNARE1	49	15	9	11	0
CDH13	95	59	9	11	0	GET4	454	99	9	11	0	MRPL15	332	71	21	11	0	RAPGEF2	49	15	9	11	0	TSNAX	119	141	26	11	0
CDH15	95	59	9	11	0	GFI1	11	4	1	11	0	MRPL16	333	71	21	11	0	RARA	472	22	6	10	1	TSPAN17	631	51	9	11	0
CDH17	95	59	9	11	0	GFI1B	11	4	1	11	0	MRPL17	334	71	21	11	0	RARG	472	22	6	10	1	TSPAN32	68	59	9	11	0
CDH19	95	59	9	11	0	GFRA1	206	106	9	11	0	MRPL18	341	71	21	11	0	RASGRP3	49	15	9	11	0	TSPAN33	631	51	9	11	0
CDH2	95	59	9	11	0	GFRA2	206	106	9	11	0	MRPL19	334	71	21	11	0	RASIP1	473	15	9	11	0	TSR1	645	79	26	11	0
CDH20	95	59	9	11	0	GFRA3	206	106	9	11	0	MRPL2	334	71	21	11	0	RASSF2	49	15	9	11	0	TTBK1	271	59	9	11	0
CDH22	95	59	9	11	0	GFRA4	284	59	9	11	0	MRPL20	334	71	21	11	0	RB1	82	13	3	11	0	TTC21A	619	99	9	11	0
CDH23	95	59	9	11	0	GGA1	123	147	49	7	1	MRPL21	334	71	21	11	0	RB1CC1	84	150	52	11	0	TTC21B	619	99	9	11	0
CDH24	95	59	9	11	0	GGA3	123	147	49	7	1	MRPL22	334	71	21	11	0	RBBP4	474	108	36	5	1	TTC26	619	99	9	11	0
CDH26	95	59	9	11	0	GH1	207	80	24	11	0	MRPL23	405	59	9	11	0	RBBP5	645	79	26	11	0	TTC30A	619	99	9	11	0
CDH3	95	59	9	11	0	GHR	208	37	9	11	0	MRPL24	334	71	21	11	0	RBBP6	129	55	28	5	1	TTC30B	619	99	9	11	0
CDH4	95	59	9	11	0	GID4	536	15	9	11	0	MRPL27	334	71	21	11	0	RBBP7	474	108	36	5	1	TTC37	619	99	9	11	0
CDH5	95	59	9	11	0	GID8	209	99	9	11	0	MRPL28	341	71	21	11	0	RBBP8	30	150	52	11	0	TTC8	619	99	9	11	0
CDH6	95	59	9	11	0	GIGYF1	210	15	9	11	0	MRPL3	332	71	21	11	0	RBCK1	23	9	9	11	0	TTF2	280	59	9	11	0
CDH8	95	59	9	11	0	GIGYF2	210	15	9	11	0	MRPL30	341	71	21	11	0	RBL1	82	13	3	11	0	TTI1	620	59	9	11	0
CDHR3	95	59	9	11	0	GINS1	211	135	26	11	0	MRPL32	339	73	47	11	0	RBL2	82	13	3	11	0	TTI2	620	59	9	11	0
CDK1	97	59	9	11	0	GINS2	211	135	26	11	0	MRPL33	605	79	26	11	0	RBM10	573	15	9	11	0	TTN	129	55	28	5	1
CDK12	270	15	9	11	0	GINS3	211	135	26	11	0	MRPL34	335	72	22	11	0	RBM12	129	55	28	5	1	TTR	143	32	9	11	0
CDK13	270	15	9	11	0	GINS4	211	135	26	11	0	MRPL35	335	72	22	11	0	RBM12B	475	36	20	11	0	TTYH1	621	69	9	11	0
CDK14	96	59	9	11	0	GJA1	212	58	48	5	1	MRPL36	335	72	22	11	0	RBM14	129	55	28	5	1	TTYH2	621	69	9	11	0
CDK16	96	59	9	11	0	GJA3	212	58	48	5	1	MRPL37	336	73	47	11	0	RBM15	129	55	28	5	1	TTYH3	621	69	9	11	0
CDK19	96	59	9	11	0	GJA4	212	58	48	5	1	MRPL38	339	73	47	11	0	RBM15B	129	55	28	5	1	TUBB	353	49	14	5	1
CDK2	97	59	9	11	0	GJA5	212	58	48	5	1	MRPL39	341	71	21	11	0	RBM17	589	135	26	11	0	TUBG1	413	59	9	11	0
CDK3	97	59	9	11	0	GJB1	212	58	48	5	1	MRPL4	332	71	21	11	0	RBM22	547	22	6	10	1	TUBG2	413	59	9	11	0
CDK4	97	59	9	11	0	GJB2	212	58	48	5	1	MRPL40	338	73	47	11	0	RBM28	418	67	20	11	0	TUBGCP2	622	141	26	11	0
CDK5	97	59	9	11	0	GJB3	212	58	48	5	1	MRPL41	218	75	34	11	0	RBM3	110	55	28	5	1	TUBGCP3	622	141	26	11	0
CDK5R1	98	13	3	11	0	GJB4	212	58	48	5	1	MRPL42	338	73	47	11	0	RBM39	99	22	6	10	1	TUBGCP4	622	141	26	11	0
CDK5R2	98	13	3	11	0	GJB5	212	58	48	5	1	MRPL43	338	73	47	11	0	RBM41	418	67	20	11	0	TUBGCP5	622	141	26	11	0
CDK5RAP3	542	16	26	11	0	GJB6	212	58	48	5	1	MRPL44	642	79	26	11	0	RBM42	524	13	3	11	0	TUBGCP6	623	141	26	11	0
CDK6	97	59	9	11	0	GJB7	212	58	48	5	1	MRPL45	316	79	26	11	0	RBM44	58	150	52	11	0	TUSC3	308	32	9	11	0
CDK7	97	59	9	11	0	GJC1	212	58	48	5	1	MRPL46	333	71	21	11	0	RBM45	547	22	6	10	1	TXNL1	299	69	9	11	0
CDK8	96	59	9	11	0	GJC2	212	58	48	5	1	MRPL47	339	73	47	11	0	RBM4B	476	55	28	5	1	TXNL4A	151	32	9	11	0
CDK9	97	59	9	11	0	GJC3	212	58	48	5	1	MRPL48	333	71	21	11	0	RBM5	573	15	9	11	0	TXNL4B	308	32	9	11	0
CDKN1A	531	15	9	11	0	GJD3	212	58	48	5	1	MRPL49	337	71	21	11	0	RBM8A	544	132	0	11	0	TYK2	473	15	9	11	0
CDKN1B	531	15	9	11	0	GLDC	561	59	9	11	0	MRPL50	338	73	47	11	0	RBMS1	99	22	6	10	1	TYRO3	322	69	9	11	0
CDKN2D	27	69	9	11	0	GLDN	79	17	9	11	0	MRPL51	338	73	47	11	0	RBMS2	99	22	6	10	1	U2AF1	547	22	6	10	1
CDX2	555	4	1	11	0	GLE1	115	56	26	11	0	MRPL52	339	73	47	11	0	RBMS3	99	22	6	10	1	U2AF1L4	547	22	6	10	1
CEBPA	137	13	3	11	0	GLI1	374	13	3	11	0	MRPL53	338	73	47	11	0	RBMX	609	132	0	11	0	U2AF2	99	22	6	10	1
CEBPB	137	13	3	11	0	GLI2	374	13	3	11	0	MRPL54	339	73	47	11	0	RBMX2	477	144	9	11	0	UBA2	293	59	9	11	0
CEBPD	137	13	3	11	0	GLI3	374	13	3	11	0	MRPL55	339	73	47	11	0	RBMXL1	609	132	0	11	0	UBA3	293	59	9	11	0
CEBPE	137	13	3	11	0	GLIPR1L1	197	64	9	11	0	MRPL57	340	75	34	11	0	RBPJ	49	15	9	11	0	UBA52	170	59	9	11	0
CEBPG	31	12	9	11	0	GLMN	263	56	26	11	0	MRPL58	334	71	21	11	0	RBPJL	49	15	9	11	0	UBAC1	464	103	33	7	1
CEBPZ	123	147	49	7	1	GLRA1	205	59	9	11	0	MRPL9	334	71	21	11	0	RBX1	70	22	6	10	1	UBAP1	418	67	20	11	0
CECR2	524	13	3	11	0	GLRA2	205	59	9	11	0	MRPS10	333	71	21	11	0	RCC1	25	133	50	6	1	UBAP1L	136	150	52	11	0
CELF1	99	22	6	10	1	GLRA3	205	59	9	11	0	MRPS11	341	71	21	11	0	RCHY1	23	9	9	11	0	UBAP2L	301	55	28	5	1
CELF2	99	22	6	10	1	GLRB	205	59	9	11	0	MRPS12	334	71	21	11	0	RCOR1	479	70	20	11	0	UBE2A	624	142	4	10	1
CELF3	99	22	6	10	1	GLRX3	213	52	16	6	1	MRPS14	341	71	21	11	0	RCOR2	479	70	20	11	0	UBE2B	624	142	4	10	1
CELF4	99	22	6	10	1	GLRX5	213	52	16	6	1	MRPS15	339	73	47	11	0	RCOR3	479	70	20	11	0	UBE2C	625	143	48	5	1
CELF5	99	22	6	10	1	GLUL	413	59	9	11	0	MRPS16	341	71	21	11	0	REC8	478	141	26	11	0	UBE2D1	624	142	4	10	1
CELF6	99	22	6	10	1	GLYR1	59	15	9	11	0	MRPS17	334	71	21	11	0	RECK	206	106	9	11	0	UBE2D2	624	142	4	10	1
CENPA	100	23	7	1	1	GMNN	457	16	26	11	0	MRPS18A	341	71	21	11	0	RECQL5	170	59	9	11	0	UBE2D4	624	142	4	10	1
CENPF	101	24	26	11	0	GMPR	413	59	9	11	0	MRPS18B	341	71	21	11	0	REL	374	13	3	11	0	UBE2E1	624	142	4	10	1
CENPJ	84	150	52	11	0	GMPR2	413	59	9	11	0	MRPS18C	341	71	21	11	0	RELA	374	13	3	11	0	UBE2I	625	143	48	5	1
CENPO	233	75	34	11	0	GNA11	215	49	14	5	1	MRPS2	159	38	10	8	1	RELB	374	13	3	11	0	UBE2J2	625	143	48	5	1
CENPQ	233	75	34	11	0	GNA12	214	59	9	11	0	MRPS21	334	71	21	11	0	RELN	593	69	9	11	0	UBE2L3	625	143	48	5	1

CENPS	579	36	20	11	0	GNA13	214	59	9	11	0	MRPS22	338	73	47	11	0	REPIN1	270	15	9	11	0	UBE2L6	625	143	48	5	1
CENPX	100	23	7	1	1	GNA14	215	49	14	5	1	MRPS23	338	73	47	11	0	RERE	479	70	20	11	0	UBE2N	625	143	48	5	1
CEP128	395	64	9	11	0	GNA15	214	59	9	11	0	MRPS24	338	73	47	11	0	REST	30	150	52	11	0	UBE2S	625	143	48	5	1
CEP152	60	141	26	11	0	GNAI1	215	49	14	5	1	MRPS25	338	73	47	11	0	RET	322	69	9	11	0	UBE2V1	625	143	48	5	1
CEP164	187	144	9	11	0	GNAI2	215	49	14	5	1	MRPS26	339	73	47	11	0	REV3L	319	150	52	11	0	UBE2V2	625	143	48	5	1
CEP170	639	15	9	11	0	GNAI3	215	49	14	5	1	MRPS27	338	73	47	11	0	RFC1	59	15	9	11	0	UBE3A	626	82	26	11	0
CEP192	58	150	52	11	0	GNAL	215	49	14	5	1	MRPS28	179	43	19	5	1	RFC2	480	59	9	11	0	UBE3C	626	82	26	11	0
CEP250	101	24	26	11	0	GNAO1	215	49	14	5	1	MRPS30	336	73	47	11	0	RFC3	480	59	9	11	0	UBE3D	388	83	26	11	0
CEP290	395	64	9	11	0	GNAQ	215	49	14	5	1	MRPS31	338	73	47	11	0	RFC4	480	59	9	11	0	UBE4A	627	82	26	11	0
CEP76	418	67	20	11	0	GNAS	331	75	34	11	0	MRPS33	340	75	34	11	0	RFC5	480	59	9	11	0	UBE4B	627	82	26	11	0
CEP83	395	64	9	11	0	GNAT1	214	59	9	11	0	MRPS34	340	75	34	11	0	RFTN1	623	141	26	11	0	UBL4A	253	64	9	11	0
CEP89	60	141	26	11	0	GNAT2	214	59	9	11	0	MRPS35	333	71	21	11	0	RFX3	528	22	6	10	1	UBQLN1	464	103	33	7	1
CEP97	49	15	9	11	0	GNAZ	214	59	9	11	0	MRPS36	338	73	47	11	0	RFX5	30	150	52	11	0	UBQLN4	464	103	33	7	1
CETN2	354	57	9	11	0	GNB1	216	108	36	5	1	MRPS5	341	71	21	11	0	RFXANK	27	69	9	11	0	UBR1	627	82	26	11	0
CETN3	354	57	9	11	0	GNB2	216	108	36	5	1	MRPS6	501	71	21	11	0	RFXAP	524	13	3	11	0	UBR2	627	82	26	11	0
CETP	268	69	9	11	0	GNB3	216	108	36	5	1	MRPS7	334	71	21	11	0	RGP1	610	139	26	11	0	UBR3	627	82	26	11	0
CFAP206	481	64	9	11	0	GNB4	216	108	36	5	1	MRPS9	341	71	21	11	0	RGPD1	481	64	9	11	0	UBR5	627	82	26	11	0
CFAP70	45	83	26	11	0	GNB5	216	108	36	5	1	MRTO4	413	59	9	11	0	RGPD2	481	64	9	11	0	UBXN1	628	103	33	7	1
CFB	181	17	9	11	0	GNG10	217	16	26	11	0	MS4A2	291	62	18	11	0	RGPD3	481	64	9	11	0	UBXN6	629	103	33	7	1
CFDP1	605	79	26	11	0	GNG11	217	16	26	11	0	MSH2	315	59	9	11	0	RGPD4	481	64	9	11	0	UBXN7	308	32	9	11	0
CFH	102	69	9	11	0	GNG12	217	16	26	11	0	MSH3	315	59	9	11	0	RGPD5	481	64	9	11	0	UBXN8	629	103	33	7	1
CFHR1	102	69	9	11	0	GNG2	217	16	26	11	0	MSH6	59	15	9	11	0	RGPD8	481	64	9	11	0	UCHL5	633	145	49	7	1
CFHR2	102	69	9	11	0	GNG3	217	16	26	11	0	MSI1	99	22	6	10	1	RGS11	473	15	9	11	0	UEVLD	293	59	9	11	0
CFHR5	102	69	9	11	0	GNG4	217	16	26	11	0	MSI2	99	22	6	10	1	RGS4	482	59	9	11	0	UFL1	229	16	26	11	0
CFLAR	74	59	9	11	0	GNG5	217	16	26	11	0	MSL1	30	150	52	11	0	RGS7	537	125	49	7	1	UGGT1	64	17	9	11	0
CFTR	581	59	9	11	0	GNG7	217	16	26	11	0	MSL2	639	15	9	11	0	RHBDD2	481	64	9	11	0	UGGT2	64	17	9	11	0
CGA	304	80	24	11	0	GNG8	217	16	26	11	0	MSL3	330	125	49	7	1	RHEB	250	25	9	11	0	UGT3A1	596	64	9	11	0
CGB3	295	106	9	11	0	GNGT2	217	16	26	11	0	MSR1	526	17	9	11	0	RIC1	388	83	26	11	0	UGT3A2	596	64	9	11	0
CGN	101	24	26	11	0	GNL3L	645	79	26	11	0	MST1R	424	69	9	11	0	RICTOR	123	147	49	7	1	UIMC1	30	150	52	11	0
CGNL1	558	64	9	11	0	GNPTG	252	17	9	11	0	MSX1	11	4	1	11	0	RILP	473	15	9	11	0	ULK1	49	15	9	11	0
CHAF1A	558	64	9	11	0	GOLGA7	331	75	34	11	0	MSX2	11	4	1	11	0	RING1	405	59	9	11	0	UNC80	319	150	52	11	0
CHAF1B	103	133	50	6	1	GOLGA7B	457	16	26	11	0	MT-ATP6	343	74	9	11	0	RINT1	119	141	26	11	0	UNK	11	4	1	11	0
CHCHD1	370	73	47	11	0	GOLT1B	114	7	9	11	0	MT-ATP8	343	74	9	11	0	RIOK1	483	109	37	5	1	UPF1	170	59	9	11	0
CHCHD10	107	27	8	2	1	GON7	218	75	34	11	0	MT-CO1	342	59	9	11	0	RIOK2	483	109	37	5	1	UPF2	567	83	26	11	0
CHCHD3	605	79	26	11	0	GOPC	403	15	9	11	0	MT-CO2	343	74	9	11	0	RIOK3	483	109	37	5	1	UPF3A	630	144	9	11	0
CHCHD6	605	79	26	11	0	GOSR1	572	16	26	11	0	MT-CO3	342	59	9	11	0	RIOX2	59	15	9	11	0	UPF3B	630	144	9	11	0
CHD3	59	15	9	11	0	GOSR2	572	16	26	11	0	MT-CYB	342	59	9	11	0	RIPK1	253	64	9	11	0	UPK1A	631	51	9	11	0
CHD4	59	15	9	11	0	GP1BA	219	37	9	11	0	MT-ND1	343	74	9	11	0	RIPK2	484	69	9	11	0	UQCR10	125	73	47	11	0
CHD5	59	15	9	11	0	GP1BB	221	51	9	11	0	MT-ND2	343	74	9	11	0	RIPK3	485	64	9	11	0	UQCR11	125	73	47	11	0
CHD8	21	83	26	11	0	GP5	221	51	9	11	0	MT-ND3	343	74	9	11	0	RLIM	138	31	50	6	1	UQCRB	338	73	47	11	0
CHEK1	104	25	9	11	0	GP9	221	51	9	11	0	MT-ND4	343	74	9	11	0	RMI1	586	150	52	11	0	UQCRC1	632	118	9	11	0
CHM	105	26	26	11	0	GPAA1	590	59	9	11	0	MT-ND4L	343	74	9	11	0	RMI2	486	16	26	11	0	UQCRC2	632	118	9	11	0
CHML	105	26	26	11	0	GPATCH1	635	79	26	11	0	MT-ND5	343	74	9	11	0	RMND5A	209	99	9	11	0	UQCRFS1	341	71	21	11	0
CHMP1A	106	27	8	2	1	GPBAR1	221	51	9	11	0	MT-ND6	343	74	9	11	0	RMND5B	209	99	9	11	0	UQCRH	370	73	47	11	0
CHMP1B	106	27	8	2	1	GPD1	293	59	9	11	0	MTA1	344	70	20	11	0	RNASEH2A	586	150	52	11	0	UQCRHL	370	73	47	11	0
CHMP2A	106	27	8	2	1	GPD1L	293	59	9	11	0	MTA2	344	70	20	11	0	RNASEH2B	487	110	20	11	0	UQCRQ	368	73	47	11	0
CHMP2B	106	27	8	2	1	GPD2	413	59	9	11	0	MTA3	344	70	20	11	0	RNASEH2C	331	75	34	11	0	URI1	157	36	20	11	0
CHMP3	106	27	8	2	1	GPIHBP1	304	80	24	11	0	MTCP1	345	75	34	11	0	RNASEK	470	106	9	11	0	USE1	572	16	26	11	0
CHMP4A	107	27	8	2	1	GPKOW	586	150	52	11	0	MTERF4	253	64	9	11	0	RNF11	488	111	38	7	1	USF1	556	10	9	11	0
CHMP4B	107	27	8	2	1	GPN3	220	50	15	5	1	MTMR9	346	76	23	7	1	RNF111	489	112	39	11	0	USH1C	297	59	9	11	0
CHMP4C	107	27	8	2	1	GPR156	49	15	9	11	0	MTNR1A	347	51	9	11	0	RNF113A	534	79	26	11	0	USH2A	289	69	9	11	0
CHMP5	107	27	8	2	1	GPR160	347	51	9	11	0	MTOR	266	147	49	7	1	RNF125	49	15	9	11	0	USP10	49	15	9	11	0
CHMP6	107	27	8	2	1	GPR20	221	51	9	11	0	MTPN	27	69	9	11	0	RNF135	614	59	9	11	0	USP14	633	145	49	7	1
CHMP7	253	64	9	11	0	GPR37	221	51	9	11	0	MTTP	143	32	9	11	0	RNF139	23	9	9	11	0	USP22	633	145	49	7	1
CHP1	326	57	9	11	0	GPR37L1	221	51	9	11	0	MTX1	348	59	9	11	0	RNF14	23	9	9	11	0	USP28	633	145	49	7	1
CHRAC1	579	36	20	11	0	GPR61	221	51	9	11	0	MTX2	112	32	9	11	0	RNF144A	23	9	9	11	0	USP33	49	15	9	11	0
CHRFAM7A	109	59	9	11	0	GPR62	221	51	9	11	0	MTX3	112	32	9	11	0	RNF144B	23	9	9	11	0	USP39	633	145	49	7	1
CHRNA1	108	59	9	11	0	GPR63	347	51	9	11	0	MUS81	586	150	52	11	0	RNF165	489	112	39	11	0	USP47	49	15	9	11	0
CHRNA2	108	59	9	11	0	GPR84	347	51	9	11	0	MUSK	491	69	9	11	0	RNF168	49	15	9	11	0	USP7	633	145	49	7	1
CHRNA3	108	59	9	11	0	GPRC5A	222	51	9	11	0	MVB12A	349	75	34	11	0	RNF19A	212	58	48	5	1	UTP11	634	79	26	11	0
CHRNA4	108	59	9	11	0	GPRC5B	222	51	9	11	0	MVB12B	349	75	34	11	0	RNF19B	212	58	48	5	1	UTP14A	635	79	26	11	0
CHRNA5	108	59	9	11	0	GPRC5C	222	51	9	11	0	MVP	42	99	9	11	0	RNF2	405	59	9	11	0	UTP14C	635	79	26	11	0
CHRNA7	109	59	9	11	0	GPRC5D	222	51	9	11	0	MX2	280	59	9	11	0	RNF20	240	16	26	11	0	UTP18	645	79	26	11	0
CHRNA9	109	59	9	11	0	GPS1	454	99	9	11	0	MXD1	31	12	9	11	0	RNF215	400	15	9	11	0	UTP20	620	59	9	11	0
CHRNB1	109	59	9	11	0	GPS2	524	13	3	11	0	MXD3	583	11	1	11	0	RNF217	212	58	48	5	1	UTP23	194	48	13	5	1
CHRNB2	108	59	9	11	0	GPSM1	223	59	9	11	0	MXI1	31	12	9	11	0	RNF220	490	59	9	11	0	UTP3	376	79	26	11	0
CHRNB4	108	59	9	11	0	GPSM2	223	59	9	11	0	MYB	93	59	9	11	0	RNF31	212	58	48	5	1	UTP4	642	79	26	11	0
CHRND	109	59	9	11	0	GPX4	224	52	16	6	1	MYBBP1A	620	59	9	11	0	RNF40	240	16	26	11	0	UTP6	619	99	9	11	0
CHRNE	109	59	9	11	0	GRAP	225	59	9	11	0	MYBL2	93	59	9	11	0	RNF7	70	22	6	10	1	UTRN	560	59	9	11	0
CHTF18	397	86	9	11	0	GRB10	549	15	9	11	0	MYC	375	12	9	11	0	RNF8	253	64	9	11	0	UTY	93	59	9	11	0
CHTF8	157	36	20	11	0	GRB2	225	59	9	11	0	MYD88	350	69	9	11	0	RNH1	358	59	9	11	0	UVRAG	34	16	26	11	0
CHTOP	639	15	9	11	0	GRHPR	77	64	9	11	0	MYEF2	351	55	28	5	1	RNMT	49	15	9	11	0	UXS1	636	59	9	11	0
CHUK	253	64	9	11	0	GRIA1	226	59	9	11	0	MYF5	373	12	9	11	0	RNPC3	418	67	20	11	0	UXT	412	16	26	11	0
CIAO1	25	133	50	6	1	GRIA2	226	59	9	11	0	MYF6	373	12	9	11	0	RNPS1	609	132	0	11	0	VAC14	567	83	26	11	0
CIR1	477	144	9	11	0	GRIA3	226	59	9	11	0	MYH1	353	49	14	5	1	ROGDI	155	150	52	11	0	VAMP1	572	16	26	11	0
CIRBP	110	55	28	5	1	GRIA4	226	59	9	11	0	MYH10	352	59	9	11	0	ROMO1	594	136	46	13	1	VAMP2	572	16	26	11	0
CISD2	398	62	18	11	0	GRID2	491	69	9	11	0	MYH11	352	59	9	11	0	ROR1	491	69	9	11	0	VAMP3	572	16	26	11	0
CISH	552	129	9	11	0	GRIK2	226	59	9	11	0	MYH13	353	49	14	5	1	ROR2	491	69	9	11	0	VAMP4	572	16	26	11	0
CITED2	48	16	26	11	0	GRIK3	226	59	9	11	0	MYH14	352	59	9	11	0	ROS1	424	69	9	11	0	VAMP7	572	16	26	11	0
CKAP5	360	59	9	11	0	GRIK4	227	59	9	11	0	MYH15	356	59	9	11	0	RP9	477	144	9	11	0	VAMP8	572	16	26	11	0

