TEL AVIV UNIVERSITY אוניברסיטת תל-אביב

Blavatnik School of Computer Science

Cancer subtype identification using large-scale omics data analysis

THESIS SUBMITTED FOR THE DEGREE OF
"DOCTOR OF PHILOSOPHY"

by

Dvir Netanely

The work on this thesis has been carried out
under the supervision of

Prof. Ron Shamir

Submitted to the Senate of Tel-Aviv University
December 2019

Acknowledgments

Working on this thesis was both challenging and rewarding. I'm grateful for the opportunity to work on improving cancer classification, an important topic that converges my love and passion for computers, biology, and medicine. I could not have done it without help and inspiration from many people.

Firstly, I would like to thank my supervisor, Prof. Ron Shamir, for enabling me to go on this journey. I admire his high professionalism, vast knowledge, eloquent writing, and idealism. He created the high-standard and nurturing environment in which I worked and evolved during the past few years and I learned a lot just from watching him run our group and the Safra Center for bioinformatics. Most of all I appreciate his patience and guidance at times when I was overloaded with too many details and couldn't see the entire picture. Thank you, Ron.

Secondly, I would like to thank the amazing Gilit Zohar-Oren, the master of organizational skills, for her consistent support and assistance in tackling any administrative hardship. Also great thanks to Mika Shahar for helping me embark on this route, and to her successors at the Ph.D. secretariat, Anat Amirav, and Eitan Hoffmann, for pleasantly helping me get to the finish line.

I was lucky to collaborate and learn a lot from impressive researchers, including Dr. Ella Evron, Dr. Ayelet Avraham, Prof. Adit Ben-Baruch, and Prof. Carmit Levy. Special thanks also to Prof. Zohar Yakhini. In addition, I wish to thank my past and present fellow lab mates for numerous stimulating discussions and bilateral motivational talks. I had great fun working with Neta Stern and Itay Laufer on developing PROMO.

In parallel to working on my research during my Ph.D. years, I was also privileged to teach more than 2000 students how to program, and wish to thank Dr. Yael Amsterdamer and Prof. Dan Halperin for showing me how to do it professionally. Taking part in organizing Safra young researchers' forum was another source for joy and satisfaction during these years, and I wish to thank Roni Wilentzik-Müller for this opportunity, as well for her wise comments along the way.

Last but not the least, I would like to thank my family: my wife Irit, my kids – Danielle, Roy and Nadav, my parents – Eli and Yona, and my brothers – Or, Guy, and Dan, for supporting me in numerous ways throughout this long endeavor.

Declarations

The results published here are based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

This study was supported by the Israeli Science Foundation (grants 317/13, 2193/15, 1339/18), by the Israel Cancer Association (Bella Walter Memorial Fund and donation of Avraham Rotstein), by an IDEA grant of the Dotan Center in Hemato-Oncology, by the Israeli Center of Research Excellence (I-CORE), Gene Regulation in Complex Human Disease, Center No. 41/11, by grant 2016694 from the United State - Israel Binational Science Foundation (BSF) and the United States National Science Foundation (NSF), and by German-Israeli Project DFG RE 4193/1-1. D.N. was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics, Tel Aviv University.

The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Preface

This thesis is based on the following papers:

1. Netanely D, Avraham A, Ben-Baruch A, Evron E, Shamir R. **Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups**. Breast Cancer Res. 2016; 18:74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27386846

2. Netanely D, Stern N, Laufer I, Shamir R. **PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets.** BMC Bioinformatics. BioMed Central; 2019; 20:732. . Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3142-5

3. Netanely D, Leibou S, Parikh R, Stern N, Amar S, Haiat Factor R, Brenner R, Vaknine H, Levy C, Shamir R. **Classification of melanomas into subgroups using keratin, immune and melanogenesis expression patterns**. In preparation.

The following works done during the Ph.D. are not covered in the thesis:

4. Bell RE, Khaled M, Netanely D, Schubert S, Golan T, Buxbaum A, et al. **Transcription Factor/microRNA Axis Blocks Melanoma Invasion Program by miR-211 Targeting NUAK1.** J Invest Dermatol. 2014; 134:441–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23934065

5. Ohana R, Weiman-Kelman B, Raviv S, Tamm ER, Pasmanik-Chor M, Rinon A, et al. **MicroRNAs are essential for differentiation of the retinal pigmented epithelium and maturation of adjacent photoreceptors.** Development. 2015; 142:2487–98. Available from: https://www.ncbi.nlm.nih.gov/pubmed/26062936

6. Santana-Magal N, Farhat-Younis L, Rasoulouniriana D, Gleiberman A, Gutwillig A, Tal L, Netanely D, Shamir R, Blau R, Gutman H, Rider P, Carmi Y. **Melanoma-secreted lysosomes trigger monocyte-derived dendritic cell apoptosis and limit cancer immunotherapy.** *Cancer Res.* 2020; 80:1942–1956. Available from: https://pubmed.ncbi.nlm.nih.gov/32127354/

Abstract

Cancer is the second leading cause of death worldwide. It is characterized by abnormal cell proliferation, potentially followed by spreading into surrounding tissue and body organs. Cancer is challenging to treat since it is very heterogeneous: even tumors originating from the same organ can greatly vary in their biological mechanism, survival risk, and response to treatment.

Recent years have shown the emergence of large cancer genomic projects, providing detailed multi-omic profiles together with clinical information for thousands of cancer samples. In line with the vision of precision medicine, integration of omic and clinical data using statistical and algorithmic methods allows us to computationally identify clinically distinct subgroups that may have a profound impact on diagnosis, drug discovery, and treatment.

In this work, we developed a methodology for improving the classification of cancers based on high-throughput omic data and applied it to both breast and skin cancers. Our analysis of the breast cancer cohort revealed a significant heterogeneity within the luminal-A subtype and partitioned its samples into prognostic subgroups based on expression and methylation patterns. Our analysis of the skin cancer cohort identified a group of poor-prognosis melanoma samples characterized by melanogenesis genes. We also suggested a simple three-gene classifier for predicting melanoma subtypes. Lastly, we describe *PROMO,* an interactive software tool we developed for multi-omic cancer data analysis and subtyping that generalizes the methodology used in the breast and skin cancer projects.

Contents

1. Introduction

1.1. Cancer

1.1.1. Introduction to cancer

Cancer is a large group of diseases characterized by uncontrolled proliferation of the body cells, with the potential of spreading into surrounding tissues. Cancer is the second leading cause of death worldwide [1]. In 2018, about 1 in 6 deaths worldwide occurred due to cancer, with estimates of 9.6 million deaths and 18.1 million incidents of cancer occurring globally [2][3]. Cancer can occur in all body parts, but lung, breast, and colorectal cancers are the most common types of cancer worldwide (Figures 1.1-1.3) [2]. Lung cancer was the most common cancer in men worldwide. For women, breast cancer was the most frequently diagnosed cancer in most countries, as well as the most frequent cause of death from cancer. As the world population is growing and aging, global cancer incidents are on the rise and projected to increase by more than 60% by the year 2040, making cancer a significant health and economic burden worldwide [2].

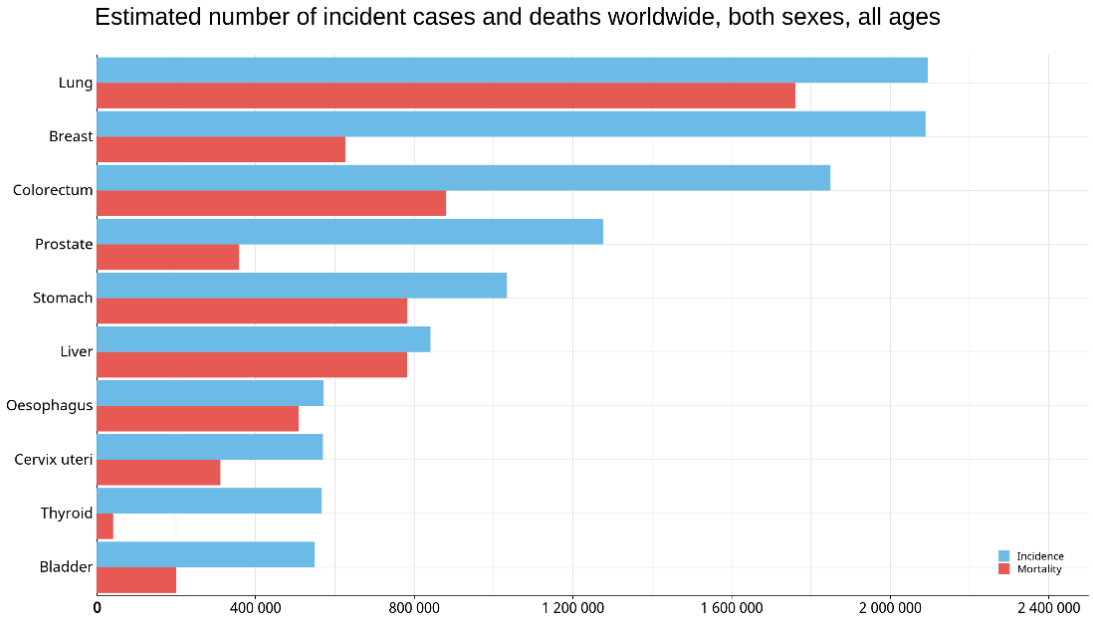


Figure 1.1: Estimated number of worldwide incidents and deaths from different types of cancer. Image source: [2]

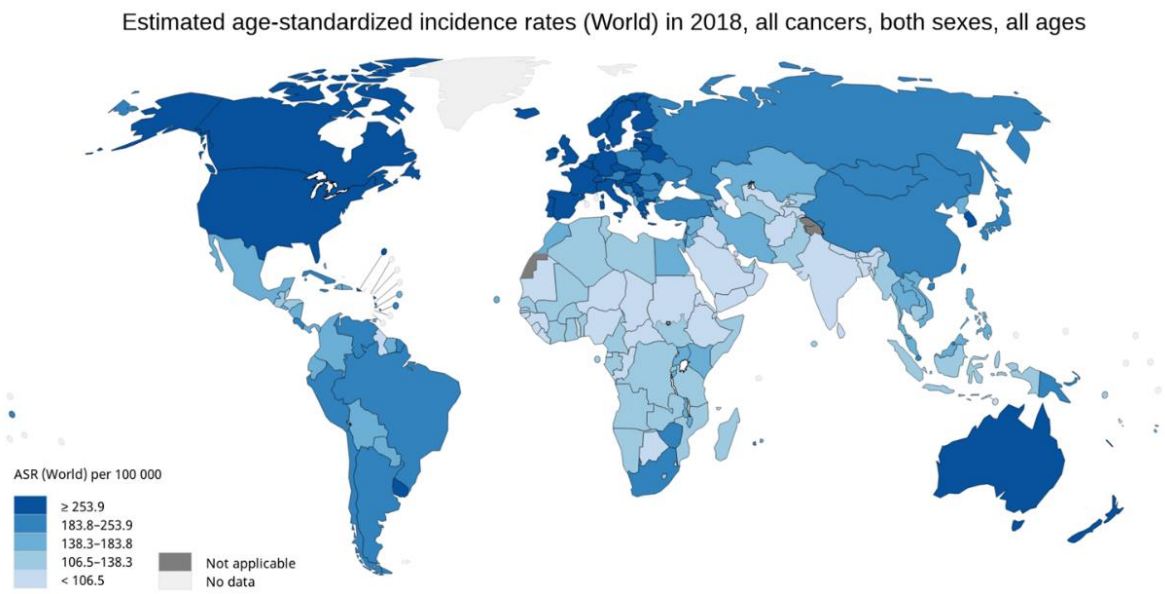


Figure 1.2: Estimated worldwide age-standardized incidence rates of cancer. Image source: [2].

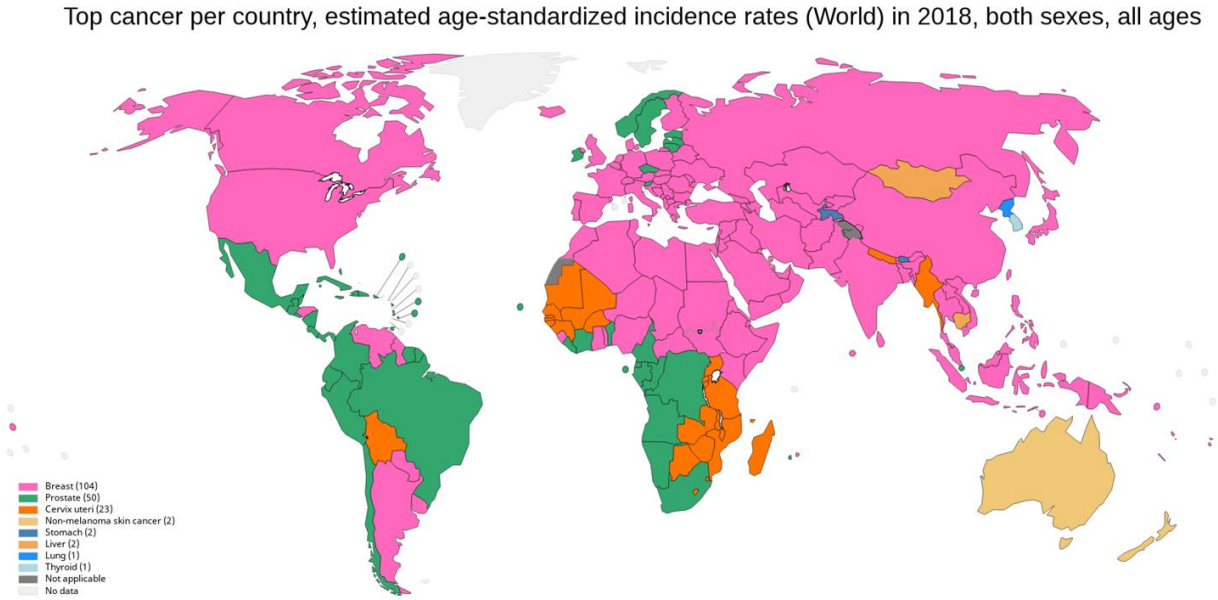


Figure 1.3: Top cancer site per country in 2018. Breast cancer is the most frequent top cancer sites, followed by prostate and cervix uteri cancers. Image source: [2]

Cancer develops through a multi-step process by which normal cells transform into malignant cells in a sequence of genetic and epigenetic changes (Figure 1.4). These changes allow the transformed cells to increase their proliferation rate and acquire new properties. The growing mass of transformed cells is initially localized to their site of origin (also called a primary or an in situ cancer). However, additional changes occurring within the proliferating tumor cells may cause them to break away from the primary tumor and invade healthy tissues or enter the blood or lymph. These invading cells may travel through the bloodstream or lymphatic system to set up new colonies of cancer in distant sites, called metastases. Most deaths associated with cancer result from metastases, as the invading cells damage healthy tissues and compromise organ functions [4].

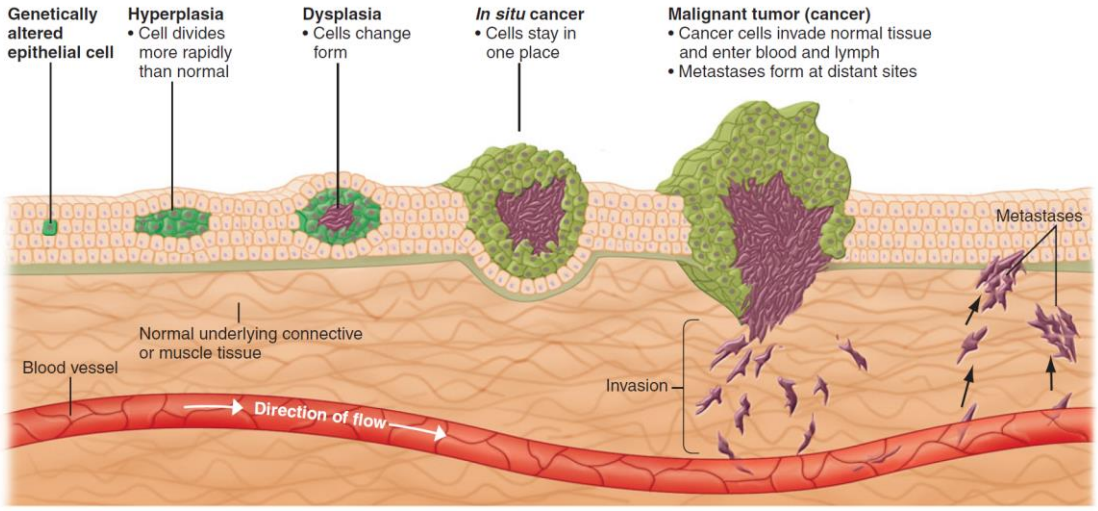


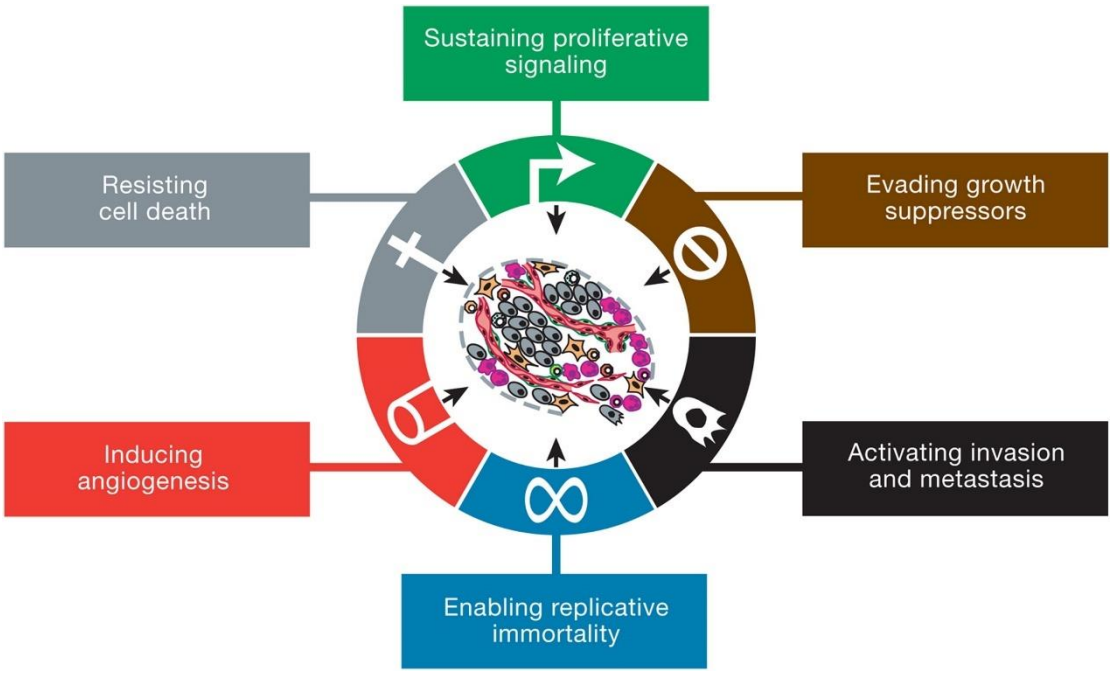
Figure 1.4: The development of a malignant tumor (cancer). Tumors develop from normal cells through a series of genetic alternations that enable them to increase their proliferation rate and acquire new properties. The localized mass of altered cells is called an in situ cancer. Additional changes in the cells may cause them to break away from the tumor and invade normal tissues or enter the blood or lymph. These invading cells may set up new colonies of cancer (called metastases) at distant sites. Source: [4].

In a seminal paper published in 2000, D. Hanahan and R. Weinberg attempted to reduce the complexity of the body of knowledge regarding the changes occurring during tumor development into six underlying principles, which they called "The hallmarks of cancer" (Figure 1.5A) [5]. The hallmarks that the authors define in the paper are (1) Cancer cells acquire the ability to stimulate their own growth ("self-sufficiency in growth signals"); (2) They become resistant to inhibitory signals that might otherwise stop their growth ("insensitivity to anti-growth signals"); (3) They evade their programmed cell death ("evading apoptosis"); (4) They acquire the ability to multiply indefinitely ("limitless replicative potential"); (5) They stimulate the growth of blood vessels to support further growth of the tumor by supplying nutrients

("sustained angiogenesis"); (6) They invade local tissue and spread to distant sites ("tissue invasion and metastasis"). Highlighting a small number of underlying principles common to many cancers is essential because it provided an organizational framework of cellular properties uncovered during tumorigenesis [6]. This framework improved the understanding of cancer biology and, in a sense, portrayed different types of cancer according to their hallmark characteristics.

A decade letter, Hanahan and Weinberg published a second, related paper that added two emerging hallmarks: reprogramming energy metabolism and evading immune response, and two enabling characteristics: genome instability and mutation, and tumor-promoting inflammation (Figure 1.5B). Of particular interest to us and of relevance to this thesis, was the emphasis on the interplay between cancer and the adaptive immune system, as an association between immune gene expression and cancer variability, as well as patient survival, was evident in our cancer data analyses. By the theory of cancer immunoediting, the interaction between the evolving tumor and its host's immune system is composed of three phases: elimination, equilibrium, and escape [7]. Whereas during the elimination phase, a competent immune system is still capable of destroying transformed cells, in the equilibrium phase, sporadic tumor cells evade destruction by the immune system and undergo immunoediting, which allows the tumor to evolve under immune selection. Finally, the immunologically sculpted tumors manage to escape the immune attack, which allows them to establish an immunosuppressive tumor microenvironment, to increase their proliferation rate, and finally also to metastasize [8]. As various routes are available through the process of immunoediting, even tumors of the same type may significantly differ in their immunogenicity, which is the ability of a substance to induce an immune response [9]. Further, the activity of the immune system in cancer patients, such as the presence of tumor-infiltrating lymphocytes (TILs) was shown to correlate with prognosis and with the response to treatment in several types of cancer [10][11].

A



B

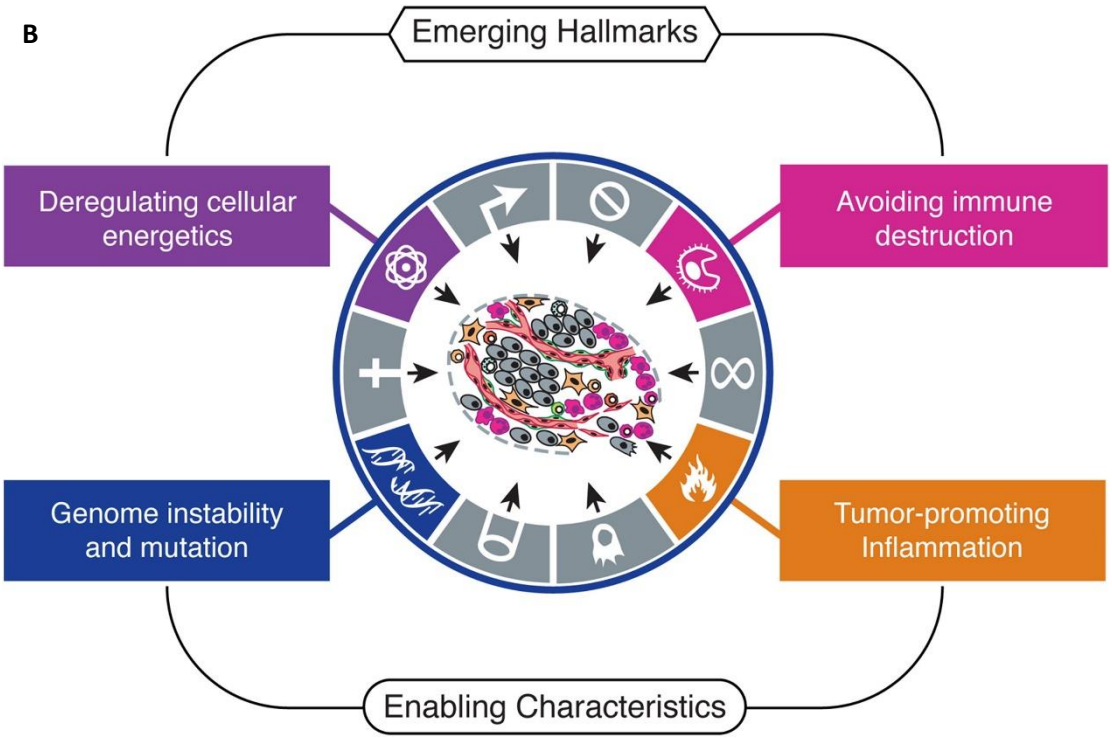


Figure 1.5: The hallmarks of cancer (A) The six hallmarks of cancer as defined by Hanahan and Weinberg, 2000 [5] **(B)** additional emerging hallmarks and enabling characteristics as defined by Hanahan and Weinberg., 2011 [12]. Image source: [5] [12]

Classic treatments for cancer include surgery, radiation, chemotherapy, or a combination of two or more of these. Recent advancements in cancer treatment introduced new treatments such as immunotherapy (boosting the responsiveness of the patient's immune system to fight the tumor more effectively), hormone therapy (slowing down hormone-dependent tumors such as breast or prostate cancers) and targeted therapies (targeting specific cancer deregulated proteins) [11]. With the advancements of our understanding of the variability within each cancer type, it is hoped that new subtype-specific treatment will be developed as part of the precision medicine approach. The development of such subtype-specific drugs depends on our ability to define clinically distinct tumor subtypes and to accurately classify tumors into subtypes based on informative biomarkers.

1.1.2. Breast cancer

Breast cancer is a heterogeneous disease exhibiting high tumor variability in terms of the underlying biological mechanisms, response to treatment, and overall survival rate [13]. Originally, therapeutic decisions in breast cancer were guided by clinicopathological parameters like tumor size, presence of lymph-node/remote metastases and histological grade. In addition, the status of three immunohistochemistry biomarkers - estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2/ERBB2) allowed the development of targeted therapies and proved predictive of treatment response [14].

With the emergence of global molecular profiling techniques, large genomic datasets became available for subtype discovery using unsupervised algorithms. By this methodology, breast samples are partitioned into subgroups using clustering algorithms, such as hierarchical clustering [15] or K-Means, and then subgroup significance is evaluated using the clinical data associated with the samples.

Initially, microarray data were used to define four molecular breast cancer subtypes (basal-like, HER2-enriched, luminal and normal-like) based on characteristic gene expression signatures in correlation with clinical data [16]. These molecular subtypes showed a reasonable correlation with the immunohistochemistry biomarker-based classification. Thus, basal-like samples are mostly triple-negative (ER-/PR-/Her2-), luminal samples are mostly ER+, and Her2 tumors are characterized by amplification and high expression of the ERBB2/HER2 gene [17][18].

Subsequent analysis conducted on a larger dataset separated the luminal subtype into two distinct subgroups named luminal-A and luminal-B. Luminal-B cancers have a higher expression of proliferation genes including Ki-67, and confer worse prognosis [19][20][21]. Moreover,

luminal-B cancers respond better to chemotherapy, while patients with luminal-A cancers benefit most from antiestrogen treatment [22].

As the partitioning of breast tumors into five molecular subtypes gained acceptance and popularity, several expression-based predictors have been developed. A central predictor is PAM50, which maps a tumor sample to one of the five subtypes based on the gene expression pattern of 50 genes [23]. Though expected to be more robust than traditional classification systems that rely only on a few biomarkers, the separation between luminal-A and luminal-B by the various predictors is not consistent, suggesting that these molecular subtypes may not represent distinct coherent sample groups [24].

Other attempts to classify breast tumors were based on other profiling technologies such as miRNA arrays [25][26], copy number variations [27] or a combination of several different technologies [28][29]. The various studies show different levels of agreement with the expression-based molecular subtypes, but taken together, they strongly indicate the existence of additional, more subtle subtypes than the PAM50 subtypes[30].

Epigenetic modifications such as DNA methylation arrays, which measure the methylation status of thousands of CpG sites across the genome [31], were also used for breast cancer classification. DNA methylation changes were shown to play a pivotal role in cancer initiation and progression [32]-[33]. Particularly, promoter hyper-methylation was associated with the silencing of tumor suppressor genes [34]. Several studies associated breast cancer molecular subtypes with specific methylation patterns [35], while others showed that methylation data might reveal additional complexity not captured on the expression level, possibly identifying finer patient groups of clinical importance [36].

Further improving the classification of breast tumors into clinically significant subtypes as well as accurate identification of the unique biological features characterizing each subtype is pivotal for improving our understanding of the disease, identifying subtype-specific biomarkers, targeted drug development and better prediction of response to treatment.

1.1.3. Cutaneous melanoma

Cutaneous melanoma is the most lethal form of skin cancer, showing a continuous rise in worldwide incidence over the past several decades [37][38][39]. Melanoma tumors develop by uncontrolled proliferation of melanocytes, the pigment-producing cells of the skin [40]. Primary melanoma tumors are regularly localized to the skin and are usually curable by excision when detected early [41]. However, melanoma tumors tend to metastasize rapidly into surrounding tissues and distant organs and are therefore considerably more challenging to cure at later stages [42].

Melanoma tumors are heterogeneous and show high diversity in their biological characteristics, metastatic potential, survival risk, and response to treatment [43]. Therefore, the stratification of melanoma tumors into clinically distinct, prognostic subtypes is crucial for accurate diagnosis, treatment guidance, and subtype-specific drug development. For the past 40 years, a clinicopathological system has been used to classify primary melanomas into four major subtypes (superficial spreading, nodular, lentigo maligna, and acral lentiginous) based on clinical and pathological features [44][45]. Although beneficial for diagnosis, this classification showed limited clinical relevance, especially for prognosis and treatment guidance [44].

With the emergence of high-throughput genomic technologies, several commonly mutated genes that play a central role in melanoma tumorigenesis and metastasis, such as BRAF, NRAS, and NF1, were identified. These findings significantly advanced the understanding of melanoma progression and led to the development of targeted therapies that have improved patient survival [46][47].

In 2015, The Cancer Genome Atlas (TCGA) reported on a study of 331 melanoma patients using six different high-throughput omic technologies [48]. The study partitioned melanoma tumors (both primary and metastatic) based on the pattern of the most prevalent mutated genes into four subtypes: BRAF, NRAS, NF1, and WT [48]. While this mutation-based classification has proven beneficial for highlighting key potential subtype-specific drug targets, it provides little prognostic value.

The same study also suggested a transcriptomics-based classification, which divided melanoma tumors (both primary and metastasis) into three prognostic groups: high-immune, keratin, and MITF-low [48]. The high-immune group showed the best 10-year survival and was characterized by the over-expression of many immune genes. The keratin group contained most of the primary tumors, conferred the worst survival (possibly due to a bias of large primary-tumor thickness in the TCGA cohort), and was characterized by over-expression of keratin, pigmentation, and

epithelial genes. Lastly, the MITF-low group showed medium survival and was characterized by the under-expression of keratin and pigmentation genes. Interestingly, these three transcriptomic sample groups showed little agreement with the mutation-based groups. Moreover, the keratin transcriptomic group showed low consistency in terms of both the expression-profiles and the clinical labels of its comprising samples, possibly suggesting the need for a more refined transcriptomic tumor classification.

For improving the survival of metastatic melanoma, a better understanding of its development as well as of its various subtypes is required, in addition to identifying informative biomarkers capable of predicting patient prognosis and response to specific treatments.

1.2. The era of omics and personalized medicine

1.2.1. High-throughput omic technologies and the multi-omics era

The "central dogma of molecular biology", stated by Francis Crick in 1957, is a framework for describing the flow of genetic information between DNA, RNA, and proteins in biological systems [49][50] (Figures 1.6 and 1.7). The framework includes three general information transfers that describe the normal flow of biological information: DNA can be copied to DNA (DNA replication), DNA information can be copied into mRNA (transcription or gene expression), and mRNA can be used as a template for synthesizing proteins (translation or protein expression). The framework also includes three special information transfers that occur only under specific conditions in case of some viruses or in a laboratory: RNA can be copied from RNA (RNA replication), DNA can be synthesized from an RNA template (reverse transcription), and proteins can be synthesized directly from a DNA template without the use of mRNA [51]. Several exceptions to the dogma have been discovered in time (such as Prions, which are self-replicating proteins[52]), but the dogma is still useful in organizing our knowledge of genetic information flow.

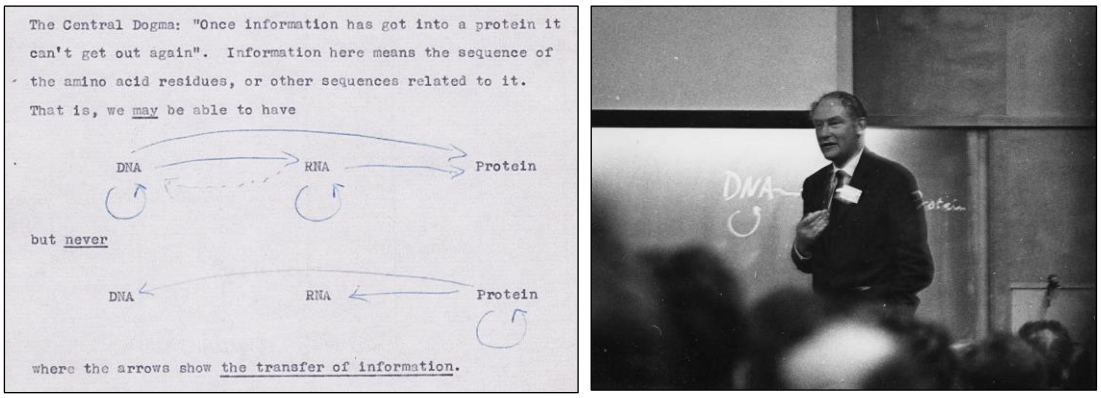


Figure 1.6: Left: Crick's first outline of the central dogma, from an unpublished note made in 1956. Source: [50], Credit: Wellcome Library, London. **Right:** Crick speaking at the 1963 Cold Spring Harbor Symposium. Source: [50], Credit: Cold Spring Harbor Laboratory.

In recent decades, several high-throughput technologies have been developed for interrogating the information captured in biological molecules such as DNA, RNA, Protein, and others [50]. These high-throughput technologies allow the simultaneous measurement of multiple biological features in a given biological sample. They are collectively called "Omic" technologies, as this suffix is common to the many types of large scale data they interrogate (genomics, transcriptomics, epigenomics, proteomics, metabolomics and others). See Figures 1.7 and 1.8.

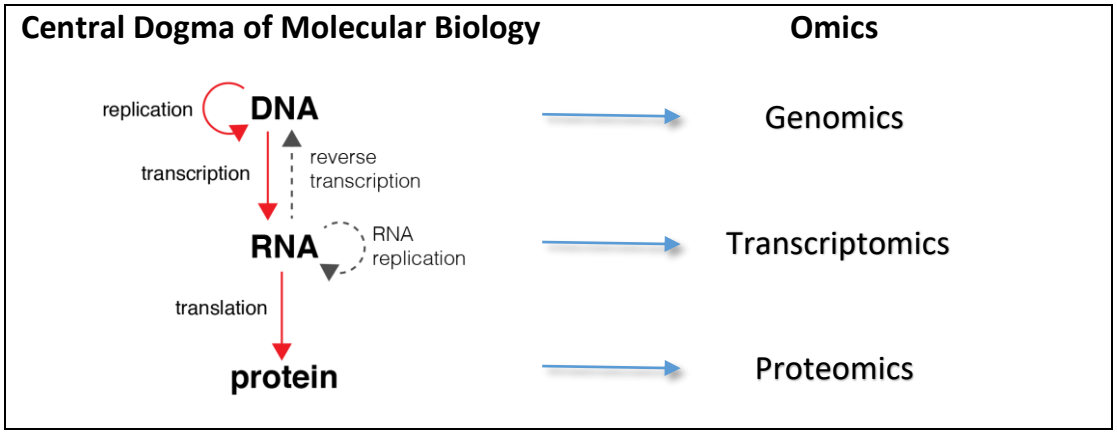


Figure 1.7: The connection between the "Central dogma of molecular biology" and the type of omic data obtained from each molecule

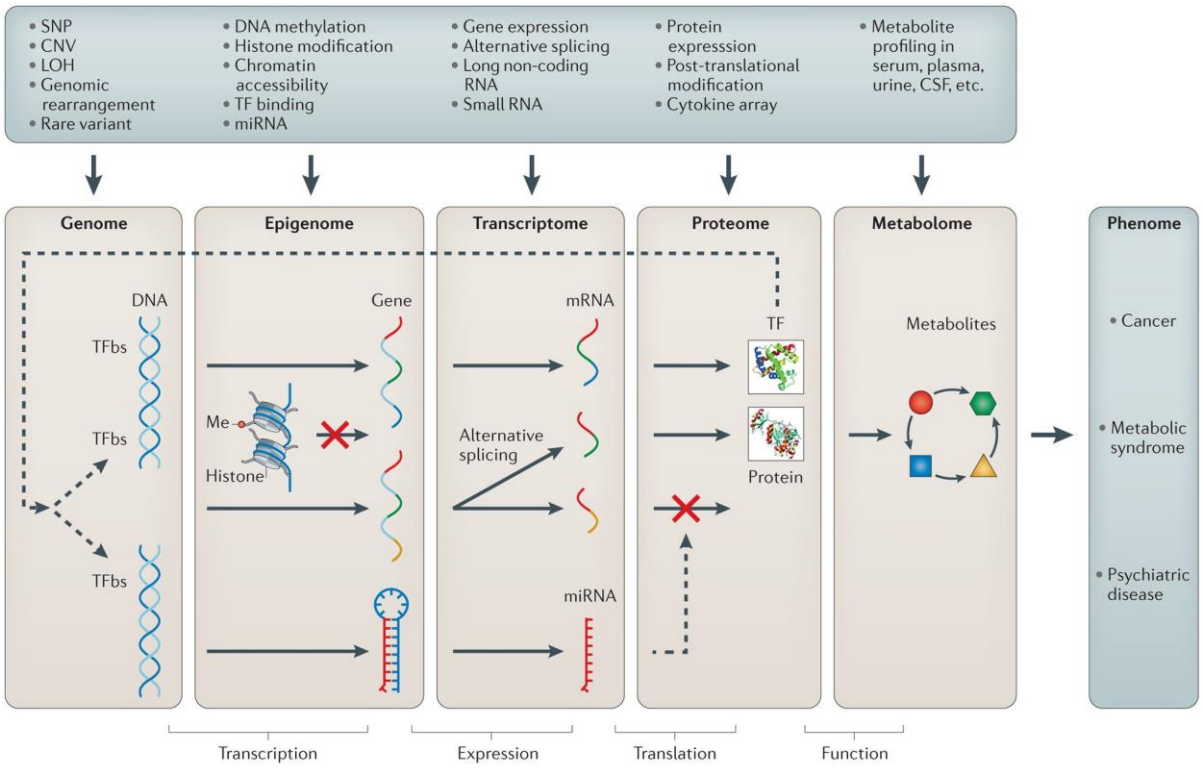


Figure 1.8: Types of omics data for interrogating the genome, epigenome, transcriptome, proteome and the metabolome. Image source: [53]

The various high-throughput omic technologies interrogate different levels of biological regulation at unprecedented speed, generating large datasets describing the examined samples in great detail, thus transforming biomedical research into an information-based field (Table 1.1). In ***Genomics***, genotype arrays [54] and next-generation sequencing (NGS) for whole-genome sequencing [55], and exome sequencing [56][57] are the main technologies currently used for interrogating DNA sequences. In ***Transcriptomics***, probe-based microarrays [58] were the first widely used method for measuring genome-wide mRNA abundance levels and allowed the generation of large scale datasets used to explore gene expression variability and dynamics in different tissues and disease states. More accurate measurement of mRNA levels became possible with the introduction of RNA-Seq technology [59][60]. Unlike microarrays, RNA-Seq profiling is not restricted to known genes, and also has a higher larger dynamic range. RNA-Seq applied NGS technologies to qualitatively and quantitatively profile all types of RNA molecules such as mRNAs, small RNAs and other non-coding RNAs [59][61][62]. In ***Epigenomics***, genome-wide characterization of DNA methylation and histone acetylation is interrogated using methylation arrays [31] or by NGS [63]. In ***Proteomics***, mass-spectrometry [64] and reverse-phase protein arrays (RPPA)[65] can be used to quantify peptide abundance in a given sample. Lastly, mass-spectrometry is also utilized to measure the abundance and relative ratios of metabolites in ***Metabolomics*** [66]. Additional omics, as well as their associated technologies, exist, and many more are expected to be developed in the next decade as the field is advancing rapidly.

The result of most omic experiments can generally be represented as a matrix whose columns represent samples, and rows represent biological features (such as genes, transcripts, CpGs, peptides or metabolites). The matrix entries indicate the existence or abundance of a specific feature in a specific sample. For convenience, we will call this matrix an "expression matrix", and will interchangeably use the terms genes and features.

An extensive array of computational methods is available for downstream analysis of large omic datasets. Several of the methods are reviewed later in this chapter. Briefly, in the context of cancer research, omic datasets can be used to identify groups of similar samples and similar features using unsupervised methods. If additional external information is available, such as clinical labels describing the samples ('Phenome'), or gene annotations describing the features, then supervised methods can be used to statistically characterize the identified sample and feature groups, to identify differentially expressed features, and to identify label-specific biomarkers [67]. Further, integrative multi-omic analysis, which combines data from more than

a single omic type (such as mRNA and miRNA, or mRNA and DNA methylation) may provide additional insights and reveal interactions between features of different types [68][69][70].

Analysis of large omic datasets might be challenging for several reasons. Firstly, the dataset can be huge in size, posing computational challenges in terms of storage, computing speed, required analysis skills, and the limited number of visualization methods suited for large-scale datasets. Secondly, the features generated by different omic technologies can significantly vary. The number of features in the resulting matrix, their biological meaning, their value distribution and the way they are correlated to features in other omics, can greatly differ between omic technologies, requiring the adjustment of the analysis workflow and the statistical methods used.

Assay	Goal	Platform	Main advantages and disadvantages
Genomics	Identify nucleotide variants (SNPs) in the whole genome associated with clinical traits (GWAS)	Genotyping arrays, whole-exome sequencing	SNP variability is stable during life; provides limited information in complex diseases due to several loci implicated
Transcriptomics	Quantify expression levels of cellular transcripts (*e.g.* mRNA)	Expression arrays, RNA sequencing	Widely used due to its high information content on cell status; differences in mRNA expression do not imply differences in proteins; does not take into account post-transcriptional modifications
Epigenomics	Determine modifications in DNA and small RNA that interfere with gene expression	DNA methylation analysis with arrays (Infinium MethylationEPIC 850K; Illumina, San Diego, CA, USA), next-generation sequencing, small RNA sequencing, arrays, *etc.*	Provides additional information to transcriptomics; related to exposures; more expensive than transcriptomics; sequencing-based approaches have computational tools in active development
Proteomics	Characterize protein expression levels of cells/samples	MS-based approaches	Expected to be closer to the phenotype; not widely used, expensive and more cumbersome analysis
Metabolomics	Characterize abundance profile of metabolites and their relative ratios	MS-based approaches	Representatives of the cellular status; applicable to many biological fluids (*i.e.* breath, blood, urine, *etc.*); not widely used

Table 1.1: Common omic data types. Source: [71]

1.2.2. Omic-based profiling and the vision of precision medicine in cancer

Tumors, even those of the same type, show great heterogeneity on the molecular level. The molecular makeup of each tumor prominently determines its proliferation rate, tendency to metastasize and its response to specific drugs. Even today, for many cancer types, subtype diagnosis is imprecise, as it is determined based on a limited number of fuzzy clinicopathological parameters. Also, traditional treatments like chemotherapy, radiotherapy and surgery, are still largely unspecific and do not take into account the concrete genetic makeup of the patient's tumor, thus often leading to treatment inefficiency, drug toxicity and significant side effects. In contrast to this "one size fits all" approach employed by traditional cancer medicine, the vision of precision medicine is that patients will receive precisely tailored treatments that target specific malfunctioning molecular pathways identified in their tumor [72][73].

The wealth of high-resolution biological data provided by large-scale omic technologies as well as their dropping costs lie at the basis of fulfilling the vision of precision medicine. Promoting precision medicine in cancer depends on the following efforts, all utilizing large-scale omics data of different types:

1. Identifying distinct groups of similar patients based on omic profiling and characterizing the prognosis, response to treatment and other clinical attributes of each group.
2. Characterizing the malfunctioning biological pathways in each patient group, and using this information to guide targeted drug development.
3. Identifying informative biomarkers that will allow classifying new patients into one of the known subtypes.

The evolution of breast cancer treatment over the past several decades demonstrates the dependency of treatment efficiency on accurate patient stratification into clinically distinct subgroups, which in turn depends on the resolution of profiling technologies (Figure 1.9). As new technologies emerged, the resolution by which tumors are interrogated increased, and finer ways to stratify the patients as well as relevant biomarkers were identified [74].

Eventually, precision medicine is envisioned to ensure that patients get the right treatment at the right dose at the right time, with minimum side-effects and maximum efficacy [75]. However, to achieve this ambitious aim, several challenges must be overcome: (1) Acquisition, storage, and analysis of even larger amounts of omics data are required for identifying even finer patient groups [76], (2) Translation of the knowledge gained from omics data analysis to practical use

into the clinic [77] (3) Development and regulatory approval of new targeted drugs aimed at treating small groups of patients [78].

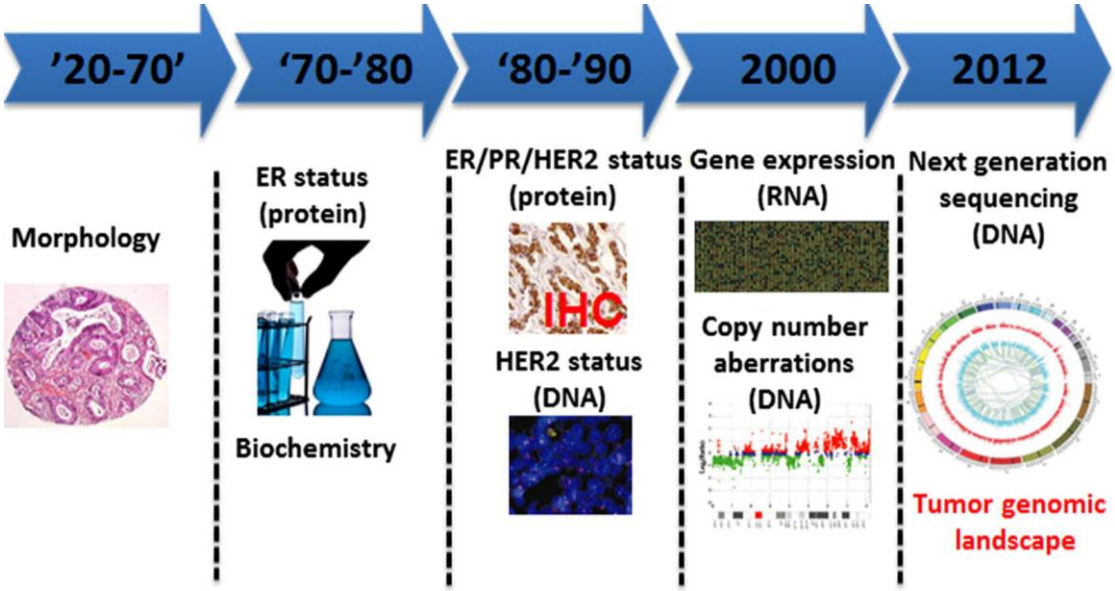


Figure 1.9: Evolution of breast cancer subtyping and personalization of treatment. Image source: [74]

1.2.3. The Cancer Genome Atlas project (TCGA)

The Cancer Genome Atlas (TCGA) [79] project is an American public-funded project, aimed to discover major cancer-causing genomic alterations and create a comprehensive "atlas" of cancer genomic profiles [80]. The project involved 20 collaborating intuitions across the US and Canada, responsible for collection and sample processing, followed by high-throughput sequencing and bioinformatics data analyses. During the years of its activity (2005-2016), the project has generated, analyzed, and made publicly available 2.5 Petabytes of genomic sequence, expression, methylation, and copy number variation data on more than 11,000 tumor samples that represent 33 different types of cancer [81].