CKS1B	341	71	21	11	0	GRIK5	227	59	9	11	0	MYH2	353	49	14	5	1	RPA1	405	59	9	11	0	VBP1	412	16	26	11	0
CKS2	341	71	21	11	0	GRIN1	227	59	9	11	0	MYH3	353	49	14	5	1	RPA2	315	59	9	11	0	VCAM1	409	37	9	11	0
CLCC1	134	16	26	11	0	GRIN2A	226	59	9	11	0	MYH4	353	49	14	5	1	RPA3	492	83	26	11	0	VCL	177	141	26	11	0
CLCF1	382	80	24	11	0	GRIN2B	226	59	9	11	0	MYH6	353	49	14	5	1	RPA4	481	64	9	11	0	VCP	453	59	9	11	0
CLCN1	111	59	9	11	0	GRIN2C	227	59	9	11	0	MYH7	353	49	14	5	1	RPAP1	620	59	9	11	0	VCPKMT	324	69	9	11	0
CLCN2	111	59	9	11	0	GRIN2D	227	59	9	11	0	MYH7B	356	59	9	11	0	RPAP2	317	150	52	11	0	VDAC1	637	118	9	11	0
CLCN7	111	59	9	11	0	GRIN3A	491	69	9	11	0	MYH8	353	49	14	5	1	RPAP3	362	15	9	11	0	VDAC2	637	118	9	11	0
CLCNKA	111	59	9	11	0	GRIN3B	227	59	9	11	0	MYH9	352	59	9	11	0	RPF1	333	71	21	11	0	VDAC3	637	118	9	11	0
CLCNKB	111	59	9	11	0	GRM1	491	69	9	11	0	MYL1	354	57	9	11	0	RPH3A	493	15	9	11	0	VDR	385	22	6	10	1
CLDN1	68	59	9	11	0	GRM7	491	69	9	11	0	MYL12A	354	57	9	11	0	RPL10	332	71	21	11	0	VEGFA	407	106	9	11	0
CLDN17	68	59	9	11	0	GRPEL1	175	59	9	11	0	MYL12B	354	57	9	11	0	RPL10A	248	59	9	11	0	VEZT	84	150	52	11	0
CLDN3	68	59	9	11	0	GRPEL2	175	59	9	11	0	MYL2	354	57	9	11	0	RPL10L	332	71	21	11	0	VIM	366	59	9	11	0
CLDN4	68	59	9	11	0	GRSF1	175	59	9	11	0	MYL3	354	57	9	11	0	RPL11	501	71	21	11	0	VIPAS39	481	64	9	11	0
CLEC12B	90	37	9	11	0	GRWD1	328	79	26	11	0	MYL4	354	57	9	11	0	RPL12	508	115	41	10	1	VIPR1	7	59	9	11	0
CLIC1	112	32	9	11	0	GSC	11	4	1	11	0	MYL5	354	57	9	11	0	RPL13	497	114	40	10	1	VLDLR	298	64	9	11	0
CLIC2	112	32	9	11	0	GSDMD	264	56	26	11	0	MYL6	354	57	9	11	0	RPL13A	501	71	21	11	0	VPS11	638	99	9	11	0
CLIC3	112	32	9	11	0	GSK3A	96	59	9	11	0	MYL6B	354	57	9	11	0	RPL14	494	113	40	10	1	VPS16	638	99	9	11	0
CLIC4	112	32	9	11	0	GSK3B	96	59	9	11	0	MYL7	354	57	9	11	0	RPL15	496	71	21	11	0	VPS18	638	99	9	11	0
CLIC5	112	32	9	11	0	GSPT1	228	59	9	11	0	MYL9	354	57	9	11	0	RPL17	496	71	21	11	0	VPS25	124	99	9	11	0
CLIC6	540	36	20	11	0	GSPT2	228	59	9	11	0	MYO10	356	59	9	11	0	RPL18	496	71	21	11	0	VPS26A	610	139	26	11	0
CLINT1	171	15	9	11	0	GSTP1	112	32	9	11	0	MYO15A	355	77	24	11	0	RPL18A	501	71	21	11	0	VPS26B	610	139	26	11	0
CLIP2	558	64	9	11	0	GTF2A1	556	10	9	11	0	MYO15B	473	15	9	11	0	RPL19	497	114	40	10	1	VPS28	119	141	26	11	0
CLMN	49	15	9	11	0	GTF2A1L	521	150	52	11	0	MYO16	356	59	9	11	0	RPL21	494	113	40	10	1	VPS29	249	99	9	11	0
CLNK	553	130	20	11	0	GTF2A2	331	75	34	11	0	MYO18A	356	59	9	11	0	RPL22	501	71	21	11	0	VPS33A	454	99	9	11	0
CLNS1A	157	36	20	11	0	GTF2B	82	13	3	11	0	MYO18B	49	15	9	11	0	RPL22L1	501	71	21	11	0	VPS33B	454	99	9	11	0
CLOCK	379	10	9	11	0	GTF2E1	639	15	9	11	0	MYO19	356	59	9	11	0	RPL23	334	71	21	11	0	VPS35	567	83	26	11	0
CLP1	113	28	15	5	1	GTF2E2	229	16	26	11	0	MYO1A	356	59	9	11	0	RPL23A	497	114	40	10	1	VPS36	124	99	9	11	0
CLPP	248	59	9	11	0	GTF2F1	605	79	26	11	0	MYO1B	356	59	9	11	0	RPL24	495	114	40	10	1	VPS37A	321	16	26	11	0
CLPX	315	59	9	11	0	GTF2F2	124	99	9	11	0	MYO1C	356	59	9	11	0	RPL26	494	113	40	10	1	VPS37B	321	16	26	11	0
CLSTN3	268	69	9	11	0	GTF2H1	60	141	26	11	0	MYO1D	356	59	9	11	0	RPL26L1	494	113	40	10	1	VPS37C	321	16	26	11	0
CLTA	605	79	26	11	0	GTF2H2	265	83	26	11	0	MYO1E	356	59	9	11	0	RPL27	494	113	40	10	1	VPS37D	473	15	9	11	0
CLTB	605	79	26	11	0	GTF2H2C	265	83	26	11	0	MYO1F	356	59	9	11	0	RPL27A	496	71	21	11	0	VPS39	638	99	9	11	0
CLTC	249	99	9	11	0	GTF2H3	265	83	26	11	0	MYO1G	356	59	9	11	0	RPL28	334	71	21	11	0	VPS41	638	99	9	11	0
CLTCL1	249	99	9	11	0	GTF2H4	124	99	9	11	0	MYO3A	356	59	9	11	0	RPL29	497	114	40	10	1	VPS4A	453	59	9	11	0
CLU	20	8	24	11	0	GTF2H5	331	75	34	11	0	MYO3B	356	59	9	11	0	RPL3	332	71	21	11	0	VPS4B	453	59	9	11	0
CLUAP1	251	16	26	11	0	GTF3C1	229	16	26	11	0	MYO5A	356	59	9	11	0	RPL30	501	71	21	11	0	VPS50	177	141	26	11	0
CNEP1R1	368	73	47	11	0	GTF3C2	578	141	26	11	0	MYO5B	356	59	9	11	0	RPL31	497	114	40	10	1	VPS51	176	15	9	11	0
CNGA1	271	59	9	11	0	GTF3C3	619	99	9	11	0	MYO5C	356	59	9	11	0	RPL32	496	71	21	11	0	VPS52	119	141	26	11	0
CNGA3	271	59	9	11	0	GTF3C4	481	64	9	11	0	MYO6	356	59	9	11	0	RPL34	341	71	21	11	0	VPS53	119	141	26	11	0
CNGA4	271	59	9	11	0	GTF3C5	320	56	26	11	0	MYO7A	356	59	9	11	0	RPL35	497	114	40	10	1	VPS54	177	141	26	11	0
CNGB1	271	59	9	11	0	GTF3C6	157	36	20	11	0	MYO7B	356	59	9	11	0	RPL35A	334	71	21	11	0	VPS72	428	79	26	11	0
CNGB3	271	59	9	11	0	GTPBP1	228	59	9	11	0	MYO9A	356	59	9	11	0	RPL36	497	114	40	10	1	VPS8	638	99	9	11	0
CNIH2	114	7	9	11	0	GUCY1A2	230	69	9	11	0	MYO9B	356	59	9	11	0	RPL36A	498	114	40	10	1	VRK2	253	64	9	11	0
CNIH3	114	7	9	11	0	HADHA	293	59	9	11	0	MYOD1	373	12	9	11	0	RPL36AL	498	114	40	10	1	VSIG4	409	37	9	11	0
CNNM4	271	59	9	11	0	HADHB	231	53	17	8	1	MYOG	373	12	9	11	0	RPL37	498	114	40	10	1	VTI1A	572	16	26	11	0
CNOT1	115	56	26	11	0	HAND1	583	11	1	11	0	MYRIP	49	15	9	11	0	RPL37A	498	114	40	10	1	VTI1B	572	16	26	11	0
CNOT10	619	99	9	11	0	HAND2	583	11	1	11	0	MYSM1	49	15	9	11	0	RPL38	501	71	21	11	0	VTN	252	17	9	11	0
CNOT11	115	56	26	11	0	HAT1	232	54	26	11	0	MZB1	470	106	9	11	0	RPL39	499	114	40	10	1	WAC	129	55	28	5	1
CNOT2	116	59	9	11	0	HAUS1	233	75	34	11	0	MZT1	542	16	26	11	0	RPL39L	499	114	40	10	1	WASF1	639	15	9	11	0
CNOT3	537	125	49	7	1	HAUS2	233	75	34	11	0	MZT2A	586	150	52	11	0	RPL3L	332	71	21	11	0	WASF2	639	15	9	11	0
CNOT4	126	22	6	10	1	HAUS3	233	75	34	11	0	MZT2B	586	150	52	11	0	RPL4	496	71	21	11	0	WASF3	639	15	9	11	0
CNOT6	123	147	49	7	1	HAUS4	233	75	34	11	0	N4BP2L2	521	150	52	11	0	RPL41	500	96	24	11	0	WASHC1	639	15	9	11	0
CNOT6L	123	147	49	7	1	HAUS5	233	75	34	11	0	N6AMT1	324	69	9	11	0	RPL5	501	71	21	11	0	WASHC2A	573	15	9	11	0
CNOT7	541	59	9	11	0	HAUS6	84	150	52	11	0	NAA10	636	59	9	11	0	RPL6	494	113	40	10	1	WASHC2C	573	15	9	11	0
CNOT8	541	59	9	11	0	HAUS7	233	75	34	11	0	NAA15	619	99	9	11	0	RPL7	501	71	21	11	0	WASHC3	542	16	26	11	0
CNOT9	649	147	49	7	1	HAUS8	233	75	34	11	0	NAA16	619	99	9	11	0	RPL7A	496	71	21	11	0	WASHC4	363	56	26	11	0
CNPPD1	82	13	3	11	0	HAX1	321	16	26	11	0	NAA20	636	59	9	11	0	RPL7L1	501	71	21	11	0	WASHC5	363	56	26	11	0
CNST	117	150	52	11	0	HBA1	234	99	9	11	0	NAA25	619	99	9	11	0	RPL8	334	71	21	11	0	WBP11	639	15	9	11	0
CNTFR	158	37	9	11	0	HBB	234	99	9	11	0	NAA30	212	58	48	5	1	RPL9	508	115	41	10	1	WBP4	640	15	9	11	0
CNTNAP2	593	69	9	11	0	HBD	234	99	9	11	0	NAA35	363	56	26	11	0	RPLP0	413	59	9	11	0	WDR12	641	146	50	6	1
CNTRL	118	64	9	11	0	HBE1	234	99	9	11	0	NAA38	302	65	19	5	1	RPLP1	107	27	8	2	1	WDR18	35	31	50	6	1
COA6	407	106	9	11	0	HBG1	234	99	9	11	0	NAA50	636	59	9	11	0	RPLP2	107	27	8	2	1	WDR19	249	99	9	11	0
COG1	119	141	26	11	0	HBG2	234	99	9	11	0	NABP1	405	59	9	11	0	RPN1	143	32	9	11	0	WDR24	328	79	26	11	0
COG2	119	141	26	11	0	HBM	234	99	9	11	0	NABP2	405	59	9	11	0	RPN2	143	32	9	11	0	WDR26	582	133	50	6	1
COG3	119	141	26	11	0	HBQ1	234	99	9	11	0	NACA	357	120	34	11	0	RPP14	162	79	26	11	0	WDR3	642	79	26	11	0
COG4	119	141	26	11	0	HCFC1	99	22	6	10	1	NACA2	271	59	9	11	0	RPP21	586	150	52	11	0	WDR33	129	55	28	5	1
COG5	119	141	26	11	0	HCFC2	285	69	9	11	0	NACAD	540	36	20	11	0	RPP25	502	79	26	11	0	WDR35	249	99	9	11	0
COG6	119	141	26	11	0	HCLS1	573	15	9	11	0	NAE1	293	59	9	11	0	RPP25L	502	79	26	11	0	WDR36	642	79	26	11	0
COG7	119	141	26	11	0	HCN1	271	59	9	11	0	NAF1	524	13	3	11	0	RPP30	642	79	26	11	0	WDR4	643	79	26	11	0
COG8	176	15	9	11	0	HCN2	271	59	9	11	0	NAIP	358	59	9	11	0	RPP38	333	71	21	11	0	WDR41	644	79	26	11	0
COL10A1	122	17	9	11	0	HCN3	271	59	9	11	0	NAPA	454	99	9	11	0	RPP40	155	150	52	11	0	WDR46	645	79	26	11	0
COL11A1	120	17	9	11	0	HCN4	271	59	9	11	0	NAPB	454	99	9	11	0	RPRD1A	568	59	9	11	0	WDR5	582	133	50	6	1
COL11A2	120	17	9	11	0	HDAC1	271	59	9	11	0	NAPG	454	99	9	11	0	RPRD1B	568	59	9	11	0	WDR59	328	79	26	11	0
COL12A1	593	69	9	11	0	HDAC10	235	59	9	11	0	NASP	310	66	9	11	0	RPRD2	129	55	28	5	1	WDR5B	582	133	50	6	1
COL13A1	121	17	9	11	0	HDAC11	235	59	9	11	0	NAT10	359	30	15	5	1	RPS10	124	99	9	11	0	WDR61	641	146	50	6	1
COL14A1	593	69	9	11	0	HDAC2	271	59	9	11	0	NAT9	324	69	9	11	0	RPS11	504	71	21	11	0	WDR74	140	79	26	11	0
COL15A1	79	17	9	11	0	HDAC3	271	59	9	11	0	NBAS	481	64	9	11	0	RPS12	501	71	21	11	0	WDR77	140	79	26	11	0
COL16A1	122	17	9	11	0	HDAC4	236	59	9	11	0	NBN	49	15	9	11	0	RPS13	497	114	40	10	1	WDR82	216	108	36	5	1
COL17A1	526	17	9	11	0	HDAC5	236	59	9	11	0	NBR1	49	15	9	11	0	RPS14	503	115	41	10	1	WDR83	646	146	50	6	1
COL18A1	79	17	9	11	0	HDAC6	235	59	9	11	0	NCALD	273	57	9	11	0	RPS15	504	71	21	11	0	WDR83OS	398	62	18	11	0

COL19A1	122	17	9	11	0	HDAC7	236	59	9	11	0	NCAPD2	360	59	9	11	0	RPS15A	497	114	40	10	1	WDR93	481	64	9	11	0
COL1A1	121	17	9	11	0	HDAC8	235	59	9	11	0	NCAPD3	360	59	9	11	0	RPS16	332	71	21	11	0	WDTC1	138	31	50	6	1
COL1A2	121	17	9	11	0	HDAC9	236	59	9	11	0	NCAPG	360	59	9	11	0	RPS17	501	71	21	11	0	WHRN	403	15	9	11	0
COL20A1	170	59	9	11	0	HDGF	329	36	20	11	0	NCAPG2	360	59	9	11	0	RPS18	497	114	40	10	1	WIPI2	642	79	26	11	0
COL21A1	121	17	9	11	0	HDLBP	404	22	6	10	1	NCAPH	586	150	52	11	0	RPS19	124	99	9	11	0	WNK1	639	15	9	11	0
COL22A1	121	17	9	11	0	HEATR1	263	56	26	11	0	NCAPH2	586	150	52	11	0	RPS2	508	115	41	10	1	WNK4	271	59	9	11	0
COL23A1	121	17	9	11	0	HES1	373	12	9	11	0	NCBP1	567	83	26	11	0	RPS20	501	71	21	11	0	WNT2	289	69	9	11	0
COL24A1	79	17	9	11	0	HES6	583	11	1	11	0	NCBP2	547	22	6	10	1	RPS21	505	75	34	11	0	WNT3	289	69	9	11	0
COL25A1	121	17	9	11	0	HEXA	268	69	9	11	0	NCBP3	361	36	20	11	0	RPS23	504	71	21	11	0	WNT3A	289	69	9	11	0
COL26A1	79	17	9	11	0	HEXB	268	69	9	11	0	NCCRP1	526	17	9	11	0	RPS24	497	114	40	10	1	WRAP53	138	31	50	6	1
COL27A1	79	17	9	11	0	HEXIM1	30	150	52	11	0	NCF1	362	15	9	11	0	RPS25	506	116	42	10	1	WRAP73	25	133	50	6	1
COL28A1	121	17	9	11	0	HEY2	373	12	9	11	0	NCF2	362	15	9	11	0	RPS26	341	71	21	11	0	WTAP	558	64	9	11	0
COL2A1	121	17	9	11	0	HFE	238	59	9	11	0	NCF4	362	15	9	11	0	RPS27	507	117	43	10	1	WTIP	270	15	9	11	0
COL3A1	121	17	9	11	0	HGS	568	59	9	11	0	NCK1	225	59	9	11	0	RPS27A	170	59	9	11	0	WWC1	49	15	9	11	0
COL4A1	122	17	9	11	0	HHEX	11	4	1	11	0	NCKAP1	363	56	26	11	0	RPS27L	507	117	43	10	1	WWOX	59	15	9	11	0
COL4A2	122	17	9	11	0	HIF1A	379	10	9	11	0	NCKAP1L	363	56	26	11	0	RPS28	504	71	21	11	0	WWP1	541	59	9	11	0
COL4A3	122	17	9	11	0	HIGD1A	125	73	47	11	0	NCL	364	55	28	5	1	RPS29	498	114	40	10	1	WWP2	541	59	9	11	0
COL4A4	122	17	9	11	0	HINT1	293	59	9	11	0	NCLN	590	59	9	11	0	RPS3	508	115	41	10	1	WWTR1	365	10	9	11	0
COL4A5	122	17	9	11	0	HIPK2	639	15	9	11	0	NCOA1	365	10	9	11	0	RPS3A	501	71	21	11	0	XAB2	619	99	9	11	0
COL4A6	122	17	9	11	0	HIRA	103	133	50	6	1	NCOA2	365	10	9	11	0	RPS4X	497	114	40	10	1	XBP1	647	123	3	11	0
COL5A1	120	17	9	11	0	HLA-A	237	59	9	11	0	NCOA3	365	10	9	11	0	RPS4Y1	497	114	40	10	1	XPA	542	16	26	11	0
COL5A2	121	17	9	11	0	HLA-B	237	59	9	11	0	NCOA6	48	16	26	11	0	RPS5	496	71	21	11	0	XPC	648	135	26	11	0
COL5A3	121	17	9	11	0	HLA-C	237	59	9	11	0	NCOR1	479	70	20	11	0	RPS6	332	71	21	11	0	XPO1	123	147	49	7	1
COL6A1	121	17	9	11	0	HLA-DMA	238	59	9	11	0	NCOR2	479	70	20	11	0	RPS6KL1	473	15	9	11	0	XPO4	649	147	49	7	1
COL6A2	121	17	9	11	0	HLA-DMB	238	59	9	11	0	NCR1	237	59	9	11	0	RPS7	605	79	26	11	0	XPO5	649	147	49	7	1
COL6A3	312	69	9	11	0	HLA-DOA	238	59	9	11	0	NCSTN	590	59	9	11	0	RPS8	496	71	21	11	0	XPO6	649	147	49	7	1
COL6A6	312	69	9	11	0	HLA-DOB	239	59	9	11	0	NDC1	275	106	9	11	0	RPS9	497	114	40	10	1	XPO7	649	147	49	7	1
COL7A1	593	69	9	11	0	HLA-DPA1	238	59	9	11	0	NDC80	251	16	26	11	0	RPSA	249	99	9	11	0	XPOT	266	147	49	7	1
COL8A1	122	17	9	11	0	HLA-DPB1	239	59	9	11	0	NDE1	366	59	9	11	0	RPTOR	123	147	49	7	1	XRCC1	586	150	52	11	0
COL8A2	122	17	9	11	0	HLA-DQA1	238	59	9	11	0	NDEL1	366	59	9	11	0	RRAGA	509	9	9	11	0	XRCC2	466	69	9	11	0
COL9A1	122	17	9	11	0	HLA-DQA2	238	59	9	11	0	NDN	306	13	3	11	0	RRAGB	509	9	9	11	0	XRCC3	466	69	9	11	0
COL9A2	122	17	9	11	0	HLA-DQB1	239	59	9	11	0	NDUFA1	340	75	34	11	0	RRAGC	509	9	9	11	0	XRCC4	49	15	9	11	0
COL9A3	122	17	9	11	0	HLA-DQB2	239	59	9	11	0	NDUFA10	299	69	9	11	0	RRAGD	509	9	9	11	0	XRCC5	265	83	26	11	0
COLEC10	526	17	9	11	0	HLA-DRA	238	59	9	11	0	NDUFA11	594	136	46	13	1	RRM1	413	59	9	11	0	XRCC6	265	83	26	11	0
COLEC11	526	17	9	11	0	HLA-DRB1	239	59	9	11	0	NDUFA12	157	36	20	11	0	RRM2	41	59	9	11	0	YAP1	365	10	9	11	0
COLEC12	526	17	9	11	0	HLA-DRB5	239	59	9	11	0	NDUFA13	125	73	47	11	0	RRM2B	41	59	9	11	0	YBX1	301	55	28	5	1
COLQ	79	17	9	11	0	HLA-E	237	59	9	11	0	NDUFA2	338	73	47	11	0	RRP1	510	59	9	11	0	YEATS2	415	10	9	11	0
COMMD1	481	64	9	11	0	HLA-F	237	59	9	11	0	NDUFA3	368	73	47	11	0	RRP15	605	79	26	11	0	YEATS4	625	143	48	5	1
COMP	289	69	9	11	0	HLA-G	237	59	9	11	0	NDUFA4	125	73	47	11	0	RRP1B	510	59	9	11	0	YKT6	572	16	26	11	0
COPA	249	99	9	11	0	HLF	200	12	9	11	0	NDUFA4L2	125	73	47	11	0	RRP36	634	79	26	11	0	YPEL5	650	148	51	10	1
COPB1	123	147	49	7	1	HLTF	280	59	9	11	0	NDUFA5	339	73	47	11	0	RRP7A	418	67	20	11	0	YWHAQ	283	59	9	11	0
COPB2	249	99	9	11	0	HM13	13	59	9	11	0	NDUFA6	339	73	47	11	0	RRP8	586	150	52	11	0	YY1	254	59	9	11	0
COPE	619	99	9	11	0	HMGA1	241	135	26	11	0	NDUFA7	338	73	47	11	0	RRP9	642	79	26	11	0	YY2	254	59	9	11	0
COPG1	123	147	49	7	1	HMGA2	241	135	26	11	0	NDUFA8	370	73	47	11	0	RRS1	634	79	26	11	0	ZAP70	491	69	9	11	0
COPG2	123	147	49	7	1	HMGB1	59	15	9	11	0	NDUFA9	367	78	9	11	0	RSF1	157	36	20	11	0	ZBTB16	617	150	52	11	0
COPS2	454	99	9	11	0	HMGB2	59	15	9	11	0	NDUFAB1	341	71	21	11	0	RSL1D1	241	135	26	11	0	ZBTB17	49	15	9	11	0
COPS3	454	99	9	11	0	HMGCL	248	59	9	11	0	NDUFAF1	338	73	47	11	0	RSPH4A	157	36	20	11	0	ZBTB8OS	541	59	9	11	0
COPS4	454	99	9	11	0	HMGXB4	524	13	3	11	0	NDUFB1	368	73	47	11	0	RTCB	413	59	9	11	0	ZC3H12A	11	4	1	11	0
COPS5	455	59	9	11	0	HNF1A	528	22	6	10	1	NDUFB10	370	73	47	11	0	RTF1	605	79	26	11	0	ZC3H13	592	36	20	11	0
COPS6	165	39	9	11	0	HNF1B	528	22	6	10	1	NDUFB11	369	73	47	11	0	RUNX1	126	22	6	10	1	ZC3H14	49	15	9	11	0
COPS7A	124	99	9	11	0	HNRNPA0	110	55	28	5	1	NDUFB2	369	73	47	11	0	RUNX2	126	22	6	10	1	ZC3H18	210	15	9	11	0
COPS7B	124	99	9	11	0	HNRNPA1	110	55	28	5	1	NDUFB3	368	73	47	11	0	RUNX3	126	22	6	10	1	ZC3H3	30	150	52	11	0
COPS8	454	99	9	11	0	HNRNPA1L2	110	55	28	5	1	NDUFB4	368	73	47	11	0	RUVBL1	453	59	9	11	0	ZC3H8	651	149	20	11	0
COPS9	331	75	34	11	0	HNRNPA2B1	110	55	28	5	1	NDUFB5	369	73	47	11	0	RUVBL2	453	59	9	11	0	ZC4H2	542	16	26	11	0
COPZ1	15	6	2	10	1	HNRNPA3	110	55	28	5	1	NDUFB6	368	73	47	11	0	RWDD1	321	16	26	11	0	ZCCHC17	652	144	9	11	0
COPZ2	15	6	2	10	1	HNRNPAB	110	55	28	5	1	NDUFB7	370	73	47	11	0	RXRA	472	22	6	10	1	ZCCHC7	653	150	52	11	0
COQ3	175	59	9	11	0	HNRNPC	129	55	28	5	1	NDUFB8	369	73	47	11	0	RXRB	472	22	6	10	1	ZCCHC8	157	36	20	11	0
COQ4	338	73	47	11	0	HNRNPD	110	55	28	5	1	NDUFB9	338	73	47	11	0	RXRG	472	22	6	10	1	ZCRB1	544	132	0	11	0
COQ5	248	59	9	11	0	HNRNPF	175	59	9	11	0	NDUFC1	369	73	47	11	0	RYBP	49	15	9	11	0	ZDHHC9	271	59	9	11	0
COQ6	248	59	9	11	0	HNRNPH1	175	59	9	11	0	NDUFC2	368	73	47	11	0	RYK	491	69	9	11	0	ZER1	299	69	9	11	0
COQ7	125	73	47	11	0	HNRNPH2	175	59	9	11	0	NDUFS1	248	59	9	11	0	RYR1	269	83	26	11	0	ZFAND2A	23	9	9	11	0
COQ9	316	79	26	11	0	HNRNPH3	175	59	9	11	0	NDUFS2	413	59	9	11	0	RYR2	269	83	26	11	0	ZFC3H1	425	36	20	11	0
CORO1A	642	79	26	11	0	HNRNPK	404	22	6	10	1	NDUFS3	341	71	21	11	0	RYR3	269	83	26	11	0	ZFHX3	200	12	9	11	0
CORO1C	249	99	9	11	0	HNRNPL	547	22	6	10	1	NDUFS4	341	71	21	11	0	S100A1	326	57	9	11	0	ZFP36	655	151	9	11	0
CORO2A	642	79	26	11	0	HNRNPLL	547	22	6	10	1	NDUFS5	370	73	47	11	0	S100A10	326	57	9	11	0	ZFP36L1	374	13	3	11	0
CORT	382	80	24	11	0	HNRNPM	351	55	28	5	1	NDUFS6	341	71	21	11	0	SAA1	20	8	24	11	0	ZFP36L2	374	13	3	11	0
COX10	316	79	26	11	0	HNRNPR	175	59	9	11	0	NDUFS7	248	59	9	11	0	SAA2	20	8	24	11	0	ZFPM1	11	4	1	11	0
COX11	56	71	21	11	0	HNRNPU	59	15	9	11	0	NDUFS8	248	59	9	11	0	SAA4	20	8	24	11	0	ZFYVE9	654	150	52	11	0
COX15	125	73	47	11	0	HNRNPUL1	59	15	9	11	0	NDUFV1	248	59	9	11	0	SAC3D1	473	15	9	11	0	ZMAT2	534	79	26	11	0
COX4I1	125	73	47	11	0	HOOK1	240	16	26	11	0	NDUFV2	371	52	16	6	1	SAE1	293	59	9	11	0	ZMAT5	157	36	20	11	0
COX4I2	125	73	47	11	0	HOOK2	240	16	26	11	0	NDUFV3	339	73	47	11	0	SALL4	200	12	9	11	0	ZMPSTE24	13	59	9	11	0
COX5A	338	73	47	11	0	HOOK3	240	16	26	11	0	NECAP1	372	36	20	11	0	SAMM50	511	118	9	11	0	ZNF10	655	151	9	11	0
COX5B	341	71	21	11	0	HOXA10	11	4	1	11	0	NECAP2	372	36	20	11	0	SAP130	48	16	26	11	0	ZNF217	617	150	52	11	0
COX6A1	369	73	47	11	0	HOXA11	11	4	1	11	0	NEDD4	553	130	20	11	0	SAP18	405	59	9	11	0	ZNF224	655	151	9	11	0
COX6A2	369	73	47	11	0	HOXA9	11	4	1	11	0	NEK10	253	64	9	11	0	SAP30	512	15	9	11	0	ZNF326	129	55	28	5	1
COX6B1	370	73	47	11	0	HOXB9	11	4	1	11	0	NEK2	43	59	9	11	0	SAP30L	512	15	9	11	0	ZNF350	655	151	9	11	0
COX6B2	370	73	47	11	0	HOXC9	11	4	1	11	0	NEK6	283	59	9	11	0	SAR1A	513	119	9	11	0	ZNF622	525	79	26	11	0
COX6C	125	73	47	11	0	HOXD10	11	4	1	11	0	NELFA	131	150	52	11	0	SAR1B	513	119	9	11	0	ZNF703	555	4	1	11	0
COX7A1	125	73	47	11	0	HOXD12	11	4	1	11	0	NELFB	263	56	26	11	0	SARNP	573	15	9	11	0	ZNF827	617	150	52	11	0
COX7A2	125	73	47	11	0	HP	181	17	9	11	0	NELFCD	263	56	26	11	0	SART1	534	79	26	11	0	ZNF830	605	79	26	11	0

COX7A2L	368	73	47	11	0	HP1BP3	241	135	26	11	0	NELFE	129	55	28	5	1	SART3	619	99	9	11	0	ZNFX1	626	82	26	11	0
COX7B	368	73	47	11	0	HPR	181	17	9	11	0	NEMF	157	36	20	11	0	SASH1	49	15	9	11	0	ZNHIT1	605	79	26	11	0
COX7C	125	73	47	11	0	HPS1	87	20	26	11	0	NEO1	593	69	9	11	0	SASS6	514	16	26	11	0	ZNHIT3	605	79	26	11	0
COX8A	369	73	47	11	0	HPS3	481	64	9	11	0	NEUROD1	373	12	9	11	0	SATB2	528	22	6	10	1	ZNHIT6	157	36	20	11	0
COX8C	368	73	47	11	0	HPS4	623	141	26	11	0	NFAT5	374	13	3	11	0	SCAF8	129	55	28	5	1	ZNRF2	405	59	9	11	0
CPEB1	126	22	6	10	1	HPS5	388	83	26	11	0	NFATC1	374	13	3	11	0	SCAP	517	59	9	11	0	ZRSR2	630	144	9	11	0
CPEB2	126	22	6	10	1	HPS6	264	56	26	11	0	NFATC2	374	13	3	11	0	SCARA3	526	17	9	11	0	ZSWIM4	656	15	9	11	0
CPEB3	126	22	6	10	1	HR	242	123	3	11	0	NFATC3	374	13	3	11	0	SCIMP	233	75	34	11	0	ZSWIM5	656	15	9	11	0
CPEB4	126	22	6	10	1	HRAS	462	49	14	5	1	NFATC4	374	13	3	11	0	SCLT1	395	64	9	11	0	ZSWIM6	656	15	9	11	0
CPLX1	605	79	26	11	0	HSD17B10	77	64	9	11	0	NFE2	375	12	9	11	0	SCLY	561	59	9	11	0	ZSWIM7	642	79	26	11	0
CPLX2	605	79	26	11	0	HSD17B12	299	69	9	11	0	NFE2L2	375	12	9	11	0	SCML2	59	15	9	11	0	ZSWIM8	656	15	9	11	0
CPLX3	605	79	26	11	0	HSD17B8	77	64	9	11	0	NFE2L3	375	12	9	11	0	SCN10A	517	59	9	11	0	ZW10	119	141	26	11	0
CPS1	413	59	9	11	0	HSF1	385	22	6	10	1	NFE4	345	75	34	11	0	SCN11A	517	59	9	11	0	ZWILCH	155	150	52	11	0
CPSF1	249	99	9	11	0	HSP90AA1	243	59	9	11	0	NFIL3	375	12	9	11	0	SCN1A	516	59	9	11	0	ZYG11A	299	69	9	11	0
CPSF2	127	30	15	5	1	HSP90AB1	243	59	9	11	0	NFKB1	105	26	26	11	0	SCN1B	515	37	9	11	0	ZYG11B	299	69	9	11	0
CPSF3	127	30	15	5	1	HSP90B1	243	59	9	11	0	NFKB2	105	26	26	11	0	SCN2A	516	59	9	11	0	ZZZ3	558	64	9	11	0
CPSF4	128	29	20	11	0	HSPA1A	244	59	9	11	0	NFKBIA	27	69	9	11	0	SCN2B	515	37	9	11	0						
CPSF4L	128	29	20	11	0	HSPA1B	244	59	9	11	0	NFRKB	49	15	9	11	0	SCN3A	516	59	9	11	0						
CPSF6	129	55	28	5	1	HSPA1L	244	59	9	11	0	NFS1	561	59	9	11	0	SCN3B	515	37	9	11	0						
CPSF7	129	55	28	5	1	HSPA2	244	59	9	11	0	NFYA	556	10	9	11	0	SCN4A	517	59	9	11	0						
CPT1C	111	59	9	11	0	HSPA5	244	59	9	11	0	NFYB	579	36	20	11	0	SCN4B	515	37	9	11	0						
CR1L	102	69	9	11	0	HSPA8	244	59	9	11	0	NFYC	556	10	9	11	0	SCN5A	517	59	9	11	0						
CR2	102	69	9	11	0	HSPA9	244	59	9	11	0	NGDN	376	79	26	11	0	SCN7A	517	59	9	11	0						
CRADD	350	69	9	11	0	HSPB1	130	59	9	11	0	NHEJ1	49	15	9	11	0	SCN8A	516	59	9	11	0						
CRB1	593	69	9	11	0	HSPB11	299	69	9	11	0	NHLH2	584	12	9	11	0	SCN9A	516	59	9	11	0						
CRB2	424	69	9	11	0	HSPB8	130	59	9	11	0	NHP2	501	71	21	11	0	SCNN1A	518	69	9	11	0						
CRB3	470	106	9	11	0	HSPD1	85	59	9	11	0	NID1	268	69	9	11	0	SCNN1B	518	69	9	11	0						
CRBN	49	15	9	11	0	HSPG2	268	69	9	11	0	NIN	60	141	26	11	0	SCNN1D	518	69	9	11	0						
CRCP	157	36	20	11	0	HSPH1	245	59	9	11	0	NIP7	501	71	21	11	0	SCNN1G	518	69	9	11	0						
CREB1	556	10	9	11	0	HTATSF1	475	36	20	11	0	NIPBL	33	150	52	11	0	SCP2	231	53	17	8	1						
CREB3	375	12	9	11	0	HTR1B	246	51	9	11	0	NKD1	30	150	52	11	0	SCRIB	292	15	9	11	0						
CREBBP	415	10	9	11	0	HTR2A	246	51	9	11	0	NKX2-1	555	4	1	11	0	SCYL1	123	147	49	7	1						
CREBZF	46	14	20	11	0	HTR2B	246	51	9	11	0	NKX2-5	11	4	1	11	0	SDC1	79	17	9	11	0						
CREG1	470	106	9	11	0	HTR2C	246	51	9	11	0	NLGN1	226	59	9	11	0	SDCBP	403	15	9	11	0						
CREM	556	10	9	11	0	HTRA2	170	59	9	11	0	NLRC4	358	59	9	11	0	SDF2	585	134	9	11	0						
CRK	225	59	9	11	0	HTT	620	59	9	11	0	NLRP1	358	59	9	11	0	SDF2L1	585	134	9	11	0						
CRKL	225	59	9	11	0	HUS1	463	102	26	11	0	NLRP3	358	59	9	11	0	SDHA	636	59	9	11	0						
CRLF1	158	37	9	11	0	HUS1B	463	102	26	11	0	NLRP6	358	59	9	11	0	SDHB	248	59	9	11	0						

Appendix C: The biological systems under significant survival pressure.