Most TCGA samples were measured using several different omic technologies, including next-generation sequencing (DNA-Seq, RNA-Seq, and microRNA-Seq) and microarray (mRNA, DNA methylation, SNP, and Protein) based technologies (Figures 1.10 and Table 1.2). TCGA also provided detailed clinical information for each sample, which included parameters like age, gender, tumor stage, results of lab tests, treatment history and follow-up data. The data were used in the past decade by both TCGA researchers and by many other researchers around the world to advance the understanding of cancer development and cancer subtyping, and to identify the aberrations in different omics that characterize different subtypes of cancer [82].

Due to its unprecedented scope, resolution and multi-dimensional nature, TCGA's database was also used to trigger multiple computational approaches and served as a playground for testing new data mining and machine learning algorithms [80][83].

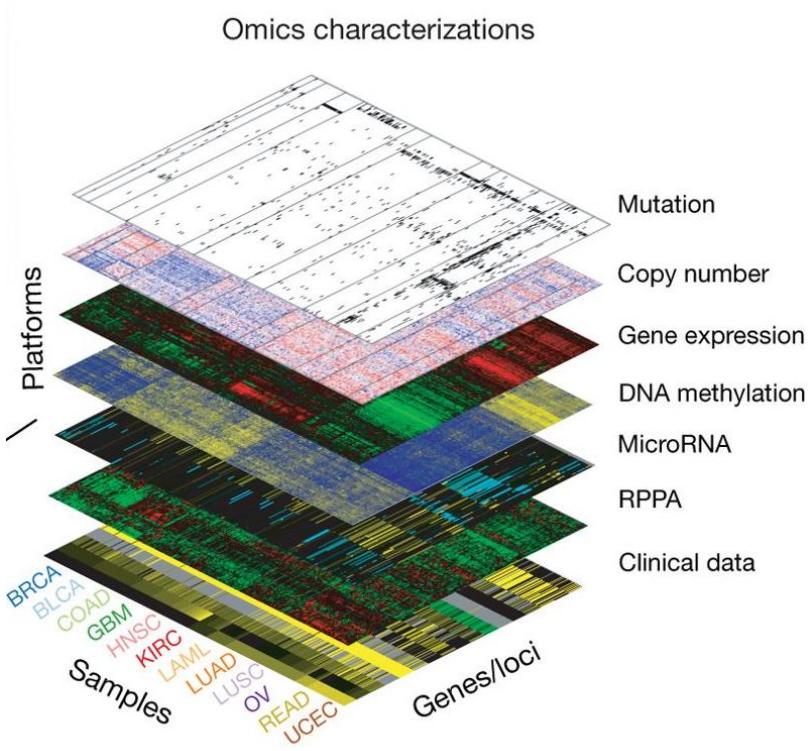


Figure 1.10: The Cancer Genome Atlas (TCGA). A multi-cancer multi-omic database. Source: [84]

"Omic" Study	Technology	Aberration
DNA sequencing	Whole-exome sequencing[a] (by massively parallel sequencing technologies)	Point mutation, small insertion/deletion (indel)
DNA copy number	Affymetrix 6.0 single nucleotide polymorphism (SNP) arrays	Deletion/amplification
DNA methylation	Illumina Infinium DNA methylation chips	Epigenetic alteration
mRNA expression	Agilent custom 244K whole genome microarrays	Gene expression
miRNA sequencing	miRNA sequencing	Aberrations in miRNA
Protein and phosphoprotein expression	Reverse-phase protein array (RPPA)	Signaling pathway activity, cell lineage marker expression

[a]Exome sequencing involves selectively sequencing the coding regions of the genome. In the human genome, coding regions comprise about 30 megabases and 180,000 exons, or 1% of the human genome. Exome sequencing enables identification of variants that affect protein sequence, but it cannot be used to identify structural or non-coding variants. miRNA = microRNA; mRNA = messenger RNA; TCGA = The Cancer Genome Atlas.

Table 1.2: Omic technologies (platforms) included in TCGA's database. Source: **[85]**

1.3. Computational methods

1.3.1. Identification of distinguishing features

Analysis of biomedical high-throughput data often aims to identify genes (or other biological features) that are differentially regulated across different sample classes [86]. Differentially expressed genes (DEGs) are characterized by having significantly different expression means on two (or more) samples classes. Therefore, they can be used as biomarkers to distinguish between the sample classes, to reveal dysregulation of biological pathways among sample classes, and also to identify informative genes for downstream analysis.

The Student's t-test and Wilcoxon rank-sum test (Mann–Whitney U test) can be used to identify genes that are differentially expressed between two sample classes (such as experiment and control or two disease subtype) [86]. The t-test is a parametric test that assumes that the data are normally distributed, and it has higher statistical power than the rank-sum test, which does not make that assumption. The rank-sum test aims to detect differences of variable values between two samples based on ranking, and therefore it is less sensitive to outliers and can also be performed when the only available data are those relative ranks [87]. For identifying differentially expressed genes among more than two sample classes (such as multiple disease subtypes or experiment time points), the parametric ANOVA (analysis of variance) test or the non-parametric, ranking-based Kruskal-Wallis ANOVA test are appropriate [88][89].

The null hypothesis made by the four tests mentioned above is that there is no difference in expression between the classes. After a test statistic was computed by one of the tests, for each gene separately, it can be converted into a p-value, which represents the probability of having observed our data (or more extreme data) when the null hypothesis is true [90]. When the p-value is below a certain cut-off (0.05 is often used), we reject the null hypothesis and the result is considered statistically significant [88]. Since typical analyses for identifying differentially expressed genes in modern high-throughput datasets may include many thousands of simultaneous hypothesis testing, we must account for the multiple testing problem [91]. The problem refers to the situation where the expected number of false discoveries becomes large relative to the number of true discoveries. The problem was originally addressed by methods to control the family-wise type I error rate (FWER), such as the Bonferroni correction method [92], and later by the less conservative method of FDR (False Discovery Rate), which controls the family-wise error rate [93][94].

Lastly, Fold-change is a simple metric for comparing gene expression levels between two sample classes. Fold-change is the ratio between expression averages in the two sample classes. Fold change is applied mainly as a measure of effect size, and nowadays it is considered inadequate inference statistic because it does not incorporate variance and offers no associated level of confidence.

Altogether, these methods allow for the generation of lists of differentially expressed genes which can then be associated with biological function by performing a gene enrichment analysis on the list genes, as described below.

1.3.2. Clustering analysis

Clustering analysis is an unsupervised method, used to discover relations between objects by grouping them into disjoint groups based on a defined similarity metric [95]. Ideally, objects assigned to each group will have markedly higher similarity to objects in the same group, compared to their similarity to objects assigned to other groups. Similarly to other unsupervised methods, clustering attempts to find previously unknown patterns in a given dataset without using any preexisting labels [96][97].

Clustering is a very useful method in the exploratory biomedical analysis of high-dimensionality data, as it enables to reveal high-level structures in large datasets [98][99][100]. Given an expression matrix representing the expression levels of F features on S tumor samples taken from patients (such as the one in Figure 1.11), clustering can be applied in two different but complementary ways:

1. Clustering the dataset samples (the columns in Figure 1.11) identifies groups of similar samples, that share a similar genomic signature and may correspond to disease subtypes [101].
2. Clustering the dataset features (the rows in Figure 1.11) identifies groups of similar features that may correspond to co-regulated genes [102].

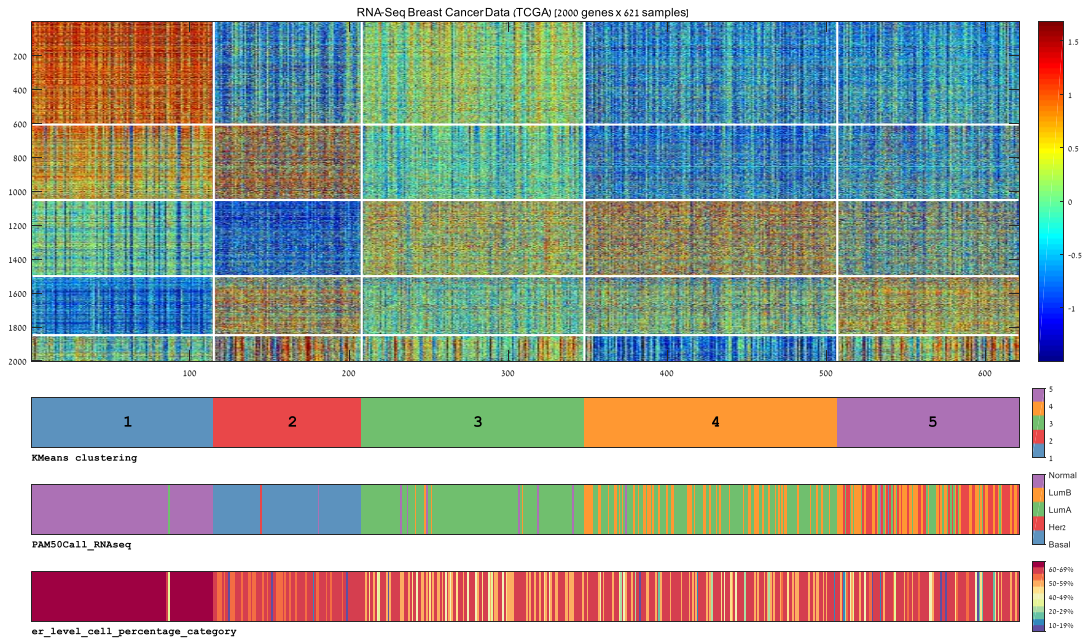


Figure 1.11: An example of a two-way clustered expression matrix. Clustering of the matrix samples (columns) identifies groups of similar samples that may correspond to disease subtypes, whereas clustering the features (rows) identifies groups of correlated features that may represent co-regulated genes.

Many different clustering algorithms have been proposed in the literature. There are several ways to categorize them based on the way they operate [103][104], including hierarchy-based (such as hierarchical-clustering [15]), partition-based (such as K-means[105]), density-based (such as DBSCAN[106]), and graph-theory-based (such as CLICK [107]). Table 1.3 lists several common clustering algorithms by category. The algorithms may greatly differ in their time complexity, sensitivity to noise or outliers, input parameters and fit to specific applications. Many clustering algorithms function based on a distance (or similarity) function by which object similarity is calculated (common distance functions are listed in Table 1.4). Other common inputs are the number of desired groups (such as K in K-means) or other parameters for determining group granularity (such as dendrogram cutoff thresholds in hierarchical clustering or the homogeneity threshold in CLICK). Virtually all the clustering formulations give rise to NP-hard problems [108][109][110]. Determining an optimal (or "true") number of clusters in a given dataset is a fundamental and unsolved problem in clustering analysis [111]. In practice, obtaining a satisfactory solution may require repeated attempts (and application of multiple algorithms) and reliance on measures for clustering goodness as described below [112].

Category	Typical algorithm
Clustering algorithm based on partition	K-means, K-medoids, PAM, CLARA, CLARANS
Clustering algorithm based on hierarchy	BIRCH, CURE, ROCK, Chameleon
Clustering algorithm based on fuzzy theory	FCM, FCS, MM
Clustering algorithm based on distribution	DBCLASD, GMM
Clustering algorithm based on density	DBSCAN, OPTICS, Mean-shift
Clustering algorithm based on graph theory	CLICK, MST
Clustering algorithm based on grid	STING, CLIQUE
Clustering algorithm based on fractal theory	FC
Clustering algorithm based on model	COBWEB, GMM, SOM, ART

Table 1.3: Categories of clustering algorithms and typical examples of specific algorithms in each category. Source: [113].

Name	Formula	Explanation
Minkowski distance	$\left(\sum_{l=1}^{d}\left\|x_{il}-x_{jl}\right\|^{n}\right)^{1/n}$	A set of definitions for distance:
		1. City-block distance when n = 1
		2. Euclidean distance when n = 2
		3. Chebyshev distance when n $\rightarrow \infty$
Standardized Euclidean distance	$\left(\sum_{l=1}^{d}\left\|\frac{x_{il}-x_{jl}}{s_l}\right\|^{2}\right)^{1/2}$	1. S stands for the standard deviation
		2. A weighted Euclidean distance based on the deviation
Cosine distance	$1-\cos\alpha=\frac{x_i^T x_j}{\|x_i\|\|x_j\|}$	1. Stay the same in face of the rotation change of data
		2. The most commonly used distance in document area
Pearson correlation distance	$1-\frac{Cov(x_i,x_j)}{\sqrt{D(x_i)}\sqrt{D(x_j)}}$	1. Cov stands for the covariance for and D stands for the variance
		2. Measure the distance based on linear correlation
Mahalanobis distance	$\sqrt{(x_i-x_j)^T S^{-1}(x_i-x_j)}$	1. S is the covariance matrix inside the cluster
		2. With high computation complexity

Table 1.4: Common distance metrics used in clustering analysis. Source: [113].

K-means and Hierarchical clustering are two of the most widely used clustering algorithms, especially in the field of biomedical research [98]. Both are simple, widely implemented and their results can be easily visualized and understood. K-Means assumes that the number of clusters is k. It starts by choosing k data objects at random as cluster centers. The distance function is then used to assign each dataset object to the closest center. This partitions all objects into k groups. Next, the cluster centers are updated to be the centroid (mean) of the groups. Iteration of this process continues until a minimal decrease in squared error is reached. K-means is well suited for identifying size-balanced disease subtypes, as using centroids has the advantage of a clear geometric and statistical meaning while keeping the algorithm insensitive to data ordering. However, k-means is sensitive to data noise and outliers, can only work with numerical features, and the number of clusters k must be specified in advance [98][100][114]. Hierarchical clustering produces a dendrogram, i.e. a rooted tree with edge lengths where all objects are leaves and all root-leaf distances are equal. The tree-distance of two objects in the tree is the length of the path between them. Given pairwise input distances of objects, the goal is to build a tree such that the tree-distances will match the input distances as much as possible. Having created the dendrogram, clusters of different granularity can later be produced by thresholding pairwise tree-distances. The algorithm can work for any type of data and does not make any assumptions about the underlying data distribution. However, the algorithm is less scalable to large datasets and performs poorly when the clusters vary considerably in shape, density, or size [114][115].

Several methods are available for validating the goodness of a clustering result [116]:

a. **Internal cluster validation methods** (such as the Silhouette coefficient or the Dunn index) use only the clustered data itself without any external information, to evaluate the tradeoff between clusters compactness (intra-cluster similarity) and separation (inter-cluster similarity).
b. **External cluster validation methods** use external information for comparing the resulting clusters to class labels using statistical tests for enrichment (such as hypergeometric test or Chi-square test).
c. **Relative cluster validation methods** explore a variation of the clustering parameters until reaching a stable cluster structure (example: testing various values for the number of clusters k).

In this study, clusters were mainly validated using external information, as described in the next section about enrichment analysis.

1.3.3. Enrichment analysis

Gene enrichment analysis

Gene enrichment analysis is often used in biomedical research for interpreting the biological meaning of gene groups [117]. The gene groups commonly originate from a clustering operation performed on the genes of an expression matrix, or from a supervised test identifying the top differentially expressed gene among two or more sample classes. For characterizing the biological meaning of each gene group, the group can be tested for enrichment for an array of known gene classes using statistical methods such as the hypergeometric test, Fisher's exact test, chi-square test and binomial probability [117]. The result of such analysis usually takes the form of a list of gene classes ranked by decreasing significance of enrichment on the gene group. Due to the large number of enrichment tests conducted in a typical gene enrichment analysis, resulting p-values must be corrected by a method such as FDR[93]. Gene annotation databases such as Gene Ontology (GO) [118], Kyoto encyclopedia of genes and genomes (KEGG) pathways [119], Wiki-Pathways [120], chromosomal location annotations and catalogs of tumor suppressor genes, are commonly used as gene classes for enrichment analysis.

The gene enrichment analyses performed in this thesis were conducted using several tools, including TANGO (which is part of the Expander tool) [121]-[122], PROMO [123], and GOrilla [124]. The GOrilla tool, as well as other tools like GSEA [125], can identify enrichment of gene classes in a list of ranked genes, preventing the need to decide on a significance cutoff for the list of differentially expressed genes. These methods perform well when genes are easily ranked but may be suboptimal when lack of information prevents reliable ranking of the genes [126].

Sample enrichment analysis

A major effort in promoting precision medicine in cancer is to stratify the patients of a certain cancer type into clinically distinct subgroups. To this end, a clustering algorithm is first applied on the dataset samples (taken from patients), based on the genomic data only, and then the clinical labels are used for external validation of the clusters. Cancer datasets often include a wealth of clinical sample-labels of various types: Numeric (e.g., age, tumor size and the number of cigarettes smoked per day), Categorical (e.g., gender, histological type, and receptor status), Ordinal (e.g., pathological stage, metastasis stage) and Survival (overall survival, recurrence-free survival, etc.). These labels can be used to statistically characterize each of the sample subgroups by employing an appropriate test for label type. The enrichment of sample clusters for categorical labels can be tested using the hypergeometric or the chi-square tests. Differences of numeric and ordinal labels between sample-subgroups can be tested using t-test and ANOVA for

normally distributed values or using the Wilcoxon rank-sum test and the Kruskal-Wallis tests, which are non-parametric [87][127][128][90]. Survival labels can be used to prognostically characterize the sample-groups using survival specific tests as described in the next section. Evaluation of the most significant enrichments for clinical labels found for each sample group allows us to clinically characterize the sample groups and determine their relation to previously known labeling of the samples.

1.3.4. Survival analysis

Survival analysis is a collection of methods for comparing the risks for an event such as death or disease recurrence, for groups of patients, where the risk changes over time [129]. The patient groups can be formed by disease subtypes, by different treatments administered to patients in a clinical trial, or by partitioning of the patients based on a certain biomarker. Survival analysis methods are suited for analyzing censored longitudinal data, which include incomplete data for patients who did not experience an event by the time their follow-up ended (either since the study ended or since they left the trial earlier). The censored longitudinal data underestimate the true (but unknown) time to event, but still, hold valuable information taken into consideration by the various survival analysis methods [130]. Kaplan–Meier (KM) plots, log-rank tests, and Cox (proportional hazards) regression are the most commonly used methods for survival analysis in cancer research [129][130].

The Kaplan-Meier (KM) method plots the empirical survival probability based on observed survival times [130][131]. The KM survival curve is a plot of the KM survival probability as a function of time, and is often used to visually compare the estimated survival function of two or more groups. The difference between curves can be tested statistically, most commonly using the log-rank test. Two issues are important when interpreting KM curves: (1) the validity of the curve depends on the assumption that censoring is unrelated to prognosis and that the survival probabilities are the same for subjects recruited at any stage of the study.(2) The statistical precision diminishes as follow-up increases because the curve is based on a smaller number of patients [132].

The log-rank (Mantel-Haenzel) test [133][134] is a nonparametric statistical test used for statistically comparing the survival curves of two or more groups. It is used to test the null hypothesis that there is no difference between the population survival curves. Several variations to the log-rank test exist [135][136][137], and all of them make the same assumptions as those for interpreting KM curves, namely independence of censoring from the outcome and time homogeneity. The tests produce a p-value indicating the significance of the difference between

the tested curves. For a more detailed examination of survival differences between survival curves, the log rank tests can be applied for comparing the groups with one another, but then the resulting p-values must be adjusted for multiple comparisons [138].

The Cox proportional hazards model is a semiparametric regression model that is commonly used to test the association between the survival time of patients and one or more explanatory variables [139][140][141]. Unlike the KM method and the log-rank test, the COX proportional hazards model supports more than a single explanatory variable and can test for either univariate or multivariate associations of both categorical and numerical variables to survival. The model calculates a Hazard Ratio (HR) for each explanatory variable where values equal, greater or lower than 1 represent no effect, increased or reduction in hazard, respectively.

In this work, we extensively used the three survival analysis methods described above for testing the clinical significance and prognostic value of patient subgroups we identified and of genomic signatures and biomarkers we suggested.

2. Breast cancer subtypes

The large breast cancer dataset developed and provided by The Cancer Genome Atlas project (TCGA [142]) includes more than a thousand breast tumor samples characterized by various modern high throughput genomic technologies. This dataset constitutes a significant leap forward compared to the older microarray-based data. mRNA abundance levels are measured in TCGA's dataset using RNA-Seq technology. This technology shows increased sensitivity and a higher dynamic range compared to microarrays [20][21]. DNA-methylation arrays applied on the same samples can help decipher biological tumor variability by epigenetic modifications not manifested on the gene expression level.

The aim of this project was to improve the classification of breast tumors based on the extensive TCGA expression and methylation data that have recently become available. We utilized these datasets to re-visit the current classification of breast tumors into biologically distinct subgroups.

Our initial question was whether unsupervised clustering of all TCGA breast samples using the RNA-Seq data would reconstruct the partition defined by PAM50. As the luminal samples showed the highest level of variability in our global clustering, we also asked how the luminal samples would cluster into two groups based on the RNA-Seq data, how the resulting sample groups would compare to PAM50's partition into luminal-A and luminal-B, and whether that partition would have a clinical advantage over PAM50's partition of the luminal samples. Looking into the internal structure of the highly variable luminal-A samples, we asked whether this PAM50 group can be further partitioned into finer subgroups showing biological distinctness and clinical significance. We then used enrichment analysis to explore the biological mechanisms underlying the new luminal-A subgroups.

We asked similar questions regarding breast tumor variability on the epigenetic level. We evaluated the methylation-based partition of all breast tumors, all the luminal samples, and the highly heterogeneous luminal-A, and compared the resulting partitions to PAM50. To examine the biological characteristics of differentially methylated CpGs (DMCs) separating the new methylation-based luminal-A subgroups, we conducted an enrichment analysis. Finally, we performed a multivariate Cox survival analysis to determine whether these subgroups have independent prognostic value. Our improved and refined classification may contribute to the precision of diagnosis and thus to more personalized treatment.

2.1. Results

2.1.1. Separation of luminal-A and luminal-B samples is not reconstructed by RNA-Seq unsupervised analysis

We started by evaluating the global sample structure within the RNA-Seq gene expression data obtained from TCGA. We applied unsupervised analysis on both tumor (n=1035) and normal (n=113) breast samples using the K-Means clustering algorithm over the top 2000 variable genes. Since our initial goal was to compare the resulting partition to the four intrinsic molecular types, we used K=5 (corresponding to the four types represented by PAM50 label classes in addition to Normal). The results are shown in Figure 2.1.

The resulting clusters exhibited moderate correspondence with PAM50 labels: Most basal-like, normal and HER2-enriched samples fell into three different clusters (numbers 4, 5, and 3 respectively, listed in decreasing levels of homogeneity), whereas the luminal samples exhibited a much greater variability. Importantly, most luminal-A samples were split between two different clusters - a homogenous luminal-A cluster (cluster 2), and a cluster composed of a mix of luminal-A and luminal-B samples (cluster 1).

Furthermore, the samples assigned to cluster 2 exhibited a very distinct expression pattern, over-expressing 1184 genes compared to cluster 1 (out of 1421 differentially expressed genes, see "Methods"). Cluster 1 samples over-expressed only 229 genes compared to cluster 2 (See Figure S1.1E for per-cluster distribution and Figure S1.1F for results of differential gene expression analysis).

According to these results, the variability within the luminal samples is not sufficiently captured by the PAM50 luminal-A and luminal-B subtypes. Specifically, they suggest that luminal-A samples can be further partitioned into finer subgroups, possibly having clinical meaning.

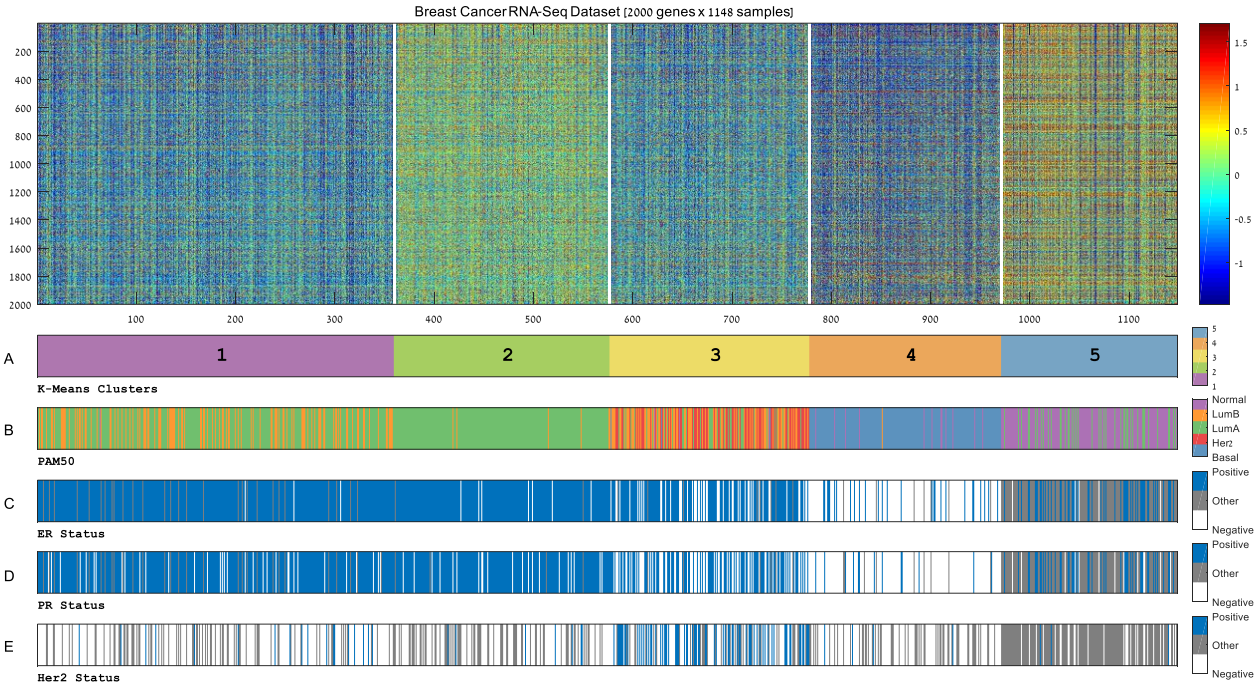


Figure 2.1: Global unsupervised clustering of 1148 breast samples using RNA-Seq data. Applying the K-Means algorithm using K=5 on the RNA-Seq dataset yielded a partition exhibiting moderate agreement with PAM50 labels and the three IHC markers. Notably, luminal-A samples were split between a rather homogenous cluster 2 and cluster 1 which is composed of luminal-A and luminal-B mix. (A) K-Means clusters (B) PAM50 calls (C) Estrogen receptor status (D) Progesterone receptor status (E) HER2 status.

2.1.2. Unsupervised partition of luminal samples predicts survival and recurrence better than PAM50

To further investigate the variability among luminal samples, we clustered the 737 luminal samples (534 luminal-A and 203 luminal-B samples based on PAM50 labels) into two groups. The results are shown in Figure 2.2A. Similar to the global analysis, the luminal-A samples were divided between a luminal-A mostly homogenous cluster (cluster 2) and a cluster composed of both luminal-A and luminal-B samples (cluster 1).

Survival analysis performed on the two luminal partitions (the PAM50 luminal-A/luminal-B partition, and the two K-Means clusters shown in Figure 2.2A) showed that the RNA-Seq-based clustering partition outperforms the luminal-A/luminal-B distinction in terms of both survival and recurrence (5-year survival plots are shown in Figure 2.2B; also see Figure S1.2A for overall survival plots). Hence, the signal identified by our unsupervised analysis of the RNA-Seq data translates into a clinically relevant partition of the luminal samples that has better predictive power than PAM50's luminal-A/luminal-B partition in terms of both survival and recurrence.

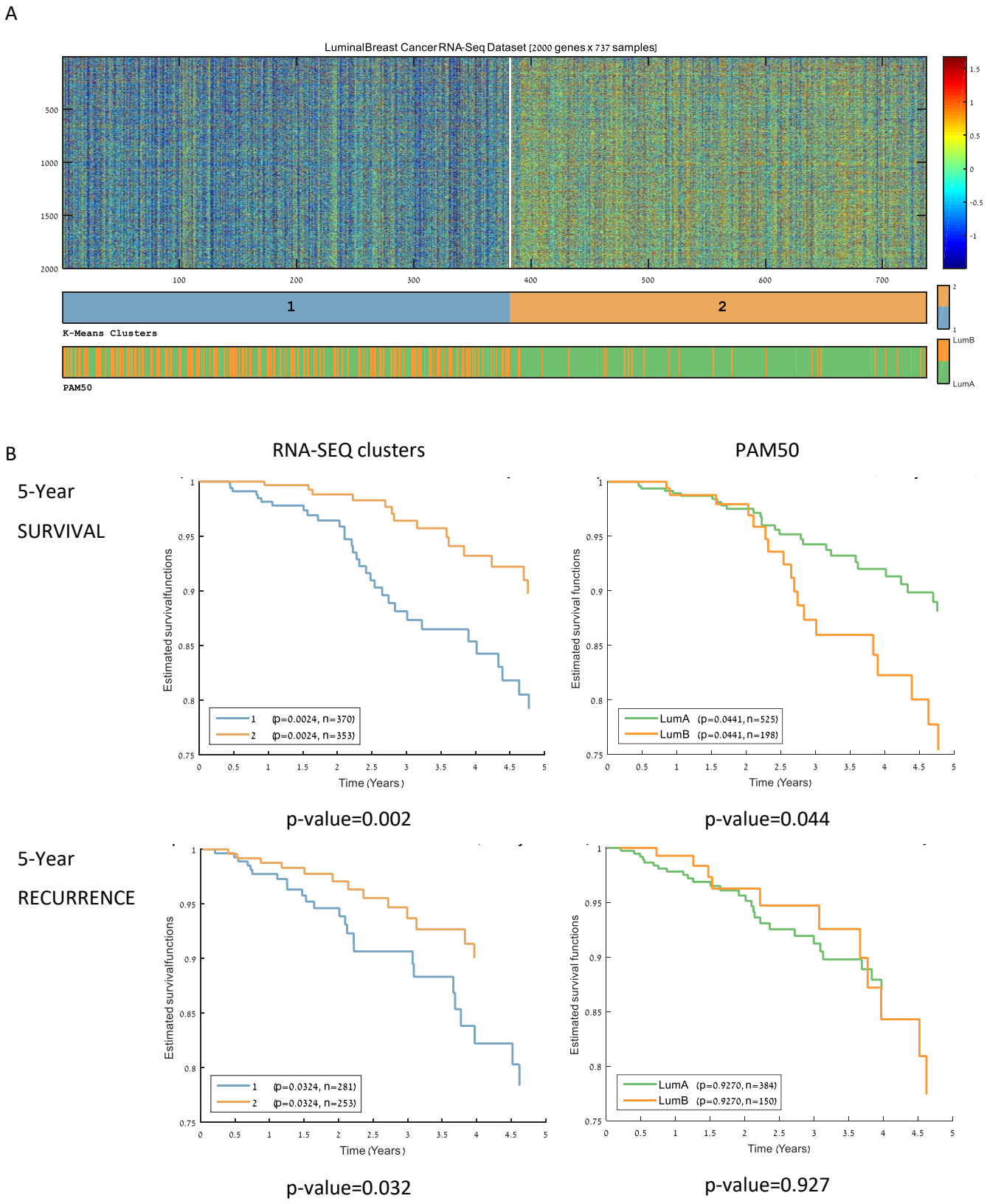


Figure 2.2: Unsupervised analysis of luminal breast samples using RNA-Seq data. (A) Applying the K-Means algorithm on the 737 luminal samples using K=2 split the samples into two subgroups exhibiting better five-year prognostic value than the PAM50's luminal-A/luminal-B partition. **(B)** Five-year survival and recurrence Kaplan-Meier plots for the two luminal breast cancer partitions. The partition into two RNA-Seq based clusters outperforms PAM50 partition of the luminal samples in both survival and recurrence. P-values were calculated using the log-rank test.

2.1.3. Luminal-A samples show two distinct classes exhibiting clinical significance

As the luminal-A samples displayed the highest level of variability by consistently falling into two major subgroups in previous steps, we focused on this PAM50 class in an attempt to explore its underlying substructures. To this end, we re-clustered only the 534 luminal-A samples into two groups (Figure 2.3A). As the resulting clusters were found to be significantly enriched for various clinical variables, we designated them as LumA-R1 (n=258) and LumA-R2 (n=276).

The most apparent property of the resulting partition was the general over-expression pattern exhibited by LumA-R2 samples compared to LumA-R1 samples. Indeed, out of the 2000 genes selected for clustering, 1276 were differentially expressed and 1068 of them were over-expressed in LumA-R2 samples (based on FDR corrected rank-sum test). A highly similar partition (Chi-Square p=1.1e-40) with a parallel over-expression pattern was identified on a microarray gene expression dataset also available from TCGA for a subset of the luminal-A samples used here (n=265). This supports the conclusion that the partition and distinct over-expression pattern we observed are not an artifact originating from RNA-Seq measurement technology or from any normalization protocols applied on the dataset (See Supplementary Information, section S1.4).

Recurrence analysis performed on these two luminal-A subgroups associated LumA-R2 samples with a significantly reduced 5-year recurrence rate (p=0.0076, Figure 2.3B). Enrichment analyses on additional clinical information available for the samples revealed that LumA-R1 and LumA-R2 subgroups are enriched with ductal (p=2.1-05) and lobular (p=9.7e-12) histological types, respectively. LumA-R1 samples were associated with a higher proliferation score (p=8.9e-25), older age (p=2.6-05), and a slight but significant decrease in normal cell percent (p=2.8e-08) accompanied by an increase in tumor nuclei percent (p=2.6e-12) compared with LumA-R2 samples (see Table 2.1).

Comparing the luminal-A partition shown in Figure 2.3A to the groups formed when clustering all the luminal samples (Figure 2.2A), we note that almost all LumA-R2 samples are contained within cluster 2 (composed of mainly luminal-A samples) whereas most LumA-R1 are contained within cluster 1 (composed of a luminal-A-luminal-B mixture). See the second label bar in Figure 2.3A. This suggests that LumA-R1 samples are more similar in their expression profile to luminal-B samples compared with LumA-R2 samples.

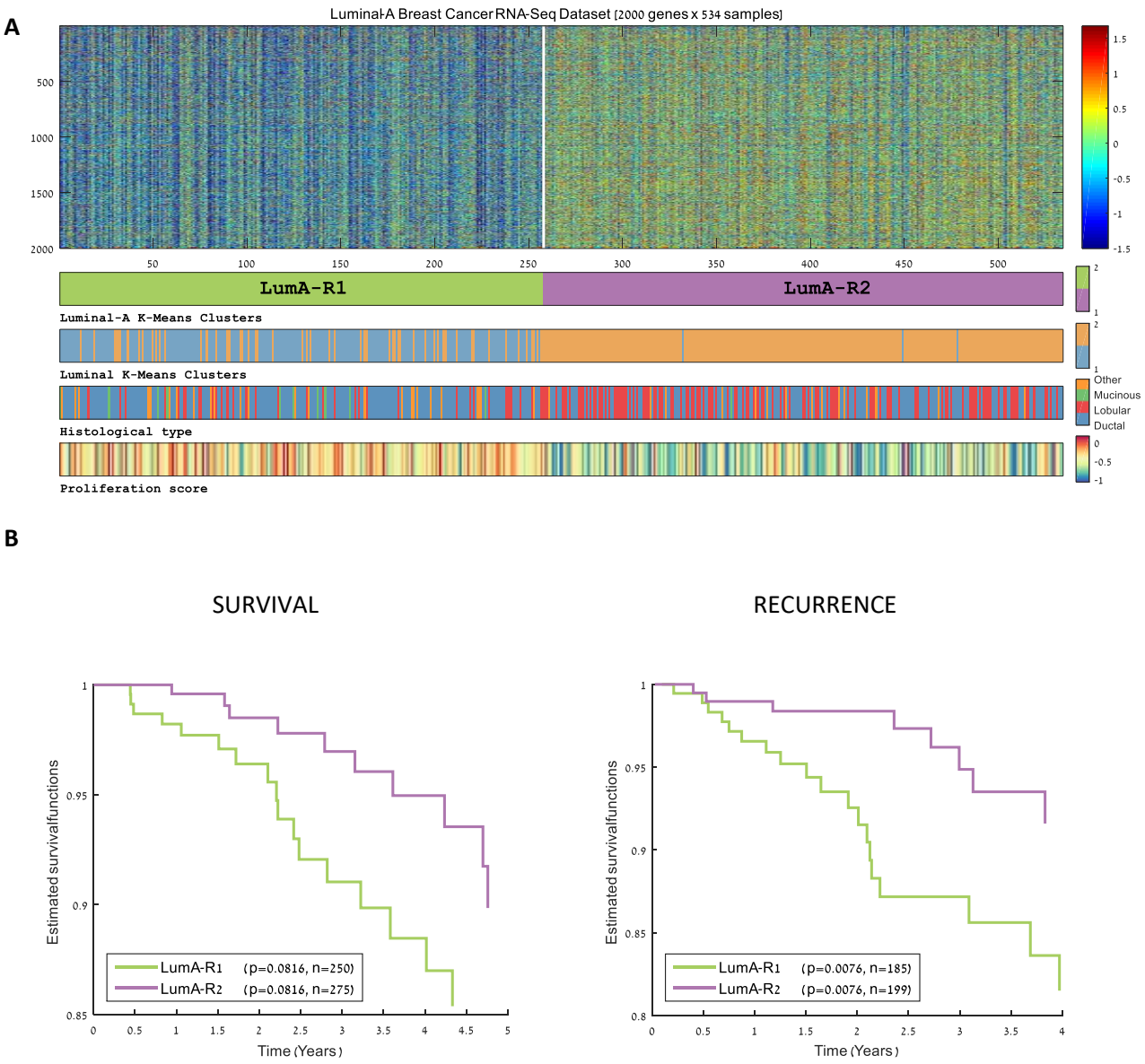


Figure 2.3: Unsupervised analysis of luminal-A breast samples. (A) Clustering of 534 RNA-Seq profiles partitions the data into two groups exhibiting distinct expression profiles. The clusters also show significant enrichment for clinical variables including recurrence, proliferation score, age and histology. The bars below the heatmap show, from top to bottom, the partition of the samples, the designation of the samples according to the clustering of all luminal samples (Figure 2.2), histological type and proliferation scores. **(B)** Five-year survival and recurrence analysis for the two luminal-A subgroups. LumA-R2 samples exhibit a significantly reduced five-year recurrence rate compared with LumA-R1.

Group Characteristic	LumA-R1	LumA-R2	p-value
Recurrence free survival	Increased recurrence	Reduced recurrence	7.6e-3
Histology Enrichment p-values for each group	Ductal (p=2.1e-05)	Lobular (p=9.7e-12)	
Age average	61.5	57.4	2.6e-05
Proliferation score	-0.4	-0.6	8.9e-25
Tumor nuclei percent	80%	73%	2.6e-12
Normal cell percent	2.9%	6.1%	2.8e-08
Gene overexpression	194	1068	

Table 2.1: The main distinguishing characteristics between the luminal-A subgroups LumA-R1 and LumA-R2. Average values are shown for each group where relevant. Gene overexpression is computed with respect to the 2000 genes used for clustering.

2.1.4. Luminal-A subgroups exhibit distinct immune system expression profiles

In order to identify genes that distinguish best between LumA-R1 and LumA-R2 samples, we created a list of the 1000 most differentially expressed genes (see "Methods"). In agreement with the general expression pattern described earlier, all genes in the list were over-expressed in LumA-R2 compared to LumA-R1 samples. The most significant categories in the enrichment analysis performed on this list were related to the immune system regulation. The more specific category of T cell receptor signaling genes appeared consistently in analyses based on various annotation databases (Gene Ontology: "T Cell activation" p=1e-05, KEGG Pathway: "T Cell receptor signaling pathway" p=3e-07, Wiki-Pathway: "T Cell receptor (TCR) Signaling Pathway" p=1.09e-07). Other enrichments of interest included the KEGG Pathways "Cytokine-cytokine receptor interaction" (p=2.13e-13), "Chemokine signaling pathway" (p= 1.14E-09) and Wiki-Pathway "B Cell Receptor Signaling Pathway" (p=1.72e-06). See Table 2.2 for a list of the most significant categories, and Supplementary Information, section S1.5 for the full list.

Careful examination of the gene list revealed that LumA-R2 samples over-express genes that are typically expressed by various immune system cells (e.g., the leukocyte marker CD45/PTPRC, T Cell marker CD3 and B-Cell marker CD19) [143] [144] [145] [146]. A significant number of over-expressed genes are related to the T Cell receptor (CD3D, CD3E, CD3G, and CD247) and the upstream part of its signaling pathway (ZAP70, LCK, FYN, LAT, PAK, ITK) [147] (Figure 2.4).

Interestingly, the over-expressed genes were related to T Cell or Natural Killer (NK)-mediated cytotoxic activities (GZMA, GZMB, GZMH, GZMM, PRF1) [148]-[149].

We also observed that the over-expression of immune receptor genes in LumA-R2 samples was accompanied by over-expression of several chemokine genes (CCL5, CCL17, CCL19+CCL21) and their corresponding receptors (CCR5, CCR4, CCR7). Topping the list of overexpressed genes in Lum-A-R2 samples (ranked by p-value) is the Interleukin-33 (IL-33) gene, which drives Th2 responses [150].

In summary, LumA-R2 samples exhibit better prognosis based on several clinical parameters while over-expressing a significant number of genes related to the immune system.

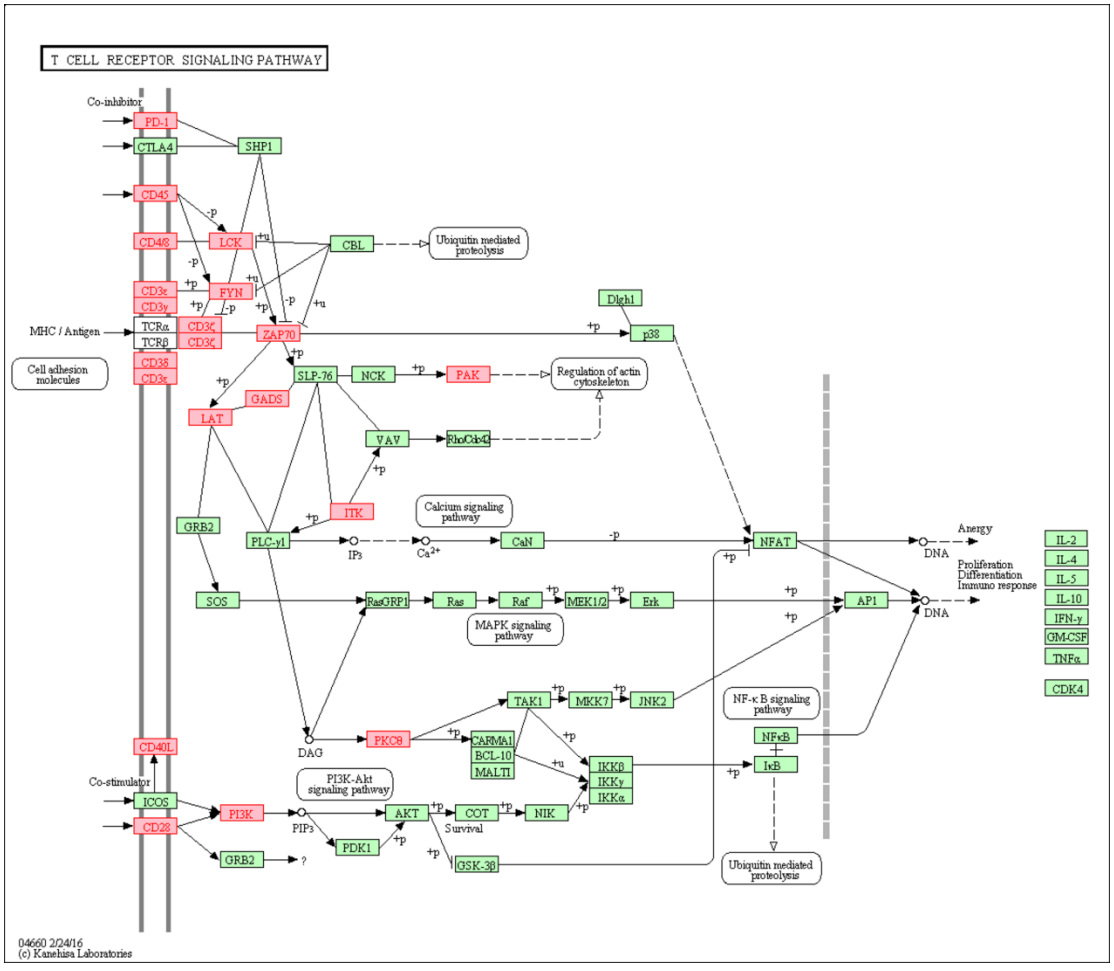


Figure 2.4: LumA-R2 over-expressed genes in the T Cell receptor signaling pathway. The list of top 1000 differentially expressed genes between LumA-R1 and LumA-R2 samples was found to be significantly enriched for the pathway genes (p=1.3e-07). Genes marked in red are over-expressed in LumA-R2 samples. Pathway and graphics were taken from the KEGG database.

Enrichment Type	Term	#Genes	P-VALUE
Gene Ontology	regulation of immune system process	152	3.74E-50
	immune system process	201	3.65E-47
	regulation of leukocyte activation	71	2.37E-28
	regulation of multicellular organismal process	183	2.89E-28
	cell activation	91	4.59E-28
	regulation of response to external	73	8.18E-27
	regulation of biological quality	218	1.82E-26
	leukocyte activation	67	1.95E-26
	positive regulation of cell activation	56	5.13E-24
	T cell activation	45	4.93E-22
	regulation of cell proliferation	128	1.83E-21
KEGG Pathways	Cytokine-cytokine receptor interaction	56	4.76E-22
	Hematopoietic cell lineage	29	1.50E-17
	Cell adhesion molecules (CAMs)	30	4.08E-13
	Primary immunodeficiency	16	8.70E-13
	Chemokine signaling pathway	31	1.14E-09
	Complement and coagulation cascades	17	1.36E-08
	T cell receptor signaling pathway	20	1.30E-07
	Allograft rejection	11	6.44E-07
	Natural killer cell mediated cytotoxicity	20	5.66E-06
	Pathways in cancer	34	1.49E-05
Wiki-Pathways	TCR Signaling Pathway	10	1.55E-09
	B Cell Receptor Signaling Pathway	10	1.72E-06
	Focal Adhesion	11	5.88E-05
	Complement Activation, Classical Pathway	6	8.38E-05
Chromosomal Location	11q23	18	1.84E-05
	Xq23	8	4.99E-05

Table 2.2: The most enriched functional categories among the 1000 genes most differentially expressed between LumA-R1 and LumA-R2 samples. All the genes on the list showed significantly higher expression on the LumA-R2 samples compared to LumA-R1 samples.

2.1.5. Analysis of DNA methylation identifies a luminal subgroup characterized by hyper-methylation and a significantly poorer outcome

The luminal-A tumors proved to be the most heterogeneous in our gene expression analysis. To further identify and characterize clinically meaningful subgroups within the luminal-A group, we explored breast tumor variability on the epigenetic level as well.

Using the Methylation 450K array dataset available from TCGA, we started our analysis as in the expression data, by clustering all 679 tumor samples into four groups, corresponding to the number of PAM50 classes. The resulting clusters (Figure 2.5A) show modest agreement with the expression based PAM50 classes; All basal-like samples were assigned to a single cluster exhibiting a distinct hypo-methylation pattern (cluster 4), whereas HER2-enriched samples were scattered over three different clusters, indicating that this subtype has reduced manifestation on the methylation level. Notably, most luminal samples were assigned to three different clusters (1-3) showing methylation level gradation on the top 2000 variable CpGs. Cluster 1 exhibited a strong hyper-methylation pattern, contained the highest ratio of luminal-B samples and was associated with significantly poorer survival compared to the three other clusters ($p=0.0001$). Cluster 3, on the other hand, exhibited opposite characteristics: lower methylation levels, the lowest ratio of luminal-B samples and a better outcome ($p=0.0129$).

Similar results were obtained when we clustered only the 513 luminal A and B samples (Figure 2.5B). Here we used the top 2000 variable genes within these samples, in order to remove the effect of the other two subtypes on the clustering. Importantly, out of the 127 samples comprising the hyper-methylated cluster 1, which was associated with reduced survival ($p=2.6e-05$), 76 samples were labeled as luminal-A, a subtype usually associated with good survival. In other words, approximately 20% of the 378 luminal-A samples (as called by the expression-based PAM50) included in the analysis, could actually be assigned to a higher risk group based on methylation data (See Supplementary Information, section S1.7 for more details).

The three-way partition by methylation levels and its association with differential survival risk also appeared when we repeated the analysis in the group of 378 luminal-A samples, using the top 2000 variable CpGs on these samples (Figure 2.5C). The three methylation-based luminal-A clusters were designated LumA-M1, LumA-M2 and LumA-M3. The 84-sample LumA-M1 cluster (composing ~22% of the luminal-A samples) was associated with significantly reduced five-year survival ($p=0.0031$).

Furthermore, the methylation-based partitioning of the luminal-A samples (LumA-M1/2/3) correlated significantly with the expression-based partitioning (LumA-R1/2, Chi-Square p = 4.4E-08). The LumA-M2 cluster was enriched for LumA-R1 samples (p = 1.4E-06) and the LumA-M3 cluster was enriched for LumA-R2 samples (p=1.6E-08), showing that the expression and the methylation-based patterns are related (See lower bar on Figure 2.5C).

Overall, we identified a poorer outcome subgroup within the luminal-A subtype, which is distinguished by robust hyper-methylation pattern.

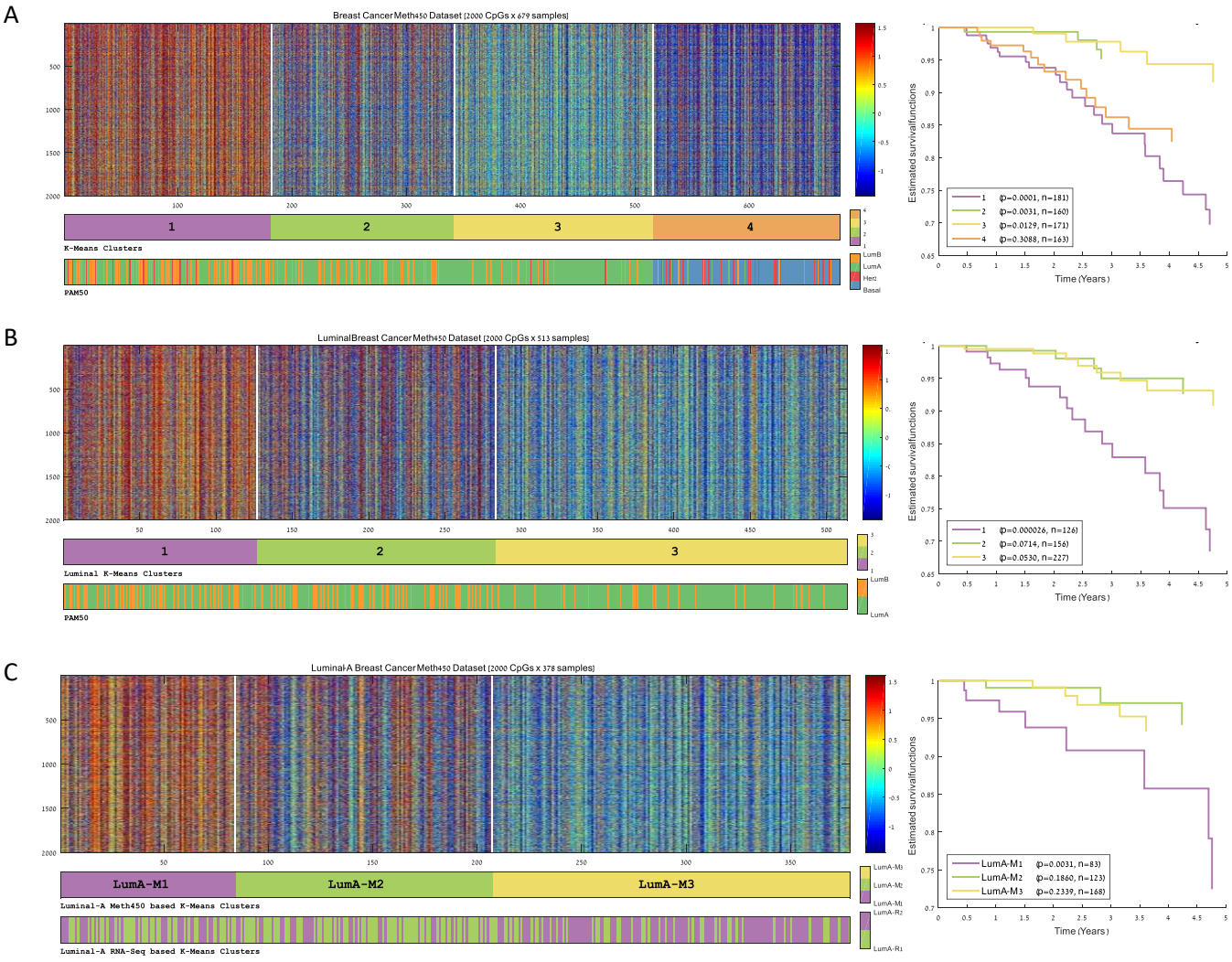


Figure 2.5: Unsupervised analysis of breast cancer tumors using DNA Methylation data. Samples were clustered by K-Means based on correlation using the top 2000 variable CpGs over each sample subset. **(A)** All 679 tumors **(B)** 579 samples identified as luminal-A and luminal-B by PAM50 classification, **(C)** 378 luminal A samples only. The first bar below each expression-matrix shows the assignment of the samples to methylation-based clusters. The second bar on A and B shows PAM50 calls for the samples. The second bar on C presents the RNA-Seq based LumA-R1/2 subgroups defined in section 3.3. The right panels show five-year Kaplan-Meier survival plots for the resulting groups.

2.1.6. Analysis of differentially methylated CpGs between the LumA-M1 and LumA-M3 subgroups and their correlation to gene expression

To uncover the biological features characterizing the distinct methylation patterns observed in the luminal-A subgroups, we examined the 1000 top DMCs (see "Methods") between the hyper-methylated LumA-M1 (n=84) and the hypo-methylated LumA-M3 (n=171). These two sample subgroups represent the two extremes of the methylation gradient observed on the luminal-A samples. Of note, all 1000 top DMCs (representing 483 genes) were hyper-methylated on the LumA-M1 samples compared to LumA-M3.

Gene enrichment analysis associated these 483 genes hyper-methylated on LumA-M1 samples with GO terms related to development, signaling, cell differentiation and transcription regulation ($p<1E-15$). The genes were also enriched for the "homeobox" InterPro term ($p=3.6E-35$), in line with previous reports describing the methylation of homeobox genes during breast tumorigenesis [151] [152] [153] . Further, the 483 genes were enriched for tumor suppressor genes according to the TSGene catalog [154] ($p=1.5E-03$), including 48 such genes. See column 1 in Table 2.3. Analysis for CpG features of the top 1000 DMCs showed significant enrichment for enhancer elements, tissue-specific promoters and cancer-specific DMRs (See column 1 in Table 2.4).

As DNA-methylation is known to regulate gene expression and as hyper-methylation of promoters was associated with gene silencing in cancer [155], we focused on LumA-M1 hyper-methylated CpGs that affect the expression of their corresponding genes. To this end, we used the RNA-Seq based expression data available from TCGA for the same 378 analyzed samples to generate a second list of CpGs that are both hyper-methylated on LumA-M1 samples (differential methylation $p<0.01$, median difference of 0.2) and whose methylation level is inversely correlated to the expression level of their corresponding gene (Spearman correlation $R < -0.2$). As can be seen in Table 2.4, the 586 CpGs that passed this filter (corresponding to 340 genes) showed significant over-representation of upstream parts of their corresponding genes (UCSC RefGene Group: TSS and 1^{st}-Exon $p<=4.4E-05$) and under-representation of gene body ($p=1.43E-16$) and 3'UTR ($p=5.83E-04$). In terms of the Regulatory Feature Group, these 586 CpGs showed over-representation of "Promoter Associated Cell type specific" elements ($p=1.40E-04$) accompanied by highly significant under-representation of "Promoter Associated" elements ($p=2.94E-31$), suggesting that the observed hyper-methylation pattern involves tissue specific promoters. Among the 340 under-expressed genes containing the 586 hyper-methylated CpGs

, there were several tumor suppressor genes whose under-expression was previously observed in breast cancer, such as L3MBTL4 [156], ID4 [157], RUNX3 [158][159], PROX1 [160], SFRP1 [161] and others. Gene and CpG level enrichments for the negative correlations are shown in column 2 of Tables 2.3 and 2.4 respectively.

Interestingly, the 212 LumA-M1 hyper-methylated CpGs that exhibited positive correlation to expression (Spearman R > 0.2) had higher enrichments of development-related GO terms compared with negatively correlated CpGs ("pattern specification process" p=1.07E-13, "embryonic morphogenesis" p=1.05E-10, "cell fate commitment" p=5.49E-10). In contrast to the negatively correlated CpGs, they showed high over-representation of "gene body" and under-representation of "TSS" regions (UCSC RefGene Group, p=9.48E-20 and p=7.28E-14 respectively). For gene and CpG level enrichments for the positive correlations see column 3 in Tables 2.3 and 2.4, respectively.

The differential methylation pattern distinguishing LumA-M1 from LumA-M3 samples could, therefore, be characterized by hundreds of CpGs that are hyper-methylated on the LumA-M1 samples. Distinct subsets of these CpGs show negative and positive correlation with the expression of developmental genes.

	(1) Hyper Meth. CpGs		(2) Neg: R < -0.2		(3) Pos: R > 0.2	
	1000 CpGs, 483 genes		586 CpGs, 340 Genes		212 CpGs, 125 Genes	
	Term	p-value	Term	p-value	Term	p-value
Gene ontology	anatomical structure development	6.1E-28	developmental process	7.8E-06	pattern specification process	1.1E-13
	developmental process	2.0E-25	single organism signaling	2.4E-05	regionalization	1.1E-12
	multicellular organismal process	9.6E-24	signaling	1.8E-05	anatomical structure development	2.2E-11
	single-multicellular organism process	1.6E-22	cellular developmental process	1.4E-05	single-organism developmental process	1.9E-11
	single organism signaling	1.7E-21	single-organism developmental process	2.3E-05	anatomical structure morphogenesis	1.8E-11
	signaling	1.9E-21	anatomical structure development	8.0E-05	developmental process	1.7E-11
	cell-cell signaling	1.7E-21	cell-cell signaling	1.8E-04	embryonic morphogenesis	1.1E-10
	neuron differentiation	1.2E-20	cell differentiation	2.2E-04	cellular developmental process	1.8E-10
	single-organism developmental process	1.4E-19	synaptic transmission	4.4E-04	organ development	5.3E-10
	regulation of transcription from RNA polymerase II promoter	1.2E-16	anatomical structure morphogenesis	6.1E-04	single-multicellular organism process	5.6E-10
INTERPRO	Homeobox	3.6E-35	Homeobox	1.1E-04	Homeobox	2.1E-31
Tumor Suppressor Genes (TSGene 2.0)	AHRR, AKR1B1, BMP2, C2orf40, CDH4, CDO1, CDX2, CNTNAP2, CSMD1, DLK1, DSC3, EBF3, EDNRB, FAT4, FOXA2, FOXC1, GALR1, GREM1, GRIN2A, ID4, IRF4, IRX1, LHX4, MAL, MIR124-2, MIR124-3, MIR125B1, MIR129-2, MIR137, MIR9-3, ONECUT1, OPCML, PAX5, PAX6, PCDH8, PHOX2A, PRKCB, PROX1, PTGDR, RASL10B, SFRP1, SFRP2, SHISA3, SLIT2, SOX7, TBX5, UNC5D, ZIC1	1.5E-03 (48 genes)	AKR1B1, ASCL1, BIN1, BMP4, CCDC67, CDK6, CDO1, EBF3, GSTP1, ID4, IRX1, L3MBTL4, LRRC4, MAP4K1, MME, NTRK3, PCDH10, PDLIM4, PROX1, PTGDR, RUNX3, SCGB3A1, SFRP1, SLC5A8, SLIT2, UBE2QL1, UNC5B, VIM, WT1	9.7E-02 (29 genes)	AMH, GATA4, HOPX, HOXB13, LHX4, LHX6, MAP4K1, ONECUT1, PAX5, RASAL1, TBX5, TP73, WT1, ZIC1	5.5E-02 (14 genes)

Table 2.3: Gene enrichments on the various subsets of differentially methylated CpGs between LumA-M1 and LumA-M3 subgroups. GO, INTERPRO and TSG 2.0 databases were used to test the hyper-methylated genes for enrichments. Group 1 is composed of the 1000 top DMCs with a mean difference of at least 0.2. All the CpGs on this list showed significant hyper-methylation on the LumA-M1 samples compared to LumA-M3 samples. Group 2 is composed of the 586 CpGs which a differential methylation p-value<0.01, methylation mean difference>0.2 and spearman based correlation with expression that is lower than 0.2. Group 3 is composed of 212 CpGs with a differential methylation p-value<0.01, methylation mean difference>0.2 and spearman based correlation with expression that is higher than 0.2.

		(1) Hyper Meth. CpGs		(2) Neg: R < -0.2		(3) Pos: R > 0.2	
Label	**Term**	Over-representation p-value	Under-representation p-value	Over-representation p-value	Under-representation p-value	Over-representation p-value	Under-representation p-value
UCSC RefGene Group	1stExon	**1.E-04**	1.E+00	**1.E-07**	1.E+00	1.E+00	3.E-02
	3'UTR	1.E+00	**2.E-03**	1.E+00	**6.E-04**	2.E-02	1.E+00
	5'UTR	1.E+00	8.E-01	3.E-01	1.E+00	1.E+00	2.E-02
	Body	1.E+00	**7.E-05**	1.E+00	**1.E-16**	**9.E-20**	1.E+00
	TSS	2.E-02	1.E+00	**4.E-05**	1.E+00	1.E+00	**7.E-14**
Regulatory Feature Group	Gene Associated	1.E+00	2.E-01	1.E+00	5.E-01	1.E+00	1.E+00
	Gene Associated Cell type specific	1.E+00	5.E-02	1.E+00	2.E-01	2.E-01	1.E+00
	NonGene Associated	1.E+00	3.E-01	1.E+00	1.E-01	1.E+00	8.E-01
	NonGene Associated Cell type specific	**3.E-03**	1.E+00	5.E-01	1.E+00	2.E-01	1.E+00
	Promoter Associated	1.E+00	**2.E-146**	1.E+00	**3.E-31**	1.E+00	**4.E-34**
	Promoter Associated Cell type specific	1.E+00	5.E-02	**1.E-04**	1.E+00	1.E+00	7.E-02
	Unclassified	1.E+00	4.E-01	**6.E-04**	1.E+00	1.E+00	1.E+00
	Unclassified Cell type specific	**9.E-35**	1.E+00	**4.E-06**	1.E+00	**1.E-10**	1.E+00
	Unassigned	**7.E-52**	1.E+00	**5.E-06**	1.E+00	**2.E-09**	1.E+00
Differentially Methylated Region (DMR)	CDMR (Cancer-DMR)	**2.E-16**	1.E+00	**4.E-03**	1.E+00	**1.E-13**	1.E+00
	DMR	**9.E-183**	1.E+00	**2.E-75**	1.E+00	**1.E-15**	1.E+00
	RDMR (Reprogramming-DMR)	**2.E-04**	1.E+00	2.E-01	1.E+00	**2.E-11**	1.E+00
	Unassigned	1.E+00	**2.E-205**	1.E+00	**2.E-75**	1.E+00	**5.E-40**
Enhancer		**1.E-09**	1.E+00	**8.E-06**	1.E+00	**2.E-04**	1.E+00
DHS (DNAse hypersensitive site)		**1.E-07**	1.E+00	**2.E-03**	1.E+00	**2.E-05**	1.E+00

Table 2.4: Feature enrichments on the various subsets of differentially methylated CpGs between LumA-M1 and LumA-M3 subgroups. CpG enrichment tests show that hyper-methylated CpGs exhibiting negative correlation to gene expression are enriched for upstream gene parts, while positively correlated CpGs are enriched for gene body. All three hyper-methylated CpG groups are enriched for informatically determined enhancer elements and experimentally determined differentially methylated regions and DNAse hypersensitive sites. The p-values represent hypergeometric based over or under-representation and are FDR corrected (significant p-values are marked in bold).

2.1.7. Cox Survival Analysis

In previous sections, we presented two different partitions of luminal-A tumors based on genomic profiles, with prognostic value: The LumA-R2 group (characterized by high expression of immune-related genes) was associated with a reduced chance of five-year recurrence, while the LumA-M1 group (characterized by hyper-methylation of CpGs located in developmental genes) was associated with poorer survival. To determine the prognostic contribution of the two partitions while adjusting for other relevant explanatory variables, we performed multivariate Cox survival analysis on both LumA-R and LumA-M partitions (see Table 2.5). Patients belonging to the LumA-M1 group exhibited 6.68 fold higher estimated five-year death hazard compared with the other groups in the COX multivariate model, after adjustment for age, pathological stage, ER status, PR status and HER2 status. Patients belonging to the LumA-R2 group had a decreased recurrence hazard of 0.06 (that is, 94% decrease) compared with LumA-R1 patients, after similar adjustment. The results reaffirm the independent prognostic value of the LumA-R2 and the LumA-M1 classes (see Supplementary Information, section S1.10 for univariate analysis).

	Survival		Recurrence	
Variable	**HR**	**p-value**	**HR**	**p-value**
LumA-R (1 vs 2)	0.56	0.36991	**0.06**	**0.00693**
LumA-M (2,3 vs 1)	**6.68**	**0.00484**	3.04	0.07028
Age (<60 vs.>=60 years)	**11.20**	**0.0037**	1.03	0.96530
Pathologic stage (I,II vs. III,IV)	2.12	0.25519	1.93	0.26992
ER Status	7.17	0.18095	0.00	0.99575
PR Status	0.47	0.50039	0.29	0.29092
Her2 Status	1.48	0.72659	0.64	0.68789

Table 2.5: Multivariate Cox analysis of luminal-A subgroups for five-year survival and five-year recurrence. Significant p values are marked in boldface. ER estrogen receptor, PR progesterone receptor, Her2 human epidermal growth factor receptor 2.

2.2. Methods

2.2.1. Data acquisition and preprocessing

TCGA data on invasive carcinoma of the breast were downloaded from UCSC Cancer Browser web site [162] together with accompanying clinical information. The downloaded RNA-Seq gene expression dataset (Illumina HiSeq platform, gene level RSEM-normalized [163], log2 transformed) included 1215 samples of which 11 male, 8 metastatic and 30 unknown tissue source samples were filtered out. PAM50 calls (obtained directly from UNC, including PAM50 proliferation scores) were available for 1148 of the filtered samples, and distributed as follows: 183 basal-like, 78 Her2, 534 luminal-A, 203 luminal-B and 150 normal-like.

We also downloaded DNA methylation profiles (Illumina Infinium Human Methylation 450K platform, beta values [31]) containing 872 samples of which 8 male, 5 metastatic and 19 unknown tissue source samples were filtered out. We used only 679 tumor samples for which PAM50 calls were available, including 124 basal-like, 42 Her2, 378 luminal-A, and 135 luminal-B samples. Our analysis used only the 107,639 probes of the Infinium-I design type for which a gene symbol was available. This allowed us to bypass the bias of the two probe designs included on the array, to focus on differentially methylated sites that are associated with known genes and also to reduce the number of analyzed features.

2.2.2. Unsupervised analysis of the tumor samples

Unsupervised analysis of the various sample subsets was executed by clustering the samples based on the 2000 features (genes or CpGs) showing the highest variability over the samples included in each analysis. Clustering was performed using correlation distance. We used the k-means clustering algorithm implementation in Matlab (release 2015a). This implementation uses the k-means++ algorithm by David and Vassilvitskii [164], which improves the initialization of the cluster seeds. To improve the quality of the resulting clustering solution, we generated 100 clustering replicates and selected the replicate that minimized the sum of point-to-centroid distances as the final clustering solution. Due to the high variability among sample subgroups in the breast cancer datasets, reselecting the top variable genes for the analysis of each sample set (and renormalizing accordingly) is crucial to ensure the use of the features most relevant to that set. Each feature was independently centered and normalized over the analyzed samples prior to clustering.

Cohort descriptions for the samples used in each analysis appear in the Supplementary Information (Tables S1.1C, S1.2A, S1.3A for the RNA-Seq analyses and Tables S1.6B, S1.7A and S1.8A for the DNA methylation analysis).

2.2.3. Sample cluster enrichment and survival analysis

To evaluate the clinical relevance of the sample clusters obtained in each unsupervised analysis, we used the extensive clinical information available from TCGA for each sample. Enrichment significance of sample clusters for categorical variables (such as PAM50 subtype or histological type) was calculated using false discovery rate (FDR) corrected hypergeometric test. For numeric variables (such as age, tumor nuclei percent and others) difference between sample groups was evaluated using the Wilcoxon rank-sum test (Mann–Whitney U test).

Survival and recurrence-free survival curves were plotted using the Kaplan-Meier estimator [131] and p-values for the difference in survival for each group versus all other groups were calculated using the log-rank (Mantel-Haenzel) test [133][134]. Cox univariate and multivariate survival analyses were conducted using Matlab implementation; p-values were corrected using FDR. The analysis and visualization scripts are publicly available as an interactive graphical tool named PROMO [165][123] (thoroughly presented in Results, section 4).

2.2.4. Analysis of differentially expressed genes and gene enrichment

A list of genes that have the highest differential expression between the two RNA-Seq-based sample groups LumA-R1 and LumA-R2 was generated by applying the Wilcoxon rank-sum test on all dataset genes exhibiting non-zero variance (n=19913) after flooring all dataset values to 1 and ceiling to 14. We selected the 1000 genes exhibiting the most significant p-value that also have a median difference of at least 0.5 (log2 transformed RSEM expression values). All genes on the list showed significantly higher expression on the LumA-R2 sample group (lowest p-Value was 8.1e-28).

Gene enrichment tests were performed on these 1000 genes against a background all genes included in the rank-sum test. The Expander software suite [121]-[122] was used to detect significant enrichments for Gene Ontology (GO) [118], Kyoto encyclopedia of genes and genomes (KEGG) pathways [119], Wiki-Pathways [120] and chromosomal location enrichments. GO tests were also performed using the GOrilla tool [124].

2.2.5. Analysis of differentially methylated CpGs, correlation to expression and CpG enrichment

To identify CpGs that are differentially methylated between LumA-M1 and LumA-M3 samples we applied the rank-sum test on all CpGs that survived our preprocessing and also had non-zero variability on the relevant samples (n=93,880). We then selected the 1000 CpGs showing the highest significance and having a minimal median difference of 0.2 (in Beta-values). All selected CpGs had significantly higher mean methylation on group LumA-M1 compared to the LumA-M3 group.

To focus on DMCs whose genes show concomitant expression changes, we calculated for each CpG its Spearman correlation with the expression profile of its associated gene based on Illumina's probe-set annotation. The correlation values enabled the identification of 586 DMCs (rank-sum p-value<0.01, median difference>0.2) negatively correlated to expression ($R < -0.2$) and a second smaller group of 212 DMCs showing positive correlation ($R > 0.2$) to expression.

We used the array CpG annotations provided by Illumina to calculate enrichments of each one of the three CpG lists (top 1000 DMCs, 586 negatively correlated DMCs, and 212 positively correlated DMCs) for features like differentially methylated regions (DMRs), Enhancer regions, UCSC RefGene Groups and Regulatory Feature Groups. Gene enrichment analysis was performed on the unique genes composing each CpG list, using the Expander and Gorilla tools as described above. Enrichment for InterPro [166] terms was calculated using David [167]. Enrichment for tumor suppressor genes was calculated by the hypergeometric test based on the TSGene [154] catalog.

3. Skin cancer subtypes

In this study, we set out to explore whether the transcription-based subtype classification can be improved based on the larger number of 469 melanoma samples currently available from TCGA. We reasoned that larger datasets might allow for the identification of new prognostic subtypes or improve the characterization of previously described subtypes. We also aimed at identifying a minimal set of informative prognostic biomarkers that can be used to stratify patients into clinically relevant subtypes. Finally, we performed a set of experimental tests on human melanoma specimens to validate our computational discoveries.

3.1. Results

3.1.1. Unsupervised analysis identifies four distinct melanoma subgroups

In order to identify groups of similar melanoma tumors, we applied unsupervised analysis on 469 RNA-Seq expression profiles obtained from TCGA's melanoma dataset. The dataset contained a mixture of primary (n=104) and metastasis samples (n=365). The clustering of the samples based on the 2000 most variable genes resulted in four distinct sample clusters showing significantly different 5-year survival rates (see Figure 3.1A, 3.1B, and Table S2.1). Gene ontology enrichment analysis identified active gene signatures that were used to characterize each sample group (Figure S2.1). Finally, we used the clinical information available for the samples in order to clinically characterize each sample group (see Figure 3.1C and Figure S2.2).

Cluster 2, with the lowest survival rate, was mainly composed of primary melanomas showing significantly high Breslow depths and high pathologic T values. This cluster was associated with over-expression of cornification, epidermis development, and keratin related genes, all of which are characteristic of differentiated keratinocytes that form the outermost skin barrier[168]. We attributed the poor survival in this cluster to the bias in the TCGA cohort for thick primary tumors [48]. The other three clusters were mainly composed of metastatic melanomas. Cluster 1, which conferred the highest survival rate, was enriched for lymph node metastases and showed significantly high values for several immune scores that correlate with lymphocyte infiltration. Cluster 1 was also associated with the overexpression of adaptive immune response genes. Cluster 3 showed relatively good survival and was found to be marginally enriched for regional cutaneous tissue sites, whereas cluster 4 showed relatively poor survival and was found to be marginally enriched for metastasis to distant tissue sites. Interestingly, what distinguished the

relatively poor prognosis cluster 4 from the relatively good prognosis cluster 3 was an expression pattern enriched for melanin biosynthesis genes (gene cluster 1) whose over-expression was correlated with poor survival.

We compared our four-cluster partition to TCGA's three transcriptomic subtypes (Figures 3.1A3 and S2.3). We found that the two partitions largely correspond (Chi-Square p-value=1.6e-79) - sample clusters 1 and 3 were significantly enriched for TCGA's Immune and MITF-Low transcriptomic subtypes, respectively. TCGA's keratin subtype was split into two distinct clusters – the primary-enriched worst outcome cluster 2 and the bad outcome metastasis-enriched cluster 4. Overall, our analysis revealed a partition of the metastatic samples into the high-immune, best survival (cluster 1), low-melanogenesis good survival (cluster 3, corresponding to TCGA's MITF-low subtypes), and a new metastasis enriched subgroup, characterized by poor survival and by significant overexpression of melanogenesis genes (cluster 4). We named the four identified melanoma subgroups accordingly: 1: "Immune", 2: "Keratin", 3: "Melanogenesis-low" and 4: "Melanogenesis-high". Table 3.1 summarizes the characteristics of the four subgroups and their relation to TCGA's subgroups.

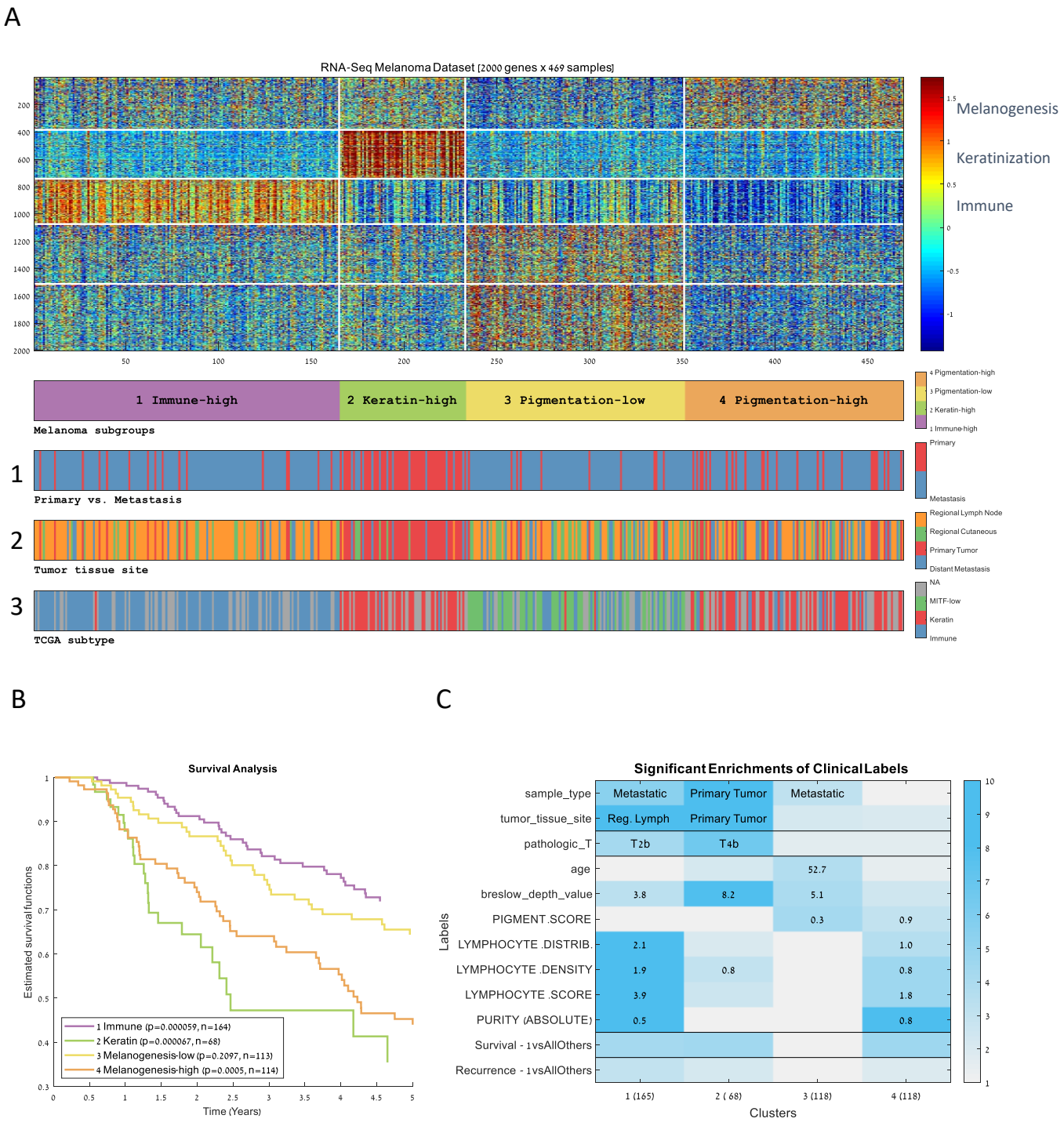


Figure 3.1: Clustering of TCGA's RNA-Seq melanoma dataset. (A) A heat map representing the clustering of 469 melanoma samples (matrix columns) into four groups based on the 2000 genes with the most variable expression profiles (matrix rows). Each sample cluster represents a group of similar melanoma tumors. Genes were also clustered in order to identify groups of co-expressing genes. Both samples and genes were clustered using the K-means algorithm (using k=4 for the samples and k=5 for the genes). The bars below the matrix display sample labels: (1) Cluster ID, (2) Primary vs. Metastasis, (3) Tissue site, (4) TCGA transcriptomic subtype. **(B)** Kaplan Meier curves for the four sample clusters. Log-rank p-values appear in the legend. **(C)** Summary of the significant enrichments of sample clusters for clinical labels. Colors indicate the significance of the enrichment.

Cluster	Cluster name	TCGA Transcriptomic subtype enrichment	Survival	Tumor tissue type enrichment	Gene ontology enrichment of highly expressed genes
1	**Immune**	Immune	Best	Regional lymph node	Immune response
2	**Keratin**	Keratin	Worst	Primary	Cornification
3	**Melanogenesis-low**	MITF-Low	Good		Nervous system development
4	**Melanogenesis-high**	Keratin	Bad		Melanin biosynthetic process

Table 3.1: Summary of the main sample cluster characteristics

3.1.2. Over-expression of melanogenesis genes characterizes a poor-survival melanoma subtype

In order to further characterize the poor-survival cluster 4 ("Melanogenesis-high"), we looked at the genes that were over-expressed in this cluster (gene cluster 1). They were enriched for genes related to the synthesis of the melanin pigment ("Melanin biosynthetic process", p-value<1.76E-08, See Figure S2.1). Additional gene sets significantly enriched in that gene cluster were the *"Melanogenesis"* KEGG-pathway (p-value<0.005, 9 genes: GNAO1, DCT, KIT, TYRP1, FZD9, ADCY2, ADCY1, TYR, WNT4) and *the "Melanosome membrane"* GO term (p-value<0.0004, 6 genes: OCA2, SLC45A2, GPR143, DCT, TYRP1, TYR). See Tables S2.2 and S2.3 for the complete enrichment results.

Interestingly, these results suggest that the "Melanogenesis-high" samples differ from the "Melanogenesis-low" samples by over-expression of genes that are specific to the melanosome organelle (see Figure S2.4). The melanosome organelle is the hallmark of melanocytes, which are the melanoma cell of origin [169]. In normal skin, melanosomes are responsible for melanin production, storage, and transport from melanocytes to surrounding keratinocytes [170] [171]. However, the reason melanoma cells retain this function of their cell of origin, and the function of the melanosome itself in melanoma cells, have only recently begun to be revealed [172][173]. The melanin biosynthesis genes OCA2, TYRP1, DCT, and PMEL (SILV) also appeared on the list of top genes over-expressed in "Melanogenesis-high" samples in comparison to all other samples (see Table S2.4).

For testing the independent prognostic value of those melanosome related genes, we partitioned all of the dataset samples into two groups based on the expression levels of each gene and calculated the difference in the survival plots of the two groups using the log-rank p-value. For OCA2, KIT, GPR143, and TYRP1, overexpression was significantly correlated with

poorer 10-year survival as well as with increased recurrence risk (Figure 3.2). These results may suggest a mechanistic link between the melanosome organelle and the increased lethality of melanoma.

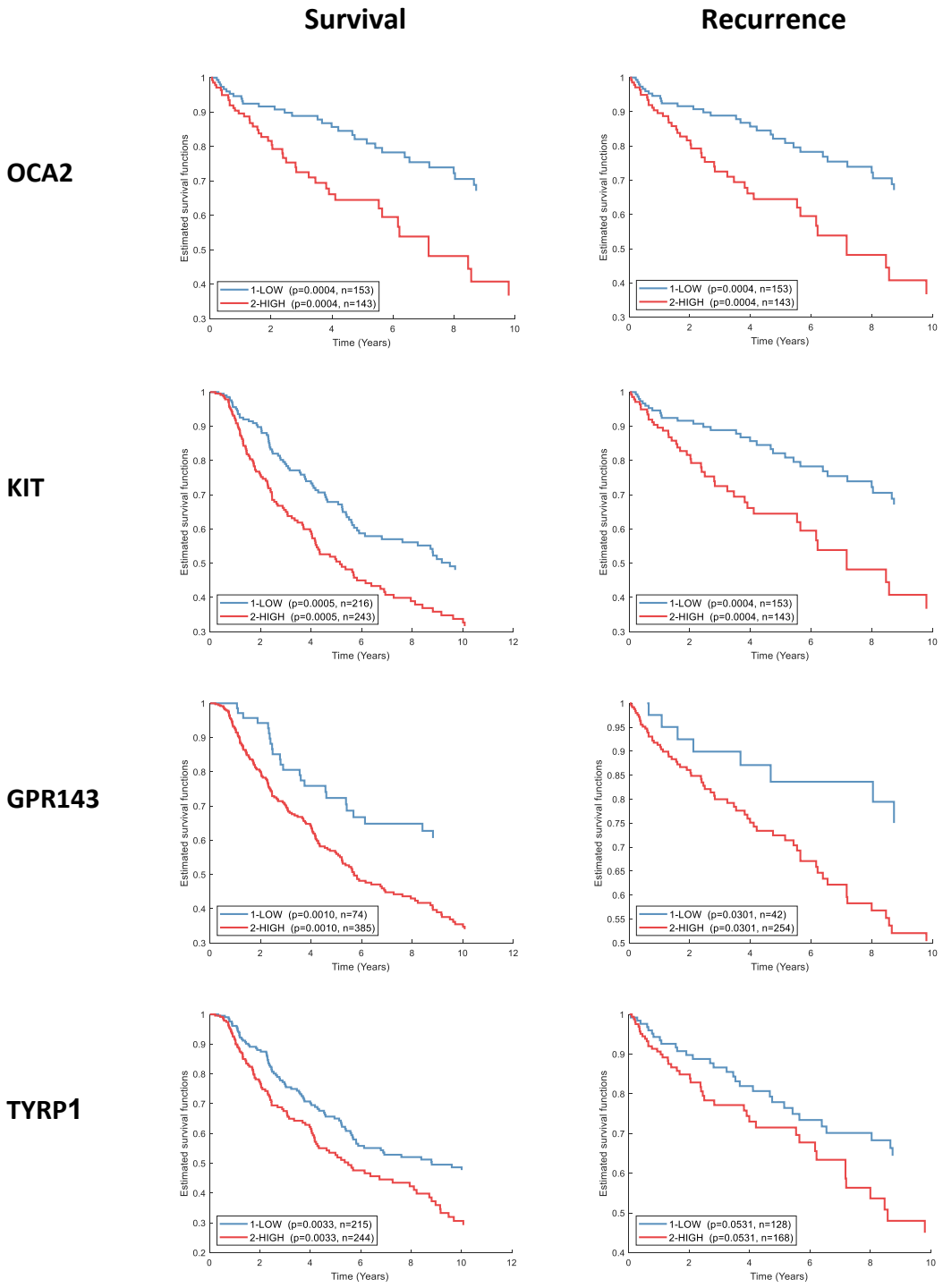


Figure 3.2: 10-year survival and recurrence risk estimates for Melanosome related genes, calculated over all dataset samples. Dataset samples were split into two groups based on the gene expression levels of several melanosome related genes. For each gene, the threshold for splitting the samples into two groups was the mean of its 5th and 95th expression percentile.

3.1.3. Metastatic melanomas retain the ability to secrete melanosomes into surrounding tissue

In primary melanoma, melanosome secretion was shown to promote the formation of the dermal metastatic niche [172]. However, the role of melanosomes in promoting metastasis of melanoma in later stages is mostly unknown. To further explore the pigmentation/melanosome function in melanoma progression, we tested clinical melanoma specimens. Since our unsupervised analysis that identified the four-melanoma subgroups was based on mRNA expression levels, we first confirmed the expression of melanogenesis genes at the protein level. Primary in-situ melanoma tissues were immunostained for PMEL (SILV) using the HMB45 antibody. PMEL is a melanocyte-specific marker known to be a melanogenesis gene and is used in the pathological diagnosis of melanoma [174][175]. PMEL is involved in the initiation of pre-melanosome production [176], and was also found in our analysis to be overexpressed in cluster 4 (see Table S2.4). PMEL strongly stained regions of melanoma (Figure 3.3A), confirming its presence at the protein level. To further confirm whether the complete melanogenesis machinery is functional, indicated by the production of mature melanosomes, specimens were immunostained with mature melanosome marker, GPNMB [172]. Primary melanoma and the surrounding tissue clearly stained with GPNMB (Figure 3.3A Left). This indicates that not only is the melanogenesis machinery active but also that melanosomes are actively secreted from melanoma into the stroma via a gradient pattern of diffusion from the epidermis (Figure 3.3A Right).

Since our computational analysis showed that the machinery of melanin production in melanosomes highly correlated with poor prognosis, we further examined melanosome synthesis and function along a typical scheme of disease progression. In order to do this, we picked melanoma metastasis specimens in the lymph nodes, liver, and brain, all from different patients. These tissues represent different stages of aggression [177]. Metastatic specimens were subjected to immunohistochemistry for PMEL and GPNMB in order to follow melanosome production and distribution. Remarkably, metastases to the lymph, liver, and brain retained a hallmark pattern of melanosome secretion into the surrounding stroma (Figure 3.3B). This indicates that melanosome production is retained throughout the progression of melanoma and that melanosomes are actively secreted to the tumor microenvironment. Taken together, our data demonstrate, for the first time, the presence of active production and secretion of melanosomes in distant metastatic sites, suggesting an important function for the melanosome organelle in the cancer metastases.

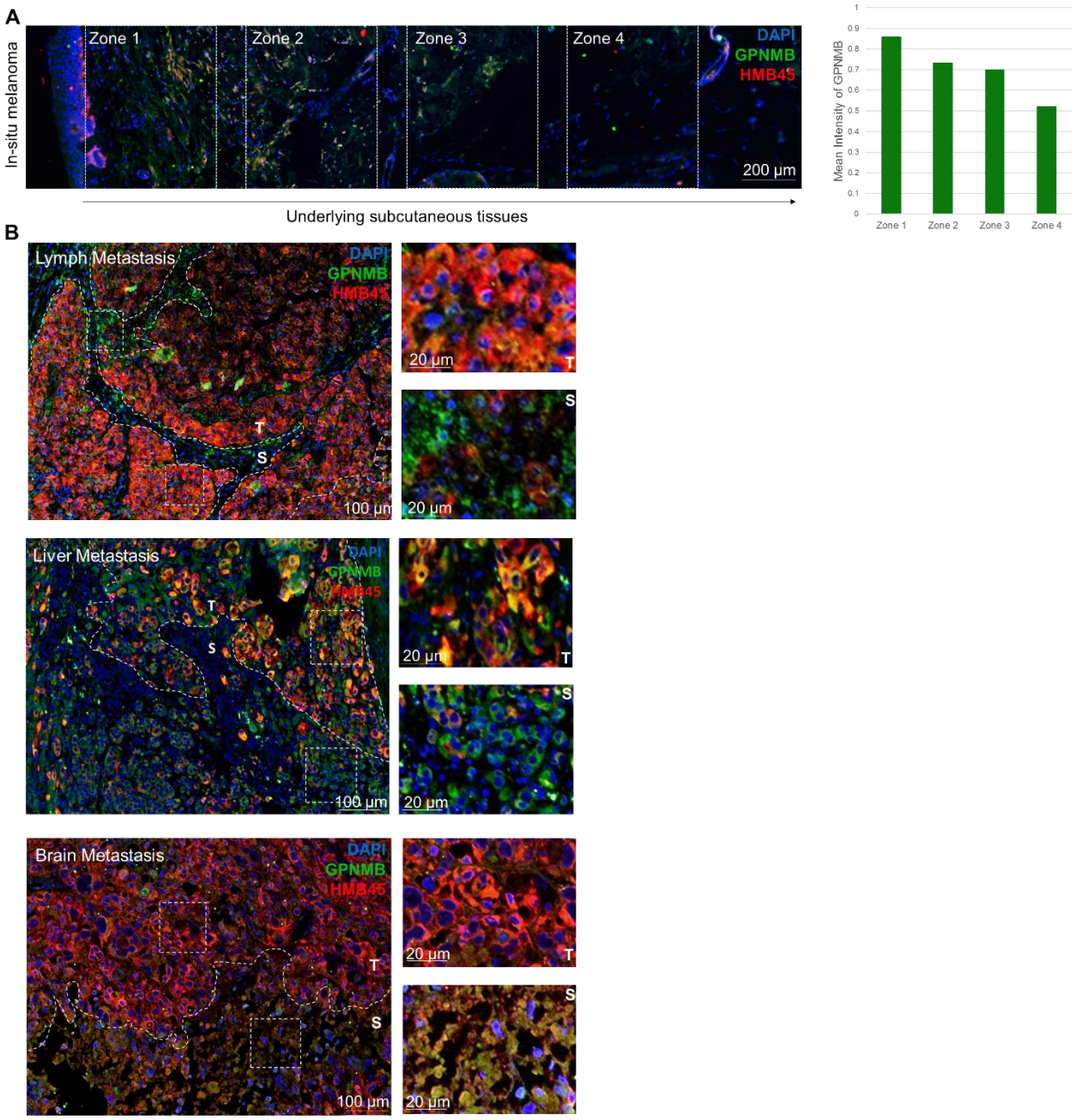


Figure 3.3: Melanosomes diffuse outward from primary and metastatic tissues. (A) Immunohistochemical (IHC) analysis of an in-situ melanoma showing mature melanosomes stained with Anti-GPNMB (Green) diffusing rightwards into the underlying subcutaneous tissues and away from the primary melanoma tumor. HMB45 (Red), an antibody for PMEL, which stains the premelanosome, shows the location of the melanoma. Nuclei were stained blue with DAPI. Equally sized, equidistant zones were delineated on the image in order to quantify differences in the intensity of GPNMB displayed by the graph to the right of the image. **(B)** IHC investigation of the metastatic sites: lymph node (top), liver (middle), and brain (bottom) showing secretion and dispersion of mature melanosomes stained with GPNMB (Green) into the stroma surrounding the tumor, stained with HMB45 (Red). Nuclei were stained blue with DAPI. Experimental validations were performed by members of Carmit Levi's lab.

3.1.4. A 3-gene classifier for predicting melanoma molecular subtype

Having identified four distinct melanoma subgroups, each bearing a different survival risk and gene expression signature, we sought to develop a simple procedure to classify a new tumor into one of the four subgroups based on a minimal number of genes. Such a procedure will be easier to interpret biologically than a 2000-gene signature and also cheaper to assay in diagnostics. We selected the decision tree classifier, which was often used in medical decision making due to its simplicity, easy interpretability, and robustness to outlier values [178]. In order to determine the number of genes to be used by our classifier, we trained a large number of variably pruned random decision trees and examined their performance as a function of the number of genes used by the tree (see Figure S2.5). Three genes gave a good tradeoff between classifier simplicity and performance. We then trained a 3-gene decision tree on the full dataset, which achieved a training error of 0.187 (Figure 3.4). Notably, the three genes selected by the tree-training algorithm, KLK8, TIGIT and TRIM63, can be viewed as representatives of the three gene expression signatures described earlier (Keratin, Immune, and Melanogenesis, respectively). The three selected genes, their corresponding thresholds and a set of 10 surrogates for each one are displayed in Table 3.2.

Remarkably, the genes identified as predictors by the decision tree have been previously associated with melanoma progression and prognosis: decrease in expression levels of kallikrein family member KLK8 was associated with the transfer from primary to metastatic melanoma [179], and its expression was linked to survival in various cancers [180][181][182]. TIGIT is a T cell immunoreceptor with Ig and ITIM domains, which was recently identified as an attractive cancer immunotherapy target due to its central role in tumor immunosurveillance [183][184]. Lastly, TRIM63 was implicated in melanoma cell migration/invasion[185]. Figure S2.6 provides a PCA visualization of the 469 melanoma samples projected to a 3-dimensional space based on the expression levels of the three genes used in the decision tree.

Interestingly, when we trained 1000 random 3-gene decision trees by resampling the dataset samples, most resulting trees had a similar configuration and contained predictors that are representatives of the three signatures (see supplemental information, section 7.2).