System_name	Genes	Selection_pressure	p	q
3	ACTL7A, ACTL10,ACTL7B,ACTL8	0.021896	0.01	0.0896
8	ADD2, ADD1	0.012874	0.01	0.0896
11	PRDM16,HOXC9,DACH1,TAF6,POU3F1,BCL11A,TBX2,INSM1,SKI,HOXD10,PROP1,BCL11B,FOXP3,ARID5A,CRX,BARX2,AKIRIN2,GSC,ZC3H12A,POU4F1,HOXB9,ZFPM1,HOXA9,GFI1B,GFI1,TAF6L,MSX2,SKOR1,OTX2,MSX1,NKX2-5,HOXA10,KANSL3,ALX4,HOXD12,PITX1,HOXA11,HHEX,PITX2,UNK,DOT1L	0.006148	0.01	0.0896
13	TRPC4,DHDDS,HM13,SLC9A1,DAPK1,SLC11A2,TRPC3,SEC61A1,TRPV3,TRPC5,TRPV6,SEC61A2,ANO2,TRPM4,ANO6,PSEN1,STT3A,KIDINS220,STT3B,TRPC6,ZMPSTE24,P4HA1,ANO1,PSEN2	0.012508	0.01	0.0896
38	ATP8B2,ATP8B4,ATP11B,ATP8A2,ABCC8,ABCC9,ATP11A,ATP8B1,ATP8A1,ATP11C	0.004192	0.01	0.0896
47	BCL7A,BCL7B,BCL7C	0.007971	0.01	0.0896
58	ATAD5,RBM44,BRCA2,ICE1,SGO2,BRCA1,CRYBG3,TET1,CEP192	0.014937	0.01	0.0896
59	BCL3,FBXL19,SET,SMARCA5,MPHOSPH8,NXF1,PAN3,BRF1,BRD9,KAT6B,BRPF1,JADE1,KDM5A,BRD1,WWOX,SRCAP,HNRNPU,KAT8,HMGB1,DPF3,DNMT3B,KMT2E,KMT2C,PIAS4,KAT2B,ERN2,DPF2,ASH2L,DGKI,ERN1,MSH6,KAT6A,SMARCA1,PHF12,ING4,ING5,HMGB2,PHF20L1,TBC1D5,ING3,ING80,TIMM50,BOD1L2,CHD4,RIOX2,MCRS1,PHF10,SMARCB1,ING2,L3MBTL1,HNRNPUL1,JADE3,PARP1,JADE2,CXXC1,DPF1,KDM2B,SCML2,DCLRE1C,CBX5,LIN28A,BOD1,DBI,CBX3,ECT2,JMJD6,MBTD1,SMARCA2,BRPF3,CHD5,KAT7,GLYR1,RFC1,BRD7,NT5DC3,PAXIP1,DNMT3A,KAT5,BRD8,CSNK2B,SMARCA4,LIG3,PRICKLE4,SETD5,ORC1,CHD3,PHF20,PBRM1,PHF14,ANP32E,KAT2A,EXOSC10	0.038839	0.01	0.0896
70	SLC25A5,CAMK2G,CAMK2B,CTBP2,CAMK2A,CTBP1,ANAPC11,RBX1,SLC25A6,PMS2,PAK1,RNF7,SLC25A4,CAMK2D	0.00523	0.01	0.0896
97	CDK7,CDK9,CDK2,CDK1,CDK5,CDK6,CDK4,CDK3	0.006805	0.01	0.0896
117	AKAP5,ELL3,TESPA1,CNST,PALB2,PPP1R15B,ENTHD1,TINF2	0.005573	0.01	0.0896
118	SPDL1,CNTRL,CCDC68,SSX2IP,AZI2	0.013179	0.01	0.0896
130	HSPB8,HSPB1,CRYAB,CRYAA	0.020847	0.01	0.0896
132	CUL4A,CUL5,CUL1,CUL4B,CUL3,CUL2	0.013913	0.01	0.0896
134	APOM,CLCC1,SSR4,LY96,AMN,SGCD,CATSPERB,SGCE,OOEP,SGCB,CATSPERG,SGCG,SGCA,DAG1	0.006138	0.01	0.0896
146	DLST,GCSH,DBT,DLAT,PDHX	0.008511	0.01	0.0896
153	DNAJC17,DNAJC9	0.012459	0.01	0.0896
156	DUOX2,DUOX1,PTGS2	0.014747	0.01	0.0896
157	EIF3D,EAF2,EIF2S2,BAZ1A,NPLOC4,MEAF6,MKRN3,SUPT6H,EAF1,GEMIN6,POLR2D,SNUPN,CCDC47,CDC73,ZMAT5,NEMF,POLR3D,POLR2F,PRIM1,LIN9,NPM1,ATG3,LIG4,PRPF18,ZCCHC8,SUB1,TAF7,DNAJC28,NDUFA12,BPTF,TAF7L,ZNHIT6,BAZ1B,RSPH4A,DPY30,CSTA,TIPIN,CAT,TRIP4,URI1,FRG1,PPP1R8,CHTF8,PPP4R2,THG1L,IFT46,RSF1,CLNS1A,BTF3,CRCP,GTF3C6	0.002515	0.01	0.0896
170	ERBB3,COL20A1,SLC27A5,MAGI2,DDR1,CAPN1,EPHB6,SERPINB6,EPHA4,RPS27A,PLCB1,DDR2,EPHB1,NPR1,FGFR3,PKN1,PKN2,DLG3,LTK,PLCG1,UBA52,EGFR,LYN,INSRR,EPHA7,FGFR4,EPHA10,EPHA3,UPF1,PRKAG2,AGO3,AGO1,FGFR1,AGO2,PLCB3,EPHA1,PLCB2,SKIV2L,AGO4,CAPN2,EPHA2,DLG4,AXL,EPHA8,HTRA2,EPHB2,EPHA6,EPHA5,ERBB2,EPHB3,PFKFB1,EPHB4,DLG1,PRKAG1,FGFR2,IGHMBP2,PARD3,IGF1R,INSR,CAPN3,RECQL5,ERBB4	0.022622	0.01	0.0896
180	EXOSC9,EXOSC5,EXOSC4,PNPT1,EXOSC8,EXOSC6,EXOSC7	0.005334	0.01	0.0896
189	FBXO32,FBXO9,FBXO31,FBXO15,TMEM183A,FBXO25	0.006185	0.01	0.0896
195	FCN1,FCN3,FCN2	0.015171	0.01	0.0896
206	GFRA2,RECK,NRXN2,GFRA3,FGA,GFRA1,APP	0.005249	0.01	0.0896
208	IL2RB,GHR,LEPR,PRLR,IL4R,CSF2RB,IL23R,IFNLR1	0.018912	0.01	0.0896
212	GJB2,GJA1,RNF217,GJB1,GJB5,GJC1,GJB7,GJB3,GJA4,GJB6,GJA3,NAA30,BEST4,RNF19B,GJC3,GJA5,GJB4,BEST3,GJC2,RNF19A,BEST2,FAM8A1,BEST1,OCLN,RNF31,GJD3	0.009037	0.01	0.0896
213	GLRX5,GLRX3,P4HB	0.005681	0.01	0.0896
217	GNG11,GNG4,GNG10,GNG12,GNG2,GNG7,GNG5,GNG8,GNG3,GNGT2	0.012031	0.01	0.0896
239	HLA-DRB5,PSG9,HLA-DQB1,HLA-DRB1,HLA-DPB1,HLA-DQB2,FCRL5,HLA-DOB	0.003432	0.01	0.0896
240	JAKMIP1,RNF20,HOOK3,DCTN1,ERC1,CCDC102A,HOOK1,RNF40,HOOK2,BICD2	0.003226	0.01	0.0896
243	HSP90AA1,HSP90AB1,PMS1,HSP90B1,MLH1	0.017458	0.01	0.0896
244	HSPA1B,HSPA1A,HSPA5,HSPA2,HSPA1L,HSPA9,HSPA8	0.013852	0.01	0.0896
248	FOXRED1,PPCDC,CLPP,SUCLA2,NDUFV1,NDUFS1,FARSA,GATB,BCKDHB,IDH3G,IDH3A,NDUFS7,RPL10A,COQ6,ADSL,NDUFS8,COQ5,FH,PDPR,SUCLG1,IDH3B,SDHB,HMGCL,DPH1,PDHB,SUCLG2	0.007173	0.01	0.0896
266	CAND1,XPOT,RANBP9,IPO7,IPO5,KPNB1,MTOR,IPO4	0.008638	0.01	0.0896
269	RYR1,ITPR3,ELMO1,ITPR1,RYR2,ITPR2,RYR3	0.028197	0.01	0.0896
276	KCNS2,KCNG2,KCNS3,KCNV2,KCNG3,KCNF1,KCNG1,KCNS1,KCNN4	0.036467	0.01	0.0896
290	LAMTOR5,LAMTOR3,LAMTOR4,LAMTOR2	0.013876	0.01	0.0896
292	ABL1,PDLIM7,PDLIM2,PSMD9,TNK2,TNKS,PDLIM3,PXN,PDLIM5,PDLIM4,SCRIB,TES,LDB3,LMO4,ESPN,LMO2,PDLIM1	0.037509	0.01	0.0896
297	LIN7B,PDZD11,LIN7A,USH1C,LIN7C	0.008754	0.01	0.0896
326	MICU1,MICU3,CHP1,PEF1,EFCAB7,MICU2,S100A1,PPP2R3B,CAPNS1,CAPNS2,S100A10	0.024211	0.01	0.0896
331	TPRKB,GOLGA7,JCHAIN,CTNNBIP1,PSMG4,ARPC3,TAF1D,CBFB,RNASEH2C,GNAS,COPS9,MMGT1,ARPC4,ARPC2,LBH,MED20,ERH,NPIPA1,GTF2A2,GTF2H5,MREG,BEX1,PMAIP1,DDA1,AHSP,LIN52,DCAF16,TCL1A,SF3B5,BBIP1,ANAPC13	0.003385	0.01	0.0896
339	MRPS15,MRPL47,NDUFA6,MRPL32,MRPL54,GADD45GIP1,MRPL55,AURKAIP1,NDUFA5,LYRM4,MRPL38,MRPL52,MRPS26,NDUFV3,PDCD7	0.014743	0.01	0.0896
340	NOP10,MRPS33,MRPL57,MRPS34,NDUFA1	0.007751	0.01	0.0896
341	MRPS14,MRPS18A,NDUFAB1,FXN,MRPS18C,MRPL30,MRPS16,UQCRFS1,MRPS18B,MRPS11,NDUFS6,RPL34,MRPS9,RPS26,NDUFS3,CKS1B,MRPL28,POLG2,MRPS5,NDUFS4,COX5B,CKS2,MRPL18,MRPL39,TBPL1	0.00703	0.01	0.0896
356	MYO1G,MYO6,MYO5A,MYO1B,MYO3B,MYO7B,MYO18A,MYO19,MYO1D,MYO3A,MYO5B,MYH15,MYO1A,MYO10,MYO7A,MYH7B,MYO1E,MYO9A,MYO1F,MYO16,MYO5C,MYO9B,MYO1C	0.002267	0.01	0.0896
362	NCF4,QPRT,SUV39H1,CD2AP,NCF2,RPAP3,FKBP4,FKBP8,NCF1,FKBP1B,SPAG1,DNAJB11,FKBP1A,AIP,STAC3	0.003499	0.01	0.0896
365	NCOA2,YAP1,NCOA1,WWTR1,NCOA3	0.020922	0.01	0.0896
369	TIMM21,NDUFC1,NDUFB2,SMDT1,NDUFB5,NDUFB11,COX6A1,COX8A,COX6A2,NDUFB8	0.012221	0.01	0.0896
370	COX6B2,NDUFB7,CHCHD1,NDUFA8,NDUFB10,UQCRHL,TRIAP1,NDUFS5,COX6B1,UQCRH	0.017029	0.01	0.0896
388	AHCTF1,RAB3GAP2,NOL11,MED16,NUP133,RIC1,NUP160,UBE3D,HPS5,NUP155	0.006546	0.01	0.0896
392	POM121C,NUP153,NUP58,NUP50,NUP54,NUP62,POM121,NUP98	0.0163	0.01	0.0896
395	ODF2,CEP83,EEA1,SCLT1,CEP290,CEP128,AKAP9,TPR,RABEP1	0.007122	0.01	0.0896
408	BCKDHA,OGDHL,OGDH,PDHA2,PDHA1	0.006327	0.01	0.0896
412	VBP1,PFDN6,PFDN2,PFDN4,UXT,PFDN1,PFDN5,PDRG1	0.008839	0.01	0.0896
416	PHKA1,PHKA2,PHKB	0.007611	0.01	0.0896
418	LARP7,CC2D2A,SRP9,PARP11,RBM41,RNPC3,PIK3C2G,PIK3C2B,RRP7A,TEP1,PEA15,UBAP1,KRIT1,CUL7,KAT14,FBXO21,FBXO8,MED18,SNAPC3,PIK3R6,LARP6,PKD1L3,PPM1E,RBM28,MKS1,PIK3C2A,ERCC6,CEP76,PIK3R5,PSMG1,CUL9	0.008552	0.01	0.0896
420	PTPRA,SH3GLB1,CAPG,PTPN11,PIK3R2,RACGAP1,DUSP22,PTPN6,PIK3R3,PTPN1,PIK3R1	0.005598	0.01	0.0896
452	PSMA2,PSMA4,PSMB7,PSMA1,PSMA6,PSMA7,PSMA3,PSMB10,PSMB5,PSMA8,PSMA5,PSMB2,PSMB6,PSMB8,PSMB9,PSMB4,PSMB3,PSMB1	0.01526	0.01	0.0896

473	VPS37D,MPP4,SAC3D1,EPS8L2,TYK2,IRAK1,TERT,ARHGAP33,NOL6,SHARPIN,RPS6KL1,SLX1A,EPS8L1,NOXA1,PIDD1,SIPA1,SPHK2,RASIP1,MON1B,IRS1,NOXO1,IRS2,EPS8,ABI3,RADIL,CARD10,MON1A,RGS11,RILP,MYO15B	0.002463	0.01	0.0896
501	RPS12,RPL7,RPL13A,ATP6V1F,RPL7L1,RPL18A,RPS17,EIF6,SRP14,RPS20,SRP19,RPL5,RPL30,SUPT4H1,SNU13,RPL38,INO80C,RPL22L1,RPL11,NIP7,RPS3A,NHP2,RPL22,MRPS6,POP4	0.007698	0.01	0.0896
504	EIF5AL1,EIF5A,EIF5A2,RPS11,RPS28,RPS15,RPS23	0.030417	0.01	0.0896
520	SEC24C,SEC23A,SEC24A,EYA3,SEC23B,SEC24B,EYA1,SEC24D	0.133703	0.01	0.0896
524	ATXN2,GPS2,LSM11,NUFIP1,TAF3,FIP1L1,HMGXB4,RFXAP,SETD1A,AEBP2,SETD1B,ELOA,RBM42,CAPRIN2,TRERF1,INO80E,CECR2,CASC3,LSR,ILDR1,NAF1	0.003097	0.01	0.0896
528	HNF1A,LHX3,PBX3,MEIS1,HNF1B,SIX6,PBX2,SIX3,PBX1,SATB2,SIX1,SIX2,RFX3,ISL1,PKNOX1	0.010366	0.01	0.0896
531	SLBP,CDKN1A,CDKN1B	0.008308	0.01	0.0896
542	KXD1,MNAT1,XPA,BRK1,BLOC1S6,BORCS5,WASHC3,ZC4H2,MZT1,BLOC1S1,SNAPIN,CDK5RAP3	0.029348	0.01	0.0896
560	IQGAP1,DMD,SYNE1,SPTA1,SPTBN1,SPTAN1,UTRN,FLNA,ACTN4,SYNE2,SPTB,SPTBN5,SPTBN4,SPTBN2	0.047516	0.01	0.0896
562	SPTSSB,SPTSSA	0.007505	0.01	0.0896
581	TAP2,ABCB6,ABCB8,CFTR,TM7SF2,TAP1,ABCB9,ABCD4	0.02473	0.01	0.0896
585	TECPR1,SDF2,SDF2L1	0.011856	0.01	0.0896
586	APOL1,POT1,MZT2B,RNASEH2A,MEN1,GPKOW,JRK,XRCC1,TERF1,RMI1,MZT2A,RPP21,AIM2,NCAPH2,TERF2IP,SNAPC4,ERCC1,EME1,TERF2,DNAAF2,FAAP24,RRP8,MUS81,PYHIN1,TSEN34,NCAPH,LRRC46	0.014327	0.01	0.0896
607	ENY2,TOMM20,FIS1	0.016301	0.01	0.0896
610	VPS26A,RGP1,VPS26B,TRAPPC13	0.00471	0.01	0.0896
619	XAB2,TTC30A,MAU2,SART3,PEX5,ANAPC7,EMC2,NAA16,CRNKL1,PEX5L,TTC37,SF3B1,UTP6,TTC21B,PRPF6,PRPF39,FNTA,ANAPC5,COPE,NAA25,IFT88,GTF3C3,CTR9,TTC30B,TTC8,NAA15,RABGGTA,TTC26,SRP72,CNOT10,TTC21A	0.002781	0.01	0.0896
627	UBR1,UBR5,TIMELESS,UBE4B,UBE4A,ASCC2,UBR2,UBR3,PEX2,PEX12	0.005037	0.01	0.0896
629	UBXN6,UBXN8,NSFL1C	0.031072	0.01	0.0896
638	TGFBRAP1,VPS18,VPS8,VPS39,VPS16,VPS41,VPS11	0.005839	0.01	0.0896
639	WASF2,CEP170,RAD21,APBB1IP,PCM1,LEO1,WASF3,INO80D,EPC2,DROSHA,EIF4G3,KANSL2,EPC1,TAF1,KANSL1L,GTF2E1,SUPT20H,WNK1,CHTOP,SYNPO2,WBP11,TAF1L,ATXN7L3,KANSL1,WASF1,EIF4G1,LMO7,SF3B2,AKAP6,TOB1,HIPK2,MSL2,WASHC1	0.025465	0.01	0.0896
649	XPO5,ATP6V1H,XPO4,CAB39L,TRRAP,CNOT9,XPO7,XPO6,RANBP17,PSME4,CAB39,TSC2,DCAF1	0.008198	0.01	0.0896
667	NCOA2,POU2F1,YAP1,LIN54,NFYA,ARNTL2,TAF4,NCOA1,GTF2A1,WWTR1,ARNT2,AHR,PHC3,HIF1A,ATF1,NPAS2,MED15,CREB1,USF1,PHC1,CREBBP,NCOA3,SP3,EP300,PHC2,AHRR,SP1,NFYC,TAF4B,EPAS1,CLOCK,CREM,ARNT,YEATS2,ARNTL	0.038895	0.01	0.0896
672	BCL3,CCDC120,VPS37D,IKBKG,ABL1,FBXL19,ERCC5,PDLIM7,SUGP1,CBL,PRPF4B,CLINT1,WASF2,SET,PDLIM2,E2F7,CEP170,MPP4,NOS1,FRMPD1,SMARCA5,RABGAP1,MPHOSPH8,PSMD9,ADD2,RBPJ,NXF1,PAN3,SARNP,BRF1,BCR,RAD21,TNK2,RAD18,BRD9,KAT6B,TFDP1,BRPF1,JADE1,KDM5A,FCER1G,KCTD13,CACTIN,RPH3A,LRRC43,SIX4,APBB1IP,TSEN2,SAC3D1,BRD1,PCM1,WWOX,P3H3,ADD1,SRCAP,LEO1,CACNB1,EPS8L2,TYK2,HNRNPU,KAT8,EXOC3L2,IRAK1,SLBP,PPP1R3G,WWC1,WASF3,TERT,ZSWIM5,KDM3A,HMGB1,TNKS,DPF3,CDKN1A,DNMT3B,KMT2E,BARD1,ARHGAP29,INO80D,E2F8,JPH1,SDCBP,PPP1R3B,PKP3,KMT2C,TJP1,TP53,PIAS4,KAT2B,EPC2,KCTD10,NCF4,QPRT,GIGYF1,DROSHA,SMAD6,EPN3,CCDC61,CTDP1,ERN2,EIF4G3,ARHGAP33,PDLIM3,APC,SMAD7,DPF2,PRPF40A,SUV39H1,WHRN,ASH2L,TRAF1,CD2AP,DGKI,KANSL2,DICER1,EPC1,ERN1,TRAF4,TAF1,NCF2,MSH6,METTL14,CACNB4,MAD1L1,REPIN1,SF1,KCTD5,KANSL1L,LMTK2,WTIP,SREBF2,NOL6,KAT6A,PXN,PDLIM5,PDLIM4,RASSF2,SOS1,GTF2E1,EPS15,ST18,SMARCA1,PHF12,SHARPIN,ING4,ING5,HMGB2,PRR5,RPAP3,SUPT20H,RPS6KL1,SNAI3,OGG1,PHF20L1,WNK1,SNTG1,CDKN1B,VPS51,SLX1A,TBC1D5,NHEJ1,TRAF5,USP10,JPH2,DCX,ARHGEF7,EPS8L1,NOXA1,USP33,FKBP4,ING3,SKIL,RNF168,DDX58,PLK4,INO80,FKBP8,TIMM50,BOD1L2,XRCC4,E2F5,ZSWIM6,SNTB2,CHD4,CDK13,RIOX2,PRDM10,MCRS1,RBM10,NCF1,PLEKHA2,AGFG1,RBPJL,TNFAIP1,E2F2,PHF10,CHTOP,NFRKB,SMARCB1,ARHGEF2,KCTD17,METTL1,ARVCF,ING2,FKBP1B,PARD6G,SIAH1,TRIM54,CACNB3,DIXDC1,PIDD1,WBP4,L3MBTL1,HNRNPUL1,CARD11,SMTNL2,BMPR2,EIF4EBP3,CARD9,JADE3,SPAG1,SCRIB,AXIN2,PRR5L,SYNPO2,SIPA1,SNTB1,PARP1,AXIN1,JADE2,SPHK2,SNTA1,APC2,SNTG2,TES,RNF215,WBP11,TAF1L,DNAJB11,CXXC1,DPF1,KDM2B,SCML2,PRPF40B,CACNB2,DCLRE1C,EXOC3L4,MBD2,CBX5,E2F1,LIN28A,TOPBP1,RASIP1,GPR156,ATXN7L3,LDB3,LMO4,MBD3,SMTN,CCM2,PARD6A,BOD1,TRAF6,EHBP1,TFDP2,KANSL1,MON1B,DBI,GRB10,IRS1,PPP1R12A,SIRT2,CEP97,SASH1,WASF1,SAP30L,EME2,TRAF2,E2F6,RNMT,E2F3,SF3A1,FKBP1A,KCTD16,CBX3,ECT2,TSNARE1,RASGRP3,JMJD6,EIF4G1,WASHC2A,FANCM,CLMN,NOXO1,AIP,EPS15L1,PKP1,LMO7,SMN1,MBTD1,SMARCA2,MID1,BRPF3,ABI2,IRS2,ZSWIM8,CHD5,ZC3H18,DZIP1,SAP30,MYSM1,TRAF3,SIRT1,EPS8,TRIM28,ABI3,USP47,WASHC2C,E2F4,KAT7,ABI1,PICALM,NBN,TAF10,GLYR1,RFC1,PDZD7,RADIL,IMPACT,PPP1R3C,ESPN,AJUBA,HCLS1,CD247,ASF1B,TBKBP1,P3H4,CARD10,BRD7,BAIAP2L2,DAB2IP,SF3B2,ZBTB17,CRBN,NT5DC3,PPP1R9B,PAXIP1,CRTAP,DNMT3A,KAT5,BRD8,PPP1R3D,MON1A,CSNK2B,SMARCA4,EXOC3L1,RAPGEF2,LIG3,GOPC,PRICKLE4,SIRT3,SETD5,ORC1,DVL1,AKAP10,CHD3,PHF20,ANAPC2,TAOK2,ASF1A,PBRM1,PRIMPOL,STAC3,ULK1,AKAP6,SMARCE1,EPN1,COG8,SLC9A3R1,ZSWIM4,TOB1,PHF14,BUB1B,MYO18B,KCTD2,RNF125,NBR1,SKAP1,AFDN,HIPK2,ANP32E,P3H1,RGS11,DNTTIP1,RILP,KSR1,GIGYF2,TBP,LIMD1,EPN2,PARD3B,LMO2,KAT2A,GID4,PDLIM1,PPP1R3E,MSL2,MYO15B,JPH3,ZC3H14,SYNRG,CTNND1,DEPDC1,MYRIP,TP63,PARD6B,JPH4,TNFAIP2,SGIP1,CDK12,PLEKHA7,EXOSC10,FRMPD4,RBM5,METTL3,RYBP,BAP1,SREBF1,MRGBP,WASHC1	0.003702	0.01	0.0896
693	RBM12B,EIF3D,EAF2,POLE3,EIF2S2,TAF13,HYPK,BAZ1A,NPLOC4,INTS12,MEAF6,NECAP1,THRAP3,TAF12,DR1,MKRN3,ARID4B,SUPT6H,EAF1,GEMIN6,TAF8,POLR2D,MLLT1,SNUPN,CCNT2,CCDC47,PNN,CDC73,ZMAT5,NEMF,CCNT1,POLR3D,POLR2F,EHBP1L1,PRIM1,POLR1D,ZC3H13,CLIC6,LIN9,NPM1,CENPS,ATG3,LIG4,PRPF18,ZCCHC8,SUB1,TAF7,TRDN,TAF9B,DNAJC28,TAF11,NDUFA12,ZFC3H1,BPTF,HTATSF1,TAF7L,APOBR,CCNK,PRKAB1,ZNHIT6,KMT2A,BAZ1B,MAP4,KDM6B,RSPH4A,NACAD,DPY30,CHRAC1,NCBP3,HDGF,CSTA,KMT2B,TIPIN,NFYB,CCDC8,MLLT3,NECAP2,BCLAF1,CAT,TRIP4,UR11,PPP1R15A,FRG1,TAF9,PPP1R8,TNKS1BP1,BAHD1,ARID4A,PRKAB2,CHTF8,PPP4R2,DRAP1,THG1L,IFT46,RSF1,CLNS1A,BTF3,CRCP,MAVS,GTF3C6,SUPT3H,SLF2,SMTNL1	0.014805	0.01	0.0896
708	GPR37,GPRC5C,GP5,UPK1A,TSPAN33,ADRB3,TSPAN17,GPRC5B,GPR63,GPR20,GP9,LHCGR,HTR2A,FSHR,GP1BB,CALCRL,GPR84,HTR2C,ADRB2,MTNR1A,GPRC5D,GPBAR1,GPRC5A,CXCR4,GPR160,GPR61,CD9,CALCR,B3GALT4,GPR37L1,IGFALS,HTR2B,LRRC55,GPR62,HTR1B,TSHR,F2RL2	0.00387	0.01	0.0896
721	SPDL1,UGT3A2,ODF2,CEP83,CUBN,RGPD3,FBXL12,EEA1,CNTRL,ITGBL1,RGPD1,TCF25,COMMD1,LRP6,AMHR2,TGS1,RGPD4,EIF2AK4,RIPK1,DFFB,CD14,ZZZ3,RGPD5,CHMP7,HPS3,FBXO33,VRK2,SWI5,MEPCE,TLR4,HSD17B8,ASCC1,RPA4,RHBDD2,SCLT1,FGL1,NBAS,SPECC1L,CCAR2,GTF3C4,JARID2,OTULIN,IKBKE,ANAPC4,CCDC68,RAG1,LENG8,SNAP47,UBL4A,SSX2IP,CEP290,INCENP,ALS2,FBXL15,LRP1,INTS8,FGG,LDLR,MTERF4,LRP2,RGPD8,LRP5,FANCG,RGPD2,CEP128,IKBKB,VIPAS39,RNF8,CHAF1A,TAF1A,LAS1L,SPATS2L,ING1,UGT3A1,MCM3AP,AKAP9,PDE3B,FBXL8,FBXO24,AZI2,CLIP2,CFAP206,TEFM,MAP7,LUM,RANBP2,TLR2,GRHPR,LRRK2,HSD17B10,INTS10,KNTC1,FGB,TRAPPC12,GLIPR1L1,ITLN1,VLDLR,DDX20,FBXO38,TLR7,TLR1,CTU2,TPR,WDR93,CBR4,TRAF3IP1,FBXO48,FBXL3,TBK1,STN1,NME8,AIMP2,SMG7,RABEP1,FGL2,CGNL1,LRP1B,FBXO39,RIPK3,TLR6,TLR10,CHUK,WTAP,MAP3K5,SPECC1,SMG6,FBXL22,NEK10,LRP8,DTX3L,SHTN1,FBXL18	0.023813	0.01	0.0896
725	BCL2L11,BAD,MBP	0.017315	0.01	0.0896
728	RPS12,RPL7,MRPS14,MRPS18A,NDUFAB1,EIF5AL1,EIF5A,RPS6,MRPL14,RPL13A,RPS5,RPL3L,EIF5A2,RPS11,MRPL24,RPL35A,ATP6V1F,FXN,RPL32,MRPL12,RPL7L1,MRPL46,MRPS18C,RPL18A,RPS17,MRPL30,MRPS16,EIF6,UQCRFS1,BOLA2,SRP14,RPL10L,RPS28,RPS20,MRPS18B,SRP19,GATC,RPL5,MRPS11,MRPL17,RPS16,NDUFS6,TRMT61B,RPL34,MRPS9,RPS26,MRPS17,TRMT10B,TRMT10C,RPL18,MRPL13,RPL8,MRPL20,RPL28,RPS15,RPL30,MRPS35,IMP4,FDX2,MRPL2,MRPL22,SUPT4H1,MRPS10,MRPS7,RPL15,NDUFS3,MRPL4,SNU13,RPL38,CKS1B,BOLA1,TRMT6,MRPL28,MRPL3,RPL3,RPS8,POLG2,RPF1,INO80C,MRPS12,RPP38,MRPS5,RPL22L1,MRPL1,MRPL15,RPL4,MRPL58,NDUFS4,NUDT21,MRPL21,MRPL11,RPL11,RPL17,NIP7,MRPL48,MRPS21,BOLA3,RPL7A,RPS3A,RPL23,COX5B,NHP2,RPL10,COX11,MRPL19,CKS2,RPL22,MRPS6,MRPL27,RPL27A,MRPL10,NSA2,MRPL49,RPS23,POP4,MRPL9,MRPL16,MRPL18,MRPL39,TBPL1	0.016397	0.01	0.0896
732	HAUS3,INIP,TPRKB,BORCS8,GOLGA7,JCHAIN,BUD31,DDIT3,NOP10,NFE4,CENPQ,CTNNBIP1,PSMG4,ARPC3,TAF1D,CBFB,MRPS33,RPS21,MRPL57,EVC,HAUS2,GON7,DFFA,BIK,PHF5A,POLR2H,RNASEH2C,GNAS,COPS9,MMGT1,ARPC4,IFT20,ARPC2,PBXIP1,ATG101,MVB12B,HAUS7,MED17,MRPS34,LBH,MED20,CDC26,TMEM249,LAMTOR1,CSF1,ERH,PKD1L1,NPIPA1,KCNE1,SKA2,GCHFR,GEMIN7,GTF2A2,MVB12A,GTF2H5,MRPL41,SUFU,MTCP1,CD34,SCIMP,THEMIS,NUPR1,TNP2,MREG,SYNE4,HAUS1,BEX1,MAGOH,NDUFA1,PMAIP1,HAUS5,MAGOHB,EVC2,CCDC103,DDA1,BORCS7,AHSP,AP5S1,LIN52,TNP1,ANAPC16,DCAF16,TCAP,HAUS8,BAG2,TCL1A,SF3B5,BBIP1,CENPO,SLC51B,SPC24,TRAT1,CAPZB,HAUS4,ANAPC13,KCNE2	0.021489	0.01	0.0896
739	UBR1,UBR5,PTPA,TIMELESS,UBE3C,UBE4B,TRPC4AP,AQR,UBE4A,NUP188,ASCC2,UBE3A,UBR2,NUP205,UBR3,PRIM2,PEX2,ZNFX1,PEX12,TSC1,ATRIP	0.010807	0.01	0.0896
756	EIF3C,XAB2,IFT80,VPS29,CLTCL1,TTC30A,EIF3K,COPS7B,VPS33A,WDR35,TGFBRAP1,MAU2,EIF3E,RMND5A,SART3,PCID2,PEX5,ANAPC7,EIF3L,EMC2,PSMD11,VPS36,NAA16,VPS18,VPS8,CRNKL1,VPS39,POLR3C,PEX5L,ATP6V1E1,FLOT1,TTC37,CLTC,PSMD12,PA2G4,PSMC3IP,COPS4,FLOT2,HBQ1,GTF2H4,COPS2,RPS19,SF3B1,UTP6,NSMCE2,ATP6V1G3,TTC21B,SNF8,COPS8,SSRP1,PTGES3,PRPF6,CACYBP,PSMD13,PRPF39,VPS16,PHB2,FNTA,POLR3F,ETF1,GPS1,ATP6V0D2,EIF3CL,RPS10,COPB2,ANAPC5,HBG2,GTF2F2,HBM,COPS7A,NAPA,HBA1,PSMD3,SRP68,GET4,COPE,CPSF1,ERLIN2,NAPB,VPS41,IFT172,ATP6V1G1,NAA25,MVP,ATP6V1G2,MAEA,IFT88,STXBP1,GTF3C3,CTR9,PSMD8,TTC30B,SF3B3,HBE1,VPS33B,NSMCE1,TTC8,ATP6V0D1,CORO1C,COPA,NAA15,RMND5B,RPSA,BBS5,IFT122,HBG1,ATP6V1E2,SUGT1,EIF3M,VPS11,HBD,HBB,RABGGTA,EIF3A,STXBP5,TTC26,VPS25,SRP72,PSMD6,CNOT10,SUPT16H,COPS3,DDB1,NAPG,WDR19,ERLIN1,TTC21A,GID8	0.020175	0.01	0.0896
771	RPL31,RPL23A,RPS18,RPL37,RPS9,RPL13,RPL39L,RPS15A,RPL39,RPS24,RPL24,RPL35,RPS4X,RPL36A,RPS29,RPL37A,RPL36,RPS4Y1,RPS13,RPL36AL,RPL19,RPL29	0.00853	0.01	0.0896
775	DNAJC17,SAMM50,PMPCB,DNAJC9,DNAJC15,SEC11A,IMMP1L,VDAC3,PAM16,PMPCA,UQCRC2,VDAC1,SEC11C,IMMP2L,DNAJC11,TOMM40,VDAC2,DNAJC19,TOMM40L,UQCRC1	0.011121	0.01	0.0896
777	NACA,SEMG1,CCNB3	0.003425	0.01	0.0896