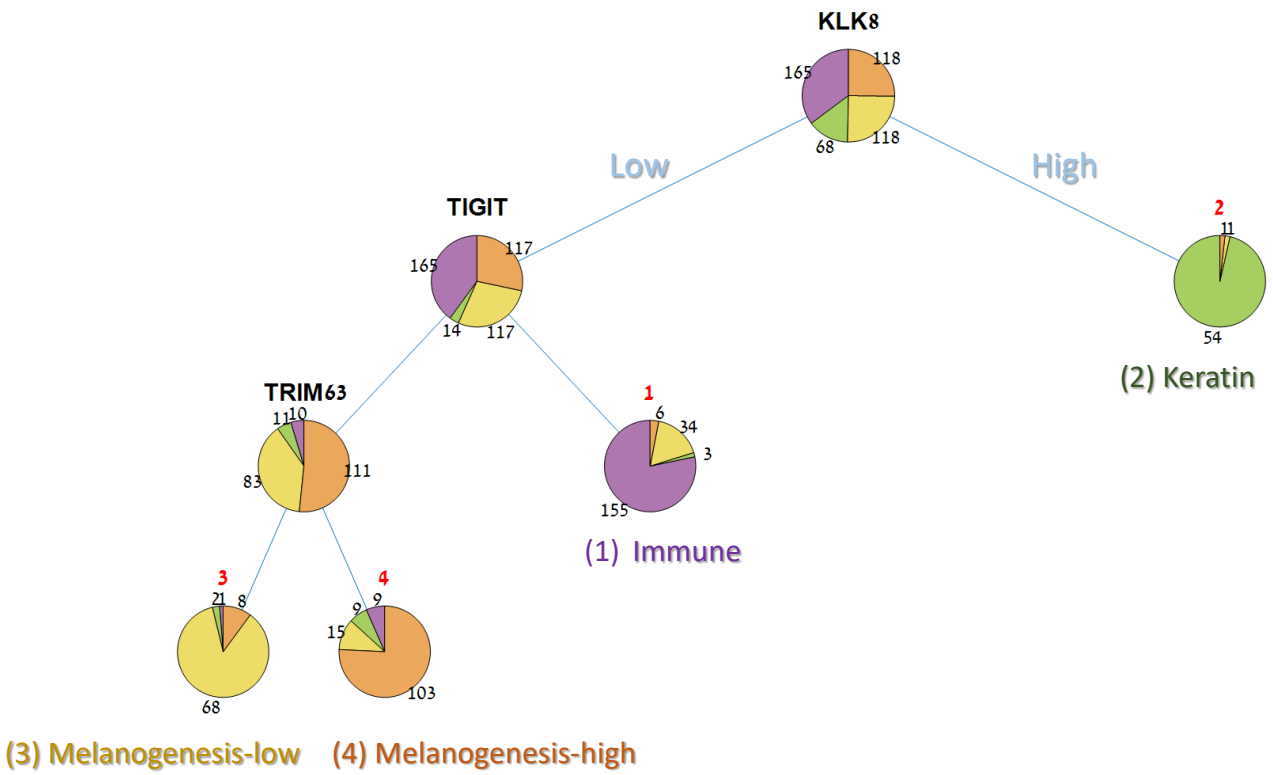


Figure 3.4: A 3-gene decision tree for classifying melanoma samples. The tree trained on the 469 TCGA samples. Classification of a new sample into one of the four subtypes is done by traversing the tree from its root to one of its leaves (representing an assignment to a subtype). Three biomarkers are used to determine the route along the tree: Over-expression of KLK8 distinguishes the "Keratin" subtype, over-expression of TIGIT distinguishes the "Immune" subtype, and finally, over-expression of TRIM63 distinguishes the "Melanogenesis-high" from the "Melanogenesis-low" subtype.

Predictor gene	*Threshold*	*Surrogate genes*
KLK8	***1.179***	***KRTDAP, FAM83C, IVL, SBSN, SPRR1B, KRT14, KRT16, WFDC5, KRT6C, SERPINB5***
TIGIT	***0.226***	***CD2, SLAMF6, LCK, SIRPG, SLA2, UBASH3A, CD3D, CD27, ITGAL, SIT1***
TRIM63	***0.155***	***TRPM1, PMEL , SLC5A10, GPR143, TSPAN10, MLPH, MLANA, MMP16, SLC45A2, GMPR***

Table 3.2: Threshold values and surrogate genes for the three decision tree predictors as identified by the algorithm. Threshold values are used to distinguish between high and low values (based on normalized expression values) during the classification procedure. For each predictor gene, 10 surrogate genes were identified by the tree-training algorithm and are displayed on the rightmost table column. The surrogate genes can be used instead of the predictor gene, with an appropriately adjusted threshold. The surrogate genes are sorted by decreasing predicted performance. The predictor gene represents the best predictor identified by the training algorithm.

3.1.5. Experimental validation of predictor genes on patient cohort

The decision tree produced consists of three informative genes (KLK8, TIGIT, and TRIM63) along with a threshold level for each gene that together provide a simple method for classifying melanoma tumors into one of the four subgroups. To classify a new tumor sample, one evaluates the sample's expression levels for three predictor genes (biomarkers): first, a keratin predictor-gene is evaluated (KLK8, or one of its keratinization surrogates such as KRT6C, IVL, SPRR1B, KRT14, KRT16), where high values would label the sample as "Keratin" and low values would lead to the next predictor. Next, an immune predictor is evaluated (TIGIT, or one of its immune surrogates such as LCK, CD2, SLAMF6, SIRPG, SLA2, UBASH3A, CD3D, CD27, ITGAL, SIT1), where high values would label the sample as "Immune" and low levels would lead to the next and final predictor. Lastly, a melanogenesis predictor is evaluated (TRIM63, or one of its melanogenesis surrogates such as SLC45A2, PMEL, GPR143) where high values would label the sample as "Melanogenesis-high" and low values as "Melanogenesis-low".

In order to validate the association between the classifier's predictor genes and outcome, we experimentally tested their expression on six samples from patients of known outcomes. Patients who survived for five years or more after initial tumor diagnosis were defined as “good survival”, and those who survived two years or less after initial diagnosis as “poor survival” (Table S2.5). In all clusters except for cluster 2 (the "Keratin" subgroup, which mostly corresponded to primary sites), lymph node tissues were identified in a substantial fraction of the samples (Fig

S2.3C). For this reason, we tested the tree predictor genes in melanoma metastases to the lymph nodes from each patient using immunohistochemistry (IHC).

We first conducted H&E staining to confirm that the metastasis was, in fact, in the lymph nodes (Figure 3.5A), and then stained each sample by the three predictor genes. Figure 3.5B shows the six tissue images per gene and Figure 3.5D shows quantification of staining levels. All lymph node specimens stained negatively for the KLK8 gene, the predictor for the primary-melanoma enriched "Keratin" subgroup in the tree (Figure 3.5B, first row), indicating that the six samples do not belong to that subgroup. Staining for the TIGIT gene, the predictor for the "Immune" subgroup, appeared positive in the lymph node specimens of patients 1 and 3, thus assigning them to the best prognosis “Immune” subgroup based on the decision tree logic, in agreement with their good survival (Figure 3.5B, second row). The specimens from patients 2, 5 and 6 stained negatively for TIGIT, excluding them from the “Immune” subgroup. The specimen from patient 4 showed borderline positive staining, making it difficult to classify. Finally, using TRIM63, the predictor for the "Melanogenesis-high" subgroup, specimens 5 and 6 were stained positively and were therefore assigned to the “Melanogenesis-high” subgroup, while specimen 3 that was stained negatively and therefore assigned to the “Melanogenesis-low” subgroup (Figure 3.5B, third row). Except for patient 4, all patients were assigned to subgroups conferring relative survival in agreement with their known outcome. The results demonstrate the utility of biomarkers in prognostication of melanoma.

In addition to verifying the expression levels of proteins identified by the decision tree and their correlation to survival, we examined three other proteins that were identified as informative predictors for general prognosis (Figure 3.5C). As an additional representative from the immune protein category we selected LCK, a Src family tyrosine kinase found on lymphocytes, that was previously identified as a biomarker for good prognosis in melanoma [48]. Indeed, patients with high LCK expression had a better prognosis. As additional representatives for the melanogenesis category, we selected GPNMB, indicative of mature melanosome presence [186], and OCA2, a transporter protein associated with melanocytes involved in melanin production and pH regulation of the melanosome [187]. Patients who had high levels of these proteins in their lymph nodes had worse outcomes associated with the “Melanogenesis-high” subgroup.

Our data demonstrate that using the expression levels of only three classifier genes (keratin, immune, and melanogenesis) in our decision tree, we can reasonably predict the patient outcome using a lymph node biopsy. Our data further suggest the involvement of melanogenesis genes and the melanosome organelle in melanoma progression and lethality.

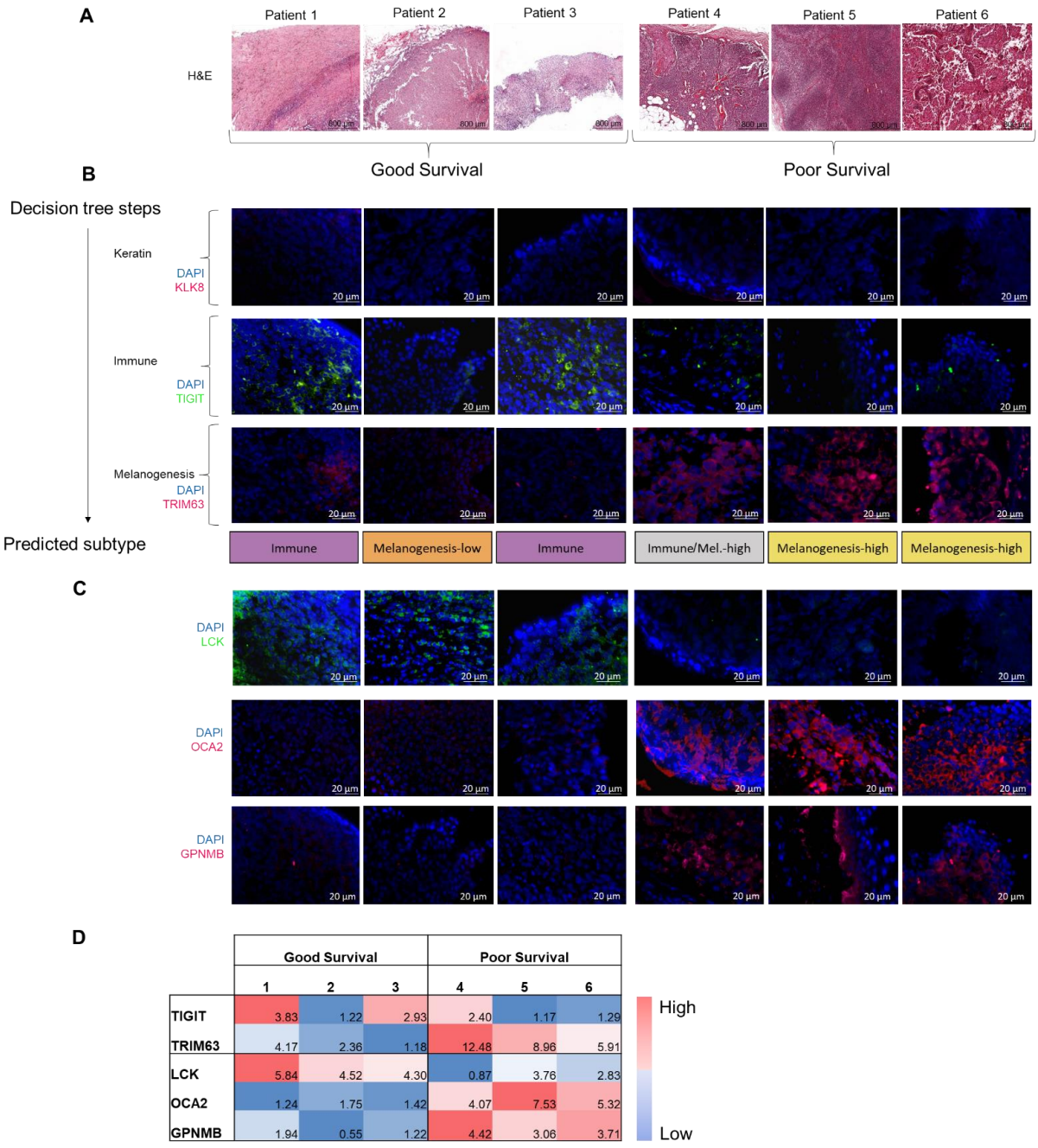


Figure 3.5: Melanogenesis and immune characteristics of melanoma metastases in good and poor prognostic outcomes. (A) Hematoxylin and Eosin (H&E) staining of lymph nodes containing melanoma metastases from six different patients taken at 20x magnification. Patients 1-3 had a good prognosis, while patients 4-6 had poor prognostic outcomes. **(B)** Immunohistochemical staining of the three proteins of the decision tree on the lymph node samples of the six patients. Nuclei were stained blue using DAPI. Row 1: Using KLK8 (pink) to validate non-primary tumor tissue. Row 2: Using TIGIT (Green) to test for immune proteins. Row 3: The expression of melanogenesis related protein TRIM63 (Pink). The assignments of the specimens from the six patients to subtypes based on the expression levels of the three predictor genes are summarized as a label at the bottom bar. **(C)** Immunohistochemical staining of additional biomarkers for general prognosis. Row 1: LCK, an immune protein indicative of good prognostic outcome. Row 2: Melanogenesis protein OCA2. Row 3: Melanogenesis protein GPNMB. **(D)** Color matrix quantifying the fluorescence intensity of immunohistochemistry across biomarkers and patients. For each protein, values were independently normalized across the samples. Experimental validations were performed by members of Carmit Levi's lab.

3.2. Methods

3.2.1. Gene expression analysis for identification of melanoma subtypes

The expression profiles of 474 samples from TCGA's melanoma RNA-Seq dataset [48] were downloaded from UCSC XENA's web site in April 2018 [188], together with their associated clinical information (213 labels). We used the PROMO software suite (release 2019.5) [123][165] for importing, preprocessing, analyzing, and visualizing the data. The downloaded RNA-Seq dataset (Illumina HiSeq platform, gene-level RSEM-normalized [27], log2 transformed) included 104 primary and 365 metastasis samples. Five samples were removed since they had inconsistent phenotype labels, and a variability-based filter was used to keep only the 2000 top variable genes. Clustering was performed on both samples and genes using the k-means algorithm with a correlation distance metric (using k=4 for the samples, and k=5 for the genes). The algorithm was run 100 times and a solution minimizing the sum of point-to-centroid distances was chosen.

We used PROMO's multi-label analysis to evaluate the enrichment of the sample-clusters for each of the clinical labels. Enrichment significance of sample-clusters for categorical variables (such as sample type) was calculated using FDR-corrected [93] hypergeometric test. For numeric variables (such as age, Breslow's depth, and pigmentation score), the difference between sample groups was evaluated using FDR-corrected Wilcoxon rank-sum test (Mann–Whitney U test). For exploring the prognostic value of the four sample-clusters based on TCGA's survival data, we used PROMO to plot 5-year survival curves using the Kaplan-Meier estimator [131], and calculated p-values for the difference in survival for each group versus all other groups using the log-rank (Mantel-Haenszel) test [133][134].

To identify active gene functions characterizing each of the sample-clusters, we applied Gene Ontology (GO) enrichment analysis [118] on the five gene-clusters using both PROMO (Figure S2.1) and the Expander software suite [121]-[122] (Table S2.2). To further characterize the biological function of the gene-clusters, we also used Expander to test each gene-cluster for enrichment for KEGG pathways [119] (Table S2.3).

Finally, to identify genes that were over-expressed on sample-cluster 4 compared to all other samples, we applied the Wilcoxon rank-sum test on all dataset genes exhibiting non-zero variance (n=20,228), and ranked all genes that were over-expressed on cluster 4 and showed p-value<1e-06 by decreasing fold-change (difference between the mean expressed in cluster 4

samples and all other samples, Table S2.4). We used the GORILLA tool[124] for identifying the melanin biosynthesis genes appearing among the top 100 differentially expressed genes.

3.2.2. Human histopathology and analysis of slides

Samples were obtained from patients at the E. Wolfson Medical Center and Tel Aviv Medical Center. The experimental study of the clinical samples was approved by the hospital ethics committee (Approval number: 0039-18WOMC). Surgeons resected the primary tumors and the metastases and confirmed clear margins on the samples. Using demographic information, tumor characteristics, and length of survival following diagnosis, patients were identified as belonging to either good or bad survival groups by a pathologist. Specimens were fixed in formalin and subsequently embedded in paraffin. Hematoxylin (HHS16, Sigma-Aldrich) and Eosin (HT110232, Sigma-Aldrich) staining (H&E) of the samples was performed according to the manufacturer instructions. H&E images were obtained at 20x using Aperio Slide Scanner. Slides were first blocked and incubated with various combinations of primary antibodies including LCK (AF3704, R&D Systems), TIGIT (A700-047, Bethyl Lab), TRIM63 (bs2539R, Bioss), OCA2 (bs15510R, Bioss), GPNMB (AF2550, R&D Systems), HMB45 (ab732, Abcam), and KLK8 (MAB1719, R&D Systems). After subsequent washes, slides were incubated with the matching combinations of secondary antibodies, including Alexa Fluor 488 (A11055, Invitrogen), Alexa Fluor 594 (A21203, Invitrogen), and or Alexa Fluor 647 (A31571, Invitrogen). 4',6-diamidino-2-phenylindole (DAPI; Vector Laboratories) was then added dropwise to adequately visualize cell nuclei in the stained specimens. Images of slides were taken using fluorescence microscopy (Nikon) at 40x magnification, split into the individual color channels, and mean intensity of representative areas from each image was measured using ImageJ software. The mean intensity values recorded were then used to generate a color matrix demonstrating the level of expression of each protein in each patient's sample.

For the analysis of melanosome spread and secretion, samples of human in-situ melanoma, as well as metastases from different patients including brain, lymph, and liver were obtained from E. Wolfson Medical Center. Immunohistochemical staining as described above was performed using GPNMB (AF2550, R&D Systems) and HMB45 (ab732, Abcam) as primary antibodies, and Alexa Fluor 488 (A11055, Invitrogen), Alexa Fluor 594 (A21203, Invitrogen) as secondary antibodies, with 4',6-diamidino-2-phenylindole (DAPI; Vector Laboratories) added at the end. Images of the slides were taken at 20x magnification using a Nikon fluorescent microscope. The image of in-situ melanoma was then broken into its component color channels using ImageJ

software, and four equally sized, equidistant frames were cut out and measured for the mean intensity of GPNMB to quantify the gradient of its diffusion from the primary tumor.

3.2.3. Training of a gene-expression based decision tree classifier

To train a molecular classifier for predicting melanoma subgroups, we used the expression levels of the 2000 most variable genes on the set of 469 melanoma samples. We used Matlab's implementation (R2019a) (accessed through PROMO [123]) to grow a classification tree using a curvature test as the method for splitting predictors [189][190]. The training procedure consisted of two steps. First, we assessed the best number of predictor genes to be included in the decision tree, by training many trees on randomly selected subsets of the dataset samples (90% of the samples were included in each iteration) while varying the number of allowed predictor genes and the pruning level. The average training error was calculated for each tree size (Figure S2.5). Next, having determined the number of predictor genes, we used the entire dataset samples (n=469) to train the final decision tree.

4. PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets

In recent years, a growing number of high-throughput genomic technologies have become available for biomedical research and are jointly providing high-resolution genomic data that fuel the revolution of personalized medicine [75][72]. These technologies (collectively named omics) allow the simultaneous quantification of a large number of features at various biological levels. The features include gene expression (mRNA and miRNA abundance levels measured by microarrays or RNA-Seq), protein expression (measured by mass spectroscopy or reverse-phase protein arrays), DNA methylation (methylation arrays), copy number variation (SNP arrays), and others [191][192]. The technologies vary broadly in the number of features they measure as well as in the distribution of measured values [50]. However, they can typically be summarized as a numeric matrix where columns represent samples and rows represent biological features (often correlating to genes). Bioinformatic analysis of such genomic matrices has been extensively used for identifying biologically distinct sample groups, and for revealing groups of correlated biological features [69][193].

The number of tumor samples and measured features that are included in a typical cancer genomic dataset have grown dramatically in the last few years, owing to increasing resolution and reduced costs of array and sequencing technologies. Modern repositories comprise thousands of patient samples and many thousands of features. Investigation of such large datasets is computationally challenging as it requires robust software tools for supporting the analysis of both samples and features in high dimensional data [60].

In addition to genomic data, modern cancer datasets can include extensive medical information (labels) describing each sample, such as clinical properties or assignment to a predefined phenotype. These clinical labels make it possible to fuse genomic and clinical data in various ways in order to discover new insights based on feature-phenotype associations. Common clinical labels in cancer datasets include disease subtypes, pathological stages, survival and recurrence follow-up information, as well as response to treatment. Identification of genomic features that are correlated with significant clinical parameters (biomarkers) is expected to play a significant role in the field of personalized medicine, by which the status of multiple biomarkers may improve subtype diagnosis and guide therapeutic decisions [194][195].

The Cancer Genome Atlas (TCGA) is an example of a revolutionary multi-label multi-omic genomic database [79]. It includes more than 11,000 samples from 33 types of cancer, where each sample was measured using multiple omic technologies and was described by dozens of clinical labels [80]. Many studies have already analyzed TCGA data, improving the subtyping of cancers and shedding light on the biological mechanisms underlying the development of various cancer types [84][142][196]. Such analyses are typically time-consuming, computationally challenging, and entail team effort, as they require applying a diverse array of methods, statistical tools, and algorithms, and often also require writing extensive computer code to perform and interweave the various steps of the analysis [197]. Hence, to effectively extract clinically meaningful insights from such multi-omic multi-label databases, specialized agile integrative tools are required.

To address this challenge, we developed **PROMO** (**PRO**filer of **M**ulti **O**mic data), a fully interactive software suite capable of quickly importing, preprocessing, visualizing, analyzing and reporting the results on cancer datasets in a seamless fashion, without writing a single line of computer code. PROMO includes an extensive array of bioinformatic methods for performing major common analysis types including exploration, visualization, identification of clinically significant disease subtypes, revealing co-regulated feature groups, biomarker discovery, simple classification and integrative multi-omic analysis. Table 4.1 presents an overview of the fundamental analysis types available in PROMO.

An early version of PROMO was developed as part of a study where we identified distinct prognostic subgroups in luminal-A breast tumors based on expression and methylation data (Results, section 2) [198]. The analysis workflow in that project provides an example of the key steps in a typical application of PROMO (Figure 4.1): Data are imported, filtered and preprocessed. Tumor samples are clustered into groups that are then assessed for clinical significance using survival analysis and statistical tests on the clinical labels. Clustering of the genes followed by gene enrichment analysis associates sample clusters with active gene functions. The analysis is summarized visually in a genomic matrix clearly showing the identified sample clusters and their association to important clinical labels (Figure 4.1, step 4), in addition to downstream analysis methods (Figure 4.1, steps 5-7).

In this chapter, we describe PROMO's main features and demonstrate its use in a study of a breast cancer cohort [142].

	Analysis type	Biomedical goal	Relevant PROMO features
1	**General exploration and visualization**	Explore the genomic dataset vis-a-vis the clinical labels Prepare the dataset for downstream analysis, test its consistency and visualize its properties	• Variance-based feature filtering • Label-based sample filtering • Normalization • Sorting by samples label or mean expression • Visualizing data distribution • PCA, t-SNE
2	**Focus on genes of interest**	Explore the expression profiles of specific genes vis-à-vis multiple clinical labels Identify co-expressed genes	• Filter features based on gene symbols • Rank genes by correlation to a given gene symbol • Multi-label matrix visualization
3	**Disease subtype identification**	Look for clinically significant sample clusters	• Sample clustering • Label enrichment analysis • Survival analysis • Classification
4	**Co-regulated feature group identification**	Identify groups of similar features, characterize each group by function	• Feature clustering • GO Enrichment analysis
5	**Biomarker discovery**	Find features that distinguish among sample groups, correlate groups with survival and other clinical data	• Statistical tests for identifying differentially expressed genes • Biomarker-based survival analysis • Rank genes by survival prediction
6	**Integrative multi-omic analysis**	Stratify patients and identify coherent feature groups by integrating data from different omics	• Multi-omic sample clustering • Inter-omic feature correlation

Table 4.1: PROMO's main analysis types

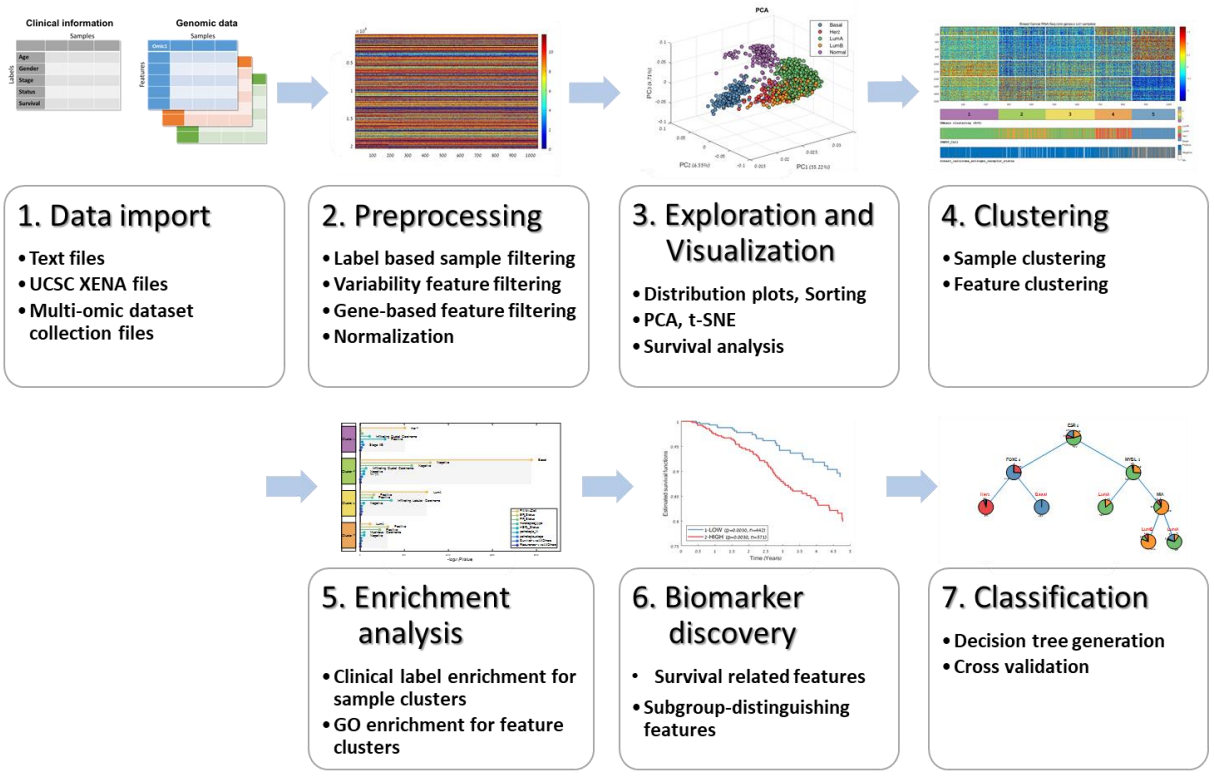


Figure 4.1: PROMO's subtype discovery workflow – From data import to subtype classifier. This figure outlines the complete workflow by which PROMO can be used for identifying and characterizing clinically distinct cancer subtypes: **(1)** Importing genomic data together with clinical information in one of several available formats. **(2)** Preprocessing the data and preparing it for downstream analysis. **(3)** Verifying the integrity of the data, characterizing its distribution and exploring dataset properties with respect to the available clinical labels. **(4)** Employing clustering algorithms partition both samples and features (genes) into groups. **(5)** Applying enrichment tests to identify clinically significant sample subtypes and groups of co-regulated genes and to characterize their function. **(6)** Statistical tests identify features that distinguish between different sample subtypes as well as survival-related features. **(7)** Decision tree classifiers can be generated for formulating a set of rules by which a new sample can be classified.

4.1. Results

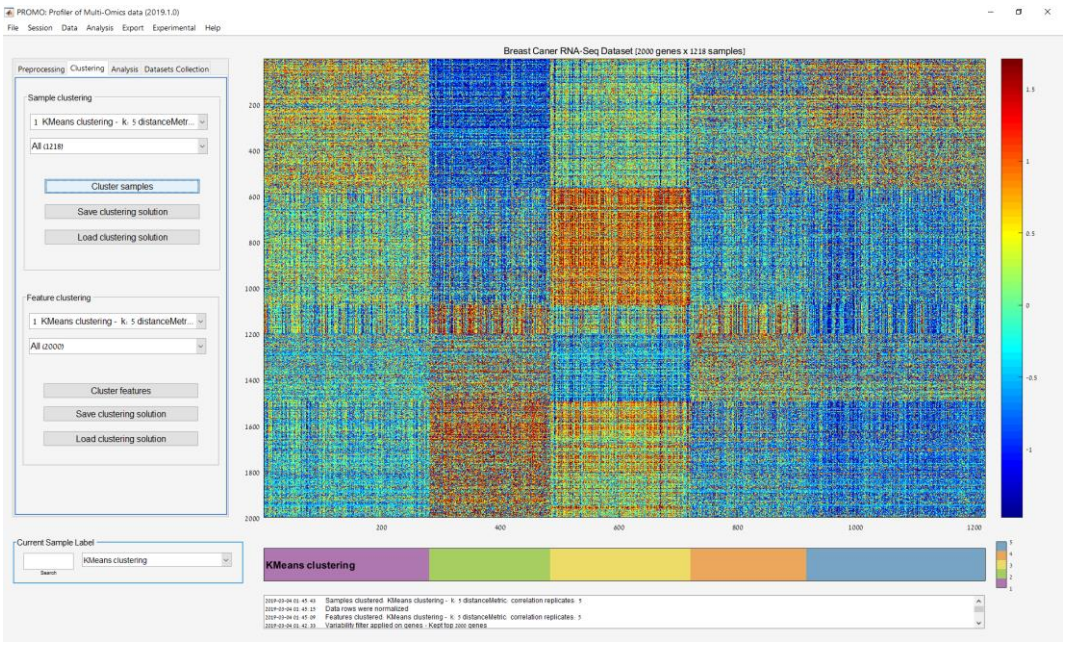
We now describe PROMO’s main features, organized by analysis steps. The described features can be accessed using PROMO's menus or graphical user interface (Figure 4.2). The dataset used was TCGA's breast cancer gene expression profiles (1218 samples downloaded from UCSC's XENA website in May 2018). It is also available on the datasets page of PROMO's website.

4.1.1. Data import and preprocessing

In all analysis types, the first steps are to import the required data from local files into PROMO and prepare it for the analysis. PROMO enables the integration of data of different types and from multiple sources by importing genomic matrices, sample labels, and sample or gene partition files. Genomic matrices accompanied by complementary phenotypic information (clinical labels) can be loaded in the following formats: tabular text files, Gene Expression Omnibus (GEO)[199] series files (including direct download from within PROMO), UCSC's XENA[200][201] file formats (available for many public datasets including all TCGA's data), and PROMO's DSC files. The latter are precompiled multi-omic datasets available at PROMO's dataset download page for selected TCGA cohorts. PROMO also allows separate loading of additional clinical labels and sample partition files to be used in the subtype discovery workflow.

After import, the loaded dataset can be 'cleaned' by filtering out samples based on clinical label values, and also by removing certain features (e.g., removing low variability genes or keeping only specific genes). Additional available common preprocessing steps include flooring, ceiling, and row normalization.

(A) PROMO's main screen



(B) Preprocessing (C) Clustering (D) Analysis (E) Dataset collection

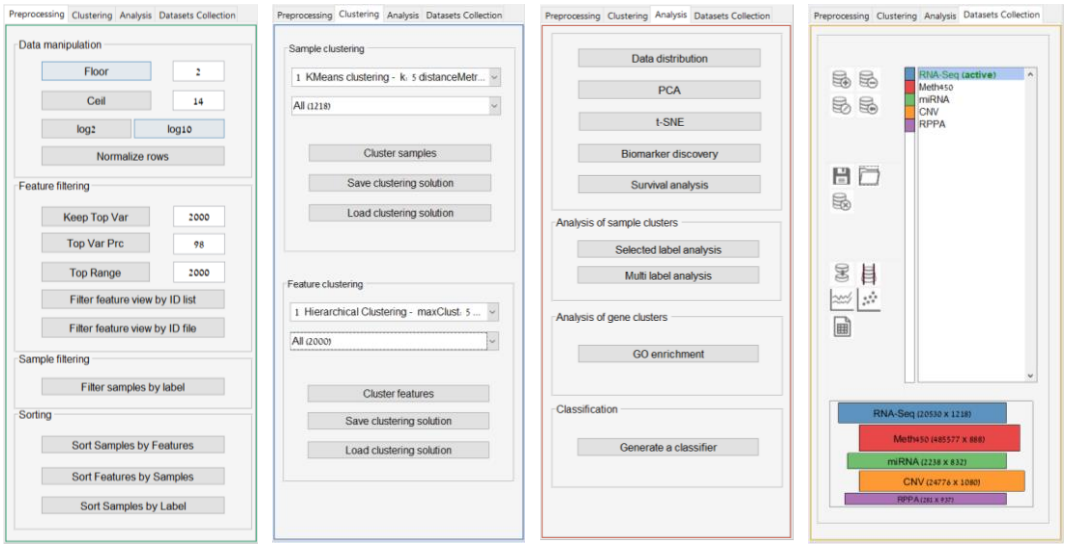


Figure 4.2: PROMO's graphical user interface (A) PROMO's main screen. The genomic matrix is in the center with columns corresponding to samples and rows to features. Colors represent feature values according to the scale on the right. The colorful label bar beneath the matrix displays the currently selected sample label. Analysis steps are documented in the textbox on the bottom of the screen. Key commands are available on the tabbed panels on the left of the screen. **(B)** The *Preprocessing* panel allows filtering, normalization, and sorting of the genomic data. **(C)** Clustering the dataset's samples and features using various algorithms and distance functions is available through the *Clustering* panel. Resulting clustering solutions are aggregated for future review and filtering. **(D)** The *Analysis* panel provides access to several visualization and exploratory tools like PCA, t-SNE, survival analysis, biomarker discovery, GO enrichment and automatic classifier generation. **(E)** In the *Dataset Collection* panel, several genomic matrices can be assembled into a multi-omic dataset collection, and then analyzed together.

4.1.2. Data exploration and visualization

Once a genomic matrix is loaded to PROMO, its properties can be explored with respect to any selected clinical label (Figure 4.3A). The samples (columns) in the matrix can be reordered based on any clinical label or by their mean expression. Basic dataset properties like value distribution (4.3B), clinical label distribution (4.3C), and sample variation (4.3D) can be studied and displayed graphically in various ways including PCA [202][203] and t-SNE [204]. For ease of interpretation, all displays consistently use the same colors to represent the various sample subgroups.

4.1.3. Clustering and enrichment analyses

A major effort in promoting precision medicine is to identify disjoint groups of similar patients and characterize each group using its distinct genomic profile, survival data, and clinical information. To reveal the similarities among patients, clustering is often performed on both samples and features [99]. Clustering the samples can reveal patient groups corresponding to disease subtypes [101] while clustering the features reveals groups of co-regulated genes [102]. PROMO provides various clustering algorithms such as K-means [105], hierarchical clustering [15], and Click [107] (PROMO's clustering panel is shown in Figure S3.1). To explore the resulting clusters, the reordered matrix can be visualized in comparison to multiple sample labels (Fig 4.4A).

After the genes have been clustered, the built-in Gene Ontology tool can help interpret the biological meaning of gene clusters using enrichment analysis (Fig 4.4B) [118]. Likewise, the clinical labels on the samples can be used to statistically characterize each sample cluster. A comprehensive analysis can be applied to each sample cluster using all clinical labels available for the cohort (numeric, ordinal, categorical, or survival labels). The result is a characterization of each cluster, together with FDR corrected p-values [93][134] in a unified report (Fig 4.4C). Enrichment tests for the sample clusters can also be performed using any selected single clinical label (Fig 4.4D). Finally, survival analysis performed on the sample clusters can test their prognostic value using Kaplan-Meier plots [131] and log-rank (Mantel–Haenszel) test [133] (Fig 4.4E). Taken together, PROMO’s clustering and automatic multi-label enrichment analysis can quickly partition both samples and features into distinct groups and assess their biological meaning using the clinical labels.

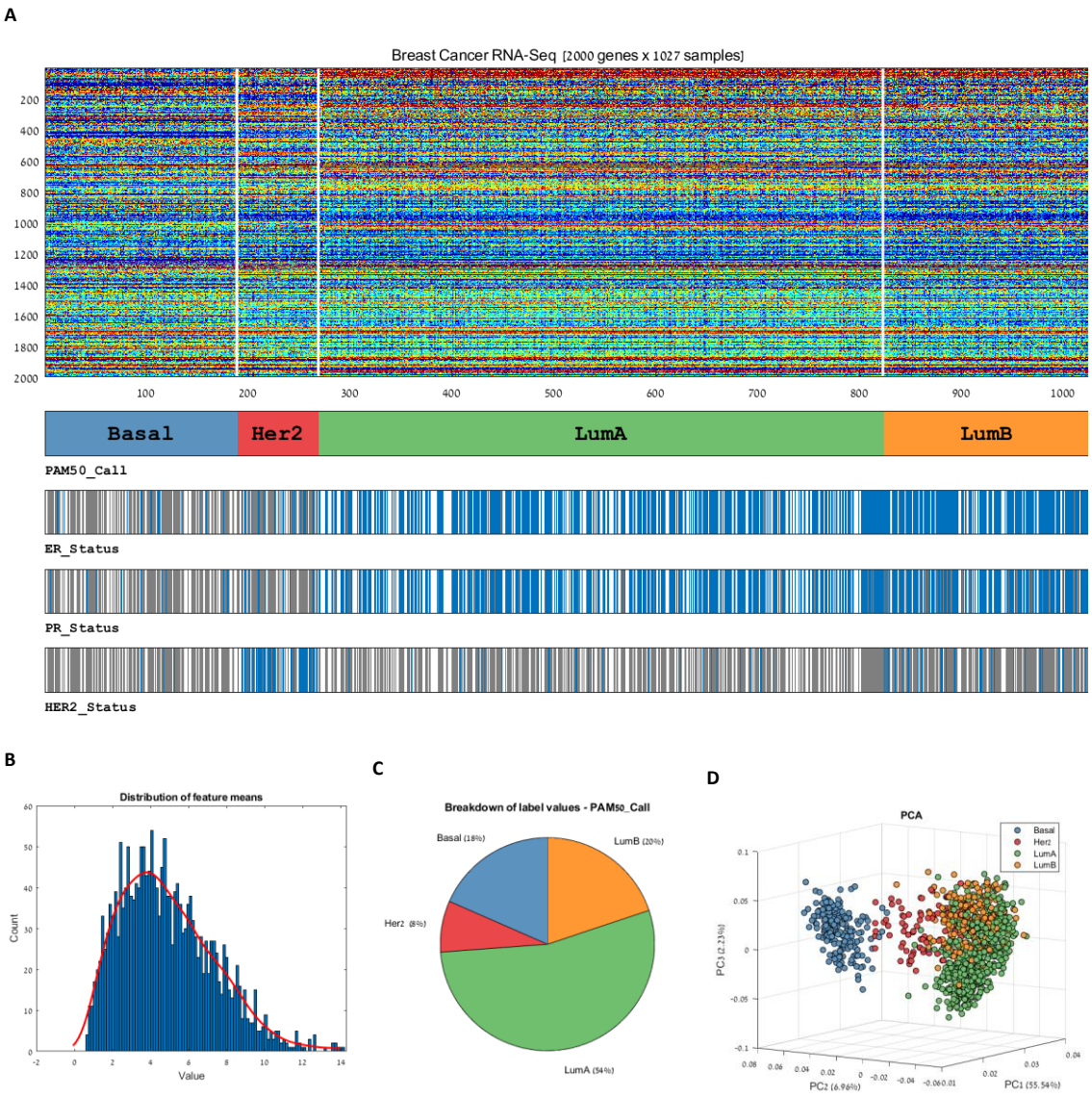


Figure 4.3: Visualization of multi-label genomic data. PROMO provides a variety of methods for visualizing a genomic dataset together with its associated clinical information. (A) A multi-label expression matrix plot. The plot is composed of a heat-map representation of the genomic matrix and several label bars beneath it showing different clinical labels that the user interactively selected. The colors in each label bar show the label value of each sample according to the legend on the right. The label appears below the lower-left corner of the bar. Here, breast cancer patient profiles were grouped according to their PAM50 category (shown in the top label bar). By observing the distribution of values in other bars, relations between the groups and the labels can be observed. For example, the ER, PR and HER2 status of most samples in the 'basal' group are negative, while the HER2 status of most 'HER2' group is positive. **(B)** Data distribution and **(C)** Clinical label distribution can be explored and visualized separately, or in combination using plots such as **(D)** PCA and others. These figures show that the basal tumor samples are mainly characterized by Negative ER, PR and HER2 labels (A) and markedly differ from all other subtypes in their gene expression pattern (D), in accordance with the literature [142].

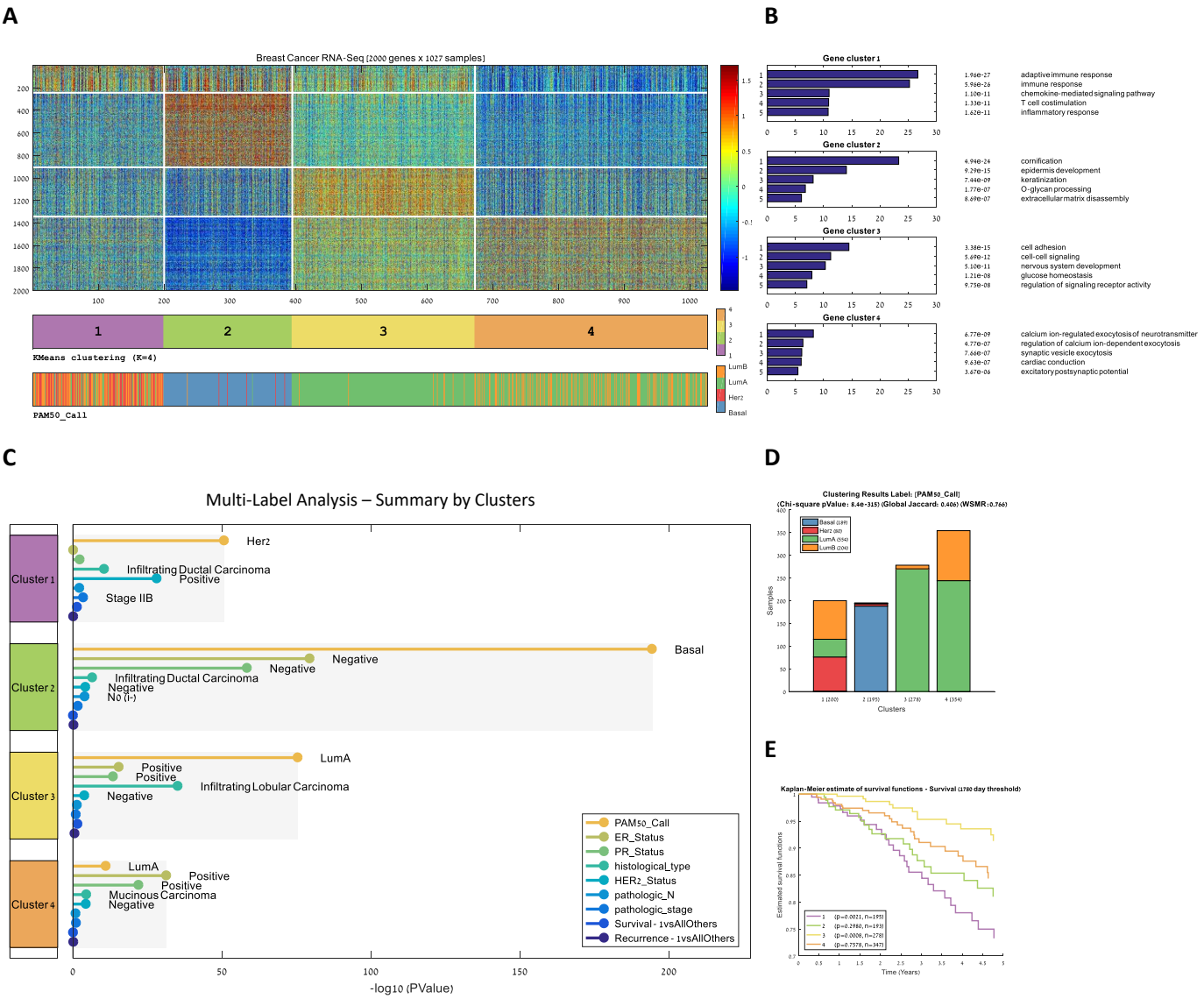


Figure 4.4: Identification and characterization of cancer subtypes. Unsupervised analysis followed by enrichment analysis is performed on both samples and features for identifying clinically significant samples groups, and for biologically characterizing them based on the functions of co-expressed gene groups. **(A)** The RNA-Seq expression matrix of TCGA's breast cancer cohort after clustering both samples (columns) and genes (rows) into four clusters using the K-means algorithm. Clustering is based on the top 2000 variable genes. White lines separate clusters in each dimension. The bars below the matrix show selected sample labels (here: the clustering and PAM50). Matrix and bars were created using PROMO's multi-label matrix drawing. **(B)** Gene clusters were characterized using PROMO's gene ontology enrichment tool. The figure shows the five most significant GO terms for every gene cluster. (C-E) Sample clusters were characterized using the sample clinical labels: **(C)** PROMO's multi-label analysis tool automatically tests the clinical labels of different types (numeric, ordinal, categorical or survival) for enrichment on the sample clusters. FDR correction is performed over all clinical labels of the same type but separately for different types. **(D)** The various sample clusters can also be characterized for a single label by showing its value distribution in each cluster and by calculating enrichment. **(E)** Survival functions for each cluster. The p-values are the significance of the separation of each cluster from the rest using the log-rank test.

4.1.4. Identification of distinguishing genes and features (Biomarker discovery)

Having obtained patient subgroups of interest, either by sample clustering or using a predefined sample label, we may wish to identify distinguishing genes and features that differ significantly among sample groups. Such differentially expressed genes can shed light on the biological difference between sample clusters, and act as biomarkers for classifying a new sample to a sample class.

After selecting the label and the groups that will be compared, PROMO enables the application of various statistical tests for identifying genes that are differentially expressed among the groups. The p-values obtained by the tests can be used for gene sorting, filtering and for clustering the genes into up-regulated and down-regulated groups. PROMO's Gene Ontology enrichment analysis can be executed on the resulting gene groups for characterizing the function of up-regulated and down-regulated genes. FDR correction and fold-change based filtering are also supported. PROMO's biomarker discovery panel and an example of its output are shown in Figure S3.3 and Table S3.1.

For detecting survival biomarkers, PROMO can rank all genes by their association to survival, based on Cox regression analysis [139]. In addition, the user can use the expression levels of selected genes to generate a new sample label (for example, HER2_Low and HER2_High). Kaplan-Meier plots can then be used to estimate the significance of survival differences between sample groups defined by the new label.

Lastly, PROMO can help in finding genes that are functionally related to a given gene of interest by ranking all genes based on their correlation to it. Altogether, the various techniques described here and implemented in PROMO can quickly identify genes that take part in the biological differences between sample groups and may serve as biomarkers for the selected label.

4.1.5. Automatic generation of a simple molecular classifier

After having partitioned the dataset samples, characterized the sample groups and their genes, and established the clinical relevance of the groups, PROMO can build an algorithm to classify a new sample into one of the groups. Such a classifier, especially if based on a small number of genes (rather than the thousands used to identify the subgroups) can serve as a significant step towards translating the analysis results into a diagnostic biomarker for clinical use.

Of the many possible classifier types, decision trees have the advantages of being easy to understand, highly interpretable biologically and easily visualized [189]. Furthermore, they allow

for controlling the tradeoff between accuracy and simplicity. For predicting any selected sample label, PROMO can generate a simple decision tree with a single click (Fig 4.5). The generated decision tree can be visualized graphically, specified textually, and saved to a Matlab file as a function. Automatic cross-validation and parameter optimization make it easy for the user to come up with a simple decision tree that may be in future subtype classification kits. It is also possible to generate a large number of random trees and rank the genes by the frequency of their appearance in the trees, thus identifying informative features for subtype classification.