790	WDR26,HIRA,ARPC1B,RCC1,ARPC1A,CIAO1,THOC3,AAAS,WDR5B,PAFAH1B1,CSTF1,TBL1XR1,SMU1,CHAF1B,WRAP73,MPP7,TBL1X,EIF2A,EIF3B,SEC13,ELP2,WDR5	0.00857	0.01	0.0896
803	WDR83,RAE1,TAF5L,TAF5,MLST8,WDR61,PRPF4,DCAF13,PRPF19,KATNB1,STRAP,WDR12,CDC40,EIF3I,PLRG1,ATG16L1,BUB3,SNRNP40	0.011198	0.01	0.0896
811	AP3M1,AP2S1,AP3S2,ARCN1,COPZ2,AP4M1,AP1M1,AP1S1,COPZ1,AP1S3,AP1S2,DYNLRB1,DYNLRB2,AP3M2,AP3S1,AP1M2,AP4S1,AP2M1	0.097001	0.01	0.0896
828	SNRPD3,LSM1,LSM8,LSM2,EXOSC2,EIF2S1,LSM7,LSM4,MRPS28,POLR2E,LSM6,POLR3H,LSM5,LSM10,SNRPF,SNRPD2,EXOSC3,POLR2G,POLR1C,NAA38,SNRPE,AIMP1,POLR2C,GAR1,LSM3,EXOSC1,SNRPD1,SNRPG	0.003436	0.01	0.0896
856	TIMM21,MRPS15,MRPL47,NDUFC1,MRPS31,NDUFA7,NDUFB4,APOOL,NDUFA6,SDHC,NDUFB6,COX6B2,COX7A2L,MRPL32,NDUFB7,TOMM5,COX4I2,TOMM22,UQCR10,NDUFB2,CHCHD1,MRPL54,NDUFA8,NDUFB10,NDUFAF1,TOMM6,MRPL53,GADD45GIP1,TOMM7,SMDT1,UQCRHL,MRPL51,PSMG3,TMEM147,NDUFB9,MRPL42,MRPS23,MRPL55,DIABLO,NDUFB5,TRIAP1,NDUFB11,AURKAIP1,NDUFA5,MPC2,LYRM4,COX6A1,COX7B,COX6C,COX7A1,COX8A,MRPL50,UQCRB,MRPS27,UQCR11,COQ4,TIMM29,PTCD3,COQ7,NDUFS5,COX7A2,PLN,NDUFB1,PSENEN,MRPL38,TMEM258,MRPL52,CYBA,NDUFB3,MRPS22,CNEP1R1,COX5A,NDUFA2,MRPL40,MRPS26,COX6A2,SDHD,NDUFA3,NDUFA4L2,COX4I1,COX6B1,NDUFC2,MRPS30,NDUFA4,MRPS36,COX8C,LRPPRC,NDUFB8,MPC1,OST4,MRPS25,APOO,HIGD1A,MRPS24,COX7C,NDUFV3,NDUFA13,PDCD7,COX15,UQCRQ,MRPL37,UQCRH,MRPL43	0.004546	0.01	0.0896
873	4252 genes including: LY6G6C,P2RX6,ENY2,ATP8B2,RBM12B,BCL3,SPDL1,PIGK,TRPC4,CCDC120,PRKCI,HMGA2,MICU1,UGT3A2,STX4,VPS37D,WASHC4,KCNK1,FCGR3A,ERBB3,ABCA2,TIMM21,IKBKG,ABL1,GTF2F1,CRADD,RPN1,SLX4,MAP1A,FBXL19,RNASEH2B,VPS52,NCOA2,APOM,KXD1,UBR1,PRKACA,ITGA4,KCNH2,APOL1,EIF3C,ERCC5,PDLIM7,DNAJC17,MAPKAPK5,MYO1G,HLA-DOA,COG5,CAPZA2,SEC61G,MRPS15,MRPL47,GNG11,MSR1,HSP90AA1,MYL6B,SPG7,PRDM16,PSMD10,UBR5,CUL4A,FOXRED1,TBX18,SUGP1,RPS12,ATAD5,MAML1,BRIP1,KPNA4,RPL7,NDUFC1,MYO6,TAP2,VSIG4,COLQ,RYR1,RBM44,SNX31,CD74,XAB2,GFRA2,GRWD1,SCARA3,KCNS2,ATP8B4,OGT,DLST,VBP1,NACA,GATA6,SECISBP2,HAUS3,MRPS31,IFT80,HES1,TNNI2,SAMM50,UXS1,ZNF622,INIP,CSNK1E,LARP7,LIN7B,POU2F1,ATF7,BCKDHA,CBL,TPRKB,ATP10D,PRPF4B,MTA1,LRWD1,F13A1,BCL7A,EIF3D,TGFBR1,VPS29,STK3,IFITM1,UBA2,MRPS14,SOX12,PTPA,BORCS8,MRPS18A,WDR46,MNAT1,ERCC3,CLU,FSHB,KMT2D,C9orf78,TNNC2,KLK8,DNHD1,NDUFAB1,HOXC9,ITPR3,CDK7,CLINT1,SOST,HYOU1,ITGA7,DERL1,PRELID1,EIF5AL1,NTN4,PAFAH1B2,DCTN2,CLTCL1,EAF2,TELO2,WASF2,TEAD3,GOLGA7,IQGAP1,FBXO32,SET,ODF2,MED6,TTC30A,CDH6,ABCB6,NDUFA7,ASCL1,MCM3,KCNG2,CASP5,PRKRIP1,COG3,NDUFB4,TBL3,PMPCB,APOOL,GOSR2,GPR37,CUL5,F3,CNOT2,GABBR1,ASB12,HSPA1B,TRA2B,PDLIM2,NCCRP1,CEP83,E2F7,UGGT2,CEP170,PTPRA,GPRC5C,IFITM2,EEF1E1,ESAM,CUBN,THOC5,NDUFA6,FCN1,PLXNB3,HCN4,SRP54,MPP4,EIF5A,CACNA1E,SDHC,JAKMIP1,JMJD1C,RANGAP1,NABP2,IL2RA,NCK1,STX11,NOS1,NDUFB6,YY1,RPS6,MRPL14,FRMPD1,EIF3J,YAP1,COX6B2,MFN1,C9,COL20A1,SUDS3,CD4,CLCC1,RPL13A,KCNB1,POLE3,SMARCA5,CCDC113,RABGAP1,MTA3,CACNA2D1,CR2,MPHOSPH8,DCUN1D2,POM121C,FOSL1,CC2D2A,KLHL20,ORC2,EIF3K,OGDHL,MEP1B,EFL1,EIF3G,ZFP36L2,GP5,DMD,SALL4,CYBB,EIF2S2,CCNE2,PSMD9,TAF13,POP5,COPS7B,TOMM20,RFTN1,GTPBP1,UTP3,STRADB,GCSH,CLCN2,DACH1,GRM7,TECPR1,NTNG2,POLG,JCHAIN,HSPB8,ADD2,XPA,KIF6,PPCDC,RBPJ,CANX,MTX1,SPTSSB,NXF1,SH3GLB1,RGPD3,RPS5,DDX41,CYB5R3,ETFA,GPSM1,HSP90AB1,ATF6B,BRCA2,PAN3,INTS5,RPL3L,NIN,MAFF,SRP9,HYPK,FBXL12,SLC27A5,RANBP3,ACTL7A,PFDN6,SFTPD,STRADA,GRIN2A,CNIH2,PYGO2,CDK5R1,PANX1,RPP25L,ATP6V1B1,COX7A2L,CRB2,SARNP,MAGI2,MAT2A,BRF1,CT45A3,SOCS7,PEX19,VPS33A,HNRNPH1,XPC,WDR35,GMNN,EEA1,TMED3,NUP88,SYNE1,CNTRL,CFAP70,SAP18,NEK2,BUD31,POT1,BCR,RIPK2,SPTA1,GRIN2B,TAF6,LY6G6D,RAD21,LIN54,TUBGCP3,DHDDS,TNK2,XRCC5,TIMELESS,ACTL10,FAM98C,BAZ1A,KIF21B,APOH,CLPP,ICE1,ENG,DNAH14,NUP214,RAD18,NPLOC4,TOP1,EIF5A2,NFYA,DDIT3,ADGRA2,RPS11,BATF3,CENPJ,ATXN2,ATP1B4,GRPEL2,BRD9,EIF4E,RPLP0,C1QTNF1,SCNN1B,AFTPH,MRPL32,TGFBRAP1,IL11RA,PARP11,MPV17L2,TNFRSF1A,MRPL24,MZT2B,DDX23,KLHL8,DUOX2,DNAJC9,FLT1,INTS12,MAU2,RPL35A,KAT6B,TFDP1,SDF2,KCNS3,ERCC6L2,PMS1,BRPF1,LAMC2,FANCE,TOP2A,LDHA,CAPG,UPF3A,EIF3E,LGALS1,NUP153,NOC2L,VPS53,RMND5A,HSPG2,CLIC3,JADE1,KDM5A,SYK,SART3,MYL4,FCER1G,PTPN11,MYO5A,ZFP36L1,PCID2,RAD9B,ITGBL1,SCNN1A,APOA1,NOP10,UBE3C,NDUFB7,ANKIB1,TOMM5,C1QTNF2,PEX5,INTS2,MICU3,RGPD1,DDR1,NFE4,RNASEH2A,KLF4,KCTD13,COX4I2,CAV2,NFKB1,CACTIN,MATN3,CENPQ,STX8,HM13,ATP6V1F,POLR3G,CASP2,ABCB8,SPTBN1,LAMTOR5,ADAM8,KCNQ5,TRAPPC5,CHP1,TCF25,CTNNBIP1,TOMM22,COL3A1,CHRNB2,RBM41,KIF11,RPH3A,ANAPC7,STAU2,SUCLA2,EIF3L,POU3F1,RNPC3,CACNA2D4,CD80,EPPIN,RGS4,MCU,STX17,MED13,LRRC43,BCL2L11,BAD,EML2,TGM3,MEAF6,NECAP1,CAPN1,LDB2,CWC22,EXOC5,THRAP3,SIX4,EMC2,PDZD11,APBB1IP,MBL2,PSMG4,SPTLC3,GPS2,INSIG1,COMMD1,TSEN2,CBX6,AHCTF1,TAF12,SCN10A,EMC3,FXN,LRRC8C,LRP6,RBL2,PIK3R2,RPL32,KATNAL1,NLGN1,SSR4,SAC3D1,BRD1,AMHR2,ARPC3,UPK1A,UQCR10,FCN3,PCM1,TGS1,KCNA4,WWOX,P3H3,IL13RA2,HSPB1,RGPD4,HNRNPH3,CCNL2,MED30,SUPV3L1,NDUFB2,MEN1,NDUFA9,DDX6,BCL11A,MIER1,HLA-DRB5, SPTAN1, TARBP2, RAB3GAP2, ADD1,BRK1,COL23A1,CHCHD1,SRCAP,EIF2AK4,EPHB6,LEO1,MRPL12,PSMD11,TAF1D,RNF20,DR1,KCNMB4,DTNA,CCDC9,LAMTOR3,SOX17,CBFB,RNF139,CACNB1,HSPA1A,MERTK,SUN2,GPATCH1,HSPD1,MRPS33,EPS8L2,SUMO1,RPS21,KCNK6,TYK2,RIPK1,HNRNPU,BCL10,DFFB,KAT8,NDUFV1,NPRL3,SERPINB6,RPL7L1,DMAP1,CD14,VPS26A,KCNAB1,NAA35,MRPL54,DDOST,ARNTL2,MRPL46,SLC9A1,EXOC3L2,TSPAN33,RBL1,DNALI1,CDC5L,FBXO9,VCPKMT,KIF5A,DNAJC15,STX10,TFAM,EPHA4,GPKOW,MKRN3,NDUFA8,TBX2,ZZZ3,IRAK1,VPS36,IL13RA1,NDUFB10,WWP2,VAMP7,RPS27A,NAA16,CNGA1,MRPS18C,SNX12,VPS18,TP53BP1,INSM1,PLCB1,RPL18A,PIK3C3,RPS17,SLBP,AKAP5,MRPL57,RGPD5,PYCARD,DAPK1,CHMP7,PIGU,PPP1R3G,LSM11,LAMTOR4,ADRB3,IL2RB,DDR2,CCBE1,MKLN1,ESRP1,GHR,EVC,NOP58,CCNI,C1QL2,SEC11A,CHRNE,WWC1,SAR1A,UTRN,WASF3,TSPAN17,NDUFAF1,RRM2B,TAF4,LIN7A,C2,COL15A1,MRPL30,ACTR10,GRIN3B,C1QTNF7,ARID4B,FANCD2,MDM4,MYO1B,ELP4,ARIH1,GTF2E2,PIK3C2G,HPS3,CORO2A,JRK,VPS8,CRNKL1,UBE4B,FLT3,ACTL6A,SUPT6H,VPS39,POLR3C,NCOA1,ALB,PEX5L,XRCC1,SLC11A2,GTF2A1,TERT,NUFIP1,HAUS2,ZSWIM5,EAF1,GOSR1,KDM3A,ELL3,HCN1,COL7A1,CEBPA,TERF1,SKI,KIF25,PFDN2,HMGB1,TAF3,BMI1,FBXO31,MTA2,HGS,PIK3C2B,ELP3,MCM2,CARD8,ARID2,MRPS16,TFR2,CNOT8,GRPEL1,RANBP3L,PCCA,TRPC3,MED12,EIF6,NDUFS2,MCUB,TOMM6,LEPR,NUP58,MCM7,COL4A3,PEX11B,TNKS,FANCA,OGDH,POLR2I,GON7,ITGB8,NCOR2,GEMIN6,RMI1,UQCRFS1,GLE1,GCH1,TCF4,DPF3,MXD3,SOCS2,FBXO33,PDCD6,HOXD10,DFFA,CDKN1A,PROP1,PSMF1,CCNA1,ATP6V1E1,PCNA,TAF2,SLC12A7,CUL1,ATP10A,NELFA,TBC1D7,CFTR,MZT2A,DHX15,MYL6,PEF1,WWTR1,DAP3,GRIK5,ARNT2,BIK,ATP11B,DNMT3B,KLHDC10,ESRP2,APOC4,TFIP11,TAS1R3,TRPC4AP,BCL11B,PYM1,BOLA2,PHF5A,DUOX1,NDUFS1,NUP50,TCTN1,POLR2H,EPHB1,MRPL53,ELL,KMT2E,MYB,BARD1,SRP14,GSK3B,PTN,SFTPA2,IMMP1L,ARHGAP29,NOL11,EMILIN2,FLOT1,FABP1,TAF8, VRK2, LAMA5,SMC6,POLR2D,TTC37,INO80D,KRTCAP2,MLLT1,PRLR,RELN,POLR1E,DYNC2LI1,HLA-B,ATF6,LY96,RGP1,SNUPN,ORMDL3,SEC61A1,NPR1,EPB41,PLXNC1,FARSA,SWI5,IFT74,RECK,PSG9,MEPCE,FGFR3,CDH5,NRXN2,KIF2A,RTF1,PPP2R5B,RPL10L,E2F8,NFIL3,EFTUD2,CDH12,CRLF2,TLR4,CHML,BET1,JPH1,SENP2,HSP90B1,CLTC,STX2,HSD17B8,PKN1,TRMT112,COL11A2,CHCHD3,PTGS2,CYC1,SDCBP,GNL3L,RNASEH2C,GATB,PPP1R3B,GTF2B,GBA3,MLH3,AHR,PSMD12,BLOC1S6,ELMO1,CREG1,PIGM,FBXW8,TRPV3,CNTFR,RPS28,MYO3B,CBLL1,SEMG1,PA2G4,MRPL45,SCN9A,PKP3,HLA-DQB1,APAF1,GADD45GIP1,KMT2C,TJP1,ZNRF2,ANKRD1,WNT2,TP53,CYFIP1,ITGA1,CTHRC1,SPTSSA,TRPC5,PIAS4,NDC80,IFT22,HCFC2,KAT2B,EPC2,PKN2,RPS20,CCNT2,CCDC47,ARL6,KCTD10,ASCC1,BORCS5,DLG3,RHEB,LTK,RRP7A,TESPA1,CDX2,NRP2,CRKL,ATP6V1A,CXADR,MBP,TXNL1,ZBTB8OS,MRPS18B,SIX5,POLE,PRKCA,BCKDHB,ZWILCH,TAF1C,GNAS,PNN,BACH2,GNA15,GABRR2,HSPA5,CDC73,CRYAB,RFC2,INTS6,NCF4,KATNAL2,CDC23,RBMX2,GPRC5B,QPRT,COL26A1,CACHD1,RPA4,TOMM7,PSMC3IP,GLI3,GPR63,COL9A3,KIF7,ACVR1C,PLCG1,SRP19,GIGYF1,DROSHA,MTX2,ZFAND2A,CLCF1,CEBPE,PIK3CA,IDH3G,COPS9,SMAD6,FLT4,SGO2,UBA52,GUCY1A2,MYOG,RHBDD2,VDAC3,KIF22,EPN3,CEBPB,TRAPPC2L,SUZ12,NVL,IL12A,COL4A4,IL6R,ZMAT5,KCNA5,NEMF,LGALS3,CCDC61,FOXP3,KPNA5,GNA13,CPSF4,SCLT1,CISD2,ITFG2,HOOK3,ELOC,GABRA2,VAMP1,COPS4,OLR1,FIP1L1,FAM86B2,KLHL25,ATP1B1,FGL1,SCN1B,CTDP1,DCAF4,STRA6,PSMC4,EFCAB7,COA6,WDR24,CT45A6,RPAP2,ERN2,CDK9,IDH3A,BAG5,EIF4G3,KIF26A,ARHGAP33,NFE2L2,VPS26B,ATP6V1C1,MMGT1,BNIP1,MAP2,RRP15,PDLIM3,BLOC1S3,AMN,HDAC5,AQR,APC,MCCC1,FEM1A,SMAD7,HLA-E,INTS13, CCNT1, ATP6V1B2, TEP1, SCNN1G, PHC3, GNG4, JUNB, GFRA3, DPF2, MED16,XRCC6,POLR3D,JUND,PRPF40A,NBAS,MYO7B,NFATC4,CHRNA3,DCTN1,RPP21,SPECC1L,SUV39H1,SOCS5,ARPC4,NCOA6,SMDT1,MYO18A,CHTF18,GNG10,RAB3GAP1,RRM1,FBXO11,FLOT2,ACTL7B,GCN1,VPS50,HLA-C, CNST, SS18L1, CLCN7, NPHS2, GATC,EBNA1BP2,ITPR1,ANAPC10,CRB3,CDK2,WHRN,MYF5,ZMAT2,VCP,FCN2,EEF2KMT,ARID5A,CCAR2,SRA1,GNG12,NOMO2,GATA1,DYRK2,GPR20,CNIH3,WASHC3,ASH2L,PLXDC1,SCN4B,BMPR1A,TRAF1,FBXO27,TSN,ITGA10,NEK6,PHAX,GANAB,GABRB3,CCDC115,GNG2,IFT20,NUP133,CENPF,CHRND,CD2AP,SOCS6,ARPC2,HBQ1,EGFR,POLR2F,PBXIP1,GRSF1,KIF18A,NCALD,ZRSR2,HCN2,COL24A1,COL19A1,ACTR1A,VPS37A,DGKI,UQCRHL,BAZ2A,KANSL2,HUS1B,RPL5,SSB,AKAP7,RRS1,CFH,GTF2H4,CREBZF,MRPL51,DICER1,SNRPA1,KCND1,EPC1,TRPV6,HES6,SSU72,PRKDC,ASCL4,TRAPPC3,PSMG3,ERN1,PRPF38B,GTF3C4,CDK1,DCAF17,COPS2,CCNJL,KCNK4,KPNA7,SLC1A6,RPS19,ATG101,EHBP1L1,HEATR1,CUL4B,JARID2,LYN,TNC,TMEM147,BOD1L1,TRAF4,CASP4,OTULIN,TRMT61A,PRIM1,DRG1,NUP210L,POLR1D,IKBKE,CGA,ZER1,PPP2R5C,CLTB,NPR3,CRX,ETV3,INSRR,ENC1,TAF1,HMGXB4,CEBPG,NDUFB9,CETP,ZC3H13,NCF2,NOL10,TMEM17,CACNA1H,ANAPC4,TIMM8A,LAMA1,IL4R,MSH6,CLIC6,CAV3,NELFCD,RPN2,RELB,CCDC68,RAD52,METTL14,DYNC1I2,CACNA1C,CCNB2,RAMP3,BRAP,HSPA2,IL23A,MRPS11,CT45A2,APOBEC1,COL25A1,KCND3,MRPL42,GNG7,FBXO17,SLC7A1,VPS4B,HSD17B12,FEM1B,GRIN1,RAG1,METTL18,EPHA7,CACNB4,DLD,PLA2R1,UBE4A,EED,MAD1L1,GP9,LENG8,TEAD4,CLDN4,REPIN1,FLNA,GSPT2,MRPL17,RTCB,MCM9,EP400,FGFR4,SF1,KCTD5,MYBL2,TUBGCP5,EPHA10,EXOC2,KANSL1L,VCAM1,ZNF350,MRPS23,LIN9,HCN3,NPM1,MAFK,LMTK2,BCL7B,WTIP,ASXL3,CLDN1,ORMDL1,PAM16,MICU2,MTPN,PIGS,B2M,STXBP6,CENPS,JAM3,PES1,SREBF2,BIRC5,SNRNP48,KLC3,S100A1,MEF2B,PALB2,SNAP47,FAS,UBL4A,NOL6,KPTN,GTF3C1,PCGF3,NUP35,SGCD,GCLC,NUP54,STX19,KAT6A,STAT2,PTPRB,PXN,MED26,LHCGR,ZNF10,CFB,PFDN4,MRPL55,INTS1,AP5M1,PDLIM5,PDLIM4,SF3B1,CDH1,KLHL3,BMP2,ITGB3,SSX2IP,CSNK1A1,KIF27,BARX2,KCNV2,CEP290,ARPC5L,UTP6,TTYH1,PDIA3,REL,NDUFS7,HTR2A,RASSF2,C1QTNF5,SOX8,ATG3,RB1,SOS1,AKT2,BABAM1,DIABLO,COL8A2,SYMPK,LIG4,MVB12B,NSMCE2,CATSPERB,INO80B,PTPRN2,RPS16,GTF2E1,ATP6V1G3,TOMM70,HAUS7,FSHR,NDUFB5,FANCL,NDUFS6,SSPN,BTBD1,KCNH1,ELL2,EPS15,GEMIN5,ACR,ACTL8,ST18,SMARCA1,GSTP1,HINT1,TNFRSF11B,NCR1,RFXAP,KCNA1,RRAGA,PLXNA3,WDR74,PHF12,CAPZA1,DISC1,GNG5,SHARPIN,TTC21B,GP1BB,RIC1,NABP1,APOC3,GDI1,PEA15,GEMIN4,MAP1S,INCENP,MYO19,NUP188,ING4,ING5,CALCRL,GLRA1,KLHL24,ZW10,LOX,GLI2,PRKAA1,CNOT7,SPOPL,WDR3,BCL9,HIF1A,ASCC2,MASP1,UXT,CRYAA,ATF1,RERE,SETD1A,TRAPPC4,HMGB2,EXOC4,FBXO7,PRR5,RPAP3,MED17,WDR36,CNNM4,PRPF18,GNAT1,MLXIPL,NDUFA10,ALS2,TFAP4,FGA,GINS2,SUPT20H,GPR84,CSNK2A1,AFF1,PIGR,NAA10,RPS6KL1,IFNL1,FBXO10,MED31,TBX15,KCNQ3,HDAC1,CLIC4,SNAI3,VEZT,OGG1,DNAL4,PHF20L1,MRPS34,CR1L,ACVR2A,SNF8,DCUN1D5,KCNMB3,TRIAP1,RRAGC,MRPL35,GFRA4,INTS3,KLHL12,KCNG3,RPL10A,PPP2R5E,SOX9,AURKB,MPHOSPH10,KIF21A,NUP62,KIF1A,WNK1,FBXL15,DYNC2H1,TRMT61B,UTP14A,C2CD6,DCTN5,SNTG1,OSMR,HTR2C,LRP1,BUD13,DDX21,NPAS2,MYL12A,TRAPPC8,STAT4,PPP4C,LBH,NUP160,UNC80,CDKN1B,FANCI,UBAP1,EPHA3,TPM2,IL18R1,CIR1,MED15,NDUFB11,RPL34,COPS8,VPS51,IGFBP5,PAGR1,SLX1A,CFHR5,TBC1D5,MRPS9,GPAA1,TNNT3,HDAC7,BCCIP,INTS8,ZCCHC8,IMP3,KBTBD6,ARFGEF1,CPLX2,EEF2,TM7SF2,NHEJ1,UPF1,NOD2,THBD,COL22A1,BLOC1S5,TRAF5,CNOT1,AKIRIN2,TAB1,NSUN3,RNPS1,RPS26,TTBK1,MRPS17,METTL22,NAA50,KIF23,CFHR2,USP10,LRRC8D,METTL21A,AIM2,BCS1L,PPP1R7,SNW1,KDM1A,L1TD1,HLA-DQA1, PRMT1, MAK16, TRMT10B, JPH2, GPD2, PRKAG2,DCX,CREB1,FGG,PPAN,PMPCA,MED20,CASP12,ARHGEF7,MYC,MUSK,TXNL4B,EPS8L1,MCM4,ZFYVE9,APOC2,ATP8A2,SCN1A,CUL3,NTRK1,NOXA1,SSRP1,BLOC1S4,SCLY,TOP2B,ATG14,SYVN1,ELOB,GAN,PTPRQ,DNAJC10,SMC1B,PDE4B,KRIT1,SUB1,NOP14	0.035506	0.01	0.0896