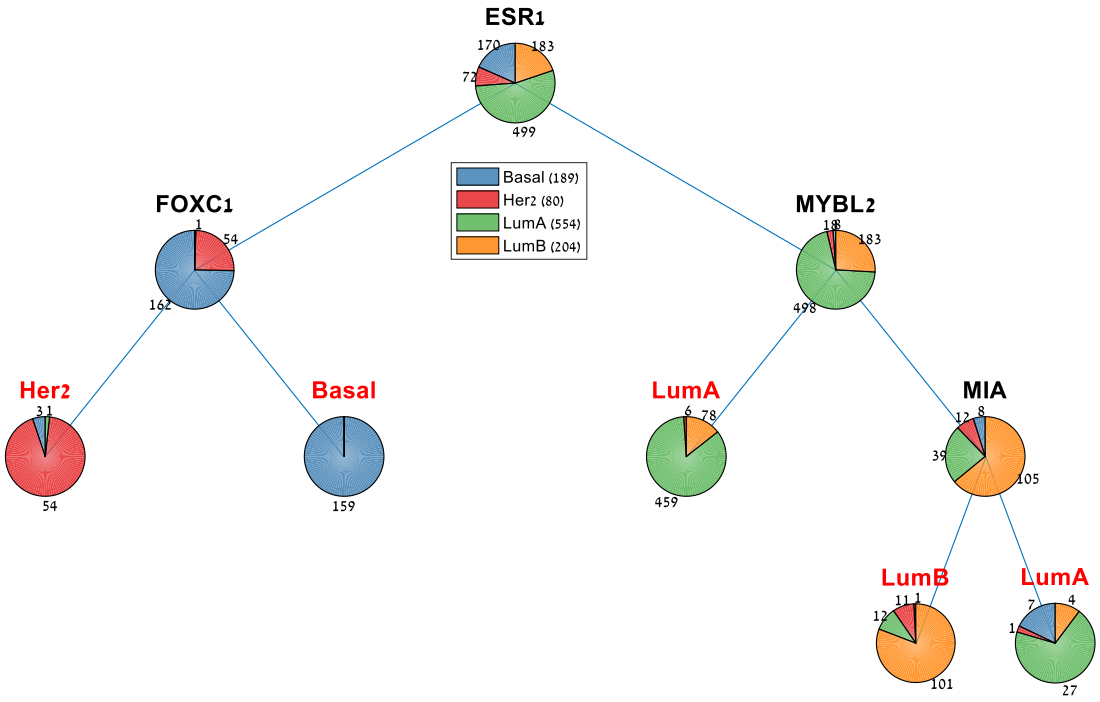


Figure 4.5: Automatically generated decision-tree for classifying breast tumors into the four PAM50 classes. PROMO can generate a cross-validated decision tree for any selected sample label using the currently loaded matrix as training data. In this figure, a four-gene molecular classifier for breast cancer subtypes is presented, showing a 7.77% loss on the training data, and a 15% averaged loss on 10-fold cross-validation.

4.1.6. Integrative multi-omic analysis

In multi-omic datasets, each sample is characterized by several omic profiles (e.g., gene expression, methylation, copy number). Integrative analysis of multi-omic cancer datasets has the potential of revealing biological regulatory patterns that are missed in single omic analysis, and tools for performing such analyses are currently in great demand [205][68].

PROMO provides several features for handling and analyzing multi-omic datasets. The profiles composing a multi-omic dataset can be imported from repositories into a 'Multi-Omic Dataset Collection' in PROMO (Figure 4.2E). The user can navigate between the matrices, edit them independently, and select a subset of the datasets for downstream integrative analysis. Precompiled dataset collections for several TCGA cancer type cohorts are available on PROMO's download page.

After setting up a multi-omic collection, the "inter-omic correlation identification" feature helps to detect correlations between features in two selected omics. This feature allows the identification of correlations between features from different biological levels. For instance, anti-correlation between mRNA expression and DNA methylation levels can pinpoint biological regulation.

The "Multi-omic clustering" feature can be used to cluster the dataset samples based on several omic matrices simultaneously. To this end, PROMO provides implementations of the multi-omic algorithms SNF [206], NEMO [83], and Consensus Clustering [207] modified for multi-omic data. Figure 6 demonstrates the application of a multi-omic clustering algorithm on three different omics of the TCGA's breast cancer cohort.

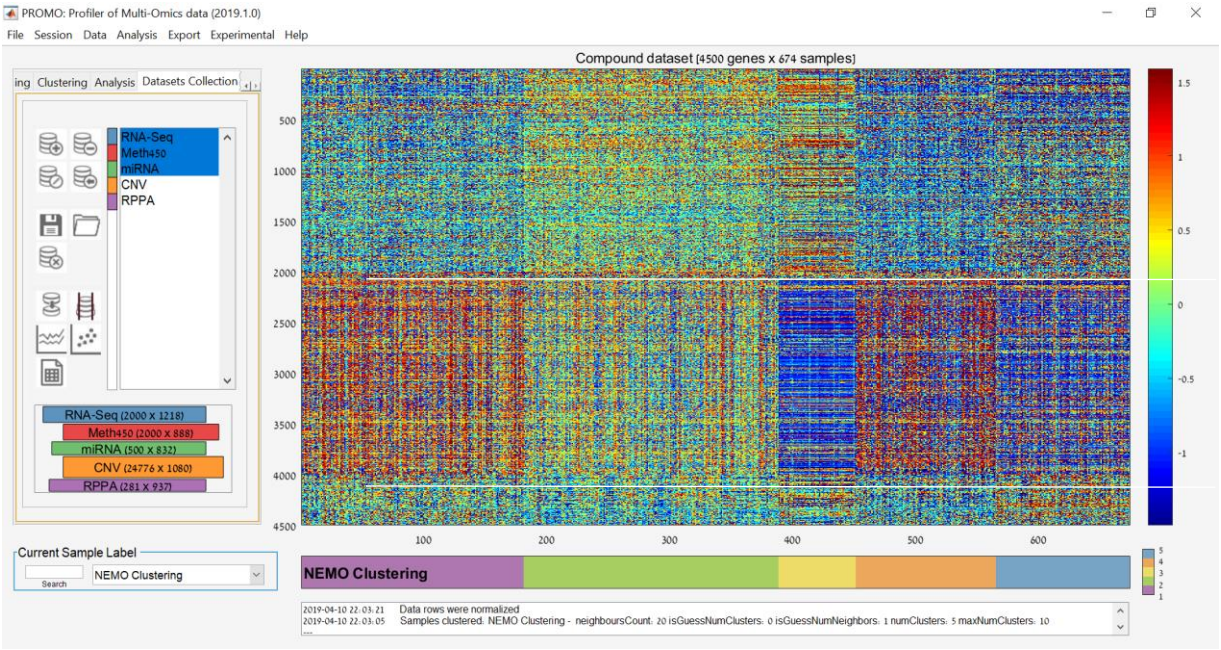


Figure 4.6: Multi-omic sample clustering Screenshot of PROMO's main screen after applying multi-omic clustering on 674 breast tumor samples from TCGA. The 'Dataset Collection' panel on the left was used to select the three omics to be used in the clustering. Here features from three different omics were used: (A) RNA-Seq (2000 features), (B) DNA methylation arrays (2000 features) and (C) miRNA arrays (500 features). Algorithm NEMO [83] was applied on the subset of samples appearing in the three omics into 5 groups, shown on the label bar below the matrix. The genomic matrix displays concatenation of the 4500 features included in the analysis after row normalization, with samples grouped by cluster. The 1st and 4th clusters from the left have high methylation signals, while the second and third have higher gene expression signals. The clustering of tumor samples using a multi-omic algorithm integrates data from different biological levels and thus has the potential of revealing biological regulatory patterns that are missed in a single omic analysis.

4.2. Methods

PROMO is a standalone Windows application that can support huge datasets and has a fast fully interactive graphical user interface. PROMO was written in MATLAB, and it runs over the freely available Matlab runtime environment, taking advantage of its strong computational engine and editable graphical outputs. PROMO is freely available for download at http://acgt.cs.tau.ac.il/promo/.

PROMO's main screen (Figure 4.2A) includes several key graphic elements: A large heatmap representing the currently analyzed genomic matrix is located at the center of the screen (heatmap colors correspond to the matrix values as indicated by the color scale on the right). Beneath the heatmap, a color-bar displays the currently selected sample labels. The same sample label colors will consistently be used by PROMO in all displays. The user can scroll down the list of clinical labels and explore their distribution over the samples. The panel on the left provides access to common commands and parameters. A text log that documents the analysis steps appears at the bottom of the screen. Figures 2B-F show the various panels that can be directly opened from the tab menu on the left of the screen, providing quick access to PROMO's most useful features.

4.3. Summary

PROMO aims to fill in a gap in available analysis tools for large genomic and clinical cancer datasets. It is an interactive tool that is freely available and supports a rich collection of analysis methods and facilitates useful workflows for data exploration and visualization, cancer subtype identification, biomarker discovery and integrative multi-omic analysis. (See Table 4.2 for a list of the key features). PROMO's support for large sample size in addition to features like survival analysis and interrogation of the clinical data on sample clusters makes it especially suitable for analyzing modern cancer datasets. While many of PROMO's features are also available in other tools (Table 4.3), PROMO is unique in its comprehensiveness, support for large sample dimension and the spectrum of tools it provides.

CATEGORY	KEY FEATURES
DATA IMPORT	▪ Importing genomic data from tabular CSV files ▪ Importing UCSC's XENA genome matrix and phenotype files ▪ Importing GEO series files ▪ Adding clinical labels from file
PREPROCESSING	▪ Flooring, ceiling and row normalization ▪ Filtering of samples by clinical labels ▪ Filter features by range, variance, gene symbols or by an external list
DATA EXPLORATION AND VISUALIZATION	▪ PCA, t-SNE ▪ Data distribution plots ▪ Survival Analysis (Kaplan Meier, Log rank) ▪ Multi-label expression matrix figures
SORTING	▪ Sorting samples and features based on genomic data ▪ Sorting samples based on clinical labels
CLUSTERING	▪ Clustering both samples and features using K-means [105], hierarchical clustering [15], and Click [107] ▪ Browsing clustering history and zooming into specific clusters
SAMPLE CLUSTER ANALYSIS	▪ Automated multi-label enrichment test for detecting enrichment of clinical labels
FEATURE CLUSTER ANALYSIS	▪ Gene ontology enrichment analysis
BIOMARKER DISCOVERY	▪ Applying statistical tests for detecting differentially expressed genes/features ▪ Filter results by FDR corrected p-value and fold change ▪ Rank genes based on survival prediction (COX regression)
CLASSIFIER GENERATION	▪ Automatic generation of decision tree classifiers for selected sample labels
INTEGRATIVE MULTI-OMIC ANALYSIS	▪ Assembly of dataset collection ▪ Multi-omic clustering using SNF [206], NEMO [83] or Consensus Clustering [207] ▪ Inter-omic correlation identification

Table 4.2: PROMO's key features

Function	**PROMO** [123]	**Expander** [122]	**XENA** [201]	**Perseus** [208]	**KnowEng** [209]	**O-Miner** [210]
Precompiled datasets	*V*	*X*	*V*	*X*	*V*	*V*
Preprocessing	*V*	*V*	*X*	*V*	*X*	*V*
Data Visualization	*V*	*V*	*V*	*V*	*V*	*V*
Sample clustering	*V*	*V*	*X*	*V*	*V*	*V*
Feature clustering	*V*	*V*	*V*	*V*	*X*	*V*
Sample clusters enrichment tests (clinical data)	*V*	*X*	*V*	*X*	*V*	*X*
Feature clusters enrichment tests	*V*	*V*	*X*	*V*	*V*	*V*
Survival analysis	*V*	*X*	*V*	*X*	*V*	*V*
Biomarker discovery	*V*	*V*	*X*	*V*	*X*	*V*
Automatic decision tree generation	*V*	*X*	*X*	*X*	*X*	*X*
Inter-omic correlation identification	*V*	*X*	*X*	*X*	*V*	*X*
Integrative multi-omic sample clustering	*V*	*X*	*X*	*X*	*X*	*X*

Table 4.3: Comparison of the main functions provided by PROMO and by other tools

5. Discussion

Cancer is a common and heterogeneous group of diseases, which poses significant health and economic burden on the world's population. The basis of cancer is genetic, and indeed a large number of genetic aberrations have already been linked to various types of cancer. Our understanding of cancer biology and its underlying molecular principles is rapidly advancing. However, we are still far from understanding the full extent of tumor variability and therefore in many cancers, treatment is still guided by coarse clinical parameters such as tumor size, histological grade, and lymph node status, and in many cases, only traditional non-specific treatments, such as surgery, chemotherapy and radiation therapy, are available.

For several decades, stratifying patients into clinically distinct subgroups has been a leading strategy for promoting cancer diagnosis and treatment. By this approach, for each cancer type, patients with similar clinical characteristics were grouped together into a designated subtype, which served as a focal point for treatment development. Over time, each subtype was further characterized by its distinguishing properties, prognosis, and response to treatment. However, since only a small number of mainly phenotypic properties were available for each patient, the subgroups defined for many cancer types were crude and did not necessarily reflect a unique underlying genetic makeup that could be used for the development of targeted drugs.

With the emergence of high-throughput omic technologies, a wealth of biological data became available for characterizing tumor samples in much greater detail. Cancer projects such as TCGA [79], GDC [211], ICGC [212] as well as the GEO[199] database, provide many thousands of omic profiles and extensive clinical information on cancer patients [213]. The increased number of samples, combined with the large number of features provided by the new omic technologies, started fueling the revolution of precision medicine, by (1) allowing the definition of more accurate, molecular-based classification for each cancer type, (2) identification of subtype-specific informative biomarkers for improving diagnostics and prognosis, and (3) suggesting subtype-specific targets for drug development.

Our aims in this thesis were to take part in advancing precision medicine by utilizing the currently available multi-omic cancer data for improving the classification of breast and skin cancers into clinically distinct subtypes, as well as to create a software tool for assisting others in carrying a similar task on other cancer datasets. To achieve these aims, we used a strategy that integrates omic and clinical information for identifying clinically significant subgroups. The strategy includes the following principal steps:

- Cluster the tumor samples into groups based on the top variable features. Since our goal was to improve the currently accepted classification of cancers, the number of clusters was selected to be larger than the number of currently accepted subtypes by 1 or 2.
- Cluster the top variable features into a small number of profiles.
- Clinically characterize the resulting sample clusters by performing enrichment and survival analysis using the available clinical information for each sample.
- Use gene enrichment analysis on the feature clusters for identifying the active gene functions characterizing each sample cluster.
- Identify distinguishing features that can be used as subtype biomarkers.

This workflow took shape during our breast cancer study, and it was eventually implemented in the software tool PROMO. When applied to several cancer datasets, it seemed to identify the main structure rapidly. During the course of the study on the breast and melanoma dataset, we have encountered several issues and challenges:

- The large dataset size, ambiguity in some of TCGA's clinical labels, the unknown effect of preprocessing on clustering results and the initial absence of adequate tools for visualizing and analyzing large genomic datasets at the beginning of the work posed a technical challenge in analyzing the data.
- The notion of a subtype was not well defined and required further clarification in our context – eventually, we converged into a definition by which a cancer subtype is a group of patients sharing a distinct genomic profile and distinct survival risk, which must also be large enough to serve as a target population for drug development. We acknowledge that this definition is far from perfect, but we found it of practical utility.
- In each of the analyses of the breast and skin cohorts, we had to decide on which omic to focus, out of the several omics provided by TCGA. In both cases, we selected to focus on RNA-Seq gene expression data. The data were available for a larger number of samples, better corresponded with previously known subtyping of the cohort, and enabled a better separation of the samples based on survival analysis. Perhaps the better match to survival was because gene expression data capture better the signal of immune activity, which appears to be associated with survival in the two cancer cohorts we investigated.
- The limited number of patients in the cohort that had long-range follow-up data was the main limiting factor in identifying finer cancer prognostic subtypes.

- Validation of identified subtypes on a second dataset is challenging, due to differences in value distribution between omic technologies and due to differences in patient distributions between cohorts, which might cause over/under-representation for certain cancer subgroups.

Overall, it is expected that future datasets containing larger sample size, feature resolution and follow-up time would enable us to significantly identify even finer cancer subgroups with a distinct molecular profile.

Interestingly, in the two cancer datasets that we analyzed, breaking the samples into finer subgroups also highlighted specific biological signals whose expression or methylation pattern distinguished patients of different outcomes. In the breast cancer dataset, the luminal-A samples were divided into finer subgroups by gene expression pattern of T cell activation, and by methylation pattern of developmental genes. In the melanoma dataset, TCGA's Keratin subgroup was divided into finer subgroups based on a melanogenesis expression pattern. This demonstrates the power of our strategy.

Still, the approach we utilized in this thesis for cancer subtyping also has limitations that are important to recognize. Firstly, our analyses rely on TCGA's bulk transcriptomic and epigenetic data. Bulk data measure the averaged expression (methylation) levels of genes (CpGs) across a large population of sample cells [59]. It is very efficient in identifying a global, dominant genomic signature in a mixture of sample cells, but cannot capture differences between the subpopulation of cells that compose the sample. Intra-tumor heterogeneity has been shown to play an important role in cancer subtyping and treatment, and should also be accounted for in future studies, perhaps by utilizing single cell sequencing technologies [59][214][215]. Intra-tumor heterogeneity can also explain in part the discrepancies between our results and commonly used classifications such as PAM50 Further, stratifying a large collection of highly variable tumors into a small number of distinct mutually-exclusive subtypes is crucial for simplifying diagnosis and treatment. However, some tumors cannot be directly assigned to one subtype as they bear characteristics of more than a single subtype [216].

We focused in our analysis on gene expression and methylation profiles, but additional types of omic data have been shown to reveal cancer subtypes. In particular, genomic alternations and their use in cancer subtyping were thoroughly explored in the two TCGA papers on breast cancer[142] and melanoma[48] on which our studies relied. Our analysis using expression and methylation data managed to extract novel clinically relevant insights out of the data. We hope that extending the analysis to additional omics and to joint analysis of multiple omics can reveal

stronger and clearer insights in the future. We now turn to discuss each one of the three thesis projects independently.

5.1. Breast cancer subtypes

Gene expression profiling has become a useful tool for breast cancer classification and for the direction of treatment [217]. Whereas the HER2-enriched and the basal-like subgroups are well defined and indicative for anti-Her2 and chemotherapy treatment respectively, the ER-positive luminal subgroup still presents a clinical challenge. In general, all luminal tumors are candidates for anti-hormonal therapy. However, some tumors within this class, often with more proliferative potential and conferring poorer outcome, are considered for additional therapy. Accordingly, the common classification based on the molecular intrinsic subtypes divides the luminal tumors into the better outcome luminal-A and the more proliferative, worse outcome luminal-B subgroups. However, this classification is sub-optimal for clinical decisions because the luminal tumors present a phenotypic and prognostic range rather than an exact partition to either group.

In our study, we applied unsupervised analysis on breast tumor samples using both expression and methylation profiles in order to reveal new genetic and epigenetic patterns that correlate with a clinical outcome, and compared them to the PAM50 subtypes. Overall, our analyses showed that the separation between luminal-A and luminal-B (as represented by PAM50 labels) is not clear-cut, but rather represents a phenotypic continuum (as previously observed [24] [218]-[77]). In fact, each of the gene expression and methylation datasets used in our analysis separately enabled partitioning of the luminal samples into groups showing better prognostic value than that of PAM50.

Furthermore, when we focused on the PAM50-designated luminal-A samples only, the RNA-Seq expression profiles could split the luminal-A samples into two subgroups (Figure 2.3A). The Lobular-enriched LumA-R2 sample group, characterized by a distinct gene over-expression pattern, was associated with significantly reduced recurrence risk compared with the more proliferative LumA-R1 subgroup. Interestingly, genes constituting that over-expression pattern were significantly enriched for functions related to the immune system, including the more specific enrichment of chemokines and genes of up-stream T cell receptor signaling pathways. We postulate that the significantly elevated mRNA levels of immune-related genes in LumA-R2 samples are indicative of increased infiltration levels of immune system cells into these tumors.

Typically, chemokines serve as ligands that by binding to their corresponding receptors, attract immune system cells to the site where they are secreted [219] [220]. LumA-R2 samples over-expressed several chemokines and their corresponding receptors. The simultaneous over-expression of both the chemokine CCL5 (previously found to be highly expressed by breast cancer cells [221]) and one of its receptors – CCR5 (expressed among others by CD8+ Cytotoxic T Cells), suggests that tumor cell-derived CCL5 attracts CD8+ cytotoxic T lymphocytes (CTLs) to LumA-R2 tumors. Similarly, the over-expressed chemokines CCL19 and CCL21 may be expressed by the tumor cells, whereas their CCR7 receptor may be expressed by licensed DC or (less typically) by naive and central memory T Cells.

In line with this possibility, the over-expressed genes in LumA-R2 samples included genes typical of CTLs (and also NK cells), which may lead to anti-tumor cytotoxic activities exerted by the granzyme (GZMA and GZMB) and perforin pathways (PRF1). Accordingly, over-expression of T cell activation genes was also detected in LumA-R2 patients. Notably, the over-expressed genes are concentrated at the upstream part of the T cell receptor signaling pathway (Figure 2.4). At this stage, it is not clear why down-stream effectors are not enriched in LumA-R2 samples, however, it is of interest to see that the alpha chain of IL-15R was over-expressed in these samples, suggesting that T cell activation processes may indeed come into effect in this sub-group of patients.

How could the over-expression of the immune genes by LumA-R2 samples be related, if at all, to reduced tumor recurrence? It is possible that only LumA-R2 tumors can release chemo-attractants that induce the migration of antigen-specific, possibly beneficial, leukocyte sub-populations to the tumor site. Despite recent reports associating tumor-infiltrating lymphocytes with a better prognosis [222] [223] [224], it is yet to be determined how enhanced immunogenic activity in the LumA-R2 tumors may improve their outcome. Possibly in the future, this LumA-R2 characteristic pattern may direct emerging immune checkpoint related therapies [225].

The role of epigenetic regulation in malignant processes is increasingly recognized. Indeed, our analysis of DNA methylation data partitioned the breast tumor samples into four clusters showing only moderate agreement with the expression based PAM50 subtypes. In line with previous studies [36][226], one cluster showed a hypo-methylation pattern and corresponded with the PAM50 basal-like subgroup that was associated with poorer outcome. However, the luminal samples did not cluster neatly into the PAM50 luminal-A and luminal-B subgroups. Instead, three luminal clusters with increasing methylation levels were obtained (Clusters 1-3 in Figure 2.5A), of which the most hyper-methylated cluster was associated with a significantly

poorer five-year prognosis. In fact, even when we clustered only the luminal-A samples (Figure 2.5C), the hyper-methylated cluster 1 (LumA-M1) still had significantly poorer survival compared to the other two clusters (LumA-M2 and LumA-M3).

Notably, the top 1000 differentially methylated CpG loci, all hyper-methylated on LumA-M1 samples, showed enrichment for genes involved in morphogenesis, differentiation, and developmental processes. Moreover, the CpG hyper-methylation correlated with under-expression of developmental genes, including various tumor suppressor genes. Indeed, hyper-methylation of developmental genes in luminal breast tumors was previously reported [227] [228], secondary to repressive histone marks, which direct de-novo methylation. Moreover, hyper-methylation was implicated in normal processes of cell aging and in tumorigenesis [61]. Taken together, the methylation-based analysis suggests poorer outcome for luminal tumors showing a characteristic hyper-methylation pattern, whether in the luminal-A or in the luminal-B subgroups. The hyper-methylation associated silencing of developmental and tumor suppressor genes may indeed explain these findings. More importantly, within the luminal-A subgroup that is generally associated with a better outcome, the hyper-methylation pattern of the LumA-M1 subgroup marks 84 samples (composing 22% of the 378 luminal-A samples) as a high-risk patient group that might benefit from more aggressive treatment.

Lastly, we showed that the sample partitions induced by the gene expression and DNA methylation patterns are related ($p = 4.4E-08$, see the lower bar on Figure 2.5C), mainly because the better outcome LumA-M3 samples are enriched for LumA-R2. However, our attempts to partition the luminal-A samples based on both patterns together did not yield a partition that is better than the separate partitions in terms of survival prediction or clustering stability. This observation was confirmed by Cox multivariate analysis showing the independent prognostic contribution of each pattern to outcome prediction (Table 2.5), suggesting that gene expression and methylation hold complementary information, reflecting different aspects of the biological complexity of breast tumors.

Recently, several novel partitions of luminal breast tumors were proposed [19][65][230]. The partitions identified in our study are reinforced by partial though significant similarity to some of the newly defined groups. LumA-R1 and LumA-R2 clusters are enriched for the Proliferative ($p=8.1e-04$) and Reactive-like ($2.4-e04$) classes respectively of ILC (Invasive Lobular Carcinoma) tumors, as defined in [229] (see Supplementary Information, section 12). Furthermore, the LumA-M1 cluster is enriched ($p=1.6e-07$) for the poorer outcome Epi-LumB group, described by Stefansson et al. [226] (named Epi-LumB, as it was largely composed of luminal-B samples, see

Supplementary Information, section 13). Additional research is needed in order to consolidate the different partitions identified using different procedures into robust and meaningful categories for prognostic and diagnostic use in clinics.

5.2. Skin cancer subtypes

Our computational analysis of the 474 melanoma expression profiles identified four clinically distinct subgroups. The identified groups (Table 3.1) showed significant correspondence to TCGA's transcriptomic classification[48]; however, TCGA's keratin subgroup was split in our analysis into a keratin subgroup, composed mainly of primary tumors (cluster 2), and a melanogenesis-high subgroup, composed mainly of high-risk metastatic melanomas (cluster 4).

Three gene expression signatures stratified the melanoma samples into the four clinically distinct subgroups: Patients in Cluster 1, characterized by high expression of immune genes, had the best survival, in agreement with previous reports in melanoma and other cancer types [48] [198].

Patients in Cluster 2, characterized by a high expression of keratin related genes, had the worst survival. That cluster contained mostly primary samples. As noted in [48], the poor survival can be attributed to the size bias of primary melanomas in the TCGA cohort.

The third expression pattern, which was of greatest interest to us, was enriched for melanogenesis and melanosome-related genes and distinguished the two, metastasis-enriched, clusters 3 (“Melanogenesis-low”) and 4 (“Melanogenesis-high”). Patients with high levels of the melanogenesis pattern were included in Cluster 4 and had a worse survival rate compared to those in Cluster 3, who had low levels. The association between over-expression of melanogenesis genes and poorer prognosis can be explained by several hypotheses: (1) Trafficking of miRNA or other agents within secreted melanosomes by melanoma cells to its environment can make it more hospitable for melanoma progression [172]; (2) Making the tumor resilient to chemotherapy, due to the drug-detoxifying properties of melanogenesis genes [173][231]; or (3) Removal of anticancer drugs from the melanoma cells by melanogenesis related transporters effluxing drugs outside of cells [232][233]. The latter hypothesis can be backed by the fact that in our analysis, samples of the Melanogenesis-high cluster overexpressed ABC transporters such as ABCB5 and ABCC2 [232] (Table S2.4). Our validation on samples from patients found that secretion of melanosomes to the surrounding tissues occurs both in primary melanoma (with clear gradient) as well as in metastatic melanoma. We, therefore, hypothesize that the reduced survival rate that characterizes the "Melanogenesis-high" subgroup is

associated with the significantly higher activation of the melanogenesis pathway in these patients, as opposed to the "Melanogenesis-low" subgroup.

The importance of keratin, immune and melanogenesis expression patterns in classifying melanoma tumors was also recognized in previous studies aimed at molecularly stratifying melanoma tumors. In 2010, Jönsson et al. identified four expression-based subgroups by analyzing 57 stage IV melanomas taken from patients[234]. These subgroups, later named 'Lund', were called 'normal-like', 'high-immune', 'pigmentation', and 'proliferative' sample subgroups. The normal-like group was characterized by over-expression of keratin genes (KRT17, KRT10, and KRT80); the high-immune group overexpressed immune genes (CCL13 and CD209), and the pigmentation group showed overexpression of melanogenesis genes (MITF, TYR, DCT, and MLANA). The proliferative group showed under-expression of the three signatures. The subgroups showed significant survival differences and were confirmed on additional patient cohorts [235][236][237]. These results support the potential utility of biomarkers for the three expression patterns in classifying melanoma tumors into clinically distinct subtypes.

We trained a simple decision tree for classifying melanoma samples into one of the four subgroups. Our tests showed that a three-gene decision tree gave a good balance between classifier simplicity and accuracy. Although inferior in accuracy to more complex classifiers like SVM, a three-gene decision tree is easier to interpret biologically, easier to translate into a useful diagnostic kit in the future, and also captures the hierarchy of biological signals we identified in the data. A drawback for using a decision tree is that its thresholds depend on the distribution of the training data, and therefore must be recalculated before the tree can be applied to other datasets.

Across multiple training runs, the trees produced tended to select one representative predictor gene from each of the three expression signatures. Key predictor genes, as well as their other signature representatives, were experimentally validated on a new cohort of melanoma taken from patients. Although limited in scope, the validation showed that the predictor genes differed in their protein expression levels among melanoma samples and confirmed the association of predictor levels with outcome. More substantial validation should be conducted.

We hope that classifiers such as the one suggested here will be translated in the near future into accurate and accessible diagnostic kits for improving the diagnosis and prognosis of melanoma tumors.

5.3. PROMO

Our vision in developing PROMO was to create a one-stop-shop for mining clinically important insights from large omic datasets, quickly and without any need for programming skills. A thorough analysis of these datasets - and larger ones expected in the future - by many researchers is crucial for improving cancer diagnosis and treatment. However, the analysis of such data is challenging and requires advanced bioinformatics, statistical, and programming skills. PROMO accelerates the analysis process and makes it more accessible for non-computational cancer researchers. Within a single short session, the user can import a cancer dataset of interest, preprocess it, cluster its samples and features, test the sample clusters for significance using survival analysis and enrichment tests on the clinical labels, test the feature clusters for GO enrichment, identify subtype distinguishing features (biomarkers) using various statistical tests and export the results using various reports and figures. The simple classification capabilities in PROMO can automatically produce a decision tree classifier for any selected label and thus act as a basis for a subtype diagnosis.

We intend to continue developing PROMO by adding features and supporting the tool's users. We hope that PROMO’s comprehensiveness and ease of use will help cancer researchers make the best use of the accumulating cancer datasets to fulfill the promises of precision medicine.

6. References

[1] "World Health Organization: Cancer." [Online]. Available: https://www.who.int/health-topics/cancer. [Accessed: 22-Dec-2019].

[2] "International Agency for Research on Cancer, GLOBOCAN 2018, World Health Organization." [Online]. Available: https://gco.iarc.fr/. [Accessed: 22-Dec-2019].

[3] J. Ferlay *et al.*, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods.," *Int. J. cancer*, vol. 144, no. 8, pp. 1941–1953, 2019.

[4] A. W. Lambert, D. R. Pattabiraman, and R. A. Weinberg, "Emerging Biological Principles of Metastasis," *Cell*, vol. 168, no. 4, pp. 670–691, Feb. 2017.

[5] D. Hanahan and R. A. R. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57–70, 2000.

[6] Y. A. Fouad and C. Aanei, "Revisiting the hallmarks of cancer.," *Am. J. Cancer Res.*, vol. 7, no. 5, pp. 1016–1036, 2017.

[7] D. Mittal, M. M. Gubin, R. D. Schreiber, and M. J. Smyth, "New insights into cancer immunoediting and its three component phases—elimination, equilibrium and escape," *Curr. Opin. Immunol.*, vol. 27, pp. 16–25, Apr. 2014.

[8] R. Kim, M. Emi, and K. Tanabe, "Cancer immunoediting from immune surveillance to immune escape," *Immunology*, vol. 121, no. 1, pp. 1–14, May 2007.

[9] T. Blankenstein, P. G. Coulie, E. Gilboa, and E. M. Jaffee, "The determinants of tumour immunogenicity," *Nat. Rev. Cancer*, vol. 12, no. 4, pp. 307–13, Apr. 2012.

[10] D. A. Quigley and V. Kristensen, "Predicting prognosis and therapeutic response from interactions between lymphocytes and tumor cells," *Mol. Oncol.*, vol. 9, no. 10, pp. 2054–2062, Dec. 2015.

[11] V. Schirrmacher, "From chemotherapy to biological therapy: A review of novel concepts to reduce the side effects of systemic cancer treatment (Review).," *Int. J. Oncol.*, vol. 54, no. 2, pp. 407–419, Feb. 2019.

[12] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation.," *Cell*, vol.

144, no. 5, pp. 646–74, Mar. 2011.

[13] N. R. Bertos and M. Park, “Breast cancer - One term, many entities?,” *J. Clin. Invest.*, vol. 121, no. 10, pp. 3789–3796, 2011.

[14] A. Goldhirsch, J. N. Ingle, R. D. Gelber, A. S. Coates, B. Thürlimann, and H.-J. Senn, “Thresholds for therapies: highlights of the St Gallen International Expert Consensus on the primary therapy of early breast cancer 2009.,” *Ann. Oncol.*, vol. 20, no. 8, pp. 1319–29, Aug. 2009.

[15] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, “Cluster analysis and display of genome-wide expression patterns,” *Proc. Natl. Acad. Sci.*, vol. 95, no. 25, pp. 14863–14868, Dec. 1998.

[16] C. M. Perou *et al.*, “Molecular portraits of human breast tumours.,” *Nature*, vol. 406, no. 6797, pp. 747–52, Aug. 2000.

[17] J. S. Reis-Filho and L. Pusztai, “Gene expression profiling in breast cancer: classification, prognostication, and prediction.,” *Lancet*, vol. 378, no. 9805, pp. 1812–23, Nov. 2011.

[18] C. Sotiriou and L. Pusztai, “Gene-Expression Signatures in Breast Cancer,” *N. Engl. J. Med.*, vol. 360, no. 8, pp. 790–800, Feb. 2009.

[19] T. Sørlie *et al.*, “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 98, no. 19, pp. 10869–10874, Sep. 2001.

[20] T. Sorlie *et al.*, “Repeated observation of breast tumor subtypes in independent gene expression data sets,” *Proc. Natl. Acad. Sci.*, vol. 100, no. 14, pp. 8418–8423, Jul. 2003.

[21] T. Sørlie, “Molecular portraits of breast cancer: tumour subtypes as distinct disease entities.,” *Eur. J. Cancer*, vol. 40, no. 18, pp. 2667–75, Dec. 2004.

[22] A. Goldhirsch *et al.*, “Strategies for subtypes--dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011.,” *Ann. Oncol.*, vol. 22, no. 8, pp. 1736–47, Aug. 2011.

[23] J. S. Parker *et al.*, “Supervised risk predictor of breast cancer based on intrinsic subtypes.,” *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–7, Mar. 2009.

[24] B. Weigelt *et al.*, “Breast cancer molecular profiling with single sample predictors: a retrospective analysis,” *Lancet Oncol.*, vol. 11, no. 4, pp. 339–349, Apr. 2010.

[25] M. Bhattacharyya, J. Nath, and S. Bandyopadhyay, “MicroRNA signatures highlight new breast cancer subtypes.,” *Gene*, vol. 556, no. 2, pp. 192–8, Feb. 2015.

[26] C. Blenkiron *et al.*, “MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype.,” *Genome Biol.*, vol. 8, no. 10, p. R214, 2007.

[27] F. Andre *et al.*, “Molecular characterization of breast cancer with high-resolution oligonucleotide comparative genomic hybridization array.,” *Clin. Cancer Res.*, vol. 15, no. 2, pp. 441–51, Jan. 2009.

[28] R. Shen, A. B. Olshen, and M. Ladanyi, “Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis.,” *Bioinformatics*, vol. 25, no. 22, pp. 2906–12, Nov. 2009.

[29] C. Curtis *et al.*, “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.,” *Nature*, vol. 486, no. 7403, pp. 346–52, Jun. 2012.

[30] S.-J. Dawson, O. M. Rueda, S. Aparicio, and C. Caldas, “A new genome-driven integrated classification of breast cancer and its implications.,” *EMBO J.*, vol. 32, no. 5, pp. 617–28, Mar. 2013.

[31] M. Bibikova *et al.*, “High density DNA methylation array with single CpG site resolution,” *Genomics*, vol. 98, no. 4, pp. 288–295, Oct. 2011.

[32] P. A. Jones and S. B. Baylin, “The fundamental role of epigenetic events in cancer.,” *Nat. Rev. Genet.*, vol. 3, no. 6, pp. 415–28, Jun. 2002.

[33] M. Esteller, “Epigenetics in Cancer,” *N. Engl. J. Med.*, vol. 358, no. 11, pp. 1148–1159, Mar. 2008.

[34] S. B. Baylin, “DNA methylation and gene silencing in cancer,” *Nat. Clin. Pract. Oncol.*, vol. 2 Suppl 1, no. August 2005, pp. S4–S11, 2005.

[35] K. Holm *et al.*, “Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns.,” *Breast Cancer Res.*, vol. 12, no. 3, p. R36, Jan. 2010.

[36] J. A. Rønneberg *et al.*, “Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer.,” *Mol. Oncol.*, vol. 5, no. 1, pp. 61–76, Feb. 2011.

[37] R. Ossio, R. Roldán-Marín, H. Martínez-Said, D. J. Adams, and C. D. Robles-Espinoza, “Melanoma: a global perspective,” *Nat. Rev. Cancer*, vol. 17, no. 7, pp. 393–394, Jul. 2017.

[38] Z. Apalla, A. Lallas, E. Sotiriou, E. Lazaridou, and D. Ioannides, “Epidemiological trends in skin cancer,” *Dermatol. Pract. Concept.*, vol. 7, no. 2, pp. 1–6, Apr. 2017.

[39] N. H. Matthews, W.-Q. Li, A. A. Qureshi, M. A. Weinstock, and E. Cho, *Epidemiology of Melanoma*. Codon Publications, 2017.

[40] A. H. Shain and B. C. Bastian, “From melanocytes to melanomas,” *Nat. Rev. Cancer*, vol. 16, no. 6, pp. 345–358, Jun. 2016.

[41] S. M. Swetter *et al.*, “Guidelines of care for the management of primary cutaneous melanoma,” *J. Am. Acad. Dermatol.*, vol. 80, no. 1, pp. 208–250, Jan. 2019.

[42] A. N. Houghton and D. Polsky, “Focus on melanoma,” *Cancer Cell*, vol. 2, no. 4, pp. 275–278, Oct. 2002.

[43] O. Kabbarah and L. Chin, “Revealing the genomic heterogeneity of melanoma,” *Cancer Cell*, vol. 8, no. 6, pp. 439–441, Dec. 2005.

[44] R. A. Scolyer, G. V Long, and J. F. Thompson, “Evolving concepts in melanoma classification and their relevance to multidisciplinary melanoma patient care,” *Mol. Oncol.*, vol. 5, no. 2, pp. 124–136, Apr. 2011.

[45] W. H. Ward, F. Lambreton, N. Goel, J. Q. Yu, and J. M. Farma, “Clinical Presentation and Staging of Melanoma,” in *Cutaneous Melanoma: Etiology and Therapy*, Codon Publications, 2017, pp. 79–89.

[46] G. Leonardi *et al.*, “Cutaneous melanoma: From pathogenesis to therapy (Review),” *Int. J. Oncol.*, vol. 52, no. 4, pp. 1071–1080, Feb. 2018.

[47] S. Rajkumar and I. R. Watson, “Molecular characterisation of cutaneous melanoma: creating a framework for targeted and immune therapies,” *Br. J. Cancer*, vol. 115, no. 2, pp. 145–155, Jul. 2016.

[48] R. Akbani *et al.*, “Genomic Classification of Cutaneous Melanoma,” *Cell*, vol. 161, no. 7, pp. 1681–1696, Jun. 2015.

[49] F. CRICK, “Central Dogma of Molecular Biology,” *Nature*, vol. 227, no. 5258, pp. 561–563, Aug. 1970.

[50] Y. Hasin, M. Seldin, and A. Lusis, “Multi-omics approaches to disease.,” *Genome Biol.*, vol. 18, no. 1, p. 83, Dec. 2017.

[51] J. M. Berg, J. L. Tymoczko, and L. Stryer, “DNA, RNA, and the Flow of Genetic Information,” in *Biochemistry*, 5th Editio., New York: W H Freeman, 2002, p. 1050.

[52] E. V. Koonin, “Does the central dogma still stand?,” *Biol. Direct*, vol. 7, no. 1, p. 27, 2012.

[53] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, “Methods of integrating data to uncover genotype–phenotype interactions,” *Nat. Rev. Genet.*, vol. 16, no. 2, pp. 85–97, Jan. 2015.

[54] J. Ragoussis, “Genotyping technologies for genetic research.,” *Annu. Rev. Genomics Hum. Genet.*, vol. 10, no. 1, pp. 117–33, Sep. 2009.

[55] D. C. Koboldt, K. M. Steinberg, D. E. Larson, R. K. Wilson, and E. R. Mardis, “The Next-Generation Sequencing Revolution and Its Impact on Genomics,” *Cell*, vol. 155, no. 1, pp. 27–38, Sep. 2013.

[56] S. B. Ng *et al.*, “Targeted capture and massively parallel sequencing of 12 human exomes,” *Nature*, 2009.

[57] S. Goodwin, J. D. McPherson, and W. R. McCombie, “Coming of age: ten years of next-generation sequencing technologies,” *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 333–351, Jun. 2016.

[58] A. Schulze and J. Downward, “Navigating gene expression using microarrays — a technology review,” *Nat. Cell Biol.*, vol. 3, no. 8, pp. E190–E195, Aug. 2001.

[59] W. W. Soon, M. Hariharan, and M. P. Snyder, “High-throughput sequencing for biology and medicine.,” *Mol. Syst. Biol.*, vol. 9, no. 1, p. 640, Jan. 2013.

[60] J. Xuan, Y. Yu, T. Qing, L. Guo, and L. Shi, “Next-generation sequencing in the clinic:

Promises and challenges," *Cancer Lett.*, vol. 340, no. 2, pp. 284–295, Nov. 2013.

[61] Y. Guo, Q. Sheng, J. Li, F. Ye, D. C. Samuels, and Y. Shyr, "Large Scale Comparison of Gene Expression Levels by Microarrays and RNAseq Using TCGA Data," *PLoS One*, vol. 8, no. 8, p. e71462, Aug. 2013.

[62] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics.," *Nat. Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[63] D. Barros-Silva, C. Marques, R. Henrique, and C. Jerónimo, "Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications," *Genes (Basel).*, vol. 9, no. 9, p. 429, Aug. 2018.

[64] B. Aslam, M. Basit, M. A. Nisar, M. Khurshid, and M. H. Rasool, "Proteomics: Technologies and Their Applications," *J. Chromatogr. Sci.*, vol. 55, no. 2, pp. 182–196, Feb. 2017.

[65] R. I. Gallagher and V. Espina, "Reverse Phase Protein Arrays: Mapping the Path Towards Personalized Medicine," *Mol. Diagn. Ther.*, vol. 18, no. 6, pp. 619–630, Dec. 2014.

[66] J.-L. Ren, A.-H. Zhang, L. Kong, and X.-J. Wang, "Advances in mass spectrometry-based metabolomics for investigation of metabolites," *RSC Adv.*, vol. 8, no. 40, pp. 22335–22350, 2018.

[67] H. Kilpinen and J. C. Barrett, "How next-generation sequencing is transforming complex disease genetics," *Trends Genet.*, vol. 29, no. 1, pp. 23–30, Jan. 2013.

[68] S. Huang, K. Chaudhary, and L. X. Garmire, "More Is Better: Recent Progress in Multi-Omics Data Integration Methods.," *Front. Genet.*, vol. 8, no. JUN, p. 84, Jun. 2017.

[69] V. Gligorijević, N. Malod-Dognin, and N. Pržulj, "Integrative methods for analyzing big data in precision medicine.," *Proteomics*, vol. 16, no. 5, pp. 741–58, Mar. 2016.

[70] K. J. Karczewski and M. P. Snyder, "Integrative omics for health and disease," *Nat. Rev. Genet.*, vol. 19, no. 5, pp. 299–310, May 2018.

[71] G. Noell, R. Faner, and A. Agustí, "From systems biology to P4 medicine: applications in respiratory medicine," *Eur. Respir. Rev.*, vol. 27, no. 147, p. 170110, Mar. 2018.

[72] N. Malod-Dognin, J. Petschnigg, and N. Pržulj, "Precision medicine — A promising, yet

challenging road lies ahead," *Curr. Opin. Syst. Biol.*, vol. 7, pp. 1–7, Feb. 2018.

[73] X. D. Zhang, "Precision Medicine, Personalized Medicine, Omics and Big Data: Concepts and Relationships," *J. Pharmacogenomics Pharmacoproteomics*, vol. 06, no. 02, 2015.

[74] A. Sonnenblick, D. Fumagalli, C. Sotiriou, and M. Piccart, "Is the differentiation into molecular subtypes of breast cancer important for staging, local and systemic therapy, and follow up?," *Cancer Treat. Rev.*, vol. 40, no. 9, pp. 1089–1095, Oct. 2014.

[75] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine," *Nat. Rev. Clin. Oncol.*, vol. 8, no. 3, pp. 184–187, Mar. 2011.

[76] X. Liu, X. Luo, C. Jiang, and H. Zhao, "Difficulties and challenges in the development of precision medicine," *Clin. Genet.*, vol. 95, no. 5, pp. 569–574, May 2019.

[77] B. Weigelt, L. Pusztai, A. Ashworth, and J. S. Reis-Filho, "Challenges translating breast cancer gene signatures into the clinic," *Nat. Rev. Clin. Oncol.*, vol. 9, no. 1, pp. 58–64, Jan. 2012.

[78] S. J. Hollingsworth, "Precision medicine in oncology drug development: A pharma perspective," *Drug Discovery Today*. 2015.

[79] "The Cancer Genome Atlas (TCGA)." [Online]. Available: http://cancergenome.nih.gov/.

[80] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.," *Contemp. Oncol. (Poznan, Poland)*, vol. 19, no. 1A, pp. A68-77, 2015.

[81] C. Hutter and J. C. Zenklusen, "The Cancer Genome Atlas: Creating Lasting Value beyond Its Data," *Cell*, vol. 173, no. 2, pp. 283–285, Apr. 2018.

[82] A. Blum, P. Wang, and J. C. Zenklusen, "SnapShot: TCGA-Analyzed Tumors," *Cell*, vol. 173, no. 2, p. 530, Apr. 2018.

[83] N. Rappoport and R. Shamir, "NEMO: cancer subtyping by integration of partial multi-omic data," *Bioinformatics*, vol. 35, no. 18, pp. 3348–3356, Sep. 2019.

[84] J. N. Weinstein *et al.*, "The Cancer Genome Atlas Pan-Cancer analysis project," *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013.

[85] C. X. Ma and M. J. Ellis, “The Cancer Genome Atlas: clinical applications for breast cancer.,” *Oncology (Williston Park).*, vol. 27, no. 12, pp. 1263–9, 1274–9, Dec. 2013.

[86] W. Pan, “A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments,” *Bioinformatics*, vol. 18, no. 4, pp. 546–554, Apr. 2002.

[87] R. McCleery, T. Watt, and T. Hart, “Nonparametric Tests,” in *Introduction to Statistics for Biology, Third Edition*, Chapman and Hall/CRC, 2007, pp. 195–218.

[88] X. Cui and G. A. Churchill, “Statistical tests for differential expression in cDNA microarray experiments,” *Genome Biology*. 2003.

[89] S. Draghici, “Statistical intelligence: effective analysis of high-density microarray data,” *Drug Discov. Today*, vol. 7, no. 11, pp. S55–S63, May 2002.

[90] D. G. Altman, *Practical Statistics for Medical Research*, 1st ed. New York: Chapman and Hall/CRC, 1990.

[91] A. Farcomeni, “A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion,” *Stat. Methods Med. Res.*, vol. 17, no. 4, pp. 347–388, Aug. 2008.

[92] S. Holm, “A simple sequentially rejective multiple test procedure,” *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.

[93] Y. Benjamini and Y. Hochberg, “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *J. R. Stat. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, Jan. 1995.

[94] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” *Ann. Stat.*, vol. 29, no. 4, pp. 1165–1188, Aug. 2001.

[95] Sergios Theodoridis and K. Koutroumbas, *Pattern Recognition (Fourth Edition)*. 2008.

[96] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[97] D. J. Hand, “Data Clustering: Theory, Algorithms, and Applications by Guojun Gan, Chaoqun Ma, Jianhong Wu,” *Int. Stat. Rev.*, vol. 76, no. 1, pp. 141–141, Apr. 2008.

[98] R. Xu and D. C. Wunsch, “Clustering Algorithms in Biomedical Research: A Review,” *IEEE Rev. Biomed. Eng.*, vol. 3, pp. 120–154, 2010.

[99] G. Kerr, H. J. Ruskin, M. Crane, and P. Doolan, “Techniques for clustering gene expression data,” *Comput. Biol. Med.*, vol. 38, no. 3, pp. 283–293, Mar. 2008.

[100] B. Andreopoulos, A. An, X. Wang, and M. Schroeder, “A roadmap of clustering algorithms: finding a match for a biomedical application.,” *Brief. Bioinform.*, vol. 10, no. 3, pp. 297–314, May 2009.

[101] S. Saria and A. Goldenberg, “Subtyping: What It is and Its Role in Precision Medicine,” *IEEE Intell. Syst.*, vol. 30, no. 4, pp. 70–75, Jul. 2015.

[102] Daxin Jiang, Chun Tang, and Aidong Zhang, “Cluster analysis for gene expression data: a survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[103] A. Saxena *et al.*, “A review of clustering techniques and developments,” *Neurocomputing*, vol. 267, pp. 664–681, Dec. 2017.

[104] D. Xu and Y. Tian, “A Comprehensive Survey of Clustering Algorithms,” *Ann. Data Sci.*, 2015.

[105] S. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[106] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996.

[107] R. Sharan and R. Shamir, “CLICK: a clustering algorithm with applications to gene expression analysis.,” *Proceedings. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 8, pp. 307–16, 2000.

[108] P. Brucker, “On the Complexity of Clustering Problems,” 1978, pp. 45–54.

[109] M. Křivánek and J. Morávek, “NP-hard problems in hierarchical-tree clustering,” *Acta Inform.*, vol. 23, no. 3, pp. 311–323, Jun. 1986.

[110] M. Mahajan, P. Nimbhorkar, and K. Varadarajan, “The planar k-means problem is NP-

hard," in *Theoretical Computer Science*, 2012.

[111] C. A. Sugar and G. M. James, "Finding the Number of Clusters in a Dataset: An Information-Theoretic Approach," *J. Am. Stat. Assoc.*, vol. 98, pp. 750–763, 2003.

[112] P. Berkhin, "A Survey of Clustering Data Mining Techniques," in *Grouping Multidimensional Data*, Berlin/Heidelberg: Springer-Verlag, 2006, pp. 25–71.

[113] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, Jun. 2015.

[114] I. Frades and R. Matthiesen, "Overview on Techniques in Cluster Analysis," Humana Press, 2010, pp. 81–107.

[115] G. Fung, "A comprehensive overview of basic clustering algorithms," *IEEE Trans. Inf. Theory*, vol. 27, no. 1, pp. 49–60, 2001.

[116] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognit.*, vol. 46, no. 1, pp. 243–256, Jan. 2013.

[117] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, Jan. 2009.

[118] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology," *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.

[119] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes.," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[120] T. Kelder *et al.*, "WikiPathways: building research communities on biological pathways.," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D1301-7, Jan. 2012.

[121] R. Shamir *et al.*, "EXPANDER--an integrative program suite for microarray data analysis.," *BMC Bioinformatics*, vol. 6, p. 232, Jan. 2005.

[122] I. Ulitsky *et al.*, "Expander: from expression microarrays to networks and functions.," *Nat. Protoc.*, vol. 5, no. 2, pp. 303–22, Mar. 2010.

[123] D. Netanely, N. Stern, I. Laufer, and R. Shamir, “PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets,” *BMC Bioinformatics*, vol. 20, no. 1, p. 732, Dec. 2019.

[124] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, “GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.,” *BMC Bioinformatics*, vol. 10, no. 1, p. 48, Jan. 2009.

[125] A. Subramanian *et al.*, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Natl. Acad. Sci.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005.

[126] P. Creixell *et al.*, “Pathway and network analysis of cancer genomes,” *Nat. Methods*, vol. 12, no. 7, pp. 615–621, Jul. 2015.

[127] F. Yan, M. Robert, and Y. Li, “Statistical methods and common problems in medical or biomedical science research.,” *Int. J. Physiol. Pathophysiol. Pharmacol.*, vol. 9, no. 5, pp. 157–163, 2017.

[128] S. Bhalerao and S. Parab, “Choosing statistical test,” *Int. J. Ayurveda Res.*, vol. 1, no. 3, p. 187, 2010.

[129] V. Bewick, L. Cheek, and J. Ball, “Statistics review 12: survival analysis.,” *Crit. Care*, vol. 8, no. 5, pp. 389–94, Oct. 2004.

[130] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses,” *Br. J. Cancer*, vol. 89, no. 2, pp. 232–238, Jul. 2003.

[131] E. L. Kaplan and P. Meier, “Nonparametric Estimation from Incomplete Observations,” *J. Am. Stat. Assoc.*, vol. 53, no. 282, p. 457, Jun. 1958.

[132] J. Ranstam and J. A. Cook, “Kaplan-Meier curve,” *Br. J. Surg.*, vol. 104, no. 4, pp. 442–442, Mar. 2017.

[133] R. I. Horwitz, “Statistical aspects of the analysis of data from retrospective studies of disease,” *J. Chronic Dis.*, vol. 32, no. 1–2, p. ii, Jan. 1979.

[134] J. M. Bland and D. G. Altman, “The logrank test.,” *BMJ*, vol. 328, no. 7447, p. 1073, May 2004.

[135] N. Mantel, “Evaluation of survival data and two new rank order statistics arising in its consideration.,” *Cancer Chemother. reports*, vol. 50, no. 3, pp. 163–70, Mar. 1966.

[136] R. Peto and J. Peto, “Asymptotically Efficient Rank Invariant Test Procedures,” *J. R. Stat. Soc. Ser. A*, vol. 135, no. 2, p. 185, 1972.

[137] M. Lucijanic, M. Skelin, and T. Lucijanic, “Survival analysis, more than meets the eye,” *Biochem. Medica*, pp. 14–18, 2017.

[138] B. R. Logan, H. Wang, and M. J. Zhang, “Pairwise multiple comparison adjustment in survival analysis,” *Stat. Med.*, 2005.

[139] D. R. Cox, “Regression Models with Life Tables,” *J. R. Stat. Soc. Ser. B*, vol. 74, pp. 187–220, 1972.

[140] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, “Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods,” *Br. J. Cancer*, vol. 89, no. 3, pp. 431–436, Aug. 2003.

[141] S. J. Walters, “What is a Cox model ?,” *Survival (Lond).*, no. May, pp. 1–8, 2009.

[142] D. C. Koboldt *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, pp. 61–70, Sep. 2012.

[143] H. Zola *et al.*, “CD molecules 2006--human cell differentiation molecules.,” *J. Immunol. Methods*, vol. 319, no. 1–2, pp. 1–5, Jan. 2007.

[144] J. M. Penninger, J. Irie-Sasaki, T. Sasaki, and A. J. Oliveira-dos-Santos, “CD45: new jobs for an old acquaintance.,” *Nat. Immunol.*, vol. 2, no. 5, pp. 389–396, 2001.

[145] M. S. Kuhns, M. M. Davis, and K. C. Garcia, “Deconstructing the Form and Function of the TCR/CD3 Complex,” *Immunity*, vol. 24, no. 2, pp. 133–9, Feb. 2006.

[146] J. H. Kehrl, A. Riva, G. L. Wilson, and C. Thévenin, “Molecular mechanisms regulating CD19, CD20 and CD22 gene expression.,” *Immunol. Today*, vol. 15, no. 9, pp. 432–6, Sep. 1994.

[147] L. Chen and D. B. Flies, “Molecular mechanisms of T cell co-stimulation and co-inhibition.,” *Nat. Rev. Immunol.*, vol. 13, no. 4, pp. 227–42, 2013.

[148] I. Voskoboinik, J. C. Whisstock, and J. A. Trapani, “Perforin and granzymes: function, dysfunction and human pathology.,” *Nat. Rev. Immunol.*, vol. 15, no. 6, pp. 388–400, Jun. 2015.

[149] S. J. F. Cronin and J. M. Penninger, “From T-cell activation signals to signaling control of anti-cancer immunity,” *Immunol. Rev.*, vol. 220, no. 1, pp. 151–168, Dec. 2007.

[150] I. P. Jovanovic *et al.*, “Interleukin-33/ST2 axis promotes breast cancer growth and metastases by facilitating intratumoral accumulation of immunosuppressive and innate lymphoid cells,” *Int. J. Cancer*, vol. 134, no. 7, pp. 1669–1682, Apr. 2014.

[151] S. Tommasi, D. L. Karm, X. Wu, Y. Yen, and G. P. Pfeifer, “Methylation of homeobox genes is a frequent and early epigenetic event in breast cancer.,” *Breast Cancer Res.*, vol. 11, no. 1, p. R14, Jan. 2009.

[152] C. Abate-Shen, “Deregulated homeobox gene expression in cancer: cause or consequence?,” *Nat. Rev. Cancer*, vol. 2, no. 10, pp. 777–785, 2002.

[153] N. Shah and S. Sukumar, “The Hox genes and their roles in oncogenesis.,” *Nat. Rev. Cancer*, vol. 10, no. 5, pp. 361–71, May 2010.

[154] M. Zhao, J. Sun, and Z. Zhao, “TSGene: a web resource for tumor suppressor genes,” *Nucleic Acids Res.*, vol. 41, no. D1, pp. D970–D976, Jan. 2013.

[155] J. G. Herman and S. B. Baylin, “Gene Silencing in Cancer in Association with Promoter Hypermethylation,” *N. Engl. J. Med.*, vol. 349, no. 21, pp. 2042–2054, Nov. 2003.

[156] L. Addou-Klouche *et al.*, “Loss, mutation and deregulation of L3MBTL4 in breast cancers.,” *Mol. Cancer*, vol. 9, p. 213, 2010.

[157] E. Noetzel *et al.*, “Promoter methylation-associated loss of ID4 expression is a marker of tumour recurrence in human breast cancer.,” *BMC Cancer*, vol. 8, p. 154, 2008.

[158] L.-F. Chen, “Tumor suppressor function of RUNX3 in breast cancer.,” *J. Cell. Biochem.*, vol. 113, no. 5, pp. 1470–7, 2012.

[159] B. Huang *et al.*, “RUNX3 acts as a tumor suppressor in breast cancer by targeting estrogen receptor α,” *Oncogene*, vol. 31, no. 4, pp. 527–534, Jan. 2012.

[160] B. Versmold *et al.*, “Epigenetic silencing of the candidate tumor suppressor gene PROX1 in sporadic breast cancer,” *Int J Cancer*, vol. 121, no. 3, pp. 547–554, 2007.

[161] E. Klopocki *et al.*, “Loss of SFRP1 is associated with breast cancer progression and poor prognosis in early stage tumors,” *Int. J. Oncol.*, vol. 25, no. 3, pp. 641–649, Sep. 2004.

[162] M. Goldman *et al.*, “The UCSC Cancer Genomics Browser: update 2015,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D812–D817, Jan. 2015.

[163] B. Li and C. N. Dewey, “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.,” *BMC Bioinformatics*, vol. 12, p. 323, Jan. 2011.

[164] D. Arthur and S. Vassilvitskii, “K-means++: The advantages of careful seeding,” in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007, vol. 07-09-Janu.

[165] D. Netanely, N. Stern, I. Laufer, and R. Shamir, “PROMO: Profiler of Multi-Omics data.” [Online]. Available: http://acgt.cs.tau.ac.il/promo/.

[166] A. Mitchell *et al.*, “The InterPro protein families database: the classification resource after 15 years,” *Nucleic Acids Res.*, vol. 43, no. D1, pp. D213–D221, Jan. 2015.

[167] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.,” *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, Jan. 2009.

[168] L. Eckhart, S. Lippens, E. Tschachler, and W. Declercq, “Cell death by cornification,” *Biochim. Biophys. Acta - Mol. Cell Res.*, vol. 1833, no. 12, pp. 3471–3480, Dec. 2013.

[169] J. Y. Lin and D. E. Fisher, “Melanocyte biology and skin pigmentation,” *Nature*, vol. 445, no. 7130, pp. 843–850, Feb. 2007.

[170] G. Raposo and M. S. Marks, “Melanosomes — dark organelles enlighten endosomal membrane transport,” *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 10, pp. 786–797, Oct. 2007.

[171] R. Lazova and J. M. Pawelek, “Why do melanomas get so dark?,” *Exp. Dermatol.*, vol. 18, no. 11, pp. 934–938, Nov. 2009.

[172] S. Dror *et al.*, “Melanoma miRNA trafficking controls tumour primary niche formation,”

Nat. Cell Biol., vol. 18, no. 9, pp. 1006–1017, Sep. 2016.

[173] K. G. Chen *et al.*, "Melanosomal sequestration of cytotoxic drugs contributes to the intractability of malignant melanomas," *Proc. Natl. Acad. Sci.*, vol. 103, no. 26, pp. 9903–9907, Jun. 2006.

[174] S. D'Mello, G. Finlay, B. Baguley, and M. Askarian-Amiri, "Signaling Pathways in Melanogenesis," *Int. J. Mol. Sci.*, vol. 17, no. 7, p. 1144, Jul. 2016.

[175] C. Levy, M. Khaled, and D. E. Fisher, "MITF: master regulator of melanocyte development and melanoma oncogene," *Trends Mol. Med.*, vol. 12, no. 9, pp. 406–414, Sep. 2006.

[176] J. F. Berson, D. C. Harper, D. Tenza, G. Raposo, and M. S. Marks, "Pmel17 Initiates Premelanosome Morphogenesis within Multivesicular Bodies," *Mol. Biol. Cell*, vol. 12, no. 11, pp. 3451–3464, Nov. 2001.

[177] R. E. Bell *et al.*, "Enhancer methylation dynamics contribute to cancer plasticity and patient mortality," *Genome Res.*, vol. 26, no. 5, pp. 601–611, May 2016.

[178] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: an overview and their use in medicine.," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–63, Oct. 2002.

[179] W. K. Martins *et al.*, "Gene network analyses point to the importance of human tissue kallikreins in melanoma progression," *BMC Med. Genomics*, vol. 4, no. 1, p. 76, Dec. 2011.

[180] X. Liu *et al.*, "Elevated expression of KLK8 predicts poor prognosis in colorectal cancer," *Biomed. Pharmacother.*, vol. 88, pp. 595–602, Apr. 2017.

[181] Y.-P. Sher *et al.*, "Human Kallikrein 8 Protease Confers a Favorable Clinical Outcome in Non-Small Cell Lung Cancer by Suppressing Tumor Cell Invasiveness," *Cancer Res.*, vol. 66, no. 24, pp. 11763–11770, Dec. 2006.

[182] C. A. Borgono, "Human Kallikrein 8 Protein Is a Favorable Prognostic Marker in Ovarian Cancer," *Clin. Cancer Res.*, vol. 12, no. 5, pp. 1487–1493, Mar. 2006.

[183] N. A. Manieri, E. Y. Chiang, and J. L. Grogan, "TIGIT: A Key Inhibitor of the Cancer Immunity Cycle," *Trends Immunol.*, vol. 38, no. 1, pp. 20–28, Jan. 2017.

[184] J.-M. Chauvin *et al.*, "TIGIT and PD-1 impair tumor antigen–specific CD8+ T cells in

melanoma patients," *J. Clin. Invest.*, vol. 125, no. 5, pp. 2046–2058, May 2015.

[185] F. Rambow *et al.*, "New Functional Signatures for Understanding Melanoma Biology from Tumor Cell Lineage-Specific Analysis," *Cell Rep.*, vol. 13, no. 4, pp. 840–853, Oct. 2015.

[186] M. Tomihari, S.-H. Hwang, J.-S. Chung, P. D. Cruz Jr., and K. Ariizumi, "Gpnmb is a melanosome-associated glycoprotein that contributes to melanocyte/keratinocyte adhesion in a RGD-dependent fashion," *Exp. Dermatol.*, vol. 18, no. 7, pp. 586–595, Jul. 2009.

[187] N. W. Bellono, I. E. Escobar, A. J. Lefkovith, M. S. Marks, and E. Oancea, "An intracellular anion channel critical for pigmentation," *Elife*, 2014.

[188] "UCSC XENA." [Online]. Available: http://xena.ucsc.edu/.

[189] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. FL: CRC Press, 1984.

[190] W. Y. Loh and Y. S. Shin, "Split selection methods for classification trees," *Stat. Sin.*, vol. 7, no. 4, pp. 815–840, 1997.

[191] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-Throughput Sequencing Technologies," *Mol. Cell*, vol. 58, no. 4, pp. 586–597, May 2015.

[192] L. E. MacConaill, "Existing and Emerging Technologies for Tumor Genomic Profiling," *J. Clin. Oncol.*, vol. 31, no. 15, pp. 1815–1824, May 2013.

[193] S. Roychowdhury and A. M. Chinnaiyan, "Translating cancer genomes and transcriptomes for precision oncology," *CA. Cancer J. Clin.*, vol. 66, no. 1, pp. 75–88, Jan. 2016.

[194] J. E. McDermott *et al.*, "Challenges in biomarker discovery: combining expert insights with statistical analysis of complex omics data," *Expert Opin. Med. Diagn.*, vol. 7, no. 1, pp. 37–51, Jan. 2013.

[195] A. Alyass, M. Turcotte, and D. Meyre, "From big data analysis to personalized medicine for all: challenges and opportunities," *BMC Med. Genomics*, vol. 8, no. 1, p. 33, Dec. 2015.

[196] "The TCGA Legacy.," *Cell*, vol. 173, no. 2, pp. 281–282, Apr. 2018.

[197] E. R. Mardis, "The $1,000 genome, the $100,000 analysis?," *Genome Med.*, vol. 2, no. 11,

p. 84, 2010.

[198] D. Netanely, A. Avraham, A. Ben-Baruch, E. Evron, and R. Shamir, “Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups,” *Breast Cancer Res.*, vol. 18, no. 1, p. 74, Dec. 2016.

[199] R. Edgar, M. Domrachev, and A. E. Lash, “Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 207–10, Jan. 2002.

[200] J. Zhu, B. Craft, M. Goldman, M. Cline, M. Diekhans, and D. Haussler, “Using the UCSC Xena Platform to integrate, visualize, and analyze your own data in the context of large external genomic datasets,” *Cancer Res.*, vol. 75, no. 22, 2015.

[201] M. Goldman *et al.*, “The UCSC Xena platform for public and private cancer genomics data visualization and interpretation,” *bioRxiv*, p. 326470, Mar. 2019.

[202] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, Jul. 2010.

[203] K. Y. Yeung and W. L. Ruzzo, “Principal component analysis for clustering gene expression data,” *Bioinformatics*, vol. 17, no. 9, pp. 763–774, Sep. 2001.

[204] C. R. García-Alonso, L. M. Pérez-Naranjo, and J. C. Fernández-Caballero, “Multiobjective evolutionary algorithms to identify highly autocorrelated areas: the case of spatial distribution in financially compromised farms,” *Ann. Oper. Res.*, vol. 219, no. 1, pp. 187–202, Aug. 2014.

[205] E. A. Vucic *et al.*, “Translating cancer ‘omics’ to improved outcomes,” *Genome Res.*, vol. 22, no. 2, pp. 188–195, Feb. 2012.

[206] B. Wang *et al.*, “Similarity network fusion for aggregating data types on a genomic scale.,” *Nat. Methods*, vol. 11, no. 3, pp. 333–7, Mar. 2014.

[207] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, “Consensus clustering: A resampling based method for class discovery and visualization of gene expression microarray data,” *Mach. Learn.*, vol. 52, no. i, pp. 91–118, 2003.

[208] S. Tyanova *et al.*, “The Perseus computational platform for comprehensive analysis of

(prote)omics data," *Nat. Methods*, vol. 13, no. 9, pp. 731–740, 2016.

[209] S. Sinha, J. Song, R. Weinshilboum, V. Jongeneel, and J. Han, "KnowEnG: a knowledge engine for genomics.," *J. Am. Med. Inform. Assoc.*, vol. 22, no. 6, pp. 1115–9, Nov. 2015.

[210] A. Sangaralingam *et al.*, "'Multi-omic' data analysis using O-miner.," *Brief. Bioinform.*, vol. 20, no. 1, pp. 130–143, Jan. 2019.

[211] "Genomic Data Commons Data Portal." [Online]. Available: https://portal.gdc.cancer.gov/.

[212] "ICGC Data Portal." [Online]. Available: https://dcc.icgc.org/.

[213] M. A. Jensen, V. Ferretti, R. L. Grossman, and L. M. Staudt, "The NCI Genomic Data Commons as an engine for precision medicine," *Blood*, vol. 130, no. 4, pp. 453–459, Jul. 2017.

[214] S. K. Yeo and J.-L. Guan, "Breast Cancer: Multiple Subtypes within a Tumor?," *Trends in Cancer*, vol. 3, no. 11, pp. 753–760, Nov. 2017.

[215] C. Swanton, "Intratumour Heterogeneity: Evolution through Space and Time," *Cancer Res.*, vol. 72, no. 19, p. 4875, 2012.

[216] N. Kumar, D. Zhao, D. Bhaumik, A. Sethi, and P. H. Gann, "Quantification of intrinsic subtype ambiguity in Luminal A breast cancer and its relationship to clinical outcomes," *BMC Cancer*, vol. 19, no. 1, p. 215, Dec. 2019.

[217] A. Prat *et al.*, "Clinical implications of the intrinsic molecular subtypes of breast cancer," *Breast*, vol. 24, pp. S26–S35, 2015.

[218] M. Alizart, J. Saunus, M. Cummings, and S. R. Lakhani, "Molecular classification of breast carcinoma," *Diagnostic Histopathol.*, vol. 18, no. 3, pp. 97–103, Mar. 2012.

[219] F. Balkwill, "Cancer and the chemokine network.," *Nat. Rev. Cancer*, vol. 4, no. 7, pp. 540–50, Jul. 2004.

[220] F. Balkwill, "Chemokine biology in cancer.," *Semin. Immunol.*, vol. 15, no. 1, pp. 49–55, Feb. 2003.

[221] G. Luboshits *et al.*, "Elevated expression of the CC chemokine regulated on activation,

normal T cell expressed and secreted (RANTES) in advanced breast carcinoma.," *Cancer Res.*, vol. 59, no. 18, pp. 4681–7, Sep. 1999.

[222] R. Salgado *et al.*, "The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014," *Ann. Oncol.*, vol. 26, no. 2, pp. 259–271, Sep. 2014.

[223] C. Denkert *et al.*, "Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer.," *J. Clin. Oncol.*, vol. 28, no. 1, pp. 105–13, Jan. 2010.

[224] C. Denkert, "The immunogenicity of breast cancer--molecular subtypes matter.," *Ann. Oncol.*, vol. 25, no. 8, pp. 1453–5, Aug. 2014.

[225] D. Bedognetti, W. Hendrickx, F. M. Marincola, and L. D. Miller, "Prognostic and predictive immune gene signatures in breast cancer.," *Curr. Opin. Oncol.*, vol. 27, no. 6, pp. 433–44, 2015.

[226] O. a. Stefansson *et al.*, "A DNA methylation-based definition of biologically distinct breast cancer subtypes," *Mol. Oncol.*, vol. 9, pp. 555–568, Nov. 2015.

[227] S. Kamalakaran *et al.*, "DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables.," *Mol. Oncol.*, vol. 5, no. 1, pp. 77–92, Feb. 2011.

[228] D. Nejman *et al.*, "Molecular rules governing de novo methylation in cancer," *Cancer Res.*, vol. 74, no. 5, pp. 1475–1483, 2014.

[229] G. Ciriello *et al.*, "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer," *Cell*, vol. 163, no. 2, pp. 506–519, Oct. 2015.

[230] M. Michaut *et al.*, "Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer," *Sci. Rep.*, vol. 6, p. 18517, Jan. 2016.

[231] K. G. Chen *et al.*, "Influence of Melanosome Dynamics on Melanoma Drug Sensitivity," *JNCI J. Natl. Cancer Inst.*, vol. 101, no. 18, pp. 1259–1271, Sep. 2009.

[232] K. G. Chen, J. C. Valencia, J.-P. Gillet, V. J. Hearing, and M. M. Gottesman, "Involvement

of ABC transporters in melanogenesis and the development of multidrug resistance of melanoma," *Pigment Cell Melanoma Res.*, vol. 22, no. 6, pp. 740–749, Dec. 2009.

[233] K. G. Chen and M. M. Gottesman, "How Melanoma Cells Evade Chemotherapy," in *From Melanocytes to Melanoma*, Totowa, NJ: Humana Press, 2007, pp. 591–603.

[234] G. Jönsson *et al.*, "Gene expression profiling-based identification of molecular subtypes in stage IV melanomas with different clinical outcome.," *Clin. Cancer Res.*, vol. 16, no. 13, pp. 3356–67, Jul. 2010.

[235] J. Nsengimana *et al.*, "Independent replication of a melanoma subtype gene signature and evaluation of its prognostic value and biological correlates in a population cohort.," *Oncotarget*, vol. 6, no. 13, pp. 11683–93, May 2015.

[236] H. Cirenajwis *et al.*, "Molecular stratification of metastatic melanoma using gene expression profiling : Prediction of survival outcome and benefit from molecular targeted therapy," *Oncotarget*, vol. 6, no. 14, pp. 12297–12309, May 2015.

[237] M. Lauss, J. Nsengimana, J. Staaf, J. Newton-Bishop, and G. Jönsson, "Consensus of Melanoma Gene Expression Subtypes Converges on Biological Entities," *J. Invest. Dermatol.*, vol. 136, no. 12, pp. 2502–2505, Dec. 2016.

7. Supplementary Information

7.1. Supplement 1: Breast cancer subtypes

7.1.1. Datasets used and global RNA-Seq dataset analysis (Normal + Tumor)

Datasets

TCGA's Breast Cancer datasets were downloaded from the UCSC cancer browser website in March 2015.

Technology	Dataset title	DatasetID	#Samples	Dataset version
Gene Expression - RNA-Seq	TCGA breast invasive carcinoma (BRCA) gene expression by RNAseq (IlluminaHiSeq)	BRCA gene expression (IlluminaHiSeq)	1215	2015-02-24
DNA-Methylation array	TCGA breast invasive carcinoma (BRCA) (HumanMethylation450)	BRCA (Methylation450k)	872	2015-02-24
Gene Expression - MicroArrays	TCGA breast invasive carcinoma (BRCA) gene expression (AgilentG4502A_07_3 array)	BRCA gene expression (AgilentG4502A_07_3)	597	2015-02-24

Table S1.1A: Properties of datasets used in the study.

Obtaining the RNA-Seq dataset and initial sample preprocessing

RSEM normalized version of TCGA's BRCA RNA-SEQ expression dataset was used in the following analyses. Updated RNA-SEQ based PAM50 calls for TCGA BRCA samples were obtained from UNC University.

Sample preprocessing: Downloaded dataset contained 1215 samples of which the following were removed based on supplied labels: 19 – Unknown tissue site, 11 Male, 7 metastatic samples, 30 Unavailable UNC_Pam50 labels. The preprocessed dataset contained 1148 samples, of which 113 are normal based on the 'sample type' field, and 150 are normal based on 'PAM50 call'.

Distribution of PAM50 calls on the preprocessed RNA-Seq expression dataset

Total number of samples after preprocessing: 1148

PAM50 label	Number of samples
Basal	183
Her2	78
LumA	534
LumB	203
Normal	150
Total	1148

Table S1.1B: distribution of PAM50 labels on TCGA's RNA-Seq dataset

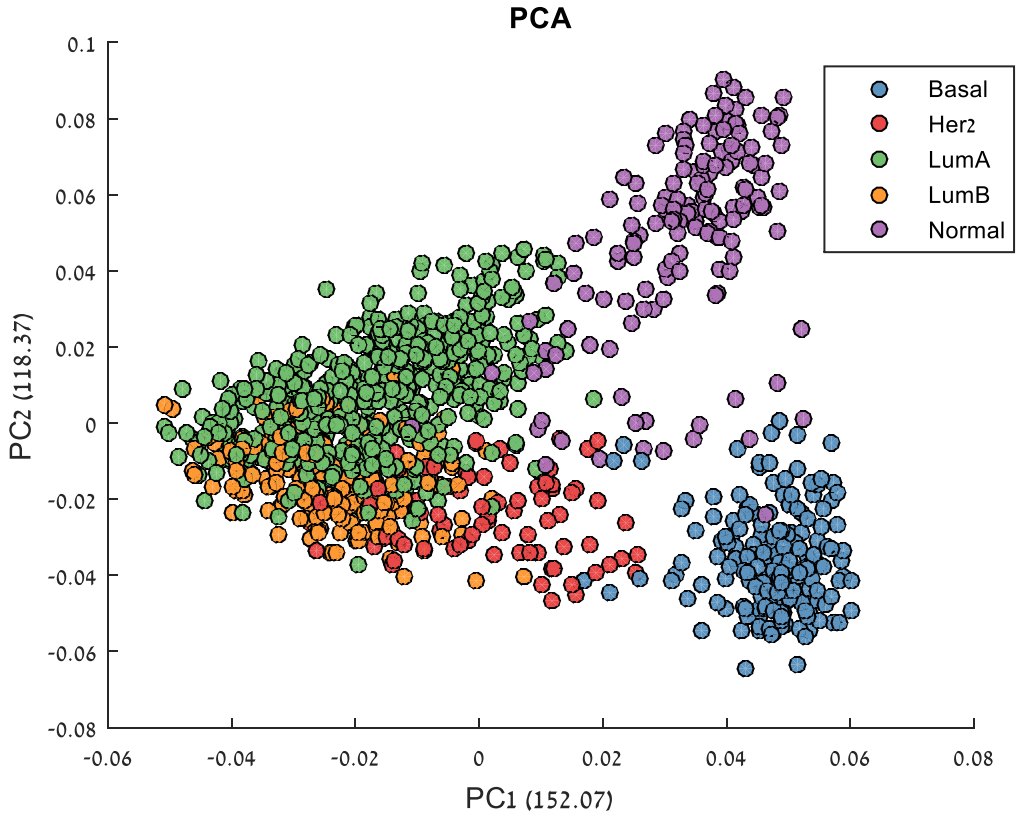


Figure S1.1A: PCA of 1148 breast samples based on 2000 top variable genes, colored by PAM50 labels

Clustering the samples based on RNA-SEQ data

The K-Means clustering algorithm was executed on the 1148 samples using the 2000 top variable genes. Matlab v8.5 implementation of the K-Means algorithm was used using correlation-based distance metric, and 100 replicates. Rows (genes) were standardized before the sample clustering.

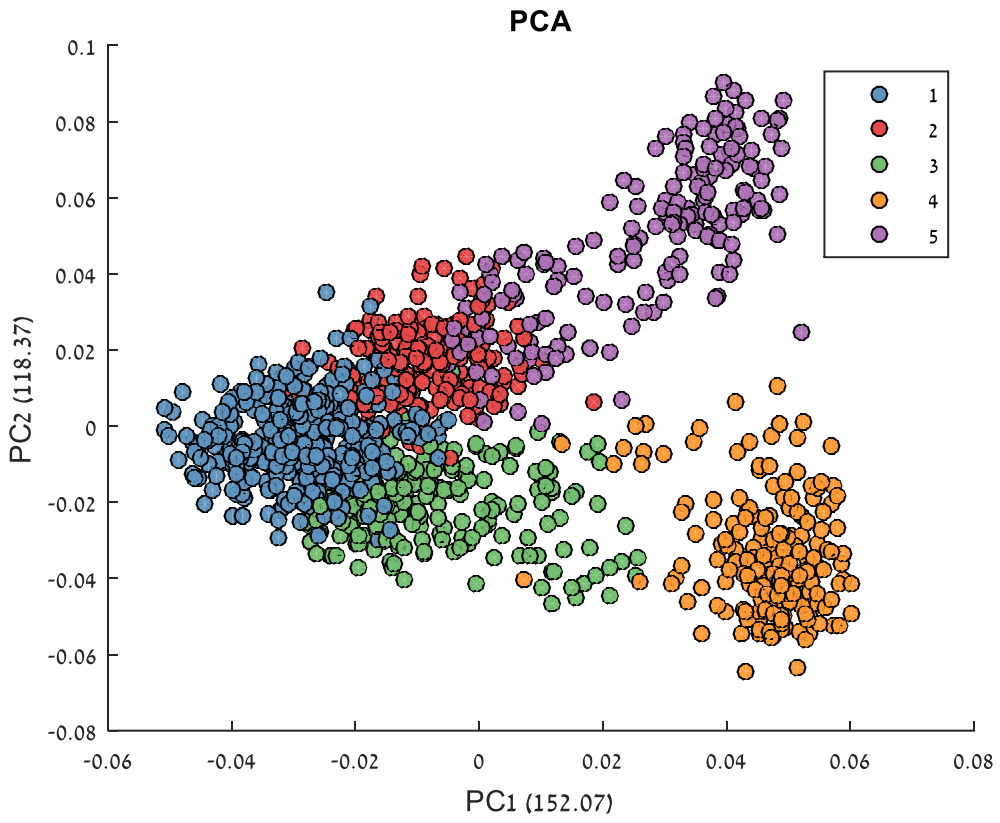


Figure S1.1B: PCA of 1148 breast samples based on 2000 top variable genes, colored by K-Means clusters.

	RNA-Seq Clusters	Total	1	2	3	4	5
		n=1148	n=360	n=217	n=201	n=193	n=177
Age (Median)		58	62	57	58	53	54
ER Status	NA	160 (14%)	28 (8%)	2 (1%)	8 (4%)	7 (4%)	115 (65%)
	Negative	227 (20%)	3 (1%)	7 (3%)	53 (26%)	159 (82%)	5 (3%)
	Positive	761 (66%)	329 (91%)	208 (96%)	140 (70%)	27 (14%)	57 (32%)
PR Status	NA	163 (14%)	29 (8%)	4 (2%)	6 (3%)	9 (5%)	115 (65%)
	Negative	326 (28%)	35 (10%)	21 (10%)	86 (43%)	171 (89%)	13 (7%)
	Positive	659 (57%)	296 (82%)	192 (88%)	109 (54%)	13 (7%)	49 (28%)
Her2 Status	NA	391 (34%)	81 (23%)	73 (34%)	37 (18%)	57 (30%)	143 (81%)
	Negative	649 (57%)	260 (72%)	137 (63%)	89 (44%)	132 (68%)	31 (18%)
	Positive	108 (9%)	19 (5%)	7 (3%)	75 (37%)	4 (2%)	3 (2%)
PAM50	Basal	183 (16%)	0 (0%)	0 (0%)	3 (1%)	180 (93%)	0 (0%)
	Her2	78 (7%)	0 (0%)	0 (0%)	78 (39%)	0 (0%)	0 (0%)
	LumA	534 (47%)	242 (67%)	212 (98%)	37 (18%)	0 (0%)	43 (24%)
	LumB	203 (18%)	117 (33%)	5 (2%)	80 (40%)	1 (1%)	0 (0%)
	Normal	150 (13%)	1 (0%)	0 (0%)	3 (1%)	12 (6%)	134 (76%)
Pathologic stage	NA	120 (10%)	1 (0%)	2 (1%)	2 (1%)	2 (1%)	113 (64%)
	Stage I	176 (15%)	64 (18%)	53 (24%)	18 (9%)	29 (15%)	12 (7%)
	Stage II	589 (51%)	202 (56%)	108 (50%)	118 (59%)	134 (69%)	27 (15%)
	Stage III	234 (20%)	81 (23%)	49 (23%)	58 (29%)	23 (12%)	23 (13%)
	Stage IV	16 (1%)	5 (1%)	1 (0%)	5 (2%)	4 (2%)	1 (1%)
	Stage X	13 (1%)	7 (2%)	4 (2%)	0 (0%)	1 (1%)	1 (1%)
Histological type	Infiltrating Ductal Carcinoma	753 (66%)	272 (76%)	107 (49%)	182 (91%)	166 (86%)	26 (15%)
	Infiltrating Lobular Carcinoma	182 (16%)	40 (11%)	95 (44%)	11 (5%)	1 (1%)	35 (20%)
	Medullary Carcinoma	5 (0%)	0 (0%)	0 (0%)	1 (0%)	4 (2%)	0 (0%)
	Metaplastic Carcinoma	4 (0%)	0 (0%)	0 (0%)	1 (0%)	3 (2%)	0 (0%)
	Mixed Histology	29 (3%)	15 (4%)	8 (4%)	3 (1%)	1 (1%)	2 (1%)
	Mucinous Carcinoma	16 (1%)	15 (4%)	0 (0%)	1 (0%)	0 (0%)	0 (0%)
	NA	159 (14%)	18 (5%)	7 (3%)	2 (1%)	18 (9%)	114 (64%)

Table S1.1C: Cohort description for the global RNA-Seq dataset analysis (Normal + Tumor)

Comparing resulting clusters to PAM50 labels

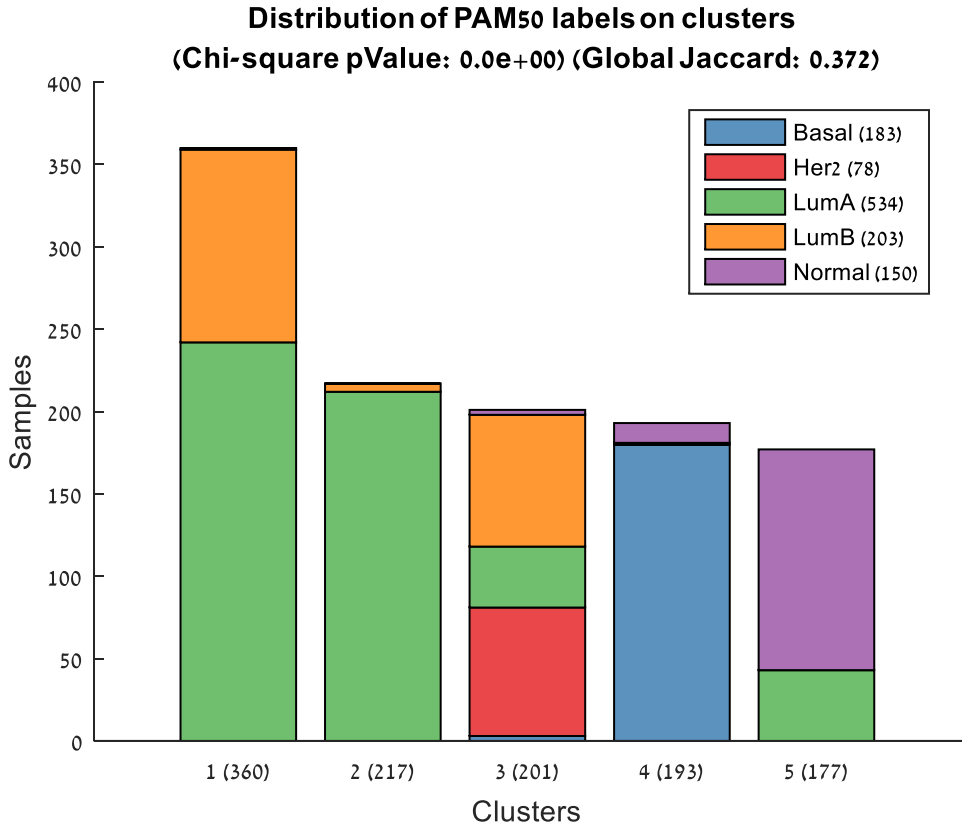


Figure S1.1C: Distribution of PAM50 labels among sample clusters

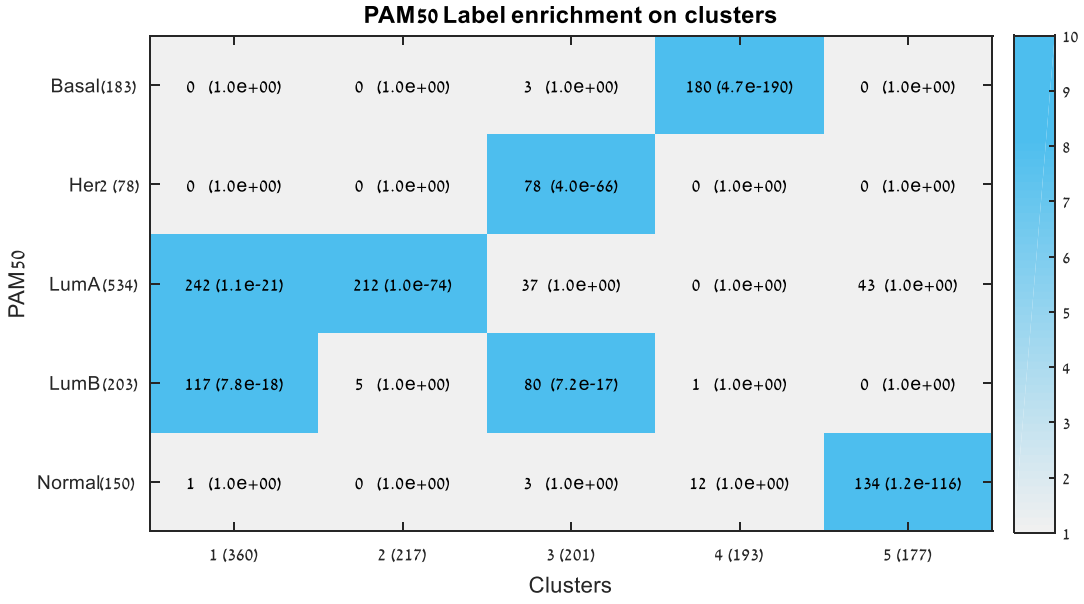


Figure S1.1D: p-values of the hypergeometric enrichment of resulting clusters for PAM50 labels

Evaluation of expression distribution in cluster 1 samples versus cluster 2 samples

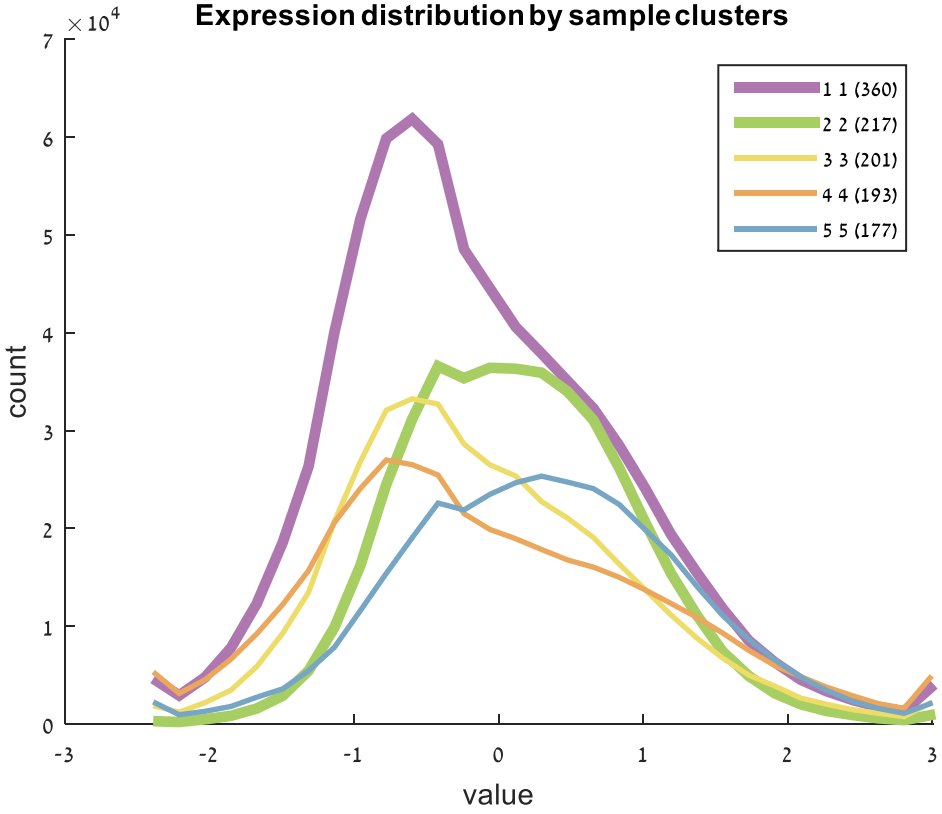


Figure S1.1E: Distribution of normalized expression values by sample cluster.

When applying rank-sum test on the top 2000 variables genes, testing for difference in means between samples of cluster 1 (n=360) and samples of cluster 2(n=217), 1421 genes out of 2000 passed the test with p-value<0.01.

All genes passing the test	1421
Genes overexpressed on cluster1 compared to cluster2	229
Genes overexpressed on cluster2 compared to cluster1	1184
Genes with FC==0	8

Table S1.1D: Analysis of differentially expressed genes

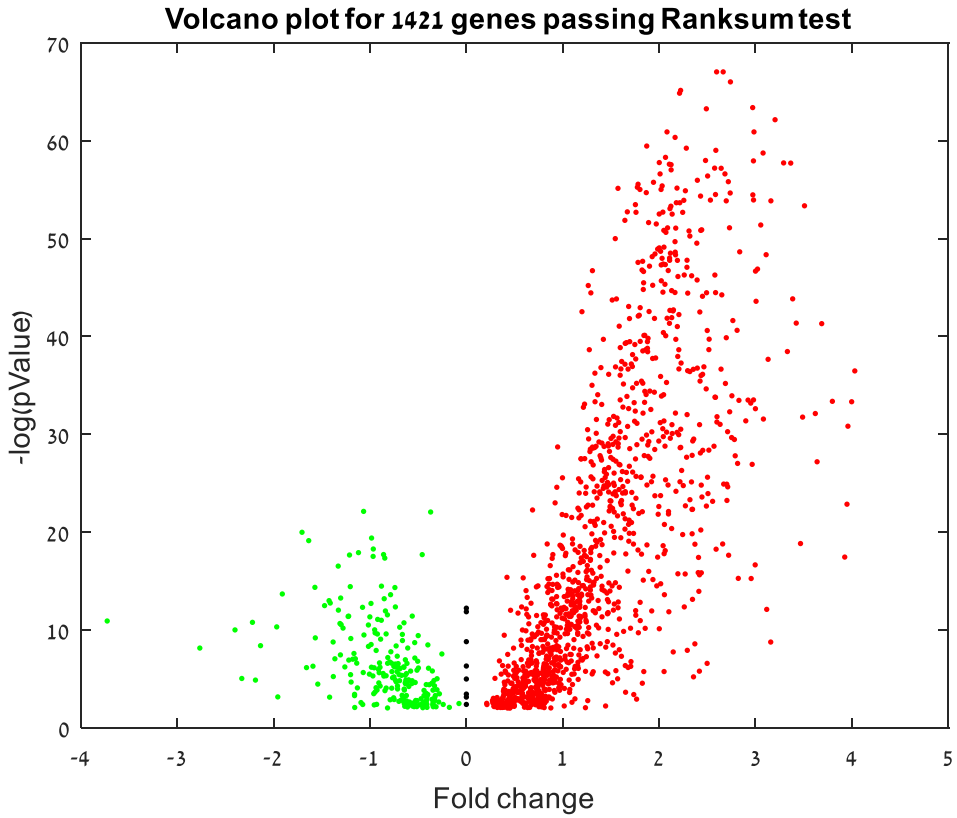


Figure S1.1F: Differentially expressed genes between cluster 1 and cluster 2

7.1.2. RNA-Seq luminal samples analysis

Zooming into the luminal samples, we applied unsupervised analysis only on samples labeled as either luminal-A or luminal-B by PAM50.

Sample preprocessing: In this step of the analysis we started with the 1215 samples included in the TCGA's BRCA RNA-Seq dataset and removed the following samples: 19 – Unknown tissue site, 11 Male, 7 metastatic, 30 Unavailable PAM50 labels, 113 normal sample type, 37 normal by PAM50. From the Remaining with 988 samples we kept only the 737 luminal samples (534 luminal-A and 203 luminal-B based on PAM50 labels).