	,KIF26B,GEMIN2,AURKAIP1,CD3D,GSC,ADAR,TMEM231,TCTN2,COX10,KLHDC3,RNF144B,TAF7,LDLR,NPPB,LAMA2,RYR2,MYO1D,CDH15,USP33,USF1,POLR3GL,MPHOSPH6,COL9A1,EML1,CASQ2,CRB1,RAD51AP1,MTERF4,NCAPD2,TCF15,APOC1,ISG15,LY6G5C,GRIA1,FKBP4,LRP2,ACTR1B,MZB1,PTGES3,ING3,HAND1,TMEM67,TRDN,IMPG2,SKIL,PHF21A,RNF168,PRPF6,TAF9B,EIF3H,DDX58,PDGFA,CDK5,ASB4,ID2,CUL7,ZAP70,ZC3H12A,CD40,PRMT5,VIM,SGCE,RGPD8,RNF14,VAMP4,TLE6,DNAJC28,KCNF1,ERCC8,PLK4,HP,PMF1,INO80,TRMT10C,MED27,COL4A5,RAD51B,DPYSL3,ZCRB1,BRCA1,PLAT,POU4F1,XRCC3,PAF1,WASHC5,LAPTM5,NUP62CL,RFXANK,LRP5,RPL18,MRPL13,ROS1,CACNA1S,PIK3CG,FKBP8,ORC5,RCHY1,TRIM40,CEP164,BYSL,SRSF1,TIMM50,KAT14,CDC26,PIGC,HPR,MARK4,EIF4G2,TOPORS,DBT,BOD1L2,POLR3E,NDUFA5,COL8A1,FOXF2,RPL8,MPC2,FAP,HSPA1L,NSRP1,HOXB9,FANCG,XRCC4,E2F5,BCL7C,KIF16B,GINS4,NCAPH2,PAFAH1B3,ACTN4,TMEM249,SNX17,ZSWIM6,LYRM4,SNTB2,UQCRC2,ZFP36,MRPL20,CCNF,USH1C,LAMTOR1,CHD4,KCNC2,CDK13,BRMS1L,IFT43,BBS2,CSF1,RIOX2,COX6A1,CACYBP,LIME1,TTYH2,DGCR8,CDC27,UTP14C,FIS1,PHC1,N4BP2L2,ZFPM1,APOE,TAF11,RGPD2,RPL28,SNAP23,PRDM10,MCRS1,BACH1,AEBP2,PROC,CNPPD1,NCAPG,GRIA2,SLU7,RPS15,LDB1,DHX38,NACA2,APPBP2,PPP1R3F,PSMD13,PLXND1,RBM10,PON1,PRPF39,PAIP1,NID1,KLHL41,SESN3,PHKG1,CDKN2D,NDUFA12,MED21,C9orf72,ITPR2,DECR1,SYNE3,NCF1,SCN2A,RPL30,HDAC6,TAP1,PLEKHA2,ASCL5,SAP130,OOEP,ERH,ZFC3H1,OS9,STAM2,MRPL34,KDM1B,NLRC4,GLMN,BCOR,KCNG1,MYO3A,EIF2B5,ITGB4,CCT7,AGFG1,CEP128,TADA2B,PAN2,MRPS35,C1QTNF9B,VPS16,CSF2RB,RBPJL,SEC61A2,TNFAIP1,ZYG11A,STAT1,ALG14,CDH2,BCL9L,IMP4,HLA-G,CPS1,SMC5,NCKAP1L,E2F2,FXYD2,PPP2R5D,COL1A2,HSPA9,ANO2,PHB2,CLTA,ADCYAP1R1,BPTF,FAF2,GABBR2,FDX2,TRAPPC2,MR1,STAT5A,PHF10,IKBKB,PRKRA,NSMCE3,FNTA,MAFG,CCND3,CHTOP,PSMC3,VIPAS39,CREB3,MAX,COQ6,IGLL1,SLIRP,SETD1B,NFRKB,AGO3,PFDN1,AFF3,TNNT1,APOB,SNRNP35,RNF8,SMC4,PPP3CA,SMARCB1,ZC4H2,HTATSF1,MRPL2,POLR3F,ETF1,TAF1B,KCNS1,OLFM4,SDF2L1,MSH2,CACNA1G,FLYWCH1,ARHGEF2,ATP6V0A2,PKD1L1,IKZF5,TIMM9,TANK,IRF4,HNRNPR,ZCCHC7,ITCH,HDAC10,AGO1,PIGP,CHAF1A,F9,SWSAP1,ZNF827,PKD2,GPS1,KCTD17,ERI1,HLA-DRB1,TEAD2,FZD1,ANAPC15,KIF1B,EIF4E2,POLRMT,TTF2,KCNIP4,AICDA,ATP6V0D2,TAF1A,RACGAP1,MAP1B,METTL1,ARVCF,DLAT,LRRC8E,ING2,CNGA3,SECISBP2L,LTV1,BBS4,KIF20A,CREBBP,FKBP1B,PDGFB,PARD6G,FGFR1,GLDN,FOSL2,RUVBL1,EIF3CL,ACTL6B,TAF7L,TRAPPC6A,FBXO21,TEK,FBXO8,PLTP,DIS3,ITGA11,AKAP4,CLEC12B,AGO2,ADRB2,RPS10,KCNN4,TERF2IP,SIAH1,EID3,SGF29,BET1L,DMXL2,LAS1L,TIMM44,IL23R,CAMK1G,SRPRB,C6,FAM98B,TRIM54,SLC17A6,NPIPA1,COPB2,CACNB3,HEXA,DPM3,DNM3,RPP40,DIXDC1,PIDD1,DUSP22,RRP36,COL14A1,ZNF224,WBP4,NAA20,CDH10,L3MBTL1,NOC4L,PIGQ,CDK6,RAG2,SPATS2L,CDH4,EEF1AKMT3,MYL1,MRPL22,GNG8,PSMC1,KCNE1,PLCB3,TMEM216,ATF7IP,DESI1,ANAPC5,SUPT4H1,MED18,CD3G,DNAH10,DNAH7,ING1,LY6G5B,PLAU,KCNMA1,APOBR,CCT3,CDH17,CARMIL1,IL27RA,HNRNPUL1,MTNR1A,CARD11,STX1B,CCNK,FXYD1,DQX1,HLA-DPB1,NUDCD3,UGT3A1,GPRC5D,HBG2,BECN1,NCOA3,SMTNL2,ASCL2,NKX2-1,TAPBP,GTF2F2,BAG6,BMPR2,PPP2R3B,QTRT2,CRK,MAGED1,TNNI3,MAGT1,MYOD1,ORMDL2,CLUAP1,SKA2,MRPS10,EIF4EBP3,FZD8,GPBAR1,CARD9,HLA-DMB,LPL,ANAPC1,ERG,UBE3A,CLDN3,MRPS7,COL4A6,MED28,JADE3,SNAPC4,EPHA1,RPL15,ADIPOQ,DEPDC5,KCNMB1,DHX33,NDUFS3,CORO1A,NHLH2,SCN8A,MCM3AP,BAK1,SP3,CLNK,MRPL44,AKAP9,MYL12B,SPAG1,HFE,ATP6AP2,SCRIB,CKAP5,KCND2,CDK4,IZUMO1,COL2A1,HBM,RBBP5,SNAPC3,SDHA,PDGFRA,DNAH6,PIK3R6,TF,IL6,MYO5B,AXIN2,ABCC8,POU2AF1,KPNA2,COG1,PRR5L,PRAME,ABCC9,IFT27,GPRC5A,SYNPO2,COPS7A,PLA2G7,ROR2,SEM1,HAUS6,PRKAR1B,COL27A1,MRPL4,SIPA1,CSNK2A3,POP1,COL6A6,SNTB1,CCT5,PARP1,PRKAB1,SRPRA,CD44,LAMTOR2,ZNHIT6,MZT1,FANCF,KPNA1,NAPA,AXIN1,XRCC2,JADE2,SPHK2,CTSB,PLCB2,PDE3B,FOXF1,UGGT1,VCL,SPAG4,SKIV2L,PEX13,ITGAV,HBA1,SDR16C5,SYNE2,ELOA,GCHFR,ROGDI,CAMLG,SNU13,COL1A1,GEMIN7,HLA-F,SLC3A2,LARP6,SNTA1,TRPS1,COLEC12,APC2,GTF2A2,RB1CC1,CRYBG3,HLA-DMA,SLC51A,PSMD3,CEP250,RNF2,MYL9,SOX2,KEAP1,SNAP25,SNTG2,CDH26,HOXA9,KIF15,PEX3,ADSL,PKD1,POP7,TES,MFAP1,RPL38,RNF215,THBS1,CHM,SRP68,MYL2,PPP3CB,RBM42,NDN,WBP11,SMG1,RYR3,GATA4,UVRAG,PPP1R15B,FBL,ATP23,ERC1,HLA-DQB2,GABRE,RFX5,TPM4,PRDM4,FBXL8,EDA,RCOR3,MVB12A,IFT57,TRPM4,CSF3R,TAF1L,THOC2,GTF2H5,COX7B,VAMP8,MRPL41,ZC3H3,DNAJB11,SHMT2,GSK3A,IGFBP6,CXXC1,CLCN1,NFATC2,ERCC2,PCGF6,GRIN2C,C1QC,DPF1,EXD1,ANO6,CKS1B,C1D,KDM2B,SCML2,FBXO15,GET4,CAPNS1,PRPF40B,COPE,KIF20B,COL6A3,VPS37B,AUP1,COX6C,STAG2,MGA,KMT2A,PPP3CC,HAT1,CCDC102A,FBXO24,AZI2,CGN,BIRC2,GABRD,ACTR3,CACNB2,CPSF1,GTF2H1,DCLRE1C,MRPL23,ERLIN2,TTI2,ATP11A,PIGT,GPSM2,ATG12,BAZ1B,MED4,COX7A1,NAPB,MALT1,IFT81,HMGA1,AMFR,STAG3,PXMP2,INHBA,EXOC3L4,MBD2,COL13A1,PSMD4,TIMM8B,SLC6A6,ATP6V0C,CHRFAM7A,CHRNA7,VPS28,GFI1B,CBX5,COX8A,VPS41,GAPDH,KCNMB2,ABT1,E2F1,AKT1,LIN28A,ERCC1,TOPBP1,CCP110,CAPNS2,CDH23,MMRN2,ORC4,CAPRIN2,BOLA1,CLIP2,MYH15,EXOC6B,TRERF1,NOTCH2,AGO4,INO80E,DCUN1D3,KCNAB2,TIMM10B,CFAP206,ZNF703,DCTN3,CXCR4,ATP1B3,BMPR1B,TRMT6,ACTR6,SEC61B,MAPKAP1,PIH1D2,TEFM,NDUFS8,MDM2,DYNC1LI2,SUFU,KCNA7,KRI1,KLC2,TUBG1,MYO1A,RASIP1,POLD1,NUP37,TMED10,STK4,KCNIP3,C5,TADA3,SPTB,MAP4,PPARGC1B,KIF5C,PKD1L3,DHX9,GABRA4,NFE2L3,NFE2,MYF6,GPR156,KIF13A,ORAI1,CDHR3,EP300,VDAC1,BRF2,LILRA4,PDCL3,GFI1,SNX5,TMEM183A,UPF3B,ATP4B,PRKAR2B,COQ5,PHC2,GATAD2A,MTCP1,BLOC1S1,UBXN7,AURKA,NEUROD1,SPTBN5,HOOK1,STAT6,NCAPD3,ATXN7L3,KCNA6,PEX11A,MRPL50,IFT172,PLOD1,ATP6V1G1,UBR2,LDB3,MEFV,BCAS2,IKZF1,LMO4,ABCB9,MFF,PSEN1,VAC14,PKD2L1,COG7,TTR,ABHD12,NAA25,RCOR1,CAPN2,UQCRB,FTH1,MRPL28,MBD3,MMS19,B9D2,FOXE3,EPB41L2,CDK8,GRIA4,DYNLT3,PRKAG3,GCM1,RNF40,PFKM,SMTN,C2orf49,MEF2A,BAX,SNRNP27,STT3A,COL16A1,SEC11C,CCM2,IFNLR1,PARD6A,FBXL4,CNTNAP2,EPHA2,GRIN3A,CACNG3,POM121,APOA4,KDM6B,BOD1,CNGB3,TAF6L,NFATC1,PCP4,MVP,CDH19,NUP205,GPR160,GPIHBP1,AHRR,GPR61,LAGE3,JAZF1,MSX2,AURKC,HR,MRPL33,SMC1A,ARID1A,TPM1,CD34,JDP2,MED13L,ENTHD1,DMXL1,IL2RG,MRPL3,MAP7,IL12RB2,MRPS27,SCIMP,SGCB,TRADD,ATF2,CD9,TRAF6,RPL3,MAFB,THEMIS,WDR41,EME1,NLRP9,RPS8,RSPH4A,KCNC1,NACAD,EHBP1,UQCR11,TFDP2,LUM,CACNG2,KIF4A,RRM2,RANBP2,MXD1,COQ4,DCHS1,ATP8B1,KANSL1,NUPR1,COMP,TIMM29,DCAF12,HAX1,ORC3,SEL1L,MON1B,DPY30,PTCD3,SKOR1,TNP2,REST,PDP1,INSC,TLR2,PDHX,PDHA2,NDE1,KLHL7,NOTCH3,COQ7,DBI,DCAF4L1,GRID2,TUSC3,SAR1B,CPSF4L,GRB10,IMMP2L,TERF2,LAT,PDCD11,CDK5R2,CLDN17,RAD50,IGF1,CHRNA1,KLHDC2,IRS1,NDUFS5,ATP8A1,CAMKMT,RPRD1A,PRM2,IL5RA,UBR3,FCRL5,DLG4,HOOK2,KLHL9,DNM2,NAE1,PPP1R12A,MYO10,ACVR2B,KCNAB3,SCN2B,UPF2,GRHPR,POLG2,NDEL1,COL5A3,ATP6V1G2,SKP1,LDHB,CCNG1,BRCC3,PLXNA1,MREG,COX7A2,SIRT2,ATP6V1D,CEP97,MTTP,RNASEK,SASH1,HSPA8,DERL3,WASF1,SAP30L,FH,NTRK3,MYO7A,LRRK2,TNNT2,GTF2A1L,NOP56,RPF1,EME2,SYNE4,DDX1,SPC25,MYL7,HUS1,PDGFRB,RUVBL2,TRAF2,MAEA,SLC7A5,E2F6,SLITRK5,RNMT,OTX2,E2F3,SLC17A7,MYH7B,SF3A1,BLM,FKBP1A,PLN,EMC6,CNOT11,PDCD10,TCF3,RNF111,KCNA2,RNF220,F13B,CPLX3,PDIA6,KIT,GLDC,AXL,ATP6AP1,CBX8,TRA2A,STK11,KIF18B,RNF113A,DBF4,CASP1,NDUFB1,TSR1,TPM3,HSD17B10,DCTN6,NELFB,STX3,ZNHIT1,CHRAC1,MED25,OLFM3,ACTR5,RAD51C,NCBP3,KCTD16,UTY,CWC15,GINS3,DEK,HSPH1,CBX3,ECT2,INO80C,ALK,TSNARE1,RBM8A,RFC5,RASGRP3,PPM1E,IFT88,NCKAP1,SLF1,TMEM30A,MRPS12,JMJD6,ITGB5,SP1,CD2BP2,APOBEC3F,DNAJC11,EIF4G1,WASHC2A,CASP8,FANCM,KCNQ4,ABCD4,CLMN,ASB1,CCT2,NOXO1,MEP1A,AIP,EPS15L1,KLF5,PTK2B,PSENEN,SASS6,MSX1,EPHA8,HDGF,CAV1,STXBP1,HDAC4,DENND4C,NCSTN,CALCR,PKP1,HAUS1,CRLF1,PSMC2,PDPR,ITGA8,MRPL38,HDAC3,SFR1,TUBGCP6,SNRNP200,MED29,CLSTN3,DCUN1D1,TMEM258,CSTA,PPM1F,MAT1A,COG6,RBM28,GTF3C3,HLA-DOB,PECAM1,ATP8B3,INTS10,CDC37,LMO7,MRPL52,MLEC,AGK,BEX1,PRKCSH,SMN1,IFITM3,ZNHIT3,CHRNA2,MBTD1,GATAD2B,RORA1,ITGB7,NIPBL,COPS6,NOX4,RPP38,CBX2,ASXL2,MRPS5,ATP6V0B,CTR9,SMARCA2,MRPL36,EIF2S3,CCNB1,MED10,TGFBR3,CFLAR,DNAAF2,PSMD8			
23	ANKIB1,DTNA,RNF139,ARIH1,POLR2I,ZFAND2A,FANCL,KDM1A,SYVN1,RNF144B,RNF14,RCHY1,DIS3,AMFR,RNF144A,RBCK1,DIS3L,DIS3L2,POLR3K,ARIH2	0.003229	0.02	0.1416
48	MAML1,PYGO2,CBLL1,NCOA6,SS18L1,EP400,BCL9,SAP130,BCL9L,ARID1A,CITED2,DAZAP2,SS18,ARID1B	0.004706	0.02	0.1416
69	CALM2,CALM1,CALM3	0.014049	0.02	0.1416
76	CAV2,CAV3,CAV1	0.001469	0.02	0.1416
102	CR2,APOH,CFH,CR1L,CFHR5,CFHR2,F13B,CFHR1	0.009177	0.02	0.1416
163	ELAVL1,PABPC3,PABPC1,PABPC5,PABPC4L,PABPC1L,ELAVL2,ELAVL3,PABPC4	0.002709	0.02	0.1416
175	HNRNPH1,GRPEL2,HNRNPH3,TFAM,ESRP1,GRPEL1,ESRP2,GRSF1,METTL18,HNRNPR,A1CF,SYNCRIP,CYB5B,HNRNPF,COQ3,HNRNPH2,TREX1	0.003392	0.02	0.1416
198	KMT2D,FLYWCH1	0.001861	0.02	0.1416
237	HLA-B,HLA-E,HLA-C,NCR1,HLA-G,HLA-F,LILRA4,HLA-A	0.003783	0.02	0.1416
251	CCDC113,IFT74,NDC80,CLUAP1,IFT57,IFT81,NUF2	0.003096	0.02	0.1416
271	KCNK1,KCNH2,HCN4,KCNB1,KCNQ5,KCNA4,KCNK6,CNGA1,HCN1,EPB41,KCNA5,HCN2,KCND1,KCNK4,KCND3,HCN3,PTPRN2,KCNH1,KCNA1,CNNM4,KCNQ3,HDAC1,PPP4C,TTBK1,PDE4B,MARK4,KCNC2,NACA2,PPP3CA,CNGA3,SLC17A6,KCNMA1,KCND2,PRKAR1B,PPP3CB,PPP3CC,KCNA7,PRKAR2B,KCNA6,EPB41L2,CNGB3,KCNC1,SLC17A7,KCNA2,KCNQ4,HDAC3,CNGB1,SORBS1,KCNC3,PRKAR1A,ZDHHC9,KCNC4,KCNN1,KCNQ1,KCNH4,WNK4,PRKAR2A,PPP5C,KCNA3,ADRA2A,KCNK2,KCNQ2,PDE4D,CNGA4,HDAC2	0.001805	0.02	0.1416
317	MDM4,RPAP2,MDM2	0.002545	0.02	0.1416
324	PAFAH1B2,VCPKMT,FAM86B2,EEF2KMT,TRMT61A,NSUN3,METTL22,METTL21A,PAFAH1B3,EEF1AKMT3,FBL,CAMKMT,NAT9,N6AMT1,NSUN4,METTL23	0.001386	0.02	0.1416
409	FCGR3A,VSIG4,VCAM1,PECAM1,CD200R1	0.007655	0.02	0.1416
463	RAD9B,PCNA,HUS1B,HUS1,RAD1,RAD9A	0.001273	0.02	0.1416
471	RANBP3,RANBP3L	0.011187	0.02	0.1416
481	RGPD3,RGPD1,COMMD1,RGPD4,RGPD5,HPS3,RPA4,RHBDD2,NBAS,GTF3C4,ANAPC4,INTS8,RGPD8,FANCG,RGPD2,VIPAS39,TAF1A,LAS1L,CFAP206,RANBP2,INTS10,KNTC1,WDR93,NME8,AIMP2,DTX3L	0.007311	0.02	0.1416
544	EIF3G,SSB,ZCRB1,SNRNP35,RBM8A,SNRNP70	0.001668	0.02	0.1416
631	UPK1A,TSPAN33,TSPAN17,CD9	0.008778	0.02	0.1416
634	PRKRIP1,EBNA1BP2,RRS1,RRP36,UTP11	0.00454	0.02	0.1416
635	GPATCH1,UTP14A,NOP14,UTP14C,KRI1	0.003316	0.02	0.1416
654	AFTPH,ZFYVE9	0.003151	0.02	0.1416
674	MSR1,COLQ,CD74,SCARA3,KLK8,NCCRP1,UGGT2,FCN1,C9,SFTPD,TMED3,ENG,C1QTNF1,C1QTNF2,COL3A1,EPPIN,MBL2,FCN3,COL23A1,SUN2,CCBE1,C1QL2,C2,COL15A1,C1QTNF7,COL4A3,PTN,SFTPA2,EMILIN2,COL11A2,CTHRC1,COL26A1,COL9A3,COL4A4,FCN2,COL24A1,COL19A1,COL25A1,PLA2R1,CFB,C1QTNF5,COL8A2,ACR,LOX,MASP1,IGFBP5,COL22A1,COL9A1,HP,COL4A5,PLAT,HPR,COL8	0.004445	0.02	0.1416

	A1,PROC,OS9,C1QTNF9B,COL1A2,OLFM4,F9,GLDN,C6,PLAU,COL4A6,ADIPOQ,COL2A1,COL27A1,CTSB,UGGT1,SPAG4,COL1A1,COLEC12,EDA,IGFBP6,C1QC,COL13A1,MMRN2,TMED10,COL16A1,COL5A3,OLFM3,ALK,MLEC,TGFBR3,COLEC11,LOXL4,C1QL1,OLFM2,ELANE,COL5A2,COL5A1,KLK3,C8A,VTN,SFTPA1,C1QTNF6,TMED7,IGFBP3,GNPTG,COL9A2,F2,SUN1,C1QL3,COL21A1,C1QA,C1QB,C7,MMRN1,COLEC10,COL18A1,COL4A2,COL10A1,F10,MARCO,COL28A1,EMID1,SDC1,COL4A1,C1QTNF3,F7,EMILIN1,C1QTNF9,IGF2R,LTBP1,C8B,COL6A1,COL6A2,COL11A1,COL17A1			
689	RPN1,EEF1E1,SPTSSB,CLIC3,DDOST,ALB,SPTSSA,MTX2,RPN2,PIGS,PDIA3,GSTP1,CLIC4,TXNL4B,DNAJC10,CASQ2,PON1,FAF2,APOB,MAGT1,PIGT,PDCL3,UBXN7,TTR,TUSC3,MTTP,PDIA6,CLIC5,EMC7,TXNL4A,CLIC1,IFT52,MTX3,SPCS3,FAF1,CLIC2	0.002847	0.02	0.1416
810	PRDM16,GATA6,SOX12,HOXC9,ASCL1,DACH1,TAF6,POU3F1,BCL11A,SOX17,TBX2,INSM1,SKI,TCF4,MXD3,HOXD10,PROP1,BCL11B,CDX2,FOXP3,ARID5A,GATA1,HES6,ASCL4,CRX,BARX2,SOX8,SOX9,AKIRIN2,GSC,TCF15,HAND1,ZC3H12A,POU4F1,FOXF2,HOXB9,ZFPM1,ASCL5,ASCL2,NKX2-1,FOXF1,SOX2,HOXA9,GATA4,GFI1B,ZNF703,GFI1,FOXE3,MEF2A,TAF6L,MSX2,MAFB,SKOR1,OTX2,TCF3,MSX1,LEF1,NKX2-5,TCF12,TCF7L1,HOXA10,HAND2,SOX18,KANSL3,TCF7,ALX4,HOXD12,MEF2C,PITX1,GATA2,SOX4,SOX15,TCF7L2,HOXA11,HHEX,PITX2,UNK,MAF,DOT1L	0.003327	0.02	0.1416
829	RBM12B,RNASEH2B,LARP7,MTA1,EIF3D,EAF2,POLE3,MTA3,CC2D2A,EIF2S2,TAF13,SRP9,HYPK,RANBP3,BAZ1A,NPLOC4,PARP11,INTS12,POLR3G,RBM41,RNPC3,MEAF6,NECAP1,THRAP3,TAF12,MIER1,DR1,BCL10,DMAP1,CDC5L,MKRN3,ARID4B,PIK3C2G,SUPT6H,EAF1,MTA2,PIK3C2B,CARD8,RANBP3L,NCOR2,GEMIN6,TAF8,POLR2D,MLLT1,SNUPN,CCNT2,CCDC47,RRP7A,PNN,CDC73,ZMAT5,NEMF,CPSF4,CCNT1,TEP1,POLR3D,SOCS5,SOCS6,POLR2F,CREBZF,EHBP1L1,PRIM1,POLR1D,ZC3H13,CLIC6,LIN9,NPM1,CENPS,ATG3,LIG4,PEA15,RERE,PRPF18,UBAP1,ZCCHC8,L1TD1,KRIT1,SUB1,TAF7,POLR3GL,MPHOSPH6,TRDN,TAF9B,CUL7,DNAJC28,PAF1,KAT14,TAF11,NDUFA12,ZFC3H1,TADA2B,BPTF,HTATSF1,ANAPC15,TAF7L,FBXO21,FBXO8,MED18,APOBR,CCNK,BAG6,BAK1,CLNK,SNAPC3,PIK3R6,SEM1,PRKAB1,ZNHIT6,PEX13,LARP6,PEX3,RCOR3,EXD1,KMT2A,BAZ1B,TADA3,MAP4,PKD1L3,RCOR1,BAX,KDM6B,PCP4,LAGE3,RSPH4A,NACAD,DPY30,CPSF4L,CHRAC1,NCBP3,PPM1E,SLF1,HDGF,CSTA,RBM28,KMT2B,PPP2R3A,TIPIN,MIER2,NXF3,NFYB,NXF5,CARD16,CCDC8,NEDD4,MKS1,MLLT3,NECAP2,BCLAF1,CAT,PIK3C2A,TRIP4,URI1,SMARCC2,PPP1R15A,FRG1,TAF9,PPP1R8,TNKS1BP1,ERCC6,BAHD1,SOCS4,ZC3H8,RCOR2,ARID4A,ANKRA2,PRKAB2,MAPK6,SMARCC1,CEP76,NCOR1,PIK3R5,TADA2A,CHTF8,MIER3,BCL2L2,PPP4R2,PSMG1,DRAP1,THG1L,EZH2,IFT46,LCP2,RSF1,CLNS1A,MCL1,BTF3,BCL2L1,CRCP,EZH1,MAVS,GTF3C6,SUPT3H,SLF2,BCL2,CUL9,SMTNL1	0.004067	0.02	0.1416
67	CACNA2D1,CACNA2D4,CACHD1,CACNA2D3,CACNA2D2	0.00147	0.03	0.2011
93	OGT,MYB,CDC23,MYBL2,TOMM70,CDC27,BBS4,SEL1L,UTY,CDC16,KDM6A	0.004878	0.03	0.2011
222	GPRC5C,GPRC5B,GPRC5D,GPRC5A	0.003586	0.03	0.2011
287	KPNA4,KPNA5,KPNA7,KPNA2,KPNA1,KPNA3,KPNA6	0.001411	0.03	0.2011
363	WASHC4,NAA35,WASHC5,NCKAP1L,NCKAP1	0.006661	0.03	0.2011
375	ATF6B,ATF6,NFIL3,BACH2,NFE2L2,MYC,BACH1,CREB3,NFE2L3,NFE2,ATF4	0.009705	0.03	0.2011
448	STAU2,TARBP2,ADAR,DGCR8,PRKRA,STAU1	0.003284	0.03	0.2011

*System_name: the name/index of the identified biological systems by the proposed MPGNN-Hilander model.

*Genes: genes contained in the biological system.

*Selection_pressure: calculated using the group-LASSO weights by HiSig.

*p: the *p-value* of the permutation test.

*q: the FDR.

Appendix D: The biological systems under significant mutation pressure.

System_name	Genes	Selection_pressure	p	q
807	BRCA2,BRCA1,NRIP1,ADNP,JMJD1C,MED1,MEN1,SLX4,PYHIN1,REV3L,MED13,ASXL2,RFX5,ASXL3,RBM44,DCAF17,MED13L,CRYBG3,AFTPH,MAP2,CEP192,XRCC1,BCOR,CNST,ENTHD1,PPARGC1B,RMI1,GPKOW,KDM3B,NSL1,TERF2IP,FAM186B,ERCC1,AIM2,RB1CC1,ATF7IP,GCM1,PPP1R15B,MDM4,AKAP4,BCL7A,ZBTB16,NCAPH,C2orf49,FANCB,MUS81,ELL3,APOF,DSN1,RNASEH2A,MIS12,LIN37,VEZT,APOL1,DYDC2,CENPJ,EME1,CT45A6,TANK,RPP40,HAUS6,ACIN1,AKAP5,AKAP7,ANAPC1,ASXL1,ATAD5,ATF7IP2,ATG13,ATXN7,BAZ2A,BCL7B,BCL7C,BDP1,BOD1L1,C2CD6,CAMLG,CCP110,CDCA8,CSF2,CT45A10,CT45A2,CT45A3,DBF4,DBF4B,DISC1,DNAAF2,DONSON,FAAP20,FAAP24,GTF2A1L,HEXIM1,ICE1,ICE2,IFT43,IKZF5,IL12A,IL6,IMPG2,JRK,KATNBL1,KNDC1,LRRC46,MBIP,MDM2,MED19,MED26,MGA,MLH3,MSL1,MZT2A,MZT2B,N4BP2L2,NCAPH2,NELFA,NIPBL,NKD1,PALB2,PHAX,POLR2M,POT1,POU2AF1,PRDM4,RAD51AP1,RBBP8,REST,ROGDI,RPAP2,RPP21,RRP8,SECISBP2,SECISBP2L,SFR1,SGO2,SKA3,SNAPC4,SRFBP1,STIL,SUPT20HL1,TERF1,TERF2,TESPA1,TET1,TINF2,TOPORS,TP53BP1,TRPS1,TSEN34,UBAP1L,UIMC1,UNC80,ZC3H3,ZCCHC7,ZFYVE9,ZNF217,ZNF827,ZWILCH	0.167341	0.02	0.1418
59	KMT2C,KMT2E,SRCAP,CHD4,SMARCA4,ORC1,GLYR1,SMARCA1,MBTD1,SCML2,PHF12,JADE3,BRD8,KAT6B,PHF20,PIAS4,BRF1,NXF1,ANP32E,HMGB2,SMARCA2,HNRNPU,NT5DC3,CBX3,TBC1D5,SMARCB1,BRD7,CBX5,ING3,PARP1,MCRS1,KAT8,JADE2,KAT6A,LIN28A,ASH2L,BCL3,BOD1,BOD1L2,BRD1,BRD9,BRPF1,BRPF3,CHD3,CHD5,CSNK2B,CXXC1,DBI,DCLRE1C,DGKI,DNMT3A,DNMT3B,DPF1,DPF2,DPF3,ECT2,ERN1,ERN2,EXOSC10,FBXL19,HMGB1,HNRNPUL1,ING2,ING4,ING5,INO80,JADE1,JMJD6,KAT2A,KAT2B,KAT5,KAT7,KDM2B,KDM5A,L3MBTL1,LIG3,MPHOSPH8,MSH6,PAN3,PAXIP1,PBRM1,PHF10,PHF14,PHF20L1,PRICKLE4,RFC1,RIOX2,SET,SETD5,SMARCA5,TIMM50,WWOX	0.153701	0.02	0.1418
716	PIK3CA,CDH1,AKT1,PIK3R1,ERBB2,MYH9,CASP8,ERBB3,KDM6A,CACNA1E,DMD,GRIN2A,KIF17,KIF4A,CPS1,MYB,MYO6,SCN10A,FCRL5,MYO7B,CHRNB2,ACTN4,CUL4B,MYO10,WWP2,SCN11A,CDH20,PMS1,KIF5B,FGFR2,KIF16B,KCNA1,DNM1,CDH12,DCHS1,DLG1,MCM7,EPHA5,MR1,PARP9,DNM3,KCNK2,LYN,SPTA1,RGS4,SAE1,KLHL13,ABCB8,KIF1A,SPTB,HSPD1,PDE4B,CFTR,MYH7B,KCND1,HNRNPH3,ACTL6B,HGS,MYO16,MCOLN1,SPOP,CNGA4,LDHC,SMC1B,GPSM2,CCT6B,LILRA4,HCN3,A1CF,GABRB2,KCNA5,CATSPER3,KPNA7,UTRN,EFTUD2,KIF5C,CCT2,TARBP2,KCNQ2,KPNA5,MYH14,KLHL41,NLRP1,EEF2,NDUFV1,VPS4B,PIK3CB,GRPEL2,BLM,HSPA2,GLDC,NDUFS2,SMC1A,TRIM21,CDH10,CHRNA5,P4HA1,PIK3R4,KCNAB1,CCT3,CACNG3,PPP2R5C,STAU1,OGDHL,TRPC1,TRPC3,GSPT2,KPNA2,PIK3R3,RECQL5,SNX6,HYOU1,CKAP5,PIK3CD,SYNCRIP,KCNS2,CSNK1A1,MYH10,PDE4D,PIWIL2,GRIK2,MLH1,HDAC8,KCNA2,PSEN2,ATP6V1A,MSH2,ACTR6,IKZF1,MYO7A,AXL,PRMT5,CACNA1F,FLNA,GABRA4,HNRNPH2,UEVLD,DYRK1A,WWP1,OGDH,B2M,SKP1,CDK5,GLUL,CLCN2,ESRP2,IVNS1ABP,DPYSL3,CAPN2,SEC61A1,PRKAR1A,PRKAR2A,PRKAR2B,SEC61A2,CLCN1,EPHB3,KIF3A,AGO4,FOXRED1,CACNG7,NAA10,PKD2,PRKACA,GABRR1,KIF22,SPTBN2,KIF1C,RNF220,CASP5,KIF26B,ZMPSTE24,CPT1C,HDAC5,TPM3,PRKX,SDHB,GRIA3,OGT,KCNH1,DDX6,METTL18,KCNA4,CDH8,KCNC2,SCN2A,PSMC5,CNGB3,KIF9,PFKP,ADSL,SLC17A6,RAD50,DGCR8,PPFIA1,MYO1A,CHRNA3,GABARAP,CUL1,RRM1,CNOT2,KLHL12,NVL,TELO2,CCT7,DDR2,ILVBL,SLC17A7,PTPRA,CAPG,CDH24,APOBEC1,AGO2,UBA52,RPA2,MAT2A,HDAC3,PPP2R5A,HSPA8,PLCG1,ATP6V0B,DPM1,HSPA9,HSPA1L,AURKA,CACNG1,HSP90AA1,CLDN17,CSNK1E,PPP2R5B,PCGF6,GCLM,UXS1,IGLL5,HLA-DQA2,AGO3,SLC6A3,RNF2,DDR1,AURKB,DUSP22,CACNG2,KLHL7,CCT8,HDAC6,CUL4A,KIF4B,GSPT1,VIM,GRIN1,DDX23,TFR2,TTI2,DYRK2,KLHL20,ACTR1A,MIF4GD,SLC1A6,GRB2,PPP3CB,SCN7A,STAM2,COQ6,DHX15,KIF20A,RNF135,SLC1A3,KCNK4,MRPL23,ERCC3,INSRR,ITCH,PCGF3,KCNA6,SLC6A6,SPTLC2,NDUFS8,KBTBD6,PCGF2,PPP4R4,EPHB2,AICDA,SAP18,EZR,COQ3,DDX21,TRPV3,GNA13,CLCNKB,GNAT2,GABRA2,MCM3,RRP1,FARSA,APOBEC3G,DDX41,LIN7C,NABP2,CDHR3,ETFB,HDAC2,CHRNB4,PPM1F,EPHA3,PTPN1,VPS4A,HNRNPR,WNK4,PIK3R2,SHMT2,EPB41,PCGF5,PDSS1,DDX46,CDK7,GCH1	0.098225	0.02	0.1418

	,GPD1,KCNAB2,KCNA7,MX2,CHRFAM7A,CLPX,DHX32,HLA-E,ESRP1,CASP1,KIF13A,SUMO1,ZDHHC9,HINT1,CATSPER2,NEK6,SCN3A,CACNG6,GRIA4,KLC3,NLGN1,KIFC3,ABAT,ABCA2,ABCB6,ABCB9,ABCD4,ABCG5,ABCG8,ABTB1,ACTL10,ACTL6A,ACTL7A,ACTL7B,ACTL8,ACTR10,ACTR1B,ACTR2,ACTR3,ACTR5,ACTR8,ADAR,ADCYAP1R1,ADRA2A,AFG3L2,AGO1,AKT2,ANO1,ANO2,ANO6,AP5Z1,APOBEC3F,ASCC3,ATG12,ATP6V0A1,ATP6V0A2,ATP6V0A4,ATP6V0C,ATP6V1B1,ATP6V1B2,AURKC,B9D1,B9D2,BBS4,BCKDHA,BCKDHB,BCS1L,BIRC2,BIRC3,BIRC5,BMI1,BRAP,BRIP1,BTBD1,C8G,CACNA1A,CACNA1B,CACNA1C,CACNA1D,CACNA1G,CACNA1H,CACNA1I,CACNA1S,CACNG4,CACNG8,CAPN1,CAPN3,CASP10,CASP12,CASP2,CASP3,CASP4,CASP9,CCT4,CCT5,CCT6A,CDC16,CDC23,CDC27,CDH11,CDH13,CDH15,CDH17,CDH19,CDH2,CDH22,CDH23,CDH26,CDH3,CDH4,CDH5,CDH6,CDK1,CDK14,CDK16,CDK19,CDK2,CDK3,CDK4,CDK6,CDK8,CDK9,CFLAR,CHRNA1,CHRNA2,CHRNA4,CHRNA7,CHRNA9,CHRNB1,CHRND,CHRNE,CLCN7,CLCNKA,CLDN1,CLDN3,CLDN4,CLPP,CNGA1,CNGA3,CNGB1,CNNM4,CNOT7,CNOT8,COL20A1,COPS5,COQ5,CRK,CRKL,CRYAA,CRYAB,CSNK2A1,CSNK2A2,CSNK2A3,CTDNEP1,CTU1,CUL2,CUL3,CUL5,CYB5B,CYCS,DAPK1,DDX1,DERL1,DERL2,DERL3,DESI1,DHDDS,DHX16,DHX29,DHX33,DHX35,DHX38,DHX8,DHX9,DLD,DLG3,DLG4,DNM1L,DNM2,DPH1,DPYSL5,DQX1,DRG1,EEFSEC,EFL1,EGFR,EIF2S3,ELOB,ELOC,ELP3,ENC1,EPB41L2,EPHA1,EPHA10,EPHA2,EPHA4,EPHA6,EPHA7,EPHA8,EPHB1,EPHB4,EPHB6,ERBB4,ERCC2,ERCC6L2,ETFA,FAU,FEN1,FGFR1,FGFR3,FGFR4,FH,FTH1,FTL,FXYD1,FXYD2,FZD1,FZD8,GABRB3,GABRD,GABRE,GABRG1,GABRP,GABRR2,GAN,GAPDH,GATB,GCLC,GCN1,GFRA4,GLRA1,GLRA2,GLRA3,GLRB,GMPR,GMPR2,GNA12,GNA15,GNAT1,GNAZ,GPAA1,GPD1L,GPD2,GPSM1,GRAP,GRIA1,GRIA2,GRIK3,GRIK4,GRIK5,GRIN2B,GRIN2C,GRIN2D,GRIN3B,GRPEL1,GRSF1,GSK3A,GSK3B,GTPBP1,HADHA,HCN1,HCN2,HCN4,HDAC1,HDAC10,HDAC11,HDAC4,HDAC7,HDAC9,HFE,HLA-A,HLA-B,HLA-C,HLA-DMA,HLA-DMB,HLA-DOA,HLA-DOB,HLA-DPA1,HLA-DPB1,HLA-DQA1,HLA-DQB1,HLA-DQB2,HLA-DRA,HLA-DRB1,HLA-DRB5,HLA-F,HLA-G,HLTF,HM13,HMGCL,HNRNPF,HNRNPH1,HSP90AB1,HSP90B1,HSPA1A,HSPA1B,HSPA5,HSPB1,HSPB8,HSPH1,HTRA2,HTT,IDH3A,IDH3B,IDH3G,IFITM1,IFITM2,IFITM3,IGF1R,IGHMBP2,IGLL1,IKZF4,INSR,INTS4,IQGAP1,ISG15,JAZF1,KATNA1,KATNAL1,KATNAL2,KBTBD7,KBTBD8,KCNA3,KCNAB3,KCNB1,KCNC1,KCNC3,KCNC4,KCND2,KCND3,KCNF1,KCNG1,KCNG2,KCNG3,KCNH2,KCNH4,KCNK1,KCNK6,KCNMA1,KCNN1,KCNN4,KCNQ1,KCNQ3,KCNQ4,KCNQ5,KCNS1,KCNS3,KCNV2,KEAP1,KIDINS220,KIF11,KIF12,KIF13B,KIF14,KIF15,KIF18A,KIF18B,KIF19,KIF1B,KIF20B,KIF21A,KIF21B,KIF23,KIF24,KIF25,KIF26A,KIF27,KIF2A,KIF2C,KIF3B,KIF3C,KIF5A,KIF6,KIF7,KIFC1,KIFC2,KLC1,KLC2,KLC4,KLHL2,KLHL21,KLHL22,KLHL24,KLHL25,KLHL3,KLHL40,KLHL42,KLHL8,KLHL9,KPNA1,KPNA3,KPNA4,KPNA6,LDHA,LDHAL6B,LDHB,LIN7A,LIN7B,LTK,MAGI2,MALT1,MARK4,MAT1A,MCCC1,MCCC2,MCM2,MCM4,MCM5,MCM6,MCM8,MCM9,MEFV,MMS19,MOCS1,MOCS2,MRTO4,MSH3,MT-CO1,MT-CO3,MT-CYB,MTX1,MYBBP1A,MYBL2,MYH11,MYH15,MYO18A,MYO19,MYO1B,MYO1C,MYO1D,MYO1E,MYO1F,MYO1G,MYO3A,MYO3B,MYO5A,MYO5B,MYO5C,MYO9A,MYO9B,NAA20,NAA50,NABP1,NACA2,NAE1,NAIP,NCAPD2,NCAPD3,NCAPG,NCAPG2,NCK1,NCLN,NCR1,NCSTN,NDE1,NDEL1,NDUFS1,NDUFS7,NEK2,NFS1,NLRC4,NLRP3,NLRP6,NLRP9,NOC2L,NOC4L,NOD2,NOP56,NOP58,NOP9,NPHS2,NPR1,ORAI1,OSGEP,PARD3,PCCA,PCCB,PCGF1,PDHA1,PDHA2,PDHB,PDP1,PDPR,PDSS2,PDZD11,PELP1,PFKFB1,PFKL,PFKM,PGM5,PIGK,PIK3C3,PIK3CG,PKD2L1,PKN1,PKN2,PLCB1,PLCB2,PLCB3,PPCDC,PPP2R5D,PPP2R5E,PPP3CA,PPP3CC,PPP4C,PPP5C,PRKAA1,PRKAA2,PRKACB,PRKAG1,PRKAG2,PRKAG3,PRKAR1B,PRKCA,PRKCB,PRKCI,PRKCZ,PRKDC,PRKRA,PRMT1,PRPF31,PSEN1,PSG9,PSMC1,PSMC2,PSMC3,PSMC4,PSMC6,PSMD14,PTH1R,PTPN11,PTPN6,PTPRN2,QRSL1,QTRT1,QTRT2,RACGAP1,RAD51,RAD54L,RFC2,RFC3,RFC4,RFC5,RING1,RNH1,RPA1,RPAP1,RPL10A,RPLP0,RPRD1A,RPRD1B,RPS27A,RRM2,RRM2B,RRP1B,RTCB,RUVBL1,RUVBL2,SCAP,SCLY,SCN1A,SCN4A,SCN5A,SCN8A,SCN9A,SDHA,SEC61B,SEC61G,SEL1L,SENP2,SENP3,SERPINB6,SH3GLB1,SHPRH,SKIV2L,SLC11A2,SLC12A7,SLC1A2,SLC26A6,SLC27A5,SLC41A1,SLC7A1,SLC7A5,SLC9A1,SMC2,SMC3,SMC4,SMC5,SMC6,SMURF2,SNRNP200,SNRNP25,SNX1,SNX2,SNX4,SNX5,SORBS1,SPG7,SPOPL,SPTAN1,SPTBN1,SPTBN4,SPTBN5,SPTLC1,SPTLC3,SRP54,SRPRA,SSU72,STAM,STAU2,STK3,STK4,STT3A,STT3B,SUCLA2,SUCLG1,SUCLG2,SUMO4,SUPV3L1,SYN2,SYNE1,SYNE2,SYNE3,SYNJ1,TAB1,TAF2,TAP1,TAP2,TCIRG1,TCP1,TFAM,TFRC,TM7SF2,TOLLIP,TOMM70,TOP2A,TOP2B,TOP3A,TOP3B,TPM1,TPM2,TPM4,TREX1,TRIM40,TRPC4,TRPC5,TRPC6,TRPM4,TRPV6,TSPAN32,TTBK1,TTF2,TTI1,TUBG1,TUBG2,UBA2,UBA3,UPF1,USH1C,UTP20,UTY,VCP,VIPR1,YWHAQ,YY1,YY2,ZBTB8OS,ZNRF2			
818	TP53,PIK3CA,CDH1,KMT2C,AKT1,PIK3R1,ERBB2,MYH9,AKAP9,CASP8,SF3B1,KMT2E,TAF1,SRCAP,ERBB3,KDM6A,CHD4,BACH2,CACNA1E,DMD,USH2A,TPR,GRID2,TAF1L,MAP1A,RANBP2,GRIN2A,LRP2,MYO18B,TLR7,CRNKL1,KIF17,KIF4A,ARHGAP29,TGS1,CPS1,TLR4,MYB,PRDM10,CDKN1B,MYO6,SCN10A,FCRL5,BAP1,NUP210L,MYO7B,CHRNB2,NOTCH2,ACVR2A,ACTN4,CUL4B,MYO10,WWP2,COL6A6,SP3,SCN11A,SMARCA4,ATP10B,FRMPD4,COL4A5,CDH20,PMS1,SGIP1,ACVR1B,KIF5B,FGFR2,KIF16B,PLXNA2,KCNA1,DNM1,DNAJC9,CDH12,COL24A1,ORC1,CREBBP,LAMA2,DCHS1,EPS8,DLG1,LAMB3,PRPF4B,MCM7,EPHA5,GBA3,MR1,COL12A1,GLYR1,TFAP4,TOB1,ROR1,PARP9,DNM3,LAMC1,C7,CASQ2,NCOA1,COL1A2,XRCC4,KCNK2,CRB1,LYN,SPTA1,RGS4,NTRK1,SAE1,KLHL13,ABCB8,ATF6,INTS8,LHCGR,SMARCA1,GPR84,IKBKB,KIF1A,SPTB,RABEP1,HSPD1,MBTD1,ATP10D,E2F1,CD19,PDE4B,SCML2,TF,WTAP,CFTR,NFE2L3,SUGP1,MYH7B,KCND1,HNRNPH3,HCFC2,SDR16C5,COL23A1,ACTL6B,METTL23,EPC1,HGS,ANAPC5,IL11RA,FAF2,FBL,IL18R1,MYO16,CALCRL,APP,ARHGEF2,MCOLN1,SPOP,ATP1B4,NEK10,CNGA4,VPS41,KANSL1,LDHC,PHF12,SMC1B,GPSM2,COL14A1,CCT6B,LILRA4,EMC9,ERG,EIF3L,HCN3,TBP,A1CF,GABRB2,KCNA5,PLXNA3,CATSPER3,KPNA7,UTRN,EFTUD2,CACNA2D2,KIF5C,CCT2,DEPDC1,TOPBP1,TARBP2,KCNQ2,SMTN,JADE3,TOX4,KPNA5,MYH14,CAMKMT,FBXO33,MMRN1,RAD21,COL11A1,PEF1,RAD51C,MST1R,IL4R,NAT9,SKIL,KLHL41,TRAF5,CR1L,IL10RB,TNR,NLRP1,EEF2,NDUFV1,VPS4B,ATP8B3,PIK3CB,GRPEL2,BLM,USF1,LEPR,HSPA2,TTC21A,PDCL3,GLDC,TRIM28,SNX3,IL3RA,NDUFS2,SMC1A,B3GALT4,BRD8,TRIM21,CD4,NAPA,NOX1,CDH10,ITGB3,CHRNA5,P4HA1,PIK3R4,POLRMT,MYF6,IL27RA,KCNAB1,C9,ZSWIM4,CCT3,SCN2B,IL6ST,TMEM67,CACNG3,CR2,PPP2R5C,SNF8,STAU1,KAT6B,OGDHL,COL4A3,RELN,ABCC9,TRPC1,ITGA9,F11R,PHF20,TRPC3,COPS3,GSPT2,KPNA2,PIK3R3,RECQL5,SNX6,HYOU1,COLEC11,CKAP5,ITGB2,PIK3CD,PDZD7,TTC30B,HEY2,SYNCRIP,KCNS2,CSNK1A1,ELOVL6,MYH10,PDE4D,PIAS4,PIWIL2,GRIK2,LIFR,MLH1,MATN2,PEX5L,THBS1,UGGT1,HDAC8,KCNA2,PSEN2,ATP4B,ATP11A,ERLIN2,HLF,VPS18,NOMO2,KLK8,BMP2,SHOC2,PIGR,ATP6V1A,MSH2,ACTR6,CEP170,CLIC1,SPI1,PSMD6,DCUN1D1,KLRC1,ELP5,IKZF1,IFT88,MYO7A,ANAPC7,CD40,CEP290,IL2RB,LZTR1,GANAB,ATP10A,AXL,PRMT5,COL1A1,EXT2,FBXO24,CACNA1F,EPS15,FLNA,DDB1,FAP,NFE2,GABRA4,GRHPR,FBXL18,HNRNPH2,UEVLD,ATP8B2,DYRK1A,ATF2,WWP1,OGDH,TGM3,B2M,ODF2,CACNB2,SKP1,CDK5,MYL6,COL8A1,GLUL,RECK,BRF1,CLCN2,ESRP2,IVNS1ABP,WDR19,DPYSL3,CIR1,CACNA2D1,DNAL1,SNTB1,PIGQ,ABI2,CAPN4,DECR1,EIF3K,RGPD3,COL10A1,RBPJL,SEC61A1,PLAT,PTGES3,COL19A1,PRKAR1A,PRKAR2A,NRP2,RAD18,ITGA10,LRRC8C,SPECC1L,PPP1R3C,MEPCE,PON1,PRKAR2B,LRP5,SEC61A2,CLCN1,EPHB3,ZNF350,CLIC2,LPA,NXF1,KIF3A,AGO4,FOXRED1,CACNG7,NAA10,PKD2,ABHD6,WWC1,ANP32E,COL17A1,PRKACA,CHUK,GABRR1,KIF22,SPTBN2,SP1,COMP,KCNIP4,KIF1C,RNF168,RNF220,CASP5,HMGB2,KIF26B,ZMPSTE24,AGK,PA2G4,CPT1C,CWC25,VPS33A,HDAC5,ATP6V1G2,TPM3,COL4A6,DIS3L2,PRKX,GPR20,FBXO27,SDHB,CNOT10,NOS1,SRPRB,GRIA3,OGT,CXCR4,KCNH1,NSMCE2,DDX6,SMARCA2,IL12RB1,METTL18,MAP3K5,C6,ADD2,KCNA4,BATF3,CDH8,CSF3R,DCUN1D4,HNRNPU,TRAF3,CD2,FBXO44,TSNARE1,FN1,KCNC2,SCN2A,PSMC5,ASF1A,ATF1,PPP1R3B,CSF2RA,CCM2,HEXB,CNGB3,SCN3B,HYAL2,KIF9,NT5DC3,PFKP,SCLT1,ADSL,EP300,LAMC2,SLC17A6,CBX3,RAD50,IGFBP5,DGCR8,KLHDC3,PPFIA1,MYO1A,TGFBR3,CHRNA3,BRCC3,GABARAP,MLLT10,STXBP1,ACVRL1,PKP1,CUL1,RRM1,SIAH1,OLR1,CLOCK,CDK12,RP9,NCOA3,CNOT2,KLHL12,MAP7,NVL,PLXNC1,SNAI3,TELO2,CCT7,DDR2,EEA1,ILVBL,NCF2,RPA4,SLC17A7,POLR3F,TBC1D5,PDLIM4,COL28A1,PTPRA,LRP6,CAPG,CDH24,TAF1A,AKAP6,APOBEC1,CTLA4,MARCO,AGO2,PLA2G7,UBA52,TAF4B,ASB12,PTGS2,E2F4,SMTNL2,ATF3,LIPC,PDCD6,F13A1,RPA2,AIP,MAT2A,MTPN,HDAC3,PPP2R5A,COL9A3,HSPA8,PLCG1,ITGA6,ATP6V0B,DPM1,GET4,SART3,SFTPD,GUCY1A2,HSPA9,SERPINA5,HSPA1L,SPCS1,STK11,TRAF1,AURKA,ALG14,KLHDC1,DDX20,ERLIN1,SMARCB1,CACNG1,HSP90AA1,ELANE,ORC3,PLP1,PSMG2,CDKN1A,CLDN17,CSNK1E,PPP2R5B,RNF215,VLDLR,CD8A,MZB1,IZUMO1,PCGF6,GCLM,KDR,MUSK,ATP6V1E2,UXS1,IGLL5,HLA-DQA2,TGFBR1,ZSWIM8,AGO3,TNFRSF11B,RYBP,SLC6A3,JUN,RNF2,STAP1,DDR1,AURKB,DUSP22,NPR3,TMED10,CACNG2,JDP2,KLHL7,CCT8,HDAC6,AKAP17A,SMARCE1,ABI1,RAD17,CUL4A,RIPK3,SWSAP1,BRD7,CACHD1,CBX5,KIF4B,MERTK,COL4A4,GSPT1,VIM,GRIN1,DNTTIP1,DDX23,HEXA,TFR2,TTI2,ETF1,PTN,RASSF2,UGT3A2,DYRK2,GIGYF1,KLHL20,SUPT16H,KCNIP1,ACTR1A,COLQ,VRK2,MAPT,MIF4GD,LMO4,MYL5,NDUFA9,SLC1A6,GRB2,ING3,ADAM10,CNIH3,PPP3CB,RPH3A,SCN7A,KCTD10,ASB1,GPRC5B,STAM2,MYL6B,PMPCA,RPS6KL1,F10,COQ6,DHX15,KIF20A,RNF135,SLC1A3,KCNK4,ACR,MRPL23,AXIN1,ARL6,ERCC3,PARP1,INSRR,LRRC43,RNF144B,ANKIB1,HTR2C,ITCH,NCCRP1,PCGF3,CCDC68,KCNA6,APBB1IP,SEC11A,SLC6A6,SPTLC2,DCTN6,IL23R,ITGA1,INO80D,NDUFS8,KBTBD6,INSIG1,RNF144A,WTIP,HBD,ADRB2,COPB2,PCGF2,PPP4R4,CD200R1,EPHB2,TNN,AICDA,CLMN,PPP1R7,FBXO48,SAP18,TNNC2,NAA15,TTC37,VDAC1,SPTSSA,LIN54,PLXND1,EZR,SARNP,WDR93,PROC,MCRS1,USP33,CLIC5,MYL12B,COQ3,DDX21,DYNC2LI1,TRPV3,GNA13,CLCNKB,RIPK1,SLC9A3R1,GNAT2,KCNMB2,MYL4,GABRA2,MCM3,PDLIM3,PRPF39,EPPIN,RRP1,GSTP1,CLEC12B,KAT8,ARVCF,FARSA,GTF2F2,IMMP1L,TMEM30A,APOBEC3G,CD3G,PPP2R3B,ETV3,DDX41,LIN7C,NABP2,CDHR3,PXK,RASIP1,SCNN1B,ETFB,HDAC2,TNFRSF1B,CHRNB4,PPM1F,RET,C1QTNF1,EMC10,EPHA3,JADE2,PTPN1,RAD51B,BBS5,E2F6,VPS4A,ARNT,GTF3C4,HNRNPR,WNK4,COA6,HPS3,PIK3R2,SNX17,CAMK1G,IFT52,KLRC2,LAMB1,ESPN,SHMT2,SPTSSB,PRAME,EPB41,PDGFB,TCF21,PCGF5,PDSS1,WBP11,GFRA3,RNF8,BATF,ARHGEF7,GTF2E1,DDX46,IL2RA,MAX,ESAM,CDK7,APPBP2,NRXN2,KAT6A,MLEC,GCH1,SLIRP,TGFBRAP1,BUB1B,GPD1,KCNAB2,LIN28A,LMO2,KCNA7,S100A1,MX2,GOPC,MAFF,CHRFAM7A,CLPX,DHX32,HLA-E,IMMP2L,ATXN7L3,ESRP1,FANCL,CASP1,KIF13A,PSMD11,SUMO1,WASF1,ZDHHC9,HINT1,CATSPER2,CD28,NEK6,SCN3A,CACNG6,GRIA4,KLC3,LAMB2,CD79A,NLGN1,TOMM40L,SNAP47,DAD1,KIFC3,ABAT,ABCA2,ABCB6,ABCB9,ABCC8,ABCD4,ABCG5,ABCG8,ABHD12,ABI3,ABL1,ABTB1,ACTL10,ACTL6A,ACTL7A,ACTL7B,ACTL8,ACTR10,ACTR1B,ACTR2,ACTR3,ACTR5,ACTR8,ACVR1,ACVR1C,ACVR2B,ADAM8,ADAR,ADCYAP1R1,ADD1,ADGRA2,ADIPOQ,ADRA2A,ADRB3,AFDN,AFG3L2,AGFG1,AGO1,AHR,AHRR,AIMP2,AJUBA,AKAP10,AKT2,ALB,ALK,ALS2,AMFR,AMHR2,AMN1,ANAPC10,ANAPC2,ANAPC4,ANKRD1,ANKRD39,ANO1,ANO2,ANO6,AP5Z1,APAF1,APC,APC2,APH1A,APH1B,APOB,APOBEC3F,APOH,ARHGAP33,ARIH1,ARIH2,ARL6IP1,ARMC8,ARNT2,ARNTL,ARNTL2,ASB2,ASB4,ASB6,ASCC1,ASCC3,ASF1B,ASH2L,ATF4,ATF6B,ATF7,ATG12,ATP11B,ATP11C,ATP1B1,ATP1B2,ATP1B3,ATP23,ATP6AP1,ATP6AP1L,ATP6AP2,ATP6V0A1,ATP6V0A2,ATP6V0A4,ATP6V0C,ATP6V0D1,ATP6V0D2,ATP6V1B1,ATP6V1B2,ATP6V1E1,ATP6V1G1,ATP6V1G3,ATP8A1,ATP8A2,ATP8B1,ATP8B4,AURKC,AXIN2,AZI2,B4GALT1,B9D1,B9D2,BACH1,BAIAP2L2,BARD1,BBS4,BCKDHA,BCKDHB,BCL3,BCR,BCS1L,BIRC2,BIRC3,BIRC5,BMI1,BMPR1A,BMPR1B,BMPR2,BOD1,BOD1L2,BRAP,BRD1,BRD9,BRIP1,BRPF1,BRPF3,BTBD1,BUD13,C1QA,C1QB,C1QC,C1QL1,C1QL2,C1QL3,C1QTNF2,C1QTNF3,C1QTNF5,C1QTNF6,C1QTNF7,C1QTNF9,C1QTNF9B,C2,C3,C5,C8A,C8B,C8G,CACNA1A,CACNA1B,CACNA1C,CACNA1D,CACNA1G,CACNA1H,CACNA1I,CACNA1S,CACNA2D3,CACNA2D4,CACNB1,CACNB3,CACNB4,CACNG4,CACNG8,CACTIN,CACYBP,CALCR,CALR,CANX,CAPN1,CAPN3,CAPNS1,CAPNS2,CARD10,CARD11,CARD9,CARMIL1,CASP10,CASP12,CASP2,CASP3,CASP4,CASP9,CATSPER1,CBL,CBR4,CCAR2,CCBE1,CCDC120,CCDC61,CCDC9,CCT4,CCT5,CCT6A,CD14,CD247,CD2AP,CD36,CD3D,CD3E,CD44,CD69,CD74,CD79B,CD80,CD8B,CD9,CDC16,CDC23,CDC27,CDH11,	0.053293	0.02	0.1418