Gene preprocessing: We kept only the top 2000 variable genes over the 737 luminal samples.

Unsupervised method: As described in the previous section, K-Means (distance metric: correlation) with K=2 applied on the 737 samples using the 2000 top variable genes (after row standardization).

We then compared the sample partition induced by our clustering, to the PAM50 luminal-A/luminal-B partition using log-rank tests and show that our RNA-Seq based partition

outperforms PAM50's partition in terms of both survival and recurrence, and in both 5-year and overall time spans.

OVERALL Survival and Recurrence

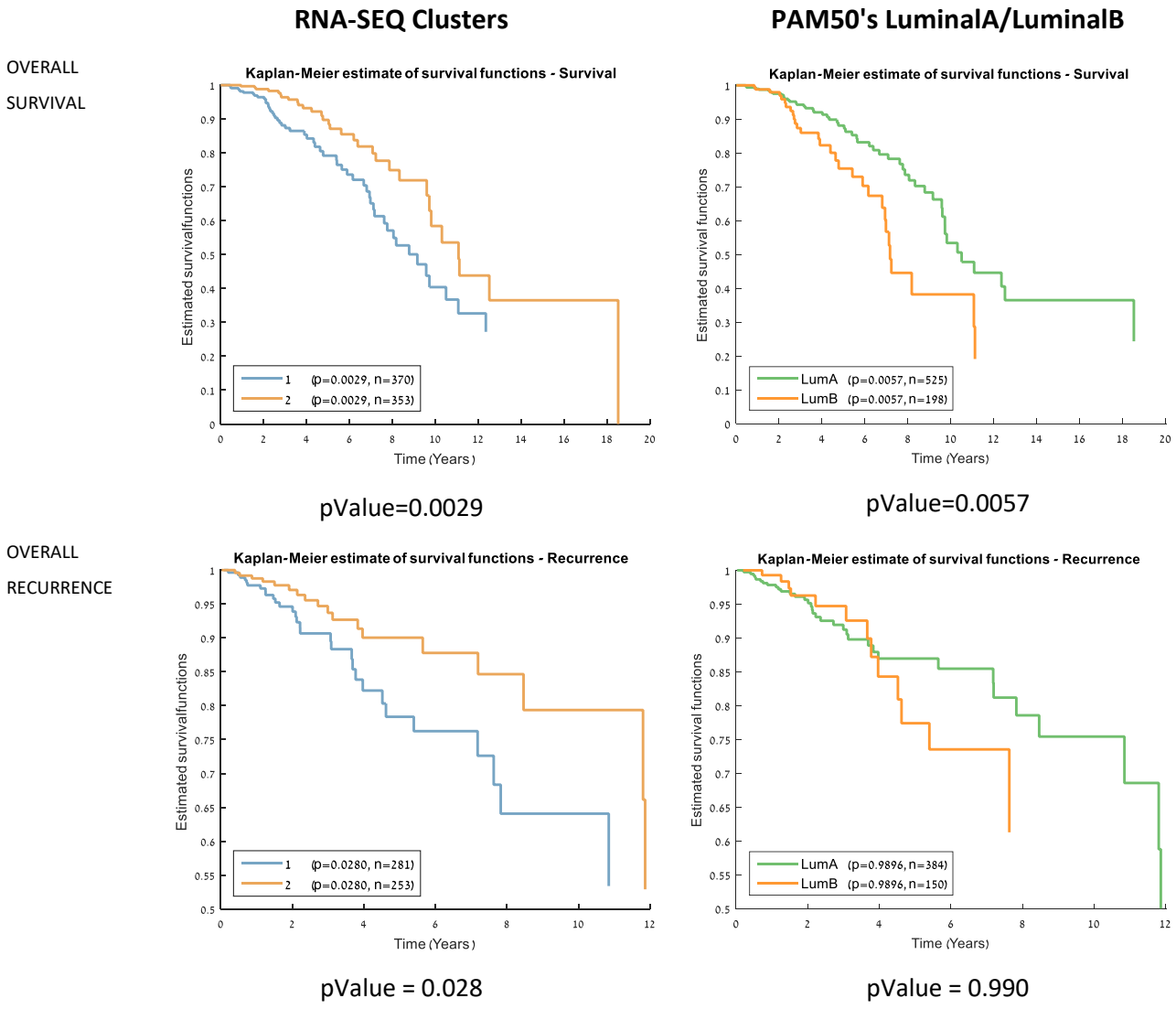


Figure S1.2A: Overall survival and recurrence plots for K-Means clusters versus PAM50 luminal-A/luminal-B classification

	RNA-Seq Clusters	Total	1	2
		n=737	n=382	n=355
Age (Median)		60	62	56
ER Status	NA	33 (4%)	27 (7%)	6 (2%)
	Negative	14 (2%)	4 (1%)	10 (3%)
	Positive	690 (94%)	351 (92%)	339 (95%)
PR Status	NA	36 (5%)	28 (7%)	8 (2%)
	Negative	87 (12%)	44 (12%)	43 (12%)
	Positive	614 (83%)	310 (81%)	304 (86%)
Her2 Status	NA	195 (26%)	77 (20%)	118 (33%)
	Negative	486 (66%)	270 (71%)	216 (61%)
	Positive	56 (8%)	35 (9%)	21 (6%)
PAM50	LumA	534 (72%)	207 (54%)	327 (92%)
	LumB	203 (28%)	175 (46%)	28 (8%)
Pathologic stage	NA	3 (0%)	1 (0%)	2 (1%)
	Stage I	137 (19%)	54 (14%)	83 (23%)
	Stage II	396 (54%)	224 (59%)	172 (48%)
	Stage III	179 (24%)	90 (24%)	89 (25%)
	Stage IV	10 (1%)	6 (2%)	4 (1%)
	Stage X	12 (2%)	7 (2%)	5 (1%)
Histological type	Infiltrating Ductal Carcinoma	504 (68%)	307 (80%)	197 (55%)
	Infiltrating Lobular Carcinoma	163 (22%)	29 (8%)	134 (38%)
	Medullary Carcinoma	1 (0%)	0 (0%)	1 (0%)
	Mixed Histology	27 (4%)	15 (4%)	12 (3%)
	Mucinous Carcinoma	16 (2%)	15 (4%)	1 (0%)
	NA	26 (4%)	16 (4%)	10 (3%)

Table S1.2A: Cohort description for the luminal RNA-Seq dataset analysis

7.1.3. RNA-Seq luminal-A sample analysis

Next, we clustered only the luminal-A samples. Similarly to the previous section, we filtered the samples further by removing the 203 luminal-B samples based on PAM50 labels. 534 luminal-A samples remained. K-Means (distance metric: correlation) with K=2 was applied to the 534 samples using the 2000 top variable genes.

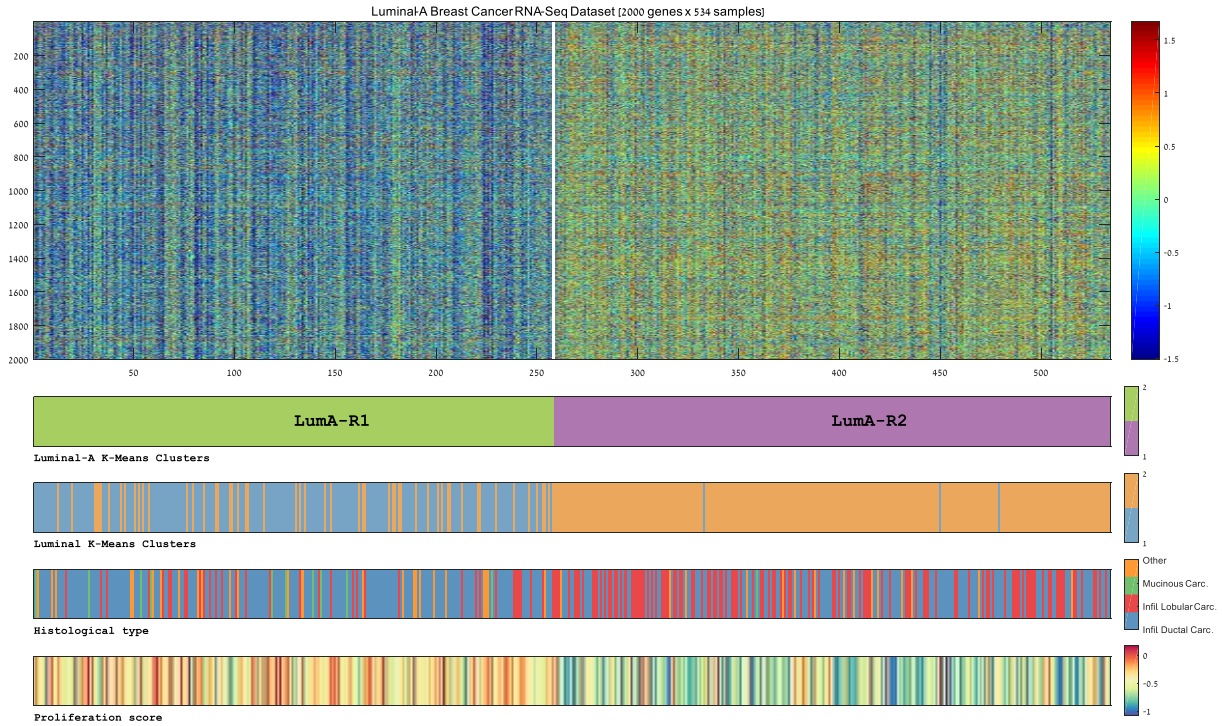


Figure S1.3A: Clustering of the luminal-A samples into two clusters

	RNA-Seq Clusters	Total	1	2
		n=534	n=258	n=276
Age (Median)		60	62	57
ER Status	NA	21 (4%)	18 (7%)	3 (1%)
	Negative	11 (2%)	3 (1%)	8 (3%)
	Positive	502 (94%)	237 (92%)	265 (96%)
PR Status	NA	24 (4%)	19 (7%)	5 (2%)
	Negative	50 (9%)	21 (8%)	29 (11%)
	Positive	460 (86%)	218 (84%)	242 (88%)
Her2 Status	NA	161 (30%)	62 (24%)	99 (36%)
	Negative	347 (65%)	181 (70%)	166 (60%)
	Positive	26 (5%)	15 (6%)	11 (4%)
PAM50	LumA	534 (100%)	258 (100%)	276 (100%)
Pathologic stage	NA	3 (1%)	1 (0%)	2 (1%)
	Stage I	113 (21%)	52 (20%)	61 (22%)
	Stage II	282 (53%)	144 (56%)	138 (50%)
	Stage III	121 (23%)	52 (20%)	69 (25%)
	Stage IV	6 (1%)	4 (2%)	2 (1%)
	Stage X	9 (2%)	5 (2%)	4 (1%)
Histological type	Infiltrating Ductal Carcinoma	331 (62%)	188 (73%)	143 (52%)
	Infiltrating Lobular Carcinoma	152 (28%)	35 (14%)	117 (42%)
	Mixed Histology	21 (4%)	12 (5%)	9 (3%)
	Mucinous Carcinoma	11 (2%)	10 (4%)	1 (0%)
	NA	19 (4%)	13 (5%)	6 (2%)

Table S1.3A: Cohort description for the luminal-A RNA-Seq dataset analysis

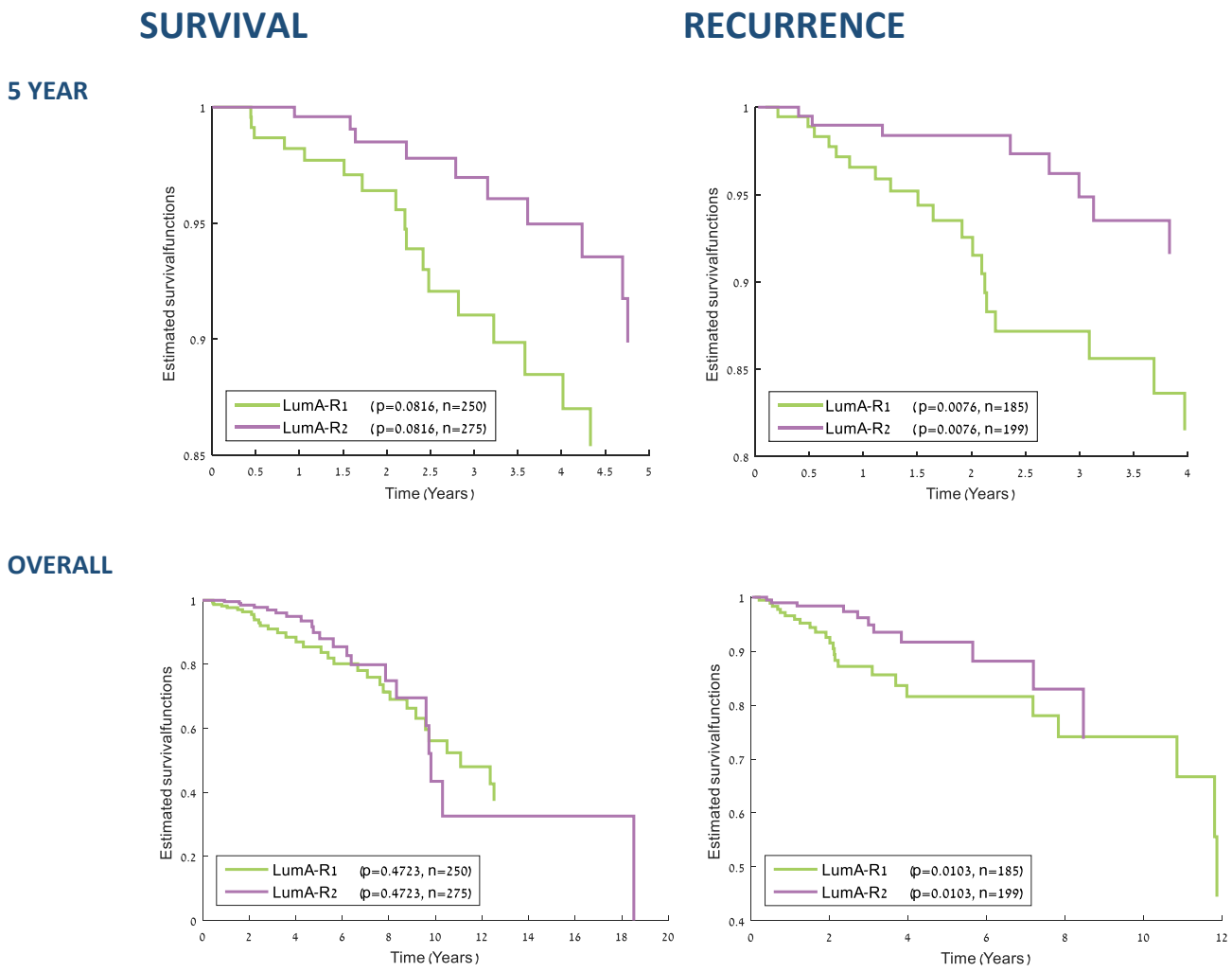


Figure S1.3B: Survival and Recurrence analysis for the 2 luminal–A subgroups

Cluster LumA-R2 samples exhibit distinct overexpression pattern

Applied rank-sum test on the top 2000 variables genes, testing for a difference in means between samples of LumA-R1 (n=258) and LumA-R2 samples (n=276). 1276 genes out of 2000 passed the test with pValue<0.01.

Total number of genes passing the rank sum test	1276
Genes over expressed on cluster1 compared to cluster2	194
Genes over expressed on cluster2 compared to cluster1	1068
Genes with zero fold change	5

Table S1.3B: Analysis of differentially expressed genes

Effect of changing the value of K in the clustering

To test the effect of choice of K in the K-means algorithm on the expression-based subtyping results, we clustered the 1027 breast cancer expression profiles (excluding the normal samples) with both k=4 and k=5. The heatmap in the following figure shows the k=5 clustering, and bar A identifies the five subgroups. Bar B shows the subgroup of each sample when clustering with k=4. Clusters 1, 4 and 5 on the K=5 clustering correspond almost perfectly to clusters 3, 1 and 2 on the K=4 clustering, respectively. Sample cluster 4 on the K=4 clustering (containing a mixture of LumA and LumB samples), was split into clusters 2 (mostly LumB) and 3 (mostly LumA) on the K=5 clustering. Panel D shows the PAM50 classification with the LumA category split into the subgroups LumA-R1 and LumA-R2 revealed in this study. We see that the leftmost sample cluster in A and B, which was identified with both k values, captures very well the 'LumA-R2' samples. This additional analysis demonstrates the stability of our clustering results: the split of the LumA samples (and especially the identification of the LumA-R2 subgroup, the orange cluster on bar D) is repeatedly reproduced when applying clustering on various sample subsets, various feature subsets (the top variable genes on each sample subsets) and various values of K.

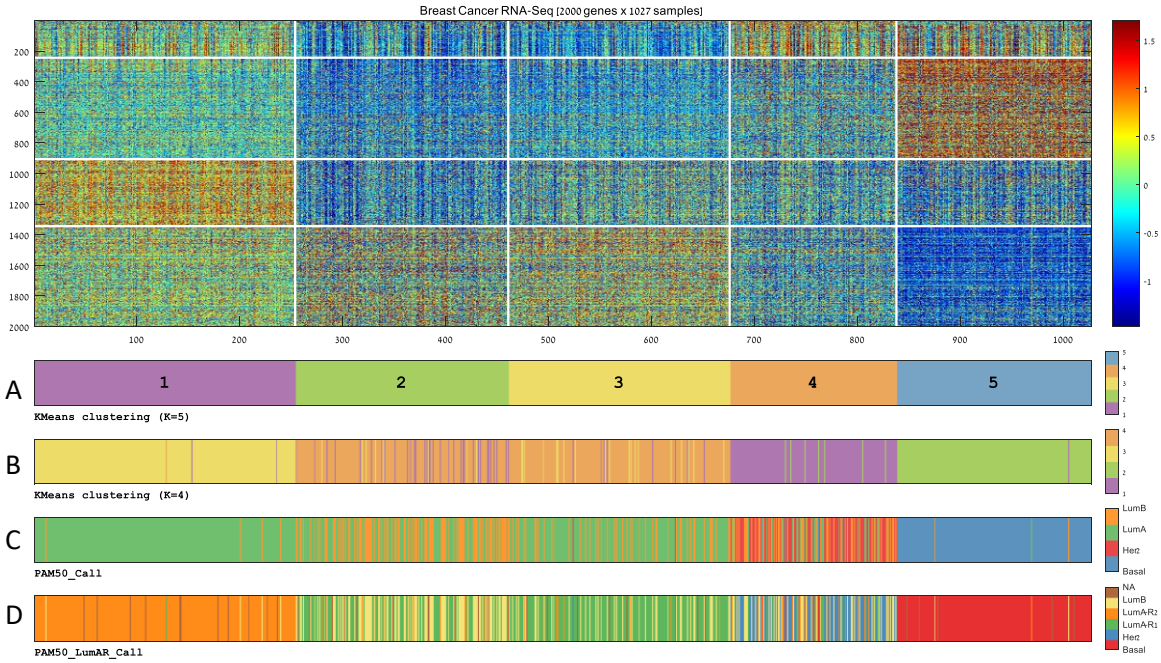


Figure S1.3C: Comparison of clustering of 1027 tumor samples using K=4 and K=5. The heatmap shows the results of K-Means using K=5 on the expression profiles of 1027 breast tumors. (A) K-means clustering with K=5. (B) K-means clustering with K=4. (C) PAM50 calls. (D) PAM50 calls with the LumA class split into the two new LumA subgroups: LumA-R1 and LumA-R2.

7.1.4. Validation of luminal-A partition on microarray gene expression data

In order to verify that the two luminal-A subgroups that were identified using the RNA-Seq data represent real biological variance rather than measurement or normalization bias, we repeated the analysis on microarray-based gene expression data.

TCGA's Microarray gene expression data were downloaded from the Cancer Browser website. The original dataset contained 597 samples x 17814 genes. We removed 11 samples (6 Male, 3 Metastatic and 2 having unknown tissue site) and remained with 586 samples.

Samples were clustered using the same protocol described for the RNA-Seq dataset (K-means algorithm applied using correlation distance after row normalization). Similarly to the Global RNA-Seq analysis. Luminal-A samples were split between a mixed luminal-A/luminal-B cluster (cluster 1) and a rather homogenous cluster (cluster 2).

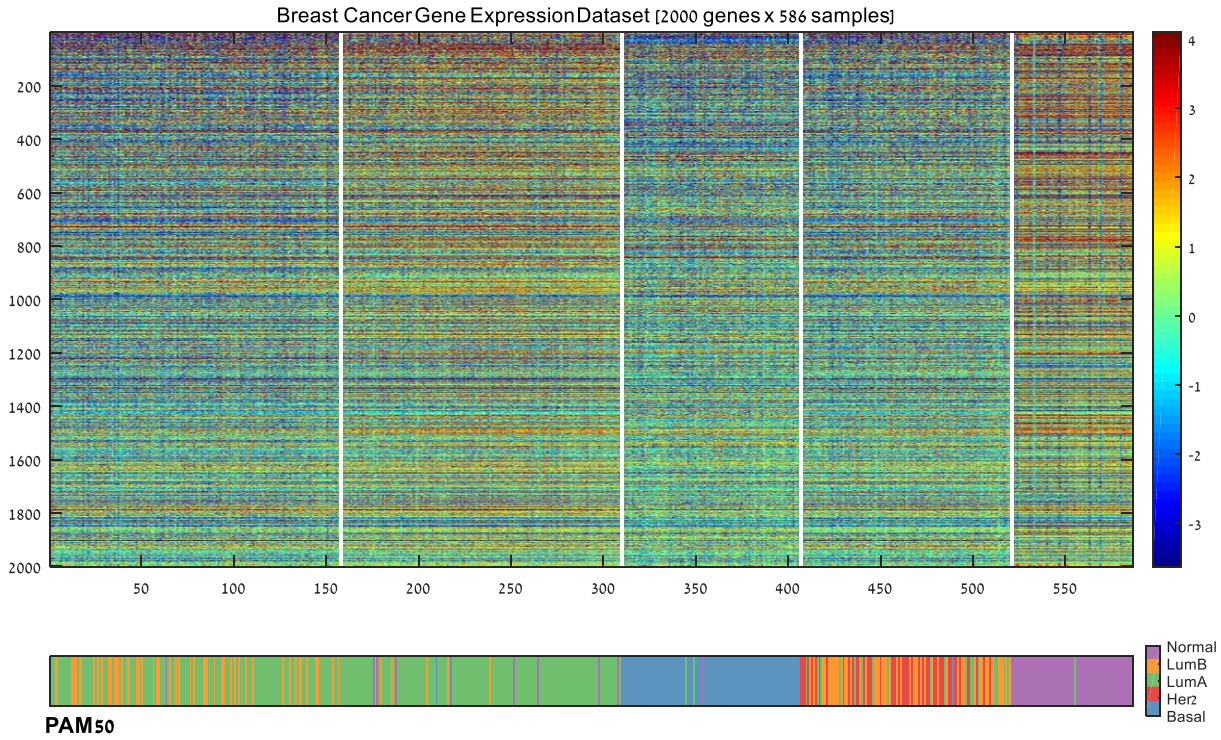


Figure S1.4A: Global unsupervised clustering of breast samples using Microarray gene expression data.

When clustering the 265 luminal-A samples in the microarray dataset into 2, the resulting partition exhibited very high similarity (Chi-square p=1.1e-40) to the luminal-A subgroups identified based on the RNA-Seq data. When comparing the top 200 genes differentially expressed on the two-microarray luminal-A subgroups to the top 200 genes differentially expressed on the two RNA-Seq luminal-A subgroups, 88 genes appeared in the intersection.

Similarly, to the RNA-Seq based list of differentially expressed genes, the 88 genes also were enriched for GO terms such as immune system process, cell differentiation and T-Cell receptor related terms.

We, therefore, conclude that the signal observed on the RNA-Seq data, splitting the luminal-A samples into two distinct subgroups is not an artifact of either the measurement technology or the normalization used by TCGA.

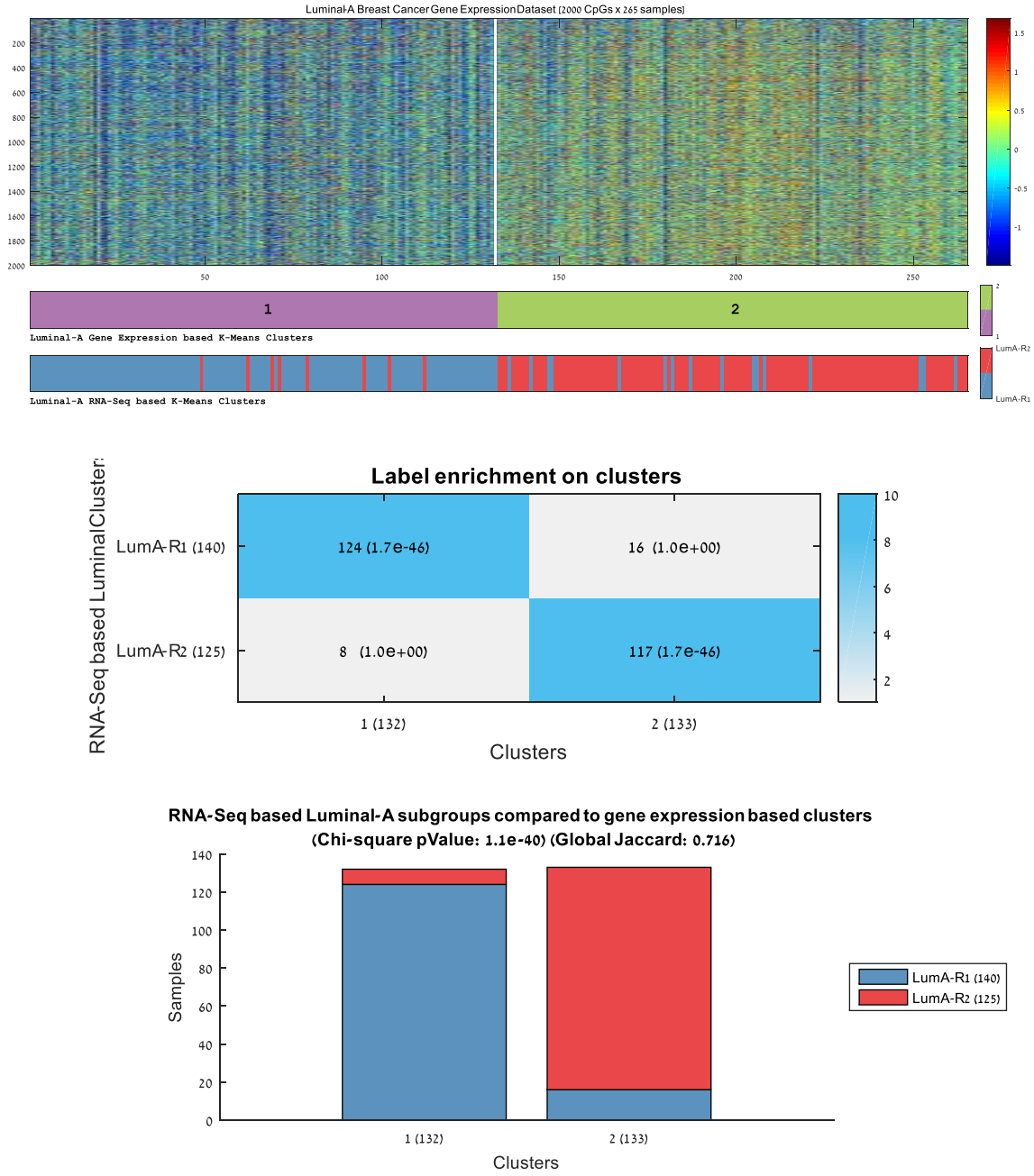


Figure S1.4B: Clustering microarray expression profiles of luminal-A samples into two and comparison to the LumA-R subgroups identified on RNA-Seq dataset.

7.1.5. Differentially Expressed Gene Analysis (LumA-R1 vs. LumA-R2)

Gene enrichment tests on the top 1000 differentially expressed genes

We started our analysis of differentially expressed genes between the two subgroups identified using RNA-Seq data within luminal-A samples, by generating a list of the top 1000 DEGs using the rank-sum test p-value, and a requirement for a minimum mean difference of 0.5. Interestingly, all 1000 genes were overexpressed in LumA-R2 compared to LumA-R1.

We then used the Expandar[122] suite to detect significant enrichments for Gene ontology terms[118], KEGG pathways[119] and Wiki-pathways[120]. The results are listed below:

Gene ontology enrichments detected using Expander TANGO on the list of 1000 DEGs

Gene Ontology Term	#Genes	Enrichment significance (pValue)	TANGO corrected pvalue
regulation of immune system process - GO:0002682	152	3.74E-50	1.00E-05
immune system process - GO:0002376	201	3.65E-47	1.00E-05
regulation of leukocyte activation - GO:0002694	71	2.37E-28	1.00E-05
regulation of multicellular organismal process - GO:0051239	183	2.89E-28	1.00E-05
cell activation - GO:0001775	91	4.59E-28	1.00E-05
regulation of response to external stimulus - GO:0032101	73	8.18E-27	1.00E-05
regulation of biological quality - GO:0065008	218	1.82E-26	1.00E-05
leukocyte activation - GO:0045321	67	1.95E-26	1.00E-05
positive regulation of cell activation - GO:0050867	56	5.13E-24	1.00E-05
T cell activation - GO:0042110	45	4.93E-22	1.00E-05
regulation of cell proliferation - GO:0042127	128	1.83E-21	1.00E-05
regulation of response to stress - GO:0080134	91	1.91E-19	1.00E-05
chemical homeostasis - GO:0048878	93	6.50E-19	1.00E-05
hemopoiesis - GO:0030097	60	6.63E-19	1.00E-05
cell migration - GO:0016477	79	9.97E-19	1.00E-05
locomotion - GO:0040011	110	1.13E-18	1.00E-05
immune response-regulating cell surface receptor signaling pathway - GO:0002768	36	2.88E-18	1.00E-05
lymphocyte differentiation - GO:0030098	37	3.09E-18	1.00E-05
leukocyte migration - GO:0050900	43	1.10E-17	1.00E-05
regulation of cytokine production - GO:0001817	58	8.45E-17	1.00E-05
positive regulation of cell proliferation - GO:0008284	79	1.46E-16	1.00E-05
cell differentiation - GO:0030154	194	2.08E-16	1.00E-05
biological adhesion - GO:0022610	90	6.97E-16	1.00E-05
response to organic substance - GO:0010033	155	8.13E-16	1.00E-05
calcium ion homeostasis - GO:0055074	43	1.11E-15	1.00E-05
cellular response to cytokine stimulus - GO:0071345	60	2.42E-15	1.00E-05

cellular response to chemical stimulus - GO:0070887	137	3.04E-15	1.00E-05
immune effector process - GO:0002252	41	1.72E-14	1.00E-05
regulation of cell migration - GO:0030334	55	3.10E-14	1.00E-05
regulation of acute inflammatory response - GO:0002673	20	4.06E-14	1.00E-05
hemostasis - GO:0007599	62	4.15E-13	1.00E-05
negative regulation of biological process - GO:0048519	213	1.29E-12	1.00E-05
positive regulation of signaling - GO:0023056	83	1.55E-12	1.00E-05
response to external stimulus - GO:0009605	104	2.20E-12	1.00E-05
blood circulation - GO:0008015	39	9.70E-12	1.00E-05
regulation of behavior - GO:0050795	28	1.21E-11	1.00E-05
positive regulation of cellular component movement - GO:0051272	37	1.54E-11	1.00E-05
regulation of adaptive immune response - GO:0002819	23	2.43E-11	1.00E-05
regulation of cell differentiation - GO:0045595	89	3.88E-11	1.00E-05
negative regulation of sequestering of calcium ion - GO:0051283	13	5.22E-11	1.00E-05
regulation of cell death - GO:0010941	107	5.51E-11	1.00E-05
regulation of secretion - GO:0051046	53	5.55E-11	1.00E-05
regulation of alpha-beta T cell activation - GO:0046634	18	9.86E-11	1.00E-05
regulation of hydrolase activity - GO:0051336	89	1.31E-10	1.00E-05
regulation of protein secretion - GO:0050708	25	1.54E-10	1.00E-05
humoral immune response - GO:0006959	22	2.04E-10	1.00E-05
positive regulation of inflammatory response - GO:0050729	19	2.82E-10	1.00E-05
regulation of lymphocyte mediated immunity - GO:0002706	19	1.28E-09	1.00E-05
cell chemotaxis - GO:0060326	21	3.53E-09	1.00E-05
regulation of protein transport - GO:0051223	36	3.98E-09	1.00E-05
positive regulation of metabolic process - GO:0009893	144	5.02E-09	1.00E-05
regulation of leukocyte chemotaxis - GO:0002688	16	5.52E-09	1.00E-05
vasculature development - GO:0001944	47	7.58E-09	1.00E-05
nervous system development - GO:0007399	127	7.73E-09	1.00E-05
positive regulation of leukocyte migration - GO:0002687	16	9.39E-09	1.00E-05
regulation of transmembrane transport - GO:0034762	45	1.00E-08	1.00E-05
positive regulation of molecular function - GO:0044093	104	1.03E-08	1.00E-05
negative regulation of multicellular organismal process - GO:0051241	41	1.11E-08	1.00E-05

Table S1.5A: Gene ontology enrichments detected using Expander TANGO on the list of 1000 DEGs

KEGG PATHWAYS

KEGG Pathway	#Genes	p-value	Enrichment	Genes
Cytokine-cytokine receptor interaction	56	4.76E-22	4.57	[ACVRL1, CD40, CXCL9, TNFRSF13B, CXCL2, CX3CL1, IL18RAP, LEPR, TNFRSF8, IL12B, CCR7, CCR5, CCR4, CCR2, PDGFRA, IL15RA, IL11RA, IL1R2, TNFRSF1B, TGFBR2, IL3RA, KIT, XCL2, XCL1, LTB, MET, CCL14, CCL13, FIGF, CXCR5, CSF2RB, CXCR6, IL2RG, EGFR, TPO, CCL5, CXCR3, TNFRSF17, CCL19, IL12RB1, CCL17, NGFR, CCL23, XCR1, CCL21, TSLP, IL10RA, IL6, BMP2, CXCL12, CD40LG, LEP, FAS, CD27, IL7R, IL18R1]
Hematopoietic cell lineage	29	1.50E-17	7.1	[CD1E, CD3G, CD1D, CD1C, CD3E, CD1B, CD3D, TPO, CD19, CD38, CD37, CD36, CD34, CR2, CR1,

				MME, IL11RA, IL1R2, CD2, FCER2, IL6, CD8B, CD5, CD8A, IL3RA, CD7, KIT, IL7R, MS4A1]
Cell adhesion molecules (CAMs)	30	4.08E-13	4.84	[CD40, ICAM2, NRXN2, SPN, CDH5, HLA-DOA, CD34, HLA-DOB, JAM2, JAM3, CADM3, PDCD1LG2, SELE, HLA-E, CD2, SELP, CLDN11, CLDN5, PTPRC, CD40LG, CD6, CD8B, SELL, CD8A, HLA-DPB1, CLDN19, PECAM1, CD28, CD226, PDCD1]
Primary immunodeficiency	16	8.70E-13	9.73	[CD40, CIITA, TNFRSF13B, IL2RG, CD3E, CD3D, CD79A, ZAP70, CD40LG, PTPRC, CD8B, LCK, CD8A, CD19, IL7R, JAK3]
Chemokine signaling pathway	31	1.14E-09	3.49	[CCL14, CCL13, ITK, CXCL9, CXCR5, ADCY4, PIK3CD, CXCR6, RASGRP2, CXCL2, CX3CL1, GNG2, CCL5, CXCR3, CCR7, CCL19, CCR5, CCL17, CCR4, JAK3, CCR2, CCL23, XCR1, CCL21, PRKCB, GNG11, FGR, CXCL12, ELMO1, XCL2, XCL1]
Complement and coagulation cascades	17	1.36E-08	5.24	[CR2, CR1, C1S, VWF, F10, CFH, C1R, PROS1, F2R, CFI, TFPI, C3, C6, C7, SERPING1, MASP1, A2M]
T cell receptor signaling pathway	20	1.30E-07	3.94	[ITK, PIK3CD, CD3G, CD3E, CD3D, ZAP70, PTPRC, CD40LG, CD8B, CD8A, LCK, GRAP2, CD28, PAK7, PRKCQ, FYN, CD247, PDCD1, PAK3, LAT]
Allograft rejection	11	6.44E-07	6.33	[CD40, CD40LG, HLA-DPB1, PRF1, CD28, GZMB, FAS, IL12B, HLA-DOA, HLA-DOB, HLA-E]
Natural killer cell mediated cytotoxicity	20	5.66E-06	3.13	[PRKCB, SH2D1A, PRF1, ICAM2, GZMB, PIK3CD, PRKCA, HLA-E, ZAP70, NCR3, KLRK1, LCK, PLCG2, FAS, CD48, FYN, CD247, HCST, LAT, CD244]
Pathways in cancer	34	1.49E-05	2.2	[FIGF, LAMA2, EPAS1, TCF7, PIK3CD, PTGS2, GLI1, ETS1, FGF2, FOXO1, EGFR, GLI2, WNT6, FGF7, ACVR1C, MECOM, PLCG2, WNT1, RUNX1T1, PDGFRA, WNT10A, PRKCB, PTCH2, PRKCA, IGF1, TRAF1, TGFBR2, IL6, BMP2, COL4A4, KIT, FAS, PPARG, MET]
PPAR signaling pathway	13	1.67E-05	4.01	[ADIPOQ, LPL, AQP7, ACSL5, ACSL4, FABP4, ACADL, FABP7, PPARG, PLIN1, CD36, PCK1, PLTP]
Autoimmune thyroid disease	11	2.39E-05	4.5	[CD40, TPO, CD40LG, HLA-DPB1, PRF1, CD28, GZMB, FAS, HLA-DOA, HLA-DOB, HLA-E]
Focal adhesion	23	6.43E-05	2.46	[PDGFRA, FIGF, TNXB, VWF, LAMA2, CAV2, PRKCB, CAV1, PIK3CD, PRKCA, IGF1, EGFR, THBS4, RELN, TNN, COL4A4, PAK7, ITGA7, COL6A6, FYN, FLNC, PAK3, MET]
Neuroactive ligand-receptor interaction	27	7.73E-05	2.25	[PTGER4, PTGFR, HTR2B, ADRB2, HTR2A, P2RY8, EDNRB, CNR2, GRM7, CNR1, S1PR1, LEPR, CTSG, S1PR2, GABRE, S1PR4, GRIA4, GABRP, GZMA, P2RY14, F2R, AVPR2, SSTR1, TACR1, P2RX1, LEP, F2RL2]

Table S1.5B: KEGG-Pathways enrichments detected using Expander TANGO on the list of 1000 DEGs

WIKI-PATHWAYS

Wiki-Pathway	#Genes	p-value	Enrichment	Genes
TCR Signaling Pathway	10	1.55E-09	11.8	[IL15RA, ITK, PSTPIP1, CD8A, GRAP2, CD3G, CD247, CD3E, CD3D, LAT]
B Cell Receptor Signaling Pathway	10	1.72E-06	6.45	[MAP4K1, BLK, KLF11, CR2, PTPRC, IRF4, INPP5D, PLCG2, HCLS1, ETS1]
Focal Adhesion	11	5.88E-05	4.11	[FGR, PDGFRA, FIGF, TNXB, RELN, TNN, TXK, COL4A4, PAK7, MET, THBS4]
Complement Activation, Classical Pathway	6	8.38E-05	7.51	[C3, C6, C7, C1S, C1R, MASP1]

Table S1.5C: WIKI-Pathways enrichments detected using Expander TANGO on the list of 1000 DEGs

Gene enrichment test using GOrilla on the gene list ranked by rank-sum test on LumA-R1 vs. LumA-R2

To verify our results with a second tool for GO enrichment analysis, we first applied a rank-sum test on all dataset genes for the testing for a difference in expression means between LumA-R1 and LumA-R2 samples. We then used the test p-values to rank the genes and applied the GOrilla[124] algorithm on the list composed of 19914 genes.

GO Term	Description	Enrichment	FDR q-value
GO:0002376	immune system process	2.18	3.44E-49
GO:0002682	regulation of immune system process	2.32	4.07E-47
GO:0022610	biological adhesion	2.47	1.99E-40
GO:0007155	cell adhesion	2.47	2.18E-40
GO:0051239	regulation of multicellular organismal process	1.86	1.60E-38
GO:0030155	regulation of cell adhesion	2.92	8.32E-38
GO:0050865	regulation of cell activation	3.25	1.01E-37
GO:0048583	regulation of response to stimulus	1.64	1.43E-37
GO:0002684	positive regulation of immune system process	2.54	8.17E-36
GO:0042127	regulation of cell proliferation	2.09	4.15E-34
GO:0006955	immune response	2.25	4.52E-34
GO:0048518	positive regulation of biological process	1.46	1.94E-33
GO:0002694	regulation of leukocyte activation	3.24	2.97E-33
GO:0007166	cell surface receptor signaling pathway	1.85	4.08E-32
GO:0007165	signal transduction	1.49	4.15E-32
GO:0051240	positive regulation of multicellular organismal process	2.09	2.36E-31
GO:0051249	regulation of lymphocyte activation	3.33	1.33E-30
GO:0050867	positive regulation of cell activation	3.69	1.37E-29
GO:0001775	cell activation	2.74	1.64E-29
GO:0034110	regulation of homotypic cell-cell adhesion	3.55	6.63E-29
GO:0048584	positive regulation of response to stimulus	1.82	1.20E-28
GO:0045785	positive regulation of cell adhesion	3.25	2.00E-28
GO:0002696	positive regulation of leukocyte activation	3.67	3.34E-28
GO:0098609	cell-cell adhesion	2.65	8.75E-28
GO:0022407	regulation of cell-cell adhesion	3.18	3.87E-27
GO:1903037	regulation of leukocyte cell-cell adhesion	3.46	4.20E-26
GO:0051251	positive regulation of lymphocyte activation	3.68	1.49E-25
GO:0050776	regulation of immune response	2.31	1.91E-25
GO:0050863	regulation of T cell activation	3.46	3.05E-25
GO:0045321	leukocyte activation	3.01	2.00E-24
GO:0016337	single organismal cell-cell adhesion	2.85	1.22E-23
GO:0050793	regulation of developmental process	1.71	1.54E-23
GO:0034112	positive regulation of homotypic cell-cell adhesion	3.91	2.40E-23
GO:0051094	positive regulation of developmental process	2.03	3.50E-23

GO:0030154	cell differentiation	1.8	3.40E-23
GO:1903039	positive regulation of leukocyte cell-cell adhesion	3.89	3.38E-23
GO:0050870	positive regulation of T cell activation	3.89	1.48E-22
GO:0008284	positive regulation of cell proliferation	2.26	1.84E-22
GO:0098602	single organism cell adhesion	2.85	1.89E-22
GO:0006952	defense response	1.96	2.37E-22
GO:0048522	positive regulation of cellular process	1.42	3.21E-22
GO:0046649	lymphocyte activation	3.14	5.71E-22
GO:0022409	positive regulation of cell-cell adhesion	3.54	1.41E-21
GO:0050896	response to stimulus	1.41	1.72E-21
GO:0016477	cell migration	2.27	2.52E-21
GO:0040011	locomotion	2.14	3.39E-21
GO:0010033	response to organic substance	1.75	7.99E-21
GO:0032101	regulation of response to external stimulus	2.14	1.51E-20
GO:2000026	regulation of multicellular organismal development	1.81	1.61E-20
GO:0002250	adaptive immune response	3.71	2.99E-20

Table S1.5D: Gene ontology enrichments identified using the Gorilla tool on the list of differentially expressed genes (LumA-R1 vs. LumA-R2)

7.1.6. DNA methylation data analysis on all tumor types

Obtaining the DNA-methylation dataset and initial preprocessing

Obtaining the data: TCGA's DNA-Methylation breast cancer dataset was downloaded from UCSC's Cancer Browser website. Samples were measured using Illumina's Infinium HumanMethylation450 BeadChip arrays.

Sample filtering: Started with 872 samples. Removed 8 gender/male, 5 sample type/metastatic, 19 tumor_tissue_site/NA, 98 sample type/normal, 33 PAM50 call/NA, 30 PAM50 Normal. Remained with 679 samples.

Distribution of PAM50 labels in preprocessed Meth450 dataset:

Total after preprocessing: 679

PAM50 label	#Samples
Basal	124
Her2	42
LumA	378
LumB	135
Total	679

Table S1.6A: Distribution of PAM50 labels in TCGA's Meth450 dataset

Probeset filtering: The Illumina Methylation 450K array contains two types of probe chemistries that may require special normalization. To avoid dealing with integrating the two probe types and in order to zoom in CpGs characterizing known genes, we used only Infinium I probes that are also associated with a Gene symbol, keeping 107,639 probes for all further analyses.

Row Normalization: Rows were standardized (centered and normalized) before clustering was applied on the columns (samples) of the methylation beta matrix.

Sample Clustering: The k-means algorithm was used to cluster the samples, using correlation as a distance metric.