	CDH13,CDH15,CDH17,CDH19,CDH2,CDH22,CDH23,CDH26,CDH3,CDH4,CDH5,CDH6,CDK1,CDK13,CDK14,CDK16,CDK19,CDK2,CDK3,CDK4,CDK6,CDK8,CDK9,CDKN2D,CEBPG,CEP128,CEP164,CEP83,CEP97,CETN2,CETN3,CETP,CFAP206,CFB,CFH,CFHR1,CFHR2,CFHR5,CFLAR,CGB3,CGNL1,CHAF1A,CHD3,CHD5,CHEK1,CHMP7,CHP1,CHRNA1,CHRNA2,CHRNA4,CHRNA7,CHRNA9,CHRNB1,CHRND,CHRNE,CHTF18,CHTOP,CISH,CLCN7,CLCNKA,CLDN1,CLDN3,CLDN4,CLIC3,CLIC4,CLINT1,CLIP2,CLPP,CLSTN3,CLTC,CLTCL1,CNGA1,CNGA3,CNGB1,CNIH2,CNNM4,CNOT7,CNOT8,CNTFR,CNTNAP2,CNTRL,COG8,COL11A2,COL13A1,COL15A1,COL16A1,COL18A1,COL20A1,COL21A1,COL22A1,COL25A1,COL26A1,COL27A1,COL2A1,COL3A1,COL4A1,COL4A2,COL5A1,COL5A2,COL5A3,COL6A1,COL6A2,COL6A3,COL7A1,COL8A2,COL9A1,COL9A2,COLEC10,COLEC12,COMMD1,COPA,COPE,COPS2,COPS4,COPS5,COPS6,COPS7A,COPS7B,COPS8,COQ5,CORO1C,CPSF1,CRADD,CRB2,CRB3,CRBN,CREB1,CREB3,CREG1,CREM,CRK,CRKL,CRLF1,CRLF2,CRTAP,CRYAA,CRYAB,CSF1R,CSF2RB,CSNK2A1,CSNK2A2,CSNK2A3,CSNK2B,CTDNEP1,CTDP1,CTHRC1,CTNND1,CTR9,CTSB,CTU1,CTU2,CUBN,CUL2,CUL3,CUL5,CXADR,CXXC1,CYB5B,CYB5R3,CYBB,CYC1,CYCS,DAB2IP,DAP3,DAPK1,DBI,DBT,DCLRE1C,DCTN5,DCUN1D2,DCUN1D3,DCUN1D5,DCX,DDOST,DDX1,DDX58,DERL1,DERL2,DERL3,DES11,DFFB,DGKI,DHDDS,DHRS7B,DHX16,DHX29,DHX33,DHX35,DHX38,DHX8,DHX9,DICER1,DIS3,DIS3L,DIXDC1,DLAT,DLD,DLG3,DLG4,DLST,DNAJB11,DNAJC10,DNAJC11,DNAJC15,DNAJC17,DNAJC19,DNM1L,DNM2,DNMT3A,DNMT3B,DPF1,DPF2,DPF3,DPH1,DPP10,DPP6,DPYSL5,DQX1,DRG1,DROSHA,DTNA,DTX3L,DUOX1,DUOX2,DVL1,DYNC1LI1,DYNC1LI2,DZIP1,E2F2,E2F3,E2F5,E2F7,E2F8,EBI3,ECT2,EDA,EEF1AKMT3,EEF1E1,EEF2KMT,EEFSEC,EFCAB7,EFL1,EGFR,EHBP1,EIF2AK4,EIF2S3,EIF3A,EIF3C,EIF3CL,EIF3E,EIF3F,EIF3H,EIF3M,EIF4EBP3,EIF4G1,EIF4G3,ELOB,ELOC,ELP3,ELP4,ELP6,EMC1,EMC2,EMC6,EMC7,EMC8,EME2,EMID1,EMILIN1,EMILIN2,ENC1,ENG,EPAS1,EPB41L2,EPC2,EPHA1,EPHA10,EPHA2,EPHA4,EPHA6,EPHA7,EPHA8,EPHB1,EPHB4,EPHB6,EPN1,EPN2,EPN3,EPS15L1,EPS8L1,EPS8L2,ERBB4,ERCC2,ERCC4,ERCC5,ERCC6L2,ERGIC2,ERGIC3,ERN1,ERN2,ETFA,EXOC3L1,EXOC3L2,EXOC3L4,EXOSC10,F13B,F2,F2RL2,F3,F7,F9,FADD,FAF1,FAM86B2,FANCG,FANCM,FAS,FAU,FBF1,FBXL12,FBXL15,FBXL19,FBXL22,FBXL3,FBXL4,FBXL5,FBXL7,FBXL8,FBXO10,FBXO11,FBXO17,FBXO2,FBXO38,FBXO39,FBXO42,FBXO6,FCER1G,FCGR3A,FCN1,FCN2,FCN3,FEM1A,FEM1B,FEM1C,FEN1,FGA,FGB,FGFR1,FGFR3,FGFR4,FGG,FGL1,FGL2,FH,FKBP1A,FKBP1B,FKBP4,FKBP8,FLOT1,FLOT2,FLT1,FLT3,FLT4,FNTA,FOS,FOSL1,FOSL2,FRMPD1,FSHB,FSHR,FTH1,FTL,FXYD1,FXYD2,FZD1,FZD8,GABBR1,GABBR2,GABRB3,GABRD,GABRE,GABRG1,GABRP,GABRR2,GAN,GAPDH,GATAD2A,GATAD2B,GATB,GCLC,GCN1,GCSH,GDF11,GFRA1,GFRA2,GFRA4,GHR,GID4,GID8,GIGYF2,GLDN,GLIPR1L1,GLRA1,GLRA2,GLRA3,GLRB,GMPR,GMPR2,GNA12,GNA15,GNAT1,GNAZ,GNPTG,GOLT1B,GP1BA,GP1BB,GP5,GP9,GPAA1,GPBAR1,GPD1L,GPD2,GPR156,GPR160,GPR37,GPR37L1,GPR61,GPR62,GPR63,GPRC5A,GPRC5C,GPRC5D,GPS1,GPSM1,GRAP,GRB10,GRIA1,GRIA2,GRIK3,GRIK4,GRIK5,GRIN2B,GRIN2C,GRIN2D,GRIN3A,GRIN3B,GRM1,GRM7,GRPEL1,GRSF1,GSK3A,GSK3B,GTF2A1,GTF2H4,GTF3C3,GTPBP1,HADHA,HBA1,HBB,HBE1,HBG1,HBG2,HBM,HBQ1,HCLS1,HCN1,HCN2,HCN4,HDAC1,HDAC10,HDAC11,HDAC4,HDAC7,HDAC9,HES1,HFE,HIF1A,HIPK2,HLA-A,HLA-B,HLA-C,HLA-DMA,HLA-DMB,HLA-DOA,HLA-DOB,HLA-DPA1,HLA-DPB1,HLA-DQA1,HLA-DQB1,HLA-DQB2,HLA-DRA,HLA-DRB1,HLA-DRB5,HLA-F,HLA-G,HLTF,HM13,HMGB1,HMGCL,HNRNPF,HNRNPH1,HNRNPUL1,HP,HPR,HSD17B10,HSD17B12,HSD17B8,HSP90AB1,HSP90B1,HSPA1A,HSPA1B,HSPA5,HSPB1,HSPB11,HSPB8,HSPG2,HSPH1,HTR1B,HTR2A,HTR2B,HTRA2,HTT,ID2,IDH3A,IDH3B,IDH3G,IFITM1,IFITM2,IFITM3,IFNLR1,IFT122,IFT172,IFT22,IFT27,IFT80,IGF1,IGF1R,IGF2R,IGFALS,IGFBP3,IGFBP6,IGHMBP2,IGLL1,IKBKE,IKBKG,IKZF4,IL12B,IL12RB2,IL13RA1,IL13RA2,IL18RAP,IL2RG,IL5RA,IL6R,IMPACT,INCENP,ING1,ING2,ING4,ING5,INHA,INHBA,INO80,INSC,INSIG2,INSR,INTS10,INTS4,IQGAP1,IRAK1,IRS1,IRS2,ISG15,ITGA11,ITGA2,ITGA2B,ITGA3,ITGA4,ITGA5,ITGA7,ITGA8,ITGAD,ITGAE,ITGAL,ITGAM,ITGAV,ITGAX,ITGB1,ITGB4,ITGB5,ITGB6,ITGB7,ITGB8,ITGBL1,ITLN1,JADE1,JAK2,JAM2,JAM3,JARID2,JAZF1,JMJD6,JPH1,JPH2,JPH3,JPH4,JUNB,JUND,KANSL1L,KANSL2,KAT2A,KAT2B,KAT5,KAT7,KATNA1,KATNAL1,KATNAL2,KBTBD7,KBTBD8,KCNA3,KCNAB3,KCNB1,KCNC1,KCNC3,KCNC4,KCND2,KCND3,KCNF1,KCNG1,KCNG2,KCNG3,KCNH2,KCNH4,KCNIP2,KCNIP3,KCNK1,KCNK6,KCNMA1,KCNMB1,KCNMB3,KCNMB4,KCNN1,KCNN4,KCNQ1,KCNQ3,KCNQ4,KCNQ5,KCNS1,KCNS3,KCNV2,KCTD13,KCTD16,KCTD17,KCTD2,KCTD5,KDM1A,KDM2B,KDM3A,KDM5A,KEAP1,KIDINS220,KIF11,KIF12,KIF13B,KIF14,KIF15,KIF18A,KIF18B,KIF19,KIF1B,KIF20B,KIF21A,KIF21B,KIF23,KIF24,KIF25,KIF26A,KIF27,KIF2A,KIF2C,KIF3B,KIF3C,KIF5A,KIF6,KIF7,KIFC1,KIFC2,KIT,KLC1,KLC2,KLC4,KLHDC10,KLHDC2,KLHL2,KLHL21,KLHL22,KLHL24,KLHL25,KLHL3,KLHL40,KLHL42,KLHL8,KLHL9,KLK3,KLRC3,KLRD1,KNTC1,KPNA1,KPNA3,KPNA4,KPNA6,KSR1,L3MBTL1,LAMA1,LAMA3,LAMA5,LAS1L,LCAT,LDB3,LDHA,LDHAL6B,LDHB,LDLR,LENG8,LEO1,LHB,LIG3,LIMD1,LIN7A,LIN7B,LMO7,LMTK2,LOX,LOXL4,LPL,LRP1,LRP1B,LRP8,LRRC55,LRRC8A,LRRC8B,LRRC8D,LRRC8E,LRRK2,LTBP1,LTF,LTK,LUM,MAD1L1,MAEA,MAFG,MAFK,MAGI2,MAGT1,MALT1,MAP1B,MAP1S,MAP3K7,MAPKAP1,MAPKAPK5,MARK4,MASP1,MAT1A,MAT2B,MATN1,MATN3,MATN4,MAU2,MBD2,MBD3,MBL2,MCCC1,MCCC2,MCM2,MCM3AP,MCM4,MCM5,MCM6,MCM8,MCM9,MDN1,MED15,MEFV,MEP1A,MEP1B,MET,METTL1,METTL14,METTL21A,METTL22,METTL3,MFN1,MICU1,MICU2,MICU3,MID1,MITF,MKLN1,MMRN2,MMS19,MOCS1,MOCS2,MON1A,MON1B,MPHOSPH8,MPP4,MRGBP,MRTO4,MSH3,MSH6,MSL2,MSR1,MT-ATP6,MT-ATP8,MT-CO1,MT-CO2,MT-CO3,MT-CYB,MT-ND1,MT-ND2,MT-ND3,MT-ND4,MT-ND4L,MT-ND5,MT-ND6,MTERF4,MTNR1A,MTTP,MTX1,MTX2,MTX3,MVP,MXD1,MXI1,MYBBP1A,MYBL2,MYC,MYD88,MYF5,MYH11,MYH15,MYL1,MYL12A,MYL2,MYL3,MYL7,MYL9,MYO15B,MYO18A,MYO19,MYO1B,MYO1C,MYO1D,MYO1E,MYO1F,MYO1G,MYO3A,MYO3B,MYO5A,MYO5B,MYO5C,MYO9A,MYO9B,MYOD1,MYOG,MYRIP,MYSM1,N6AMT1,NAA16,NAA20,NAA25,NAA50,NABP1,NACA2,NAE1,NAIP,NAPB,NAPG,NASP,NBAS,NBN,NBR1,NCALD,NCAPD2,NCAPD3,NCAPG,NCAPG2,NCF1,NCF4,NCK1,NCLN,NCOA2,NCR1,NCSTN,NDC1,NDE1,NDEL1,NDUFA10,NDUFS1,NDUFS7,NEK2,NEO1,NEUROD1,NFE2L2,NFIL3,NFKBIA,NFRKB,NFS1,NFYA,NFYC,NHEJ1,NHLH2,NID1,NLRC4,NLRP3,NLRP6,NLRP9,NME8,NOC2L,NOC4L,NOD2,NOL6,NOP56,NOP58,NOP9,NOTCH1,NOTCH3,NOX4,NOXA1,NOXO1,NPAS2,NPHS2,NPR1,NR3C1,NR3C2,NRP1,NSMCE1,NSRP1,NSUN3,NSUN4,NTN4,NTNG1,NTNG2,NTRK2,NTRK3,NUP210,NUS1,OBSL1,OGG1,OLFM2,OLFM3,OLFM4,ORAI1,ORC4,ORC5,OS9,OSGEP,OSMR,OSTM1,OTULIN,P2RX1,P2RX6,P3H1,P3H3,P3H4,PAFAH1B2,PAFAH1B3,PAM16,PAN3,PANX1,PARD3,PARD3B,PARD6A,PARD6B,PARD6G,PAXIP1,PBRM1,PCCA,PCCB,PCGF1,PCID2,PCM1,PCSK9,PDCD11,PDE3B,PDGFA,PDGFRA,PDGFRB,PDHA1,PDHA2,PDHB,PDHX,PDIA3,PDIA6,PDLIM1,PDLIM2,PDLIM5,PDLIM7,PDP1,PDPR,PDSS2,PDZD11,PECAM1,PELP1,PEX5,PFKFB1,PFKL,PFKM,PGM5,PHB2,PHC1,PHC2,PHC3,PHF10,PHF14,PHF20L1,PHF21A,PHKG1,PHKG2,PICALM,PIDD1,PIGA,PIGC,PIGH,PIGK,PIGM,PIGS,PIGT,PIGU,PIGV,PIK3C3,PIK3CG,PKD1,PKD2L1,PKN1,PKN2,PKP3,PLA2R1,PLAU,PLCB1,PLCB2,PLCB3,PLEKHA2,PLEKHA7,PLK1,PLK4,PLOD1,PLTP,PLXDC1,PLXNA1,PLXNA4,PLXNB1,PLXNB2,PLXNB3,PMPCB,POLA1,POLD1,POLE,POLG,POLR2I,POLR3C,POLR3K,PORCN,POU2F1,PPCDC,PPP1R12A,PPP1R3D,PPP1R3E,PPP1R3G,PPP1R9B,PPP2R5D,PPP2R5E,PPP3CA,PPP3CC,PPP3R1,PPP4C,PPP5C,PRELID1,PRICKLE4,PRIMPOL,PRKAA1,PRKAA2,PRKACB,PRKAG1,PRKAG2,PRKAG3,PRKAR1B,PRKCA,PRKCB,PRKCI,PRKCSH,PRKCZ,PRKDC,PRKRA,PRLR,PRMT1,PRPF31,PRPF40A,PRPF40B,PRPF6,PRR5,PRR5L,PSEN1,PSG9,PSMC1,PSMC2,PSMC3,PSMC3IP,PSMC4,PSMC6,PSMD10,PSMD12,PSMD13,PSMD14,PSMD3,PSMD7,PSMD8,PSMD9,PTH1R,PTK2B,PTPN11,PTPN6,PTPRB,PTPRN2,PTPRQ,PXN,PYCARD,PYDC1,QPRT,QRSL1,QTRT1,QTRT2,RABGAP1,RABGGTA,RACGAP1,RAD51,RAD51D,RAD54L,RADIL,RAG1,RAG2,RAMP1,RAMP2,RAMP3,RANGAP1,RAPGEF2,RASGRP3,RBCK1,RBM10,RBM5,RBMX2,RBPJ,RCHY1,REPIN1,RFC1,RFC2,RFC3,RFC4,RFC5,RFXANK,RGPD1,RGPD2,RGPD4,RGPD5,RGPD8,RGS11,RHBDD2,RHEB,RILP,RING1,RIOX2,RIPK2,RMND5A,RMND5B,RNASEK,RNF125,RNF139,RNF14,RNH1,RNMT,ROR2,ROS1,RPA1,RPAP1,RPAP3,RPL10A,RPLP0,RPN1,RPN2,RPRD1A,RPRD1B,RPS10,RPS19,RPS27A,RPSA,RRAGA,RRAGB,RRAGC,RRAGD,RRM2,RRM2B,RRP1B,RTCB,RUVBL1,RUVBL2,RYK,S100A10,SAC3D1,SALL4,SAMM50,SAP30,SAP30L,SAR1A,SAR1B,SASH1,SCAP,SCARA3,SCLY,SCN1A,SCN1B,SCN4A,SCN4B,SCN5A,SCN8A,SCN9A,SCNN1A,SCNN1D,SCNN1G,SCRIB,SDC1,SDCBP,SDF2,SDF2L1,SDHA,SEC11C,SEC61B,SEC61G,SEL1L,SENP2,SENP3,SERPINB6,SERPINE1,SERPINF2,SET,SETD5,SF1,SF3A1,SF3B2,SF3B3,SFTPA1,SFTPA2,SH3GLB1,SHARPIN,SHPRH,SHTN1,SIPA1,SIRT1,SIRT2,SIRT3,SIX4,SKAP1,SKIV2L,SKP2,SLA2,SLBP,SLC11A2,SLC12A7,SLC18A3,SLC1A2,SLC26A6,SLC27A5,SLC41A1,SLC51A,SLC7A1,SLC7A5,SLC9A1,SLITRK5,SLX1A,SMAD6,SMAD7,SMARCA5,SMC2,SMC3,SMC4,SMC5,SMC6,SMG6,SMG7,SMN1,SMURF2,SNRNP200,SNRNP25,SNRPA1,SNTA1,SNTB2,SNTG1,SNTG2,SNX1,SNX12,SNX2,SNX31,SNX4,SNX5,SOCS1,SOCS2,SOCS3,SORBS1,SOS1,SOST,SPAG1,SPAG4,SPATS2L,SPCS3,SPDL1,SPECC1,SPG7,SPHK2,SPOPL,SPTAN1,SPTBN1,SPTBN4,SPTBN5,SPTLC1,SPTLC3,SREBF1,SREBF2,SRP54,SRP68,SRP72,SRPRA,SSPN,SSRP1,SSU72,SSX2IP,ST18,STAC3,STAM,STAU2,STK3,STK4,STN1,STRA6,STRADA,STRADB,STT3A,STT3B,STXBP5,STXBP6,SUCLA2,SUCLG1,SUCLG2,SUGT1,SUMO4,SUN1,SUN2,SUPT20H,SUPV3L1,SUV39H1,SWI5,SYK,SYN2,SYNE1,SYNE2,SYNE3,SYNJ1,SYNPO2,SYNRG,SYVN1,TAB1,TAF10,TAF2,TAF4,TAL1,TAOK2,TAP1,TAP2,TAPBP,TAS1R3,TBC1D7,TBK1,TBKBP1,TBX15,TBX18,TBX5,TCF25,TCIRG1,TCP1,TCTN1,TCTN2,TECPR1,TEFM,TEK,TERT,TES,TFAM,TFDP1,TFDP2,TFRC,TGFBR2,THBD,TIE1,TIMM50,TJP1,TLR1,TLR10,TLR2,TLR6,TM7SF2,TMED3,TMED7,TMEM107,TMEM17,TMEM216,TMEM231,TMEM30B,TNC,TNFAIP1,TNFAIP2,TNFRSF1A,TNK2,TNKS,TNNC1,TNXB,TOLLIP,TOMM40,TOMM70,TOP2A,TOP2B,TOP3A,TOP3B,TP63,TPM1,TPM2,TPM4,TRA,TRAF2,TRAF3IP1,TRAF4,TRAF6,TRAPPC12,TREX1,TRIM40,TRIM54,TRMT61A,TRPC4,TRPC5,TRPC6,TRPM4,TRPV6,TSEN2,TSHR,TSPAN17,TSPAN32,TSPAN33,TTBK1,TTC21B,TTC26,TTC30A,TTC8,TTF2,TTI1,TTR,TTYH1,TTYH2,TTYH3,TUBG1,TUBG2,TUSC3,TXNL1,TXNL4A,TXNL4B,TYK2,TYRO3,UBA2,UBA3,UBL4A,UBXN7,UGGT2,UGT3A1,ULK1,UPF1,UPF3A,UPF3B,UPK1A,UQCRC1,UQCRC2,USH1C,USP10,USP47,UTP20,UTP6,UTY,VCAM1,VCP,VCPKMT,VDAC2,VDAC3,VEGFA,VIPAS39,VIPR1,VPS11,VPS16,VPS25,VPS29,VPS33B,VPS36,VPS37D,VPS39,VPS51,VPS8,VSIG4,VTN,WASF2,WASF3,WASHC1,WASHC2A,WASHC2C,WBP4,WDR35,WHRN,WNK1,WNT2,WNT3,WNT3A,WWOX,WWTR1,XAB2,XRCC2,XRCC3,YAP1,YEATS2,YWHAQ,YY1,YY2,ZAP70,ZBTB17,ZBTB8OS,ZC3H14,ZC3H18,ZCCHC17,ZER1,ZFAND2A,ZFHX3,ZFP36,ZNF10,ZNF224,ZNRF2,ZRSR2,ZSWIM5,ZSWIM6,ZYG11A,ZYG11B,ZZZ3			
58	BRCA2,BRCA1,RBM44,CRYBG3,CEP192,ATAD5,ICE1,SGO2,TET1	0.024231	0.02	0.1418
395	AKAP9,TPR,RABEP1,CEP290,ODF2,SCLT1,EEA1,CEP128,CEP83	0.020845	0.02	0.1418
95	CDH1,CDH20,CDH12,DCHS1,CDH10,CDH8,CDH24,CDHR3,CDH11,CDH13,CDH15,CDH17,CDH19,CDH2,CDH22,CDH23,CDH26,CDH3,CDH4,CDH5,CDH6	0.016878	0.02	0.1418

352	MYH9,MYH14,MYH10,MYH11	0.014251	0.02	0.1418
877	RUNX1,TTN,PHKA2,FBXW7,RBM12,EIF4A2,NCL,TBL1XR1,HNRNPA2B1,RPTOR,MYH7,G3BP2,HCFC1,DDX5,FMR1,SIN3A,WAC,EEF1A1,RGS7,GJA5,ILF2,GJB3,EIF4A3,SUPT5H,AP1G2,EIF3I,PABPC3,NUP107,ACTG2,ACTG1,SMAD4,CELF2,ESR1,G3BP1,BCKDK,AP4M1,USP28,RBM15,HNRNPK,TRAF7,DCAF5,PCBP2,PPP2CB,XPO1,PPP2CA,AP3M1,ATP2A1,HSF1,PSMB5,CNOT6,FARSB,MYH8,RNF19A,EXOSC3,CPEB1,HADHB,SIX6,DCAF11,SMAD2,HRAS,DKC1,ATG16L1,RUNX2,PRPF4,FNTB,ELOF1,NDUFV2,GNB5,RBM39,RARG,ATG16L2,RFX3,DCAF8,BTRC,CSTF2T,COPG1,PHKA1,API5,POLR1A,SNRPC,EIF2B2,UBE2D2,PSMA3,RPL26,STUB1,CPSF3,NR1H2,PPARG,YPEL5,RLIM,GNAI2,HNF1B,PPIL1,RPL12,ACTB,UBE2V1,TBL1X,DDX17,PNPT1,PSMD1,TNRC6A,POLB,AR,KATNB1,CAMK2D,WRAP73,AP3M2,PRPSAP2,RBMS1,TCEA1,SOD1,WDR82,SEC24C,HDLBP,RPL19,ARF1,AP4S1,MPP7,SNRPG,APEX1,NSFL1C,FBXL13,UBE2V2,RAD23B,XPO5,EEF1A2,RPS27L,SMAD3,RPL37A,EIF2B1,KCNJ2,RBM14,EIF4A1,RAB11A,RPL35,PAFAH1B1,RPS15A,TAF5L,ELAVL1,AP3B1,DCAF6,RPS25,FBXL2,SND1,TLE3,EIF2B4,PPIE,NR4A2,RPS18,FUBP3,CSDE1,AP2S1,LSM5,DCAF13,TIMM17A,POLR2J,CDC20,PPIL2,SPSB3,SNRPN,UBE2A,RPL23A,COPG2,AP2M1,GNB3,BUB3,GJC2,PMS2,HNRNPM,CNOT3,MORF4L1,PDK2,RABGGTB,LSM2,STIP1,RIOK2,RPL27,SIX1,PABPC4L,ACTA2,RPL9,AP1B1,POLR2L,CWF19L1,UBE2B,RPL39L,XPO6,NXT2,UBE2J2,RAB3A,UBE2D1,RBM4B,XPO7,FCF1,ELP2,TFCP2,CALM2,CAMK2A,GNAQ,RBBP7,CWC27,PSMD5,LSM4,UBE2L6,GJC3,WDR61,DYNLRB1,SCAF8,HNRNPC,MSI2,PSMB6,GGA1,PCBP4,GJA1,SATB2,CHCHD10,TAF5,MSI1,NR0B2,RPLP2,CPEB2,BEST3,ATP1A1,MYH4,GJB2,KPNB1,CDC45,KIFAP3,SNRPB,EEF1B2,RPL36,U2AF1,CAMK2B,CIRBP,SMAD9,LSM10,RPL31,RPS2,RALY,AAAS,ACTA1,ACTC1,ADRM1,AGL,AHNAK,AIMP1,AKAP8L,ALYREF,AMBRA1,ANAPC11,ANXA2,AP1G1,AP1M1,AP1M2,AP1S1,AP1S2,AP1S3,AP2A1,AP2A2,AP2B1,AP3B2,AP3D1,AP3S1,AP3S2,AP4B1,AP4E1,AQP4,ARCN1,ARPC1A,ARPC1B,ATP12A,ATP1A2,ATP1A3,ATP1A4,ATP2B4,ATP6V1H,BEST1,BEST2,BEST4,BIN1,BOP1,CAB39,CAB39L,CALM1,CALM3,CAMK2G,CAND1,CDC40,CDC42,CEBPZ,CELF1,CELF3,CELF4,CELF5,CELF6,CENPA,CENPX,CHAF1B,CHMP1A,CHMP1B,CHMP2A,CHMP2B,CHMP3,CHMP4A,CHMP4B,CHMP4C,CHMP5,CHMP6,CIAO1,CLP1,CNOT4,CNOT6L,CNOT9,COPB1,COPZ1,COPZ2,CPEB3,CPEB4,CPSF2,CPSF6,CPSF7,CSTF1,CSTF2,CTBP1,CTBP2,CTNNB1,CTNNBL1,DAZAP1,DAZL,DCAF1,DCAF10,DCAF7,DCP2,DDX19B,DDX39B,DDX3X,DTL,DYNLRB2,EEF1D,EIF1,EIF1B,EIF2A,EIF2S1,EIF3B,EIF4B,EIF4H,ELAVL2,ELAVL3,EMG1,ENO1,ENO2,ENO3,ENO4,EXOG,EXOSC1,EXOSC2,EXOSC4,EXOSC5,EXOSC6,EXOSC7,EXOSC8,EXOSC9,EYA1,EYA3,FAM8A1,FBXL14,FBXL16,FBXL17,FBXL20,FBXL6,FBXO45,FBXW11,FBXW5,FXR1,FXR2,FZR1,GAR1,GGA3,GJA3,GJA4,GJB1,GJB4,GJB5,GJB6,GJB7,GJC1,GJD3,GLRX3,GLRX5,GNA11,GNA14,GNAI1,GNAI3,GNAL,GNAO1,GNB1,GNB2,GNB4,GPN3,GPX4,HIRA,HNF1A,HNRNPA0,HNRNPA1,HNRNPA1L2,HNRNPA3,HNRNPAB,HNRNPD,HNRNPL,HNRNPLL,IGF2BP1,ILF3,INTS11,INTS7,INTS9,IPO4,IPO5,IPO7,ISCU,ISL1,JUP,KCNJ11,KCNJ14,KCNJ16,KCNJ3,KCNJ4,KCNJ5,KCNJ6,KCNJ8,KHDRBS1,KHDRBS3,KRR1,LHX3,LSM1,LSM14A,LSM14B,LSM3,LSM6,LSM7,LSM8,MEIS1,MLST8,MORF4L2,MRE11,MRPS2,MRPS28,MSL3,MTMR9,MTOR,MYEF2,MYH1,MYH13,MYH2,MYH3,MYH6,NAA30,NAA38,NAT10,NCBP2,NDUFA11,NELFE,NOB1,NONO,NR1H3,NR1H4,NR1I2,NR2E3,NR4A1,NR4A3,NR5A1,NR5A2,NR6A1,NUP85,NUP93,NUTF2,NXT1,OCLN,P4HB,PAAF1,PABPC1,PABPC1L,PABPC4,PABPC5,PABPN1,PAK1,PARP4,PAX2,PBX1,PBX2,PBX3,PCBP1,PCBP3,PCF11,PDK1,PGGT1B,PHKB,PKNOX1,PLRG1,POLA2,POLD2,POLE2,POLE4,POLR1B,POLR1C,POLR2A,POLR2B,POLR2C,POLR2E,POLR2G,POLR2J2,POLR2J3,POLR2K,POLR3A,POLR3B,POLR3H,PPIA,PPIB,PPIF,PPIH,PPIL3,PPP1CA,PPP1CB,PPP1CC,PPP1R10,PPP2R1A,PPP2R1B,PPP2R2A,PPP2R2B,PPP2R2C,PPP2R2D,PPP4R1,PPP4R3A,PPP4R3B,PRDX3,PRPF19,PRPS1,PRPS1L1,PRPS2,PRPSAP1,PSMA1,PSMA2,PSMA4,PSMA5,PSMA6,PSMA7,PSMA8,PSMB1,PSMB10,PSMB2,PSMB3,PSMB4,PSMB7,PSMB8,PSMB9,PSMD2,PSME4,PTBP2,PUF60,RAC1,RAC2,RAC3,RACK1,RAD23A,RAE1,RAN,RANBP1,RANBP10,RANBP17,RANBP9,RAP1A,RARA,RBBP4,RBBP6,RBM15B,RBM22,RBM3,RBM45,RBMS2,RBMS3,RBX1,RCC1,RICTOR,RIOK1,RIOK3,RNF11,RNF19B,RNF217,RNF31,RNF7,ROMO1,RPL13,RPL14,RPL21,RPL24,RPL26L1,RPL29,RPL36A,RPL36AL,RPL37,RPL39,RPL6,RPLP1,RPRD2,RPS13,RPS14,RPS24,RPS27,RPS29,RPS3,RPS4X,RPS4Y1,RPS9,RUNX3,RXRA,RXRB,RXRG,SCP2,SCYL1,SEC13,SEC23A,SEC23B,SEC24A,SEC24B,SEC24D,SEH1L,SF3A2,SF3B4,SF3B6,SFPQ,SGTA,SGTB,SIN3B,SIX2,SIX3,SLC25A4,SLC25A5,SLC25A6,SMAD1,SMAD5,SMARCD1,SMARCD2,SMARCD3,SMU1,SNCA,SNIP1,SNRNP40,SNRPA,SNRPB2,SNRPD1,SNRPD2,SNRPD3,SNRPE,SNRPF,SPEN,SPSB1,SPSB2,SPSB4,SRRT,SSBP3,STRAP,STRN,STRN3,STRN4,TCEA2,THOC3,THOC6,THRB,TIMM17B,TIMM22,TIMM23,TIMM23B,TLE1,TLE2,TLE4,TP53RK,TRRAP,TSC2,TSG101,TUBB,U2AF1L4,U2AF2,UBAC1,UBAP2L,UBE2C,UBE2D4,UBE2E1,UBE2I,UBE2L3,UBE2N,UBE2S,UBQLN1,UBQLN4,UBXN1,UBXN6,UBXN8,UCHL5,USP14,USP22,USP39,USP7,UTP23,VDR,WDR12,WDR18,WDR26,WDR33,WDR5,WDR5B,WDR83,WDTC1,WRAP53,XPO4,XPOT,YBX1,YEATS4,ZNF326	0.011494	0.02	0.1418
37	ATP10B,ATP10D,ATP8B3,ATP10A	0.010477	0.02	0.1418
560	DMD,ACTN4,SPTA1,SPTB,UTRN,FLNA,SPTBN2,IQGAP1,SPTAN1,SPTBN1,SPTBN4,SPTBN5,SYNE1,SYNE2	0.008456	0.02	0.1418
48	ARID1A,ARID1B,BCL9,SS18,NCOA6,CITED2,BCL9L,CBLL1,DAZAP2,EP400,MAML1,PYGO2,SAP130,SS18L1	0.007887	0.02	0.1418
193	FBXW7,TRAF7,BTRC,FBXW11	0.007578	0.02	0.1418
161	EIF4A2,DDX5,EIF4A3,DDX17,EIF4A1,DDX19B,DDX39B,DDX3X	0.00689	0.02	0.1418
386	NRIP1,BCOR	0.006693	0.02	0.1418
416	PHKA2,PHKA1,PHKB	0.004959	0.02	0.1418
592	BCLAF1,THRAP3,ZC3H13	0.004878	0.02	0.1418
319	JMJD1C,MED1,REV3L,MED13,MED13L,MAP2,KDM3B,GCM1,LIN37,ANAPC1,ATG13,ICE2,MBIP,MED19,MGA,POLR2M,POU2AF1,STIL,TP53BP1,UNC80	0.004103	0.02	0.1418
49	MYO18B,ARHGAP29,PRDM10,BAP1,FRMPD4,SGIP1,XRCC4,ARHGEF2,DEPDC1,TOPBP1,SMTN,SKIL,TRAF5,TRIM28,EPS15,RBPJL,RAD18,WWC1,RNF168,NOS1,TRAF3,TSNARE1,CCM2,SIAH1,SMTNL2,TRAF1,RYBP,RASSF2,AXIN1,LRRC43,CLMN,USP33,ARHGEF7,BUB1B,AFDN,AKAP10,AXIN2,BAIAP2L2,BARD1,BCR,BMPR2,CBL,CEP97,CRBN,CTDP1,DAB2IP,DDX58,DICER1,DZIP1,E2F7,E2F8,EHBP1,EPS15L1,ERCC5,FANCM,FRMPD1,GPR156,IMPACT,KDM3A,KSR1,LMTK2,MID1,MRGBP,MYRIP,MYSM1,NBN,NBR1,NFRKB,NHEJ1,PLEKHA2,PLEKHA7,PLK4,PPP1R12A,PRIMPOL,PRR5,PRR5L,RABGAP1,RAPGEF2,RASGRP3,RBPJ,RNF125,RNMT,SASH1,SIRT1,SIRT2,SIRT3,SIX4,SKAP1,SMN1,SOS1,ST18,TRAF2,TRAF4,TRAF6,TRIM54,TSEN2,ULK1,USP10,USP47,ZBTB17,ZC3H14	0.003113	0.02	0.1418
126	RUNX1,CPEB1,RUNX2,CPEB2,CNOT4,CPEB3,CPEB4,RUNX3	0.002993	0.02	0.1418
356	MYO6,MYO7B,MYO10,MYH7B,MYO16,MYO7A,MYO1A,MYH15,MYO18A,MYO19,MYO1B,MYO1C,MYO1D,MYO1E,MYO1F,MYO1G,MYO3A,MYO3B,MYO5A,MYO5B,MYO5C,MYO9A,MYO9B	0.001996	0.03	0.2321
419	PIK3CA,PIK3CB,PIK3CD,PIWIL2,PRMT5,PDSS1,B9D1,B9D2,FAU,ISG15,PDSS2,PIK3C3,PIK3CG	0.053764	0.03	0.2321
204	G3BP2,G3BP1	0.005864	0.03	0.2321
829	NCOR1,ARID4B,HTATSF1,MKRN3,BCLAF1,KRIT1,CDC73,THRAP3,ZCCHC8,TRDN,LARP7,CUL7,SOCS6,KMT2B,NXF5,RNPC3,MAVS,CCNT1,PPM1E,CDC5L,MTA2,BAX,RCOR2,APOBR,BAZ1B,DNAJC28,SMARCC2,RSPH4A,BCL10,POLR2F,CPSF4,NECAP1,FBXO21,SMTNL1,ZNHIT6,PAF1,POLR3GL,PIK3C2B,MAPK6,CCDC8,ZC3H13,EXD1,INTS12,POLR3G,MEAF6,SOCS4,BCL2L1,GEMIN6,PNN,UBAP1,TIPIN,KMT2A,PRKAB1,TAF12,POLE3,CARD8,LARP6,BTF3,RANBP3L,PPP4R2,GTF3C6,MTA3,PEX3,BAHD1,LIN9,LAGE3,BAZ1A,CLNS1A,PCP4,FRG1,BAK1,SNUPN,PEX13,EIF2S2,EAF2,TADA3,MED18,CARD16,CLNK,ANAPC15,ANKRA2,ARID4A,ATG3,BAG6,BCL2,BCL2L2,BPTF,CAT,CC2D2A,CCDC47,CCNK,CCNT2,CENPS,CEP76,CHRAC1,CHTF8,CLIC6	0.004204	0.03	0.2321

	,CPSF4L,CRCP,CREBZF,CSTA,CUL9,DMAP1,DPY30,DR1,DRAP1,EAF1,EHBP1L1,EIF3D,ERCC6,EZH1,EZH2,FBXO8,HDGF,HYPK,IFT46,KAT14,KDM6B,L1TD1,LCP2,LIG4,MAP4,MCL1,MIER1,MIER2,MIER3,MKS1,MLLT1,MLLT3,MPHOSPH6,MTA1,NACAD,NCBP3,NCOR2,NDUFA12,NECAP2,NEDD4,NEMF,NFYB,NPLOC4,NPM1,NXF3,PARP11,PEA15,PIK3C2A,PIK3C2G,PIK3R5,PIK3R6,PKD1L3,POLR1D,POLR2D,POLR3D,PPP1R15A,PPP1R8,PPP2R3A,PRIM1,PRKAB2,PRPF18,PSMG1,RANBP3,RBM12B,RBM28,RBM41,RCOR1,RCOR3,RERE,RNASEH2B,RRP7A,RSF1,SEM1,SLF1,SLF2,SMARCC1,SNAPC3,SOCS5,SRP9,SUB1,SUPT3H,SUPT6H,TADA2A,TADA2B,TAF11,TAF13,TAF7,TAF7L,TAF8,TAF9,TAF9B,TEP1,THG1L,TNKS1BP1,TRIP4,UR11,ZC3H8,ZFC3H1,ZMAT5			
782	FMR1,SIN3A,RGS7,SUPT5H,SMAD4,SMAD2,SMAD3,SND1,CSDE1,CNOT3,MORF4L1,TFCP2,SMAD9,FXR1,FXR2,KRR1,MORF4L2,MSL3,SIN3B,SMAD1,SMAD5,SMARCD1,SMARCD2,SMARCD3	0.003694	0.03	0.2321
375	BACH2,ATF6,NFE2L3,NFE2,ATF4,ATF6B,BACH1,CREB3,MYC,NFE2L2,NFIL3	0.002925	0.03	0.2321
837	TTN,RBM12,NCL,HNRNPA2B1,G3BP2,WAC,ILF2,G3BP1,RBM15,CSTF2T,SNRPC,TNRC6A,POLB,SEC24C,RBM14,SNRPN,HNRNPM,NXT2,RBM4B,SCAF8,HNRNPC,SNRPB,CIRBP,RALY,AHNAK,AKAP8L,ALYREF,ANXA2,CPSF6,CPSF7,CSTF2,DAZAP1,EIF4B,EIF4H,EYA1,EYA3,HNRNPA0,HNRNPA1,HNRNPA1L2,HNRNPA3,HNRNPAB,HNRNPD,ILF3,KHDRBS1,KHDRBS3,LSM14A,LSM14B,MYEF2,NELFE,NONO,NUTF2,NXT1,PCF11,PPP1R10,RBBP6,RBM15B,RBM3,RPRD2,SEC23A,SEC23B,SEC24A,SEC24B,SEC24D,SF3A2,SFPQ,SNIP1,SPEN,SRRT,UBAP2L,WDR33,YBX1,ZNF326	0.002889	0.03	0.2321
593	NUP210L,NOTCH2,COL12A1,CRB1,COL14A1,TNR,RELN,NOMO2,NRP2,FN1,LIPC,ADGRA2,ATP23,CNTNAP2,COL7A1,FLT1,FLT4,LPL,MEP1A,MEP1B,NEO1,NOTCH1,NOTCH3,NRP1,NUP210,OBSL1,TEK,TIE1,TNC,TNXB	0.00284	0.03	0.2321
2	ACTG2,ACTG1,PPP2CB,PPP2CA,ACTB,ACTA2,ACTA1,ACTC1,PPP1CA,PPP1CB,PPP1CC	0.002707	0.03	0.2321
122	COL4A5,COL4A3,COL8A1,COL10A1,COL19A1,COL4A6,COL9A3,COL4A4,COL16A1,COL4A1,COL4A2,COL8A2,COL9A1,COL9A2	0.002264	0.03	0.2321
6	ACVR2A,ACVR1B,ACVRL1,TGFBR1,ACVR1,ACVR1C,ACVR2B,BMPR1A,BMPR1B,TGFBR2	0.001627	0.03	0.2321
869	PHKA2,RPTOR,FMR1,SIN3A,RGS7,SUPT5H,AP1G2,NUP107,SMAD4,BCKDK,USP28,XPO1,CNOT6,SMAD2,FNTB,COPG1,PHKA1,API5,STUB1,PSMD1,TCEA1,APEX1,NSFL1C,RAD23B,XPO5,SMAD3,AP3B1,SND1,CSDE1,COPG2,CNOT3,MORF4L1,PDK2,RABGGTB,STIP1,AP1B1,XPO6,XPO7,TFCP2,PSMD5,GGA1,KPNB1,KIFAP3,SMAD9,ADRM1,AGL,AP1G1,AP2A1,AP2A2,AP2B1,AP3B2,AP3D1,AP4B1,AP4E1,ATP6V1H,BIN1,CAB39,CAB39L,CAND1,CEBPZ,CNOT6L,CNOT9,COPB1,CTNNB1,CTNNBL1,DCAF1,ENO4,FXR1,FXR2,GGA3,INTS7,IPO4,IPO5,IPO7,JUP,KRR1,MORF4L2,MSL3,MTMR9,MTOR,NUP85,NUP93,PARP4,PDK1,PGGT1B,PHKB,PPP2R1A,PPP2R1B,PPP4R1,PPP4R3A,PPP4R3B,PSMD2,PSME4,RAD23A,RANBP17,RANBP9,RICTOR,RNF11,SCYL1,SGTA,SGTB,SIN3B,SMAD1,SMAD5,SMARCD1,SMARCD2,SMARCD3,TCEA2,TRRAP,TSC2,UBAC1,UBQLN1,UBQLN4,UBXN1,UBXN6,UBXN8,UCHL5,USP14,USP22,USP39,USP7,XPO4,XPOT	0.001488	0.03	0.2321
669	BACH2,TFAP4,ATF6,NFE2L3,ERG,TOX4,MYF6,HEY2,HLF,SPI1,NFE2,ATF2,BATF3,MLLT10,ATF3,JUN,JDP2,ETV3,TCF21,BATF,MAX,MAFF,ATF4,ATF6B,ATF7,BACH1,CEBPG,CREB3,FOS,FOSL1,FOSL2,GATAD2A,GATAD2B,HES1,ID2,JUNB,JUND,MAFG,MAFK,MITF,MXD1,MXI1,MYC,MYF5,MYOD1,MYOG,NEUROD1,NFE2L2,NFIL3,NHLH2,NR3C1,NR3C2,PHF21A,SALL4,TAL1,TBX15,TBX18,TBX5,ZFHX3	0.001136	0.03	0.2321
517	SCN10A,SCN11A,MCOLN1,CATSPER3,PKD2,SCN7A,CATSPER2,ABCA2,PKD2L1,SCAP,SCN4A,SCN5A	0.003982	0.03	0.2477
619	SF3B1,CRNKL1,ANAPC5,TTC21A,TTC30B,PEX5L,IFT88,ANAPC7,CNOT10,SART3,NAA15,TTC37,PRPF39,COPE,CTR9,EMC2,FNTA,GTF3C3,MAU2,NAA16,NAA25,PEX5,PRPF6,RABGGTA,SRP72,TTC21B,TTC26,TTC30A,TTC8,UTP6,XAB2	0.003944	0.03	0.2477
740	NUP160,TRAPPC9,STAG2,SUZ12,POM121,CAV1,NUP62,LDB2,STAT3,AHCTF1,STAG1,TRAPPC10,INTS6,NUP54,RAB3GAP2,NUP153,ARID2,ARFGEF1,MED16,SYMPK,DCTN4,RPA3,NOL11,STAT5A,STAG3,LDB1,DMXL2,TRAPPC8,XRCC6,ELMO1,PEX14,DET1,NUP98,THOC2,DEPDC5,AHI1,BABAM1,BBS1,BBS2,BBS7,BBS9,CAV2,CAV3,CFAP70,CHD8,DCAF15,DENND4C,DMXL1,DOCK1,EIF4G2,GTF2H2,GTF2H2C,GTF2H3,HPS5,INTS13,INTS14,INTS3,INTS6L,ITFG2,ITPR1,ITPR2,ITPR3,KPTN,MED25,NCBP1,NPRL2,NPRL3,NUP133,NUP155,NUP214,NUP50,NUP58,NUP88,POM121C,PSMD4,RIC1,RYR1,RYR2,RYR3,STAT1,STAT2,STAT4,STAT5B,STAT6,TEN1,TRAPPC11,UBE3D,UPF2,VAC14,VPS35,XRCC5	0.002128	0.03	0.2477
85	HSPD1,CCT6B,CCT2,CCT3,CCT7,CCT8,CCT4,CCT5,CCT6A,TCP1	0.001982	0.03	0.2477
639	TAF1,TAF1L,TOB1,EPC1,KANSL1,RAD21,CEP170,AKAP6,APBB1IP,INO80D,WBP11,GTF2E1,ATXN7L3,WASF1,CHTOP,DROSHA,EIF4G1,EIF4G3,EPC2,HIPK2,KANSL1L,KANSL2,LEO1,LMO7,MSL2,PCM1,SF3B2,SUPT20H,SYNPO2,WASF2,WASF3,WASHC1,WNK1	0.003139	0.03	0.2477
446	AKT1,PRKACA,PRKX,AKT2,PRKACB	0.001964	0.03	0.2477
721	AKAP9,TPR,RANBP2,LRP2,TLR7,TGS1,TLR4,INTS8,IKBKB,RABEP1,WTAP,NEK10,FBXO33,CEP290,FBXO24,GRHPR,FBXL18,ODF2,RGPD3,SPECC1L,MEPCE,LRP5,CHUK,MAP3K5,SCLT1,MAP7,EEA1,RPA4,LRP6,TAF1A,DDX20,VLDLR,RIPK3,UGT3A2,VRK2,CCDC68,FBXO48,WDR93,RIPK1,GTF3C4,HPS3,RNF8,SNAP47,AIMP2,ALS2,AMHR2,ANAPC4,ASCC1,AZI2,CBR4,CCAR2,CD14,CEP128,CEP83,CFAP206,CGNL1,CHAF1A,CHMP7,CLIP2,CNTRL,COMMD1,CTU2,CUBN,DFFB,DTX3L,EIF2AK4,FANCG,FBXL12,FBXL15,FBXL22,FBXL3,FBXL8,FBXO38,FBXO39,FGB,FGG,FGL1,FGL2,GLIPR1L1,HSD17B10,HSD17B8,IKBKE,INCENP,ING1,INTS10,ITGBL1,ITLN1,JARID2,KNTC1,LAS1L,LDLR,LENG8,LRP1,LRP1B,LRP8,LRRK2,LUM,MCM3AP,MTERF4,NBAS,NME8,OTULIN,PDE3B,RAG1,RGPD1,RGPD2,RGPD4,RGPD5,RGPD8,RHBDD2,SHTN1,SMG6,SMG7,SPATS2L,SPDL1,SPECC1,SSX2IP,STN1,SWI5,TBK1,TCF25,TEFM,TLR1,TLR10,TLR2,TLR6,TRAF3IP1,TRAPPC12,UBL4A,UGT3A1,VIPAS39,ZZZ3	0.001951	0.03	0.2477
872	RUNX1,EIF4A2,HCFC1,DDX5,EIF4A3,PABPC3,ACTG2,ACTG1,CELF2,ESR1,AP4M1,HNRNPK,PCBP2,PPP2CB,PPP2CA,AP3M1,ATP2A1,HSF1,CPEB1,SIX6,RUNX2,ELOF1,RBM39,RARG,RFX3,UBE2D2,RPL26,NR1H2,PPARG,YPEL5,HNF1B,RPL12,ACTB,DDX17,AR,CAMK2D,AP3M2,PRPSAP2,RBMS1,HDLBP,RPL19,AP4S1,RPS27L,RPL37A,EIF4A1,RPL35,RPS15A,ELAVL1,RPS25,TLE3,NR4A2,RPS18,FUBP3,AP2S1,POLR2J,UBE2A,RPL23A,AP2M1,PMS2,RPL27,SIX1,PABPC4L,ACTA2,RPL9,UBE2B,RPL39L,UBE2D1,CALM2,CAMK2A,DYNLRB1,MSI2,PCBP4,SATB2,MSI1,NR0B2,CPEB2,ATP1A1,RPL36,U2AF1,CAMK2B,RPL31,RPS2,ACTA1,ACTC1,ANAPC11,AP1M1,AP1M2,AP1S1,AP1S2,AP1S3,AP3S1,AP3S2,ARCN1,ATP12A,ATP1A2,ATP1A3,ATP1A4,ATP2B4,CALM1,CALM3,CAMK2G,CELF1,CELF3,CELF4,CELF5,CELF6,CNOT4,COPZ1,COPZ2,CPEB3,CPEB4,CTBP1,CTBP2,DAZL,DDX19B,DDX39B,DDX3X,DYNLRB2,EIF1,EIF1B,ELAVL2,ELAVL3,ENO1,ENO2,ENO3,HNF1A,HNRNPL,HNRNPLL,IGF2BP1,ISL1,LHX3,MEIS1,NCBP2,NR1H3,NR1H4,NR1I2,NR2E3,NR4A1,NR4A3,NR5A1,NR5A2,NR6A1,PABPC1,PABPC1L,PABPC4,PABPC5,PAK1,PAX2,PBX1,PBX2,PBX3,PCBP1,PCBP3,PKNOX1,POLR2J2,POLR2J3,POLR2K,PPIA,PPP1CA,PPP1CB,PPP1CC,PRPS1,PRPS1L1,PRPS2,PRPSAP1,PTBP2,PUF60,RARA,RBM22,RBM45,RBMS2,RBMS3,RBX1,RNF7,RPL13,RPL14,RPL21,RPL24,RPL26L1,RPL29,RPL36A,RPL36AL,RPL37,RPL39,RPL6,RPS13,RPS14,RPS24,RPS27,RPS29,RPS3,RPS4X,RPS4Y1,RPS9,RUNX3,RXRA,RXRB,RXRG,SF3B4,SF3B6,SIX2,SIX3,SLC25A4,SLC25A5,SLC25A6,SNRPA,SNRPB2,SSBP3,THRB,TLE1,TLE2,TLE4,U2AF1L4,U2AF2,UBE2D4,UBE2E1,VDR	0.001698	0.03	0.2477
353	MYH7,EEF1A1,MYH8,EEF1A2,MYH4,MYH1,MYH13,MYH2,MYH3,MYH6,TUBB	0.001671	0.03	0.2477
392	POM121,NUP62,NUP54,NUP153,NUP98,NUP50,NUP58,POM121C	0.001197	0.03	0.2477
726	USH2A,GRID2,NUP210L,NOTCH2,COL6A6,ATP10B,PLXNA2,LAMA2,LAMB3,GBA3,COL12A1,ROR1,LAMC1,CRB1,NTRK1,ATP10D,TF,HCFC2,SDR16C5,METTL23,FBL,IL18R1,ATP1B4,COL14A1,PLXNA3,CACNA2D2,CAMKMT,RAD51C,MST1R,NAT9,CR1L,TNR,ATP8B3,NOX1,ITGB3,TMEM67,CR2,RELN,ABCC9,ITGA9,F11R,ITGB2,ELOVL6,MATN2,THBS1,ATP4B,ATP11A,NOMO2,SHOC2,ELP5,LZTR1,GANAB,ATP10A,EXT2,FAP,ATP8B2,TGM3,CACNA2D1,DNAL1,DECR1,NRP2,ITGA10,LRRC8C,ABHD6,COMP,AGK,FN1,HEXB,HYAL2,LAMC2,KLHDC3,PLXNC1,PLA2G7,ASB12,PTGS2,LIPC,F13A1,MTPN,ITGA6,GUCY1A2,ALG14,KLHDC1,PSMG2,KDR,MUSK,NPR3,SWSAP1,CACHD1,MERTK,HEXA,ADAM10,ASB1,DCTN6,ITGA1,PPP1R7,PLXND1,TMEM30A,PXK,SCNN1B,RET,RAD51B,CAMK1G,LAMB1,PRAME,APPBP2,LAMB2,ABCC8,ABHD12,ADAM8,ADGRA2,AMN1,ANAPC10,ANKRD1,ANKRD39,APAF1,APOH,ARMC8,ASB2,ASB4,ASB6,ATP11B,ATP11C,ATP1B1,ATP1B2,ATP1B3,ATP23,ATP8A1,ATP8A2,ATP8B1,ATP8B4,B4GALT1,C3,C5,CACNA2D3,CACNA2D4,CALR,CANX,CARMIL1,CD36,CDKN2D,CETP,CFH,CFHR1,	0.002728	0.04	0.2477

	CFHR2,CFHR5,CLSTN3,CNTNAP2,COL6A3,COL7A1,CRADD,CRB2,CSF1R,CXADR,CYB5R3,CYBB,CYC1,DAP3,DCTN5,DHRS7B,DPP10,DPP6,DUOX1,DUOX2,EEF1AKMT3,EEF2KMT,ELP4,ELP6,EMC1,ERGIC2,ERGIC3,F13B,FADD,FAM86B2,FBXL4,FBXL5,FBXL7,FBXO10,FBXO11,FBXO42,FEM1A,FEM1B,FEM1C,FLT1,FLT3,FLT4,GABBR1,GABBR2,GRIN3A,GRM1,GRM7,HSD17B12,HSPB11,HSPG2,IL18RAP,INSC,ITGA11,ITGA2,ITGA2B,ITGA3,ITGA4,ITGA5,ITGA7,ITGA8,ITGAD,ITGAE,ITGAL,ITGAM,ITGAV,ITGAX,ITGB1,ITGB4,ITGB5,ITGB6,ITGB7,ITGB8,JAK2,JAM2,JAM3,KIT,KLHDC10,KLHDC2,LAMA1,LAMA3,LAMA5,LCAT,LPL,LRRC8A,LRRC8B,LRRC8D,LRRC8E,LTF,MAP3K7,MAPKAPK5,MAT2B,MATN1,MATN3,MATN4,MDN1,MEP1A,MEP1B,MET,METTL21A,METTL22,MFN1,MKLN1,MYD88,N6AMT1,NDUFA10,NEO1,NFKBIA,NID1,NOTCH1,NOTCH3,NOX4,NRP1,NSUN3,NSUN4,NTN4,NTNG1,NTNG2,NTRK2,NTRK3,NUP210,NUS1,OBSL1,P2RX1,P2RX6,PAFAH1B2,PAFAH1B3,PANX1,PCSK9,PDGFRA,PDGFRB,PHKG1,PHKG2,PIGA,PIGH,PIGV,PLK1,PLOD1,PLTP,PLXNA1,PLXNA4,PLXNB1,PLXNB2,PLXNB3,PORCN,PRELID1,PRKCSH,PSMD10,PTK2B,PYCARD,PYDC1,RAD51D,RAG2,RANGAP1,RFXANK,RIPK2,ROR2,ROS1,RYK,SCNN1A,SCNN1D,SCNN1G,SKP2,SLC18A3,SLITRK5,SNRPA1,STRADA,STRADB,STXBP6,SYK,TAS1R3,TBC1D7,TEK,THBD,TIE1,TMEM30B,TNC,TNXB,TRMT61A,TTYH1,TTYH2,TTYH3,TXNL1,TYRO3,VCPKMT,WNT2,WNT3,WNT3A,XRCC2,XRCC3,ZAP70,ZER1,ZYG11A,ZYG11B			
596	TLR7,TLR4,UGT3A2,LUM,TLR1,TLR10,TLR2,TLR6,UGT3A1	0.001676	0.04	0.2477
670	GPS2,ZFP36L1,NPAS4,RB1,MAGED1,KMT2D,SHISA9,CBX8,CCNL2,LSR,CECR2,PRAM1,IRF9,ATXN2,ZFP36L2,ATF5,TEAD4,CNPPD1,TEAD1,KLF4,FLYWCH1,TEAD2,CAPRIN2,RBL1,CCNI2,NFATC4,NUFIP1,AEBP2,BRF2,CASC3,CBX2,CBX4,CBX6,CBX7,CCNA1,CCNA2,CCNB1,CCNB2,CCNC,CCND1,CCND2,CCND3,CCNE1,CCNE2,CCNF,CCNG1,CCNG2,CCNH,CCNI,CCNJ,CCNJL,CCNL1,CCNO,CCNY,CDK5R1,CDK5R2,CEBPA,CEBPB,CEBPD,CEBPE,DBP,ELL,ELL2,ELOA,FIP1L1,FOXO3,GLI1,GLI2,GLI3,GTF2B,HMGXB4,ILDR1,INO80E,IRF4,KLF5,LSM11,MAPK8IP2,MEF2B,MLXIPL,NAF1,NDN,NFAT5,NFATC1,NFATC2,NFATC3,NSMCE3,ORC6,RBL2,RBM42,REL,RELA,RELB,RFXAP,SETD1A,SETD1B,SHC1,SHISA7,SOCS7,TAF1B,TAF3,TEAD3,TRERF1	0.001402	0.04	0.2477
789	TRA2B,PRPF38B,RBM8A,RBMXL1,SRSF2,TRA2A,CWC22,RNPS1,SNRNP35,SRRM2,PRPF38A,EIF3G,RBMX,SNRNP27,SNRNP70,SREK1,SRRM1,SRSF1,SSB,ZCRB1	0.001313	0.04	0.2477
64	C7,C9,UGGT1,C6,PTN,EPPIN,ALK,C8A,C8B,IGF2R,UGGT2	0.001067	0.04	0.2477
243	PMS1,MLH1,HSP90AA1,HSP90AB1,HSP90B1	0.000998	0.04	0.2477
198	KMT2D,FLYWCH1	0.000993	0.04	0.2477
289	USH2A,LAMA2,LAMC1,THBS1,COMP,LAMB1,LAMB2,LAMA1,LAMA3,LAMA5,NTN4,NTNG1,NTNG2,WNT2,WNT3,WNT3A	0.001786	0.04	0.2477
279	KIF16B,KIF1A,KIF1C,KIF13A,KIF13B,KIF1B,KIF2A	0.001132	0.04	0.2477
694	CD19,IL11RA,IL4R,IL10RB,LEPR,IL3RA,CD4,IL27RA,SCN2B,IL6ST,LIFR,PIGR,CD40,IL2RB,IL12RB1,CSF3R,CD2,CSF2RA,SCN3B,OLR1,CTLA4,SERPINA5,CD8A,IZUMO1,TNFRSF11B,IL23R,CD200R1,CLEC12B,CD3G,TNFRSF1B,IL2RA,ESAM,CD28,CD79A,CATSPER1,CD3D,CD3E,CD69,CD79B,CD80,CD8B,CNTFR,CRLF1,CRLF2,CSF2RB,EBI3,F3,FAS,FCGR3A,GHR,GP1BA,IFNLR1,IL12B,IL12RB2,IL13RA1,IL13RA2,IL2RG,IL5RA,IL6R,OSMR,PECAM1,PKD1,PRLR,SCN1B,SCN4B,SERPINE1,SERPINF2,TAPBP,TNFRSF1A,TRA,VCAM1,VSIG4	0.001644	0.04	0.2477
123	RPTOR,AP1G2,BCKDK,XPO1,CNOT6,FNTB,COPG1,API5,PSMD1,TCEA1,APEX1,COPG2,PDK2,RABGGTB,AP1B1,PSMD5,GGA1,AP1G1,AP2A1,AP2A2,AP2B1,AP4B1,AP4E1,BIN1,CEBPZ,CNOT6L,COPB1,ENO4,GGA3,INTS7,PARP4,PDK1,PGGT1B,PPP2R1A,PPP2R1B,PSMD2,RICTOR,SCYL1,TCEA2	0.001586	0.04	0.2477
609	TRA2B,RBMXL1,SRSF2,TRA2A,RNPS1,RBMX,SRSF1	0.001048	0.04	0.2477
175	HNRNPH3,A1CF,GRPEL2,SYNCRIP,HNRNPH2,ESRP2,METTL18,COQ3,HNRNPR,ESRP1,CYB5B,GRPEL1,GRSF1,HNRNPF,HNRNPH1,TFAM,TREX1	0.000985	0.04	0.2477
185	DCAF17,FANCB	0.000853	0.04	0.2477
861	BRCA2,BRCA1,NRIP1,ADNP,JMJD1C,MED1,MEN1,SLX4,PYHIN1,REV3L,MED13,ASXL2,RFX5,ASXL3,RBM44,DCAF17,MED13L,CRYBG3,AFTPH,MAP2,CEP192,XRCC1,BCOR,CNST,ENTHD1,PPARGC1B,RMI1,GPKOW,KDM3B,NSL1,TERF2IP,FAM186B,ERCC1,AIM2,RB1CC1,ATF7IP,GCM1,PPP1R15B,MDM4,AKAP4,BCL7A,ZBTB16,NCAPH,C2orf49,FANCB,MUS81,ELL3,APOF,DSN1,RNASEH2A,MIS12,LIN37,VEZT,APOL1,DYDC2,CENPJ,EME1,CT45A6,TANK,RPP40,HAUS6,ACIN1,AKAP5,AKAP7,ANAPC1,ASXL1,ATAD5,ATF7IP2,ATG13,ATXN7,BAZ2A,BCL7B,BCL7C,BDP1,BOD1L1,C2CD6,CAMLG,CCP110,CDCA8,CSF2,CT45A10,CT45A2,CT45A3,DBF4,DBF4B,DISC1,DNAAF2,DONSON,FAAP20,FAAP24,GTF2A1L,HEXIM1,ICE1,ICE2,IFT43,IKZF5,IL12A,IL6,IMPG2,JRK,KATNBL1,KNDC1,LRRC46,MBIP,MDM2,MED19,MED26,MGA,MLH3,MSL1,MZT2A,MZT2B,N4BP2L2,NCAPH2,NELFA,NIPBL,NKD1,PALB2,PHAX,POLR2M,POT1,POU2AF1,PRDM4,RAD51AP1,RBBP8,REST,ROGDI,RPAP2,RPP21,RRP8,SECISBP2,SECISBP2L,SFR1,SGO2,SKA3,SNAPC4,SRFBP1,STIL,SUPT20HL1,TERF1,TERF2,TESPA1,TET1,TINF2,TOPORS,TP53BP1,TRPS1,TSEN34,UBAP1L,UIMC1,UNC80,ZC3H3,ZCCHC7,ZFYVE9,ZNF217,ZNF827,ZWILCH	0.00064	0.04	0.2477
228	EFTUD2,EEF2,GSPT2,GSPT1,DRG1,EEFSEC,EFL1,EIF2S3,GTPBP1	0.000482	0.04	0.2477
645	GNL3L,FTSJ3,RBBP5,WDR46,TSR1,BMS1,BYSL,IMP3,NOL10,PPAN,PPWD1,UTP18	0.000595	0.04	0.2477

*System_name: the name/index of the identified biological systems by the proposed MPGNN-Hilander model.

*Genes: genes contained in the biological system.

*Selection_pressure: calculated using the group-LASSO weights by HiSig.

*p: the *p-value* of the permutation test.

*q: the FDR.