	Meth450 Clusters	Total	1	2	3	4
		n=679	n=182	n=160	n=174	n=163
Age (Median)		58	62	58	56	54
ER Status	NA	38 (6%)	12 (7%)	7 (4%)	11 (6%)	8 (5%)
	Negative	144 (21%)	11 (6%)	3 (2%)	6 (3%)	124 (76%)
	Positive	497 (73%)	159 (87%)	150 (94%)	157 (90%)	31 (19%)
PR Status	NA	41 (6%)	12 (7%)	8 (5%)	11 (6%)	10 (6%)
	Negative	202 (30%)	43 (24%)	11 (7%)	18 (10%)	130 (80%)
	Positive	436 (64%)	127 (70%)	141 (88%)	145 (83%)	23 (14%)
Her2 Status	NA	232 (34%)	57 (31%)	41 (26%)	71 (41%)	63 (39%)
	Negative	394 (58%)	92 (51%)	115 (72%)	95 (55%)	92 (56%)
	Positive	53 (8%)	33 (18%)	4 (3%)	8 (5%)	8 (5%)
PAM50	Basal	124 (18%)	0 (0%)	0 (0%)	0 (0%)	124 (76%)
	Her2	42 (6%)	14 (8%)	0 (0%)	3 (2%)	25 (15%)
	LumA	378 (56%)	96 (53%)	119 (74%)	156 (90%)	7 (4%)
	LumB	135 (20%)	72 (40%)	41 (26%)	15 (9%)	7 (4%)
Pathologic stage	NA	3 (0%)	1 (1%)	1 (1%)	0 (0%)	1 (1%)
	Stage I	112 (16%)	28 (15%)	21 (13%)	43 (25%)	20 (12%)
	Stage II	382 (56%)	89 (49%)	93 (58%)	89 (51%)	111 (68%)
	Stage III	172 (25%)	61 (34%)	43 (27%)	40 (23%)	28 (17%)
	Stage IV	6 (1%)	2 (1%)	1 (1%)	1 (1%)	2 (1%)
	Stage X	4 (1%)	1 (1%)	1 (1%)	1 (1%)	1 (1%)
Histological type	Infil. Ductal Carcinoma	461 (68%)	123 (68%)	106 (66%)	96 (55%)	136 (83%)
	Infil. Lobular Carcinoma	140 (21%)	41 (23%)	33 (21%)	62 (36%)	4 (2%)
	Medullary Carcinoma	5 (1%)	0 (0%)	0 (0%)	1 (1%)	4 (2%)
	Metaplastic Carcinoma	2 (0%)	0 (0%)	0 (0%)	0 (0%)	2 (1%)
	Mixed Histology	24 (4%)	6 (3%)	11 (7%)	5 (3%)	2 (1%)
	Mucinous Carcinoma	14 (2%)	7 (4%)	2 (1%)	5 (3%)	0 (0%)
	NA	33 (5%)	5 (3%)	8 (5%)	5 (3%)	15 (9%)

Table S1.6B: Cohort description for the Methylation dataset analysis

Survival and Recurrence KM plots for Meth450 samples based on PAM50 labels

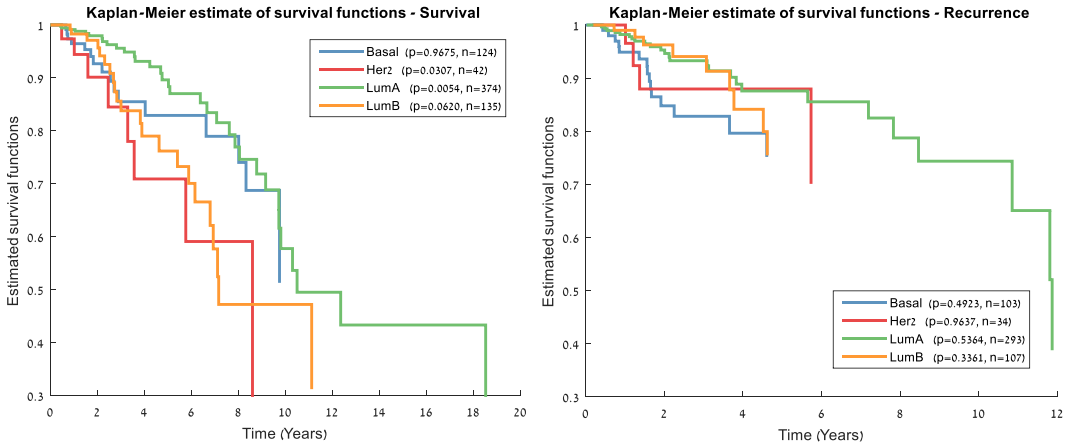


Figure S1.6A: Survival analysis for Meth450 samples based on PAM50 labels

Clustering Meth450 tumor samples to 4 using top 2000 variable CpGs (Inf I, GS only)

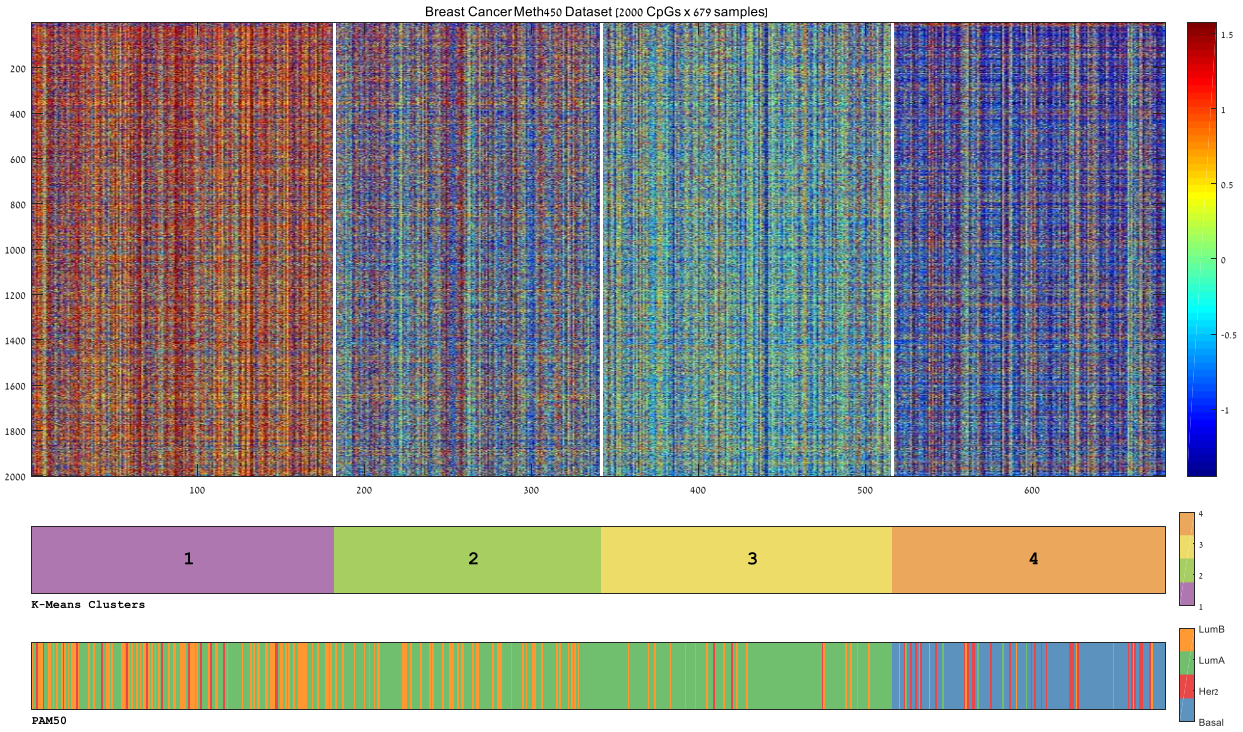


Figure S1.6B: Clustering the 469 samples of TCGA's Meth450 dataset into 4 subgroups and comparison to PAM50 labels

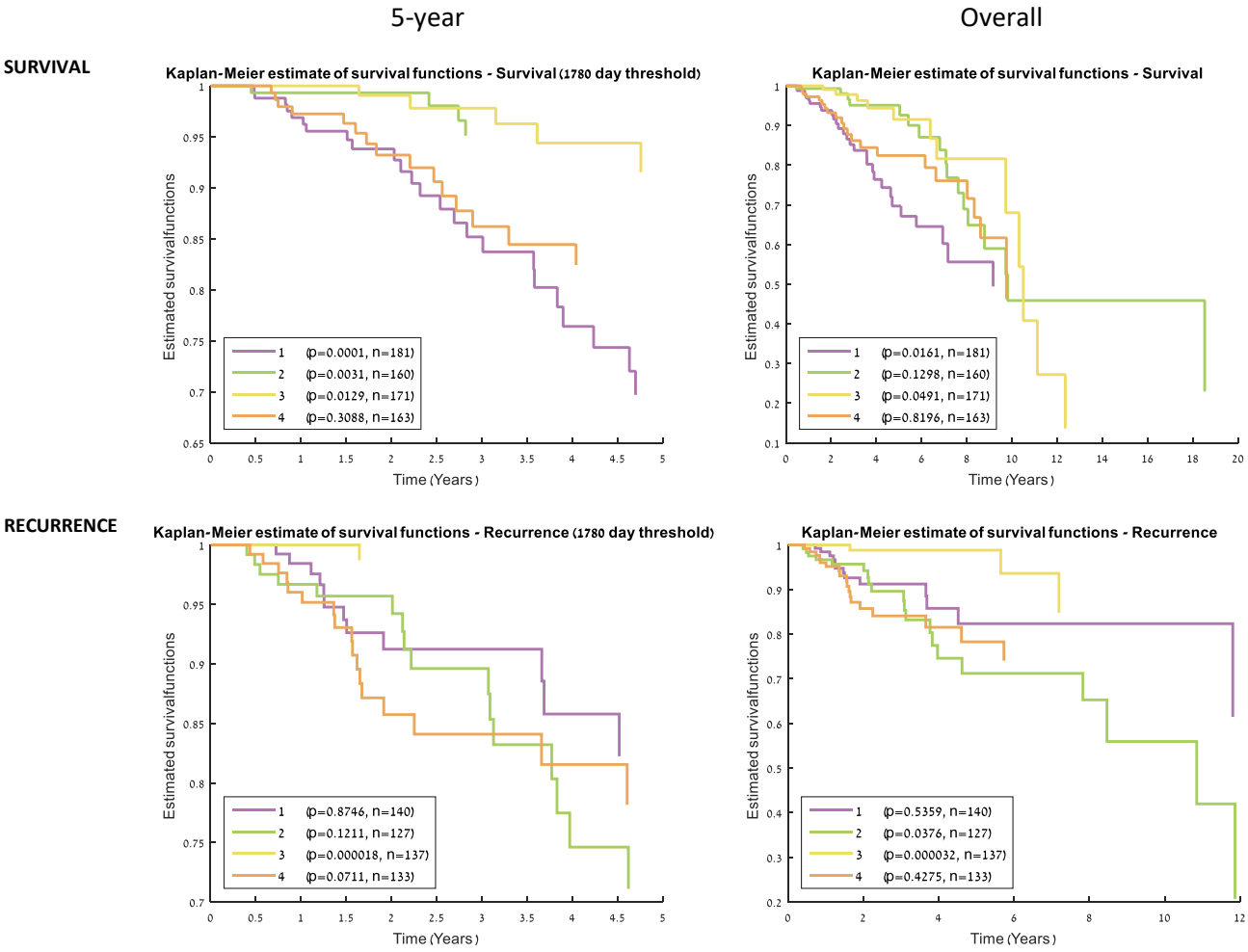


Figure S1.6C: Overall and 5-year survival analysis for the four identified Meth450 subgroups

7.1.7. Methylation luminal samples analysis

Clustering Meth450 luminal tumor samples to 3 using top 2000 CpGs (Inf I, GS only)

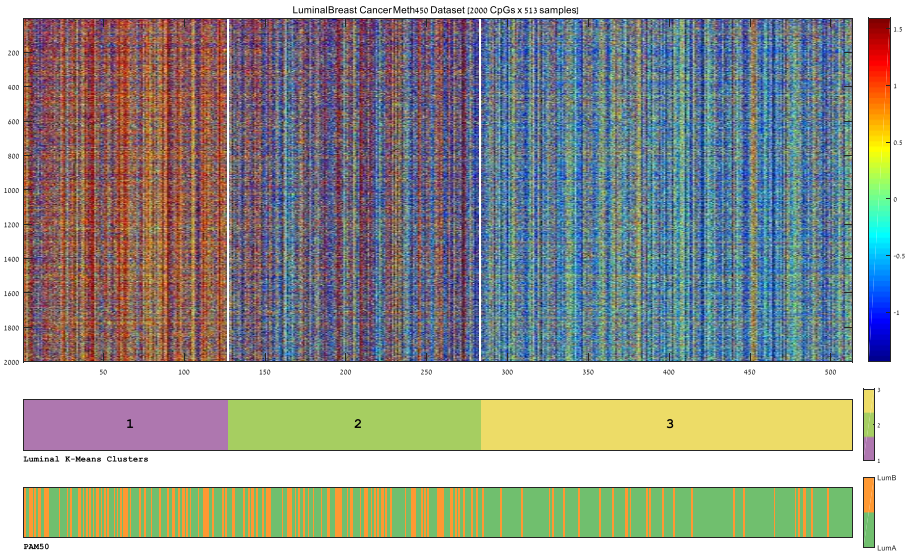


Figure S1.7A: Clustering the luminal samples into 3 based on methylation profiles.

	Meth450 Clusters	Total	1	2	3
		n=513	n=127	n=156	n=230
Age (Median)		59	63	59	56
ER Status	NA	30 (6%)	10 (8%)	7 (4%)	13 (6%)
	Negative	13 (3%)	6 (5%)	2 (1%)	5 (2%)
	Positive	470 (92%)	111 (87%)	147 (94%)	212 (92%)
PR Status	NA	31 (6%)	10 (8%)	8 (5%)	13 (6%)
	Negative	62 (12%)	25 (20%)	14 (9%)	23 (10%)
	Positive	420 (82%)	92 (72%)	134 (86%)	194 (84%)
Her2 Status	NA	171 (33%)	41 (32%)	40 (26%)	90 (39%)
	Negative	309 (60%)	71 (56%)	104 (67%)	134 (58%)
	Positive	33 (6%)	15 (12%)	12 (8%)	6 (3%)
PAM50	LumA	378 (74%)	76 (60%)	98 (63%)	204 (89%)
	LumB	135 (26%)	51 (40%)	58 (37%)	26 (11%)
Pathologic stage	NA	1 (0%)	0 (0%)	1 (1%)	0 (0%)
	Stage I	92 (18%)	22 (17%)	18 (12%)	52 (23%)
	Stage II	270 (53%)	63 (50%)	90 (58%)	117 (51%)
	Stage III	143 (28%)	39 (31%)	46 (29%)	58 (25%)
	Stage IV	4 (1%)	2 (2%)	1 (1%)	1 (0%)
	Stage X	3 (1%)	1 (1%)	0 (0%)	2 (1%)
Histological type	Infil. Ductal Carcinoma	323 (63%)	81 (64%)	112 (72%)	130 (57%)
	Infil. Lobular Carcinoma	136 (27%)	34 (27%)	26 (17%)	76 (33%)
	Medullary Carcinoma	1 (0%)	0 (0%)	0 (0%)	1 (0%)
	Mixed Histology	22 (4%)	6 (5%)	7 (4%)	9 (4%)
	Mucinous Carcinoma	14 (3%)	4 (3%)	4 (3%)	6 (3%)
	NA	17 (3%)	2 (2%)	7 (4%)	8 (3%)

Table S1.7A: Cohort description for the luminal Methylation dataset analysis.

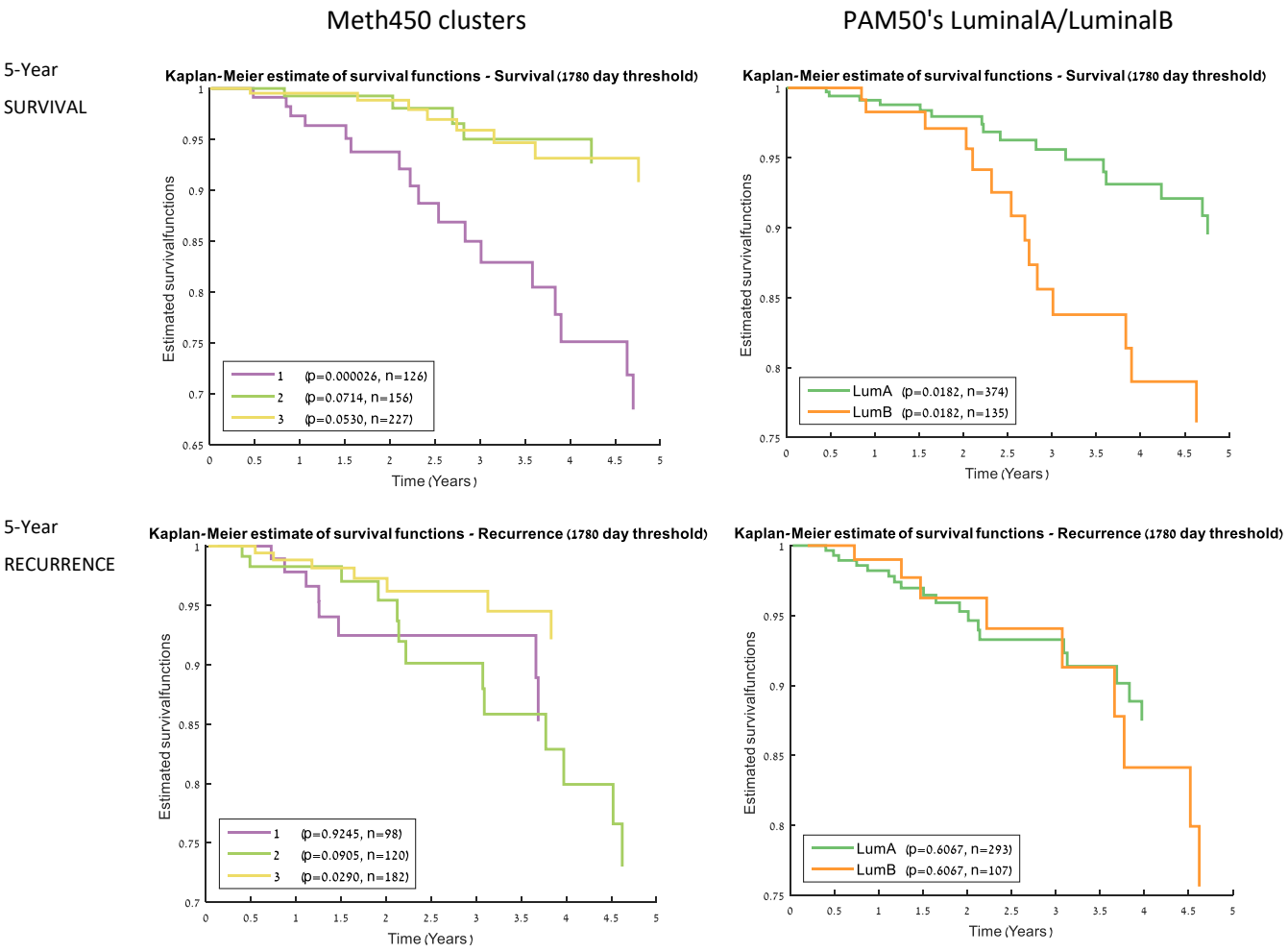


Figure S1.7B: Comparative 5-year survival and recurrence analysis for the three methylation subgroups and PAM50's luminal subgroups.

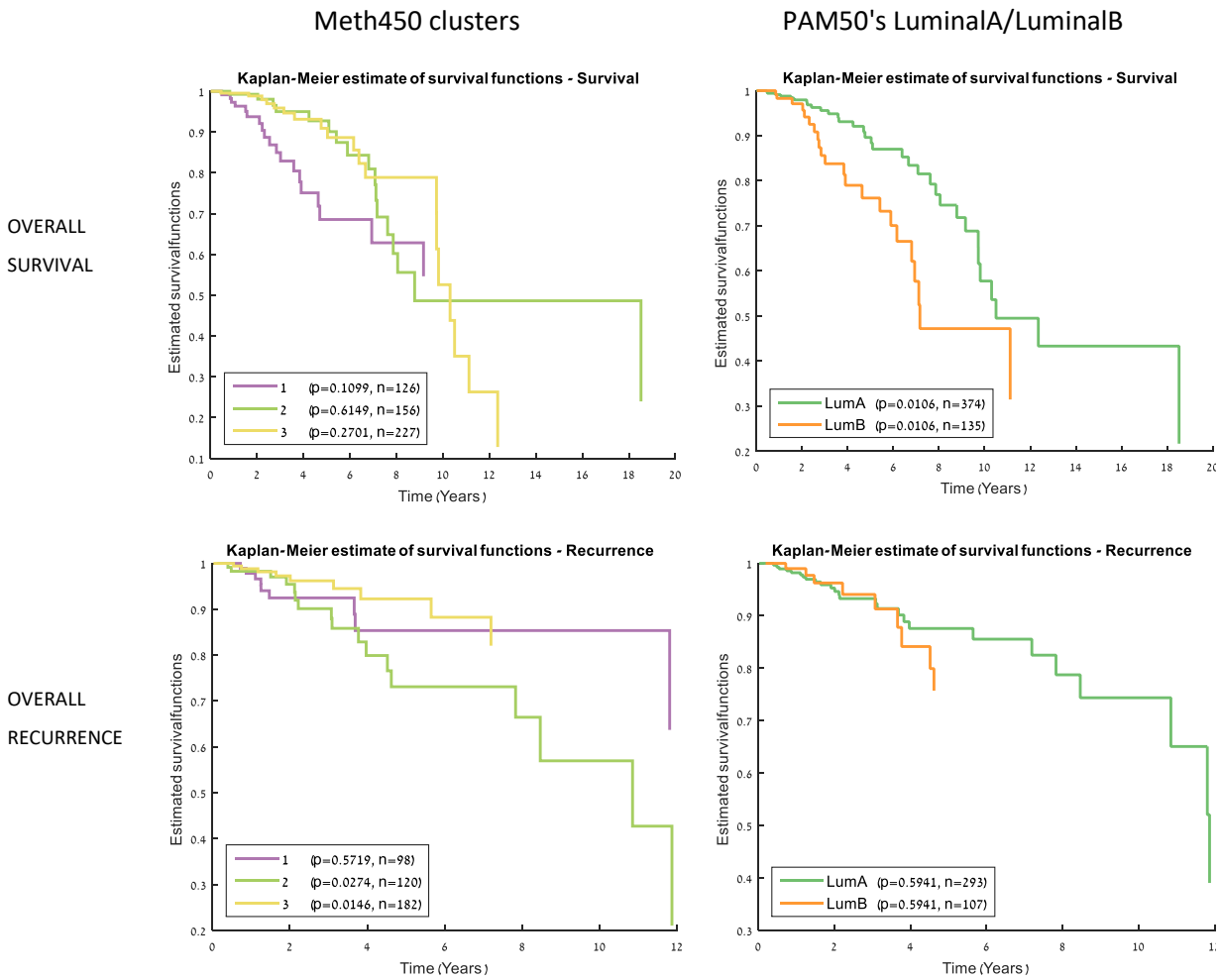


Figure S1.7C: **Comparative overall survival and recurrence analysis for the three methylation subgroups and PAM50's luminal subgroups.**

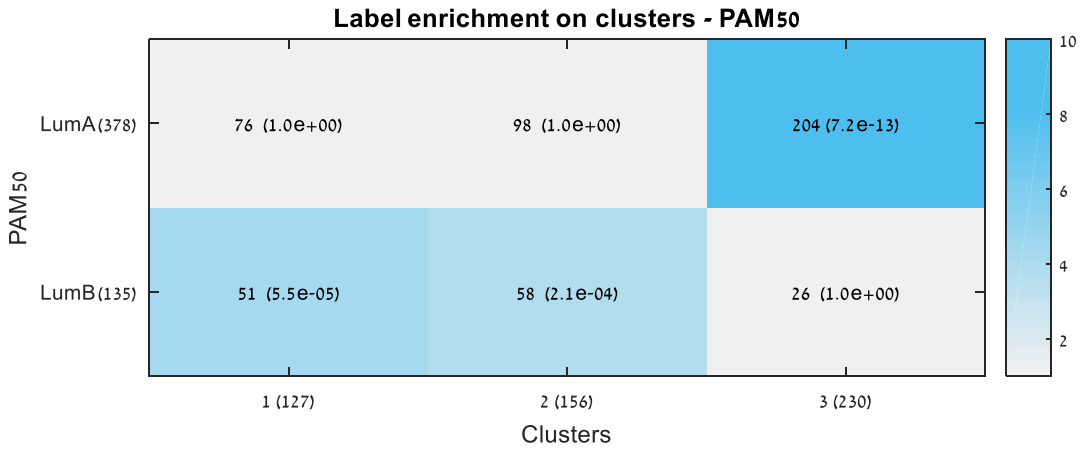


Figure S1.7D: Enrichment analysis of the 3 methylation subgroups for the two PAM50 luminal subgroups. P-values indicate hypergeometric enrichment significance.

7.1.8. Methylation luminal-A samples analysis

Clustering Meth450 378 luminal-A tumor samples to 3 using top 2000 CpGs , Inf 1, GS included only

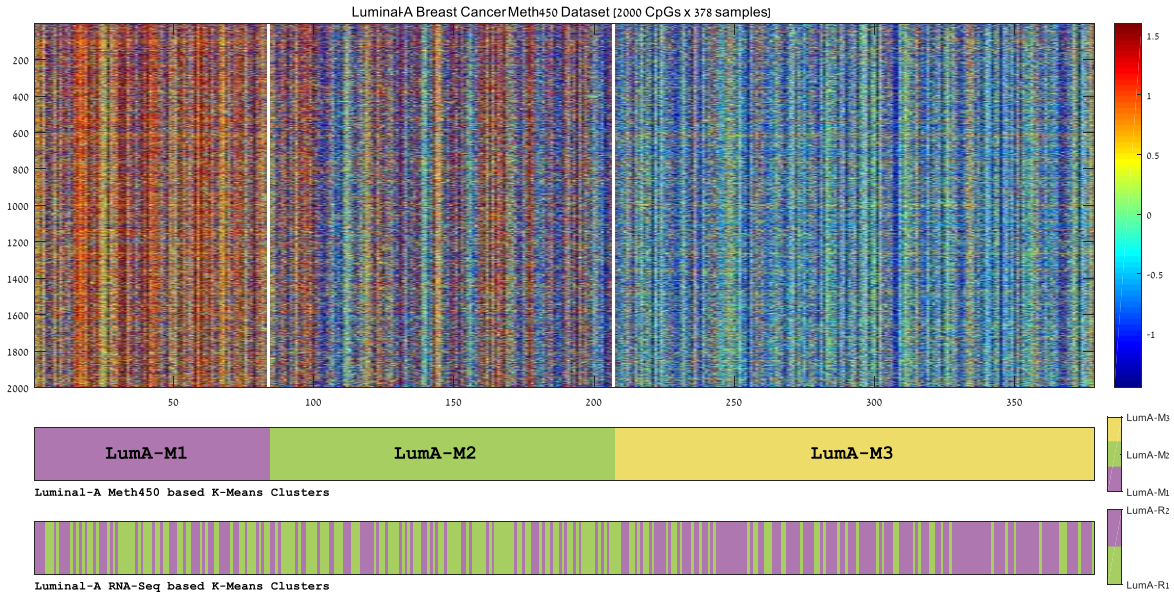


Figure S1.8A: Clustering the luminal-A samples into 3 groups using DNA-Methylation data.

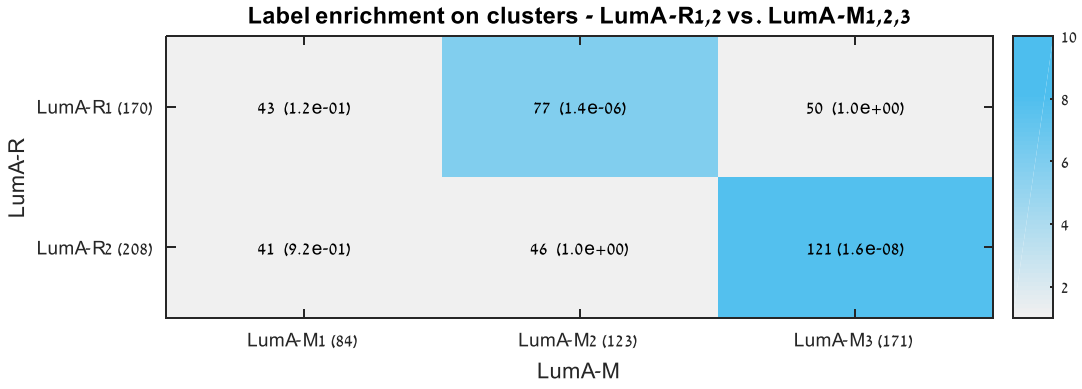


Figure S1.8B: Comparison of the RNA-Seq based partition into LumA-R1/R2 and the Methylation based partition into LumA-M1/2/3

	Meth450 Clusters	Total	1	2	3
		n=378	n=84	n=123	n=171
Age (Median)		59	62	60	56
ER Status	NA	18 (5%)	5 (6%)	4 (3%)	9 (5%)
	Negative	10 (3%)	3 (4%)	3 (2%)	4 (2%)
	Positive	350 (93%)	76 (90%)	116 (94%)	158 (92%)
PR Status	NA	19 (5%)	5 (6%)	5 (4%)	9 (5%)
	Negative	41 (11%)	18 (21%)	10 (8%)	13 (8%)
	Positive	318 (84%)	61 (73%)	108 (88%)	149 (87%)
Her2 Status	NA	143 (38%)	27 (32%)	47 (38%)	69 (40%)
	Negative	221 (58%)	50 (60%)	72 (59%)	99 (58%)
	Positive	14 (4%)	7 (8%)	4 (3%)	3 (2%)
PAM50	LumA	378 (100%)	84 (100%)	123 (100%)	171 (100%)
Pathologic stage	NA	1 (0%)	0 (0%)	1 (1%)	0 (0%)
	Stage I	77 (20%)	14 (17%)	18 (15%)	45 (26%)
	Stage II	193 (51%)	43 (51%)	66 (54%)	84 (49%)
	Stage III	103 (27%)	27 (32%)	37 (30%)	39 (23%)
	Stage IV	2 (1%)	0 (0%)	1 (1%)	1 (1%)
	Stage X	2 (1%)	0 (0%)	0 (0%)	2 (1%)
Histological type	Infil. Ductal Carcinoma	212 (56%)	42 (50%)	77 (63%)	93 (54%)
	Infil. Lobular Carcinoma	127 (34%)	32 (38%)	36 (29%)	59 (35%)
	Mixed Histology	17 (4%)	4 (5%)	5 (4%)	8 (5%)
	Mucinous Carcinoma	9 (2%)	3 (4%)	1 (1%)	5 (3%)
	NA	13 (3%)	3 (4%)	4 (3%)	6 (4%)

Table S1.8A: Cohort description for the luminal-A Methylation dataset analysis.

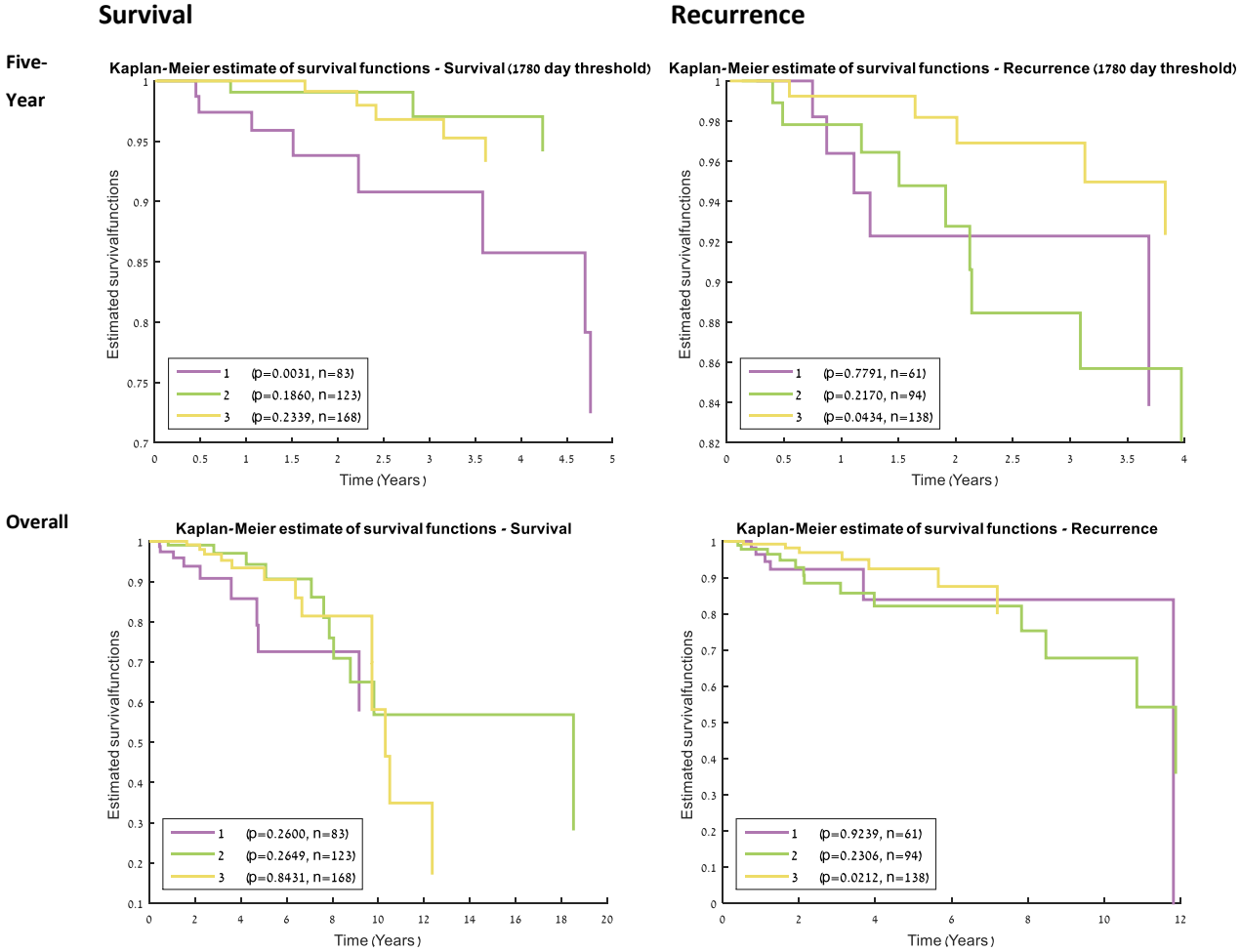


Figure S1.8C: Survival and recurrence analysis for the three methylation-based luminal-A subgroups

7.1.9. Differentially Methylated Gene Analysis (LumA-M1 vs. LumA-M2)

We have generated a list of the top 1000 differentially methylated CpGs between LumA-M1 and LumA-M3 groups using the rank-sum test having a minimal median difference of 0.2. The list represented 483 unique gene symbols for which gene enrichments were calculated using a background of 15737 genes included in the rank-sum test.

Gene Enrichment tests on the top 1000 differentially methylated CpGs

The following results were obtained by using the Expander suite on a set of 429 (unique gene symbols having Entrez ID) genes included in the set of 1000 differentially methylated CpGs:

Gene Ontology enrichments detected using Expander TANGO on the list of 1000 DMGs

Gene Ontology Term – Biological process	#Genes	Enrichment significance (pValue)	TANGO corrected pvalue
system development - GO:0048731	188	9.86E-34	1.00E-05
nervous system development - GO:0007399	132	4.38E-31	1.00E-05
system process - GO:0003008	115	3.11E-27	1.00E-05
neurological system process - GO:0050877	98	2.77E-26	1.00E-05
multicellular organismal signaling - GO:0035637	72	8.17E-24	1.00E-05
cell differentiation - GO:0030154	141	8.51E-23	1.00E-05
pattern specification process - GO:0007389	52	4.10E-21	1.00E-05
regionalization - GO:0003002	44	7.03E-21	1.00E-05
brain development - GO:0007420	56	3.47E-20	1.00E-05
neuron differentiation - GO:0030182	73	1.68E-19	1.00E-05
regulation of multicellular organismal process - GO:0051239	104	6.24E-18	1.00E-05
regulation of transcription from RNA polymerase II promoter - GO:0006357	85	2.44E-17	1.00E-05
regulation of transcription, DNA-dependent - GO:0006355	151	6.31E-17	1.00E-05
behavior - GO:0007610	45	9.98E-17	1.00E-05
anatomical structure morphogenesis - GO:0009653	105	5.02E-16	1.00E-05
central nervous system neuron differentiation - GO:0021953	26	5.79E-16	1.00E-05
positive regulation of macromolecule biosynthetic process - GO:0010557	77	1.29E-15	1.00E-05
organ morphogenesis - GO:0009887	61	1.81E-15	1.00E-05
forebrain development - GO:0030900	36	2.34E-15	1.00E-05
neuron fate commitment - GO:0048663	18	9.17E-15	1.00E-05

Table S1.9A: Top GO:Biological-procedure enrichments on the list of 1000 most differentially methylated CpGs

Gene Ontology Term – Molecular Function	#Genes	Enrichment significance (pValue)	TANGO corrected pvalue
DNA binding - GO:0003677	125	1.06E-16	0.001
regulatory region DNA binding - GO:0000975	38	6.08E-14	0.001
neuron projection - GO:0043005	48	2.27E-11	0.001
axon part - GO:0033267	19	2.73E-09	0.001

Table S1.9B: Top GO:Molecular-function enrichments on the list of 1000 most differentially methylated CpGs

KEGG PATHWAYS

KEGG Pathway	#Genes	p-value	Enrichment	Genes
Neuroactive ligand-receptor interaction	27	2.65E-10	4.2	[CHRM2, VIPR2, GPR83, GRIK2, GRM1, CRHR2, GRIN2A, EDNRB, GRM7, GRM6, GALR1, NPBWR1, P2RY1, LEPR, NTSR1, PTGDR, DRD5, GHSR, GABBR2, GABRA5, GABRA4, HTR1A, SCTR, NMBR, SSTR4, GRIN1, GRIN3A]
Maturity onset diabetes of the young	8	7.95E-08	12.8	[NEUROD1, NR5A2, ONECUT1, SLC2A2, PAX6, NEUROG3, NKX2-2, FOXA2]
Calcium signaling pathway	17	1.04E-05	3.41	[RYR1, CHRM2, RYR2, PDE1C, PRKCB, CACNA1A, CACNA1E, RYR3, GRM1, GRIN1, GRIN2A, EDNRB, GNAL, CD38, NOS1, NTSR1, DRD5]

Table S1.9C: Top enrichment of KEGG pathways on the list of 1000 most differentially methylated CpGs

Gene enrichment test using GOrilla (top 1000 CpGs + 0.2 Fold Change)

GO Term	Description	FDR q-value	Enrichment
GO:0048856	anatomical structure development	6.07E-28	2.39
GO:0032502	developmental process	1.98E-25	1.9
GO:0032501	multicellular organismal process	9.55E-24	2.17
GO:0044707	single-multicellular organism process	1.55E-22	2.15
GO:0044700	single organism signaling	1.70E-21	3.72
GO:0023052	signaling	1.89E-21	3.71
GO:0007267	cell-cell signaling	1.70E-21	3.79
GO:0030182	neuron differentiation	1.19E-20	6.57
GO:0044767	single-organism developmental process	1.43E-19	1.84
GO:0006357	regulation of transcription from RNA polymerase II promoter	1.21E-16	2.38
GO:0007610	behavior	3.46E-16	3.95
GO:0007389	pattern specification process	6.11E-16	4.49
GO:0048869	cellular developmental process	1.89E-15	2.09
GO:0021953	central nervous system neuron differentiation	1.93E-15	9.71
GO:0003008	system process	4.97E-15	2.72
GO:0007154	cell communication	5.80E-15	2.8
GO:0050877	neurological system process	6.10E-15	3.23
GO:0048731	system development	6.51E-15	3.39
GO:0003002	regionalization	6.79E-15	5.42
GO:0051239	regulation of multicellular organismal process	8.45E-15	2.1

Table S1.9D: GOrilla results for Gene Ontology enrichments on the list of top 1000 differentially methylated CpGs between LumA-M1 and LumA-M2

Gene enrichments on the various subsets of differentially methylated CpGs between LumA-M1 and LumA-M3 subgroups

	(1) Hyper Meth. CpGs		(2) Neg: R < -0.2		(3) Pos: R > 0.2	
Gene ontology	anatomical structure development	6.1E-28	developmental process	7.8E-06	pattern specification process	1.1E-13
	developmental process	2.0E-25	single organism signaling	2.4E-05	regionalization	1.1E-12
	multicellular organismal process	9.6E-24	signaling	1.8E-05	anatomical structure development	2.2E-11
	single-multicellular organism process	1.6E-22	cellular developmental process	1.4E-05	single-organism developmental process	1.9E-11
	single organism signaling	1.7E-21	single-organism developmental process	2.3E-05	anatomical structure morphogenesis	1.8E-11
	Signaling	1.9E-21	anatomical structure development	8.0E-05	developmental process	1.7E-11
	cell-cell signaling	1.7E-21	cell-cell signaling	1.8E-04	embryonic morphogenesis	1.1E-10
	neuron differentiation	1.2E-20	cell differentiation	2.2E-04	cellular developmental process	1.8E-10
	single-organism developmental process	1.4E-19	synaptic transmission	4.4E-04	organ development	5.3E-10
	regulation of transcription from RNA polymerase II promoter	1.2E-16	anatomical structure morphogenesis	6.1E-04	single-multicellular organism process	5.6E-10
	Behavior	3.5E-16	tube development	1.8E-03	cell fate commitment	5.5E-10
	pattern specification process	6.1E-16	regulation of multicellular organismal development	1.8E-03	multicellular organismal process	7.7E-10
	cellular developmental process	1.9E-15	cell development	1.7E-03	organ morphogenesis	1.8E-09
	central nervous system neuron differentiation	1.9E-15	neuron differentiation	2.0E-03	transcription, DNA-templated	6.5E-09
Tumor Suppressor Gene (TSGene 2.0)		1.5E-03		9.7E-02		5.5E-02

Table S1.9E: Gene enrichments on the various subsets of differentially methylated CpGs between LumA-M1 and LumA-M3 subgroups.

Feature enrichments on the various subsets of differentially methylated CpGs between LumA-M1 and LumA-M3 subgroups

Group		Total	(1) Hyper Meth. CpGs		(2) Neg: R < -0.2		(3) Pos: R > 0.2	
#CpGs		94880	1000		589		212	
Label	**Term**	**#Terms**	**#Terms**	**p-value**	**#Terms**	**p-value**	**#Terms**	**p-value**
UCSC RefGene Group	1stExon	9548	141	1.5E-04	104	1.4E-07	11	1.0E+00
	3'UTR	2489	11	1.0E+00	3	1.0E+00	13	1.8E-02
	5'UTR	11737	121	1.0E+00	82	3.2E-01	14	1.0E+00
	Body	32979	285	1.0E+00	111	1.0E+00	141	9.5E-20
	TSS	38127	442	1.6E-02	289	4.5E-05	33	1.0E+00
Regulatory Feature Group	Gene Associated	227	0	1.0E+00	0	1.0E+00	0	1.0E+00
	Gene Associated Cell type specific	384	0	1.0E+00	0	1.0E+00	3	1.6E-01
	NonGene Associated	472	2	1.0E+00	0	1.0E+00	0	1.0E+00
	NonGene Associated Cell type specific	40	4	2.8E-03	1	4.9E-01	1	2.2E-01
	Promoter Associated	36454	41	1.0E+00	95	1.0E+00	6	1.0E+00
	Promoter Associated Cell type specific	1676	9	1.0E+00	26	1.4E-04	0	1.0E+00
	Unclassified	7559	71	1.0E+00	73	5.8E-04	17	1.0E+00
	Unclassified Cell type specific	7962	211	8.8E-35	86	3.9E-06	50	1.3E-10
	Unassigned	40106	662	7.4E-52	308	4.9E-06	135	1.8E-09
DMR (Differentially Methylated Region)	CDMR (Cancer DMR)	855	44	1.5E-16	14	3.9E-03	20	1.1E-13
	DMR	6722	391	9.2E-18	195	1.7E-75	54	1.4E-15
	RDMR (Reprogramming DMR)	1447	33	1.9E-04	14	1.8E-01	22	2.2E-11
	Unassigned	85856	532	1.0E+00	366	1.0E+00	116	1.0E+00
Enhancer		10107	99	1.2E-09	66	8.0E-06	29	1.7E-04
DHS		17152	137	1.1E-07	86	2.1E-03	44	1.7E-05
Tumor Suppressor Gene (TSGene 2.0)		944	48	1.5E-03	29	9.7E-02	14	5.5E-02

Table S1.9F: Feature enrichments on the various subsets of differentially methylated CpGs between LumA-M1 and LumA-M3 subgroups. Group 1 is composed of the 1000 top differentially methylated CpGs exhibiting a mean difference of at least 0.2. All the CpGs on this list showed significant hyper-methylation on the LumA-M1 samples compared to LumA-M3 samples. Group 2 is composed of the 589 CpGs exhibiting differential methylation p-value<0.01, methylation mean difference>0.2 and spearman based correlation to expression that is lower than 0.2. Group 3 212 CpGs exhibiting differential methylation p-value<0.01, methylation mean difference>0.2 and spearman based correlation to expression that is higher than 0.2. All p-values represent hypergeometric based over-representation and are FDR corrected.

		Hyper Meth. CpGs		Neg: R < -0.2		Pos: R > 0.2	
Label	**Term**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**
UCSC RefGene Group	1stExon	**1.E-04**	1.E+00	**1.E-07**	1.E+00	1.E+00	3.E-02
	3'UTR	1.E+00	2.E-03	1.E+00	**6.E-04**	2.E-02	1.E+00
	5'UTR	1.E+00	8.E-01	3.E-01	1.E+00	1.E+00	2.E-02
	Body	1.E+00	**7.E-05**	1.E+00	**1.E-16**	**9.E-20**	1.E+00
	TSS	2.E-02	1.E+00	**4.E-05**	1.E+00	1.E+00	**7.E-14**
Regulatory Feature Group	Gene Associated	1.E+00	2.E-01	1.E+00	5.E-01	1.E+00	1.E+00
	Gene Associated Cell type specific	1.E+00	5.E-02	1.E+00	2.E-01	2.E-01	1.E+00
	NonGene Associated	1.E+00	3.E-01	1.E+00	1.E-01	1.E+00	8.E-01
	NonGene Associated Cell type specific	3.E-03	1.E+00	5.E-01	1.E+00	2.E-01	1.E+00
	Promoter Associated	1.E+00	**2.E-146**	1.E+00	**3.E-31**	1.E+00	**4.E-34**
	Promoter Associated Cell type specific	1.E+00	5.E-02	**1.E-04**	1.E+00	1.E+00	7.E-02
	Unclassified	1.E+00	4.E-01	6.E-04	1.E+00	1.E+00	1.E+00
	Unclassified Cell type specific	9.E-35	1.E+00	4.E-06	1.E+00	1.E-10	1.E+00
	Unassigned	7.E-52	1.E+00	5.E-06	1.E+00	2.E-09	1.E+00
Relation to UCSC CpG Island	Island	1.E+00	9.E-04	1.E+00	1.E-03	1.E+00	7.E-02
	N_Shelf	1.E+00	5.E-01	1.E+00	7.E-01	1.E+00	1.E+00
	N_Shore	6.E-02	1.E+00	4.E-02	1.E+00	7.E-01	1.E+00
	S_Shelf	8.E-02	1.E+00	8.E-02	1.E+00	1.E+00	9.E-01
	S_Shore	3.E-02	1.E+00	3.E-02	1.E+00	4.E-01	1.E+00
	Unassigned	4.E-01	1.E+00	7.E-01	1.E+00	2.E-01	1.E+00
DMR (Differentially Methylated Region	CDMR	**2.E-16**	1.E+00	4.E-03	1.E+00	1.E-13	1.E+00
	DMR	**9.E-183**	1.E+00	2.E-75	1.E+00	1.E-15	1.E+00
	RDMR	**2.E-04**	1.E+00	2.E-01	1.E+00	2.E-11	1.E+00
	Unassigned	1.E+00	**2.E-205**	1.E+00	2.E-75	1.E+00	5.E-40
Enhancer	0	1.E+00	1.E-09	1.E+00	8.E-06	1.E+00	2.E-04
	1	**1.E-09**	1.E+00	8.E-06	1.E+00	2.E-04	1.E+00
DHS	0	1.E+00	1.E-07	1.E+00	2.E-03	1.E+00	2.E-05
	1	**1.E-07**	1.E+00	2.E-03	1.E+00	2.E-05	1.E+00
Tumor Suppressor Gene Catalogue (TSG 2.0)	0	1.E+00	2.E-03	1.E+00	1.E-01	1.E+00	6.E-02
	1	**2.E-03**	1.E+00	1.E-01	1.E+00	6.E-02	1.E+00

Table S1.9G: Feature enrichment on the various subsets of differentially methylated CpGs between LumA-M1 and LumA-M3 subgroups (Including under-representation p-values)

7.1.10. Cox proportional hazards model analysis

	Survival				Recurrence			
	Univariate		*Multivariate*		Univariate		*Multivariate*	
Variable	***HR***	***pValue***	***HR***	***pValue***	***HR***	**pValue**	***HR***	***pValue***
LumA-R (1 vs 2)	0.44	0.10939	0.56	0.36991	**0.20**	**0.00421**	**0.06**	**0.00693**
LumA-M (2,3 vs 1)	**4.53**	**0.00258**	**6.68**	**0.00484**	1.64	0.34338	3.04	0.07028
Age (<60 vs.>=60 years)	**5.79**	**0.00624**	**11.20**	**0.0037**	2.18	0.10301	1.03	0.96530
Pathologic stage (I,II vs. III,IV)	1.30	0.62799	2.12	0.25519	2.09	0.11941	1.93	0.26992
ER Status	1.72	0.60363	7.17	0.18095	0.00	0.99217	0.00	0.99575
PR Status	1.03	0.96671	0.47	0.50039	0.37	0.33789	0.29	0.29092
Her2 Status	0.79	0.8208	1.48	0.72659	0.99	0.98916	0.64	0.68789

Table S1.10A: Univariate and Multivariate Cox analysis of luminal-A subgroups for five-year survival and five-year recurrence.

7.1.11. Joint clustering of luminal-A samples using both expression and DNA methylation datasets

After establishing that luminal-A samples (as labeled by PAM50) can be further divided into distinct clinically meaningful subgroups by either the RNA-Seq or the methylation datasets separately, we set out to generate a single robust luminal-A partition that would leverage from the complementary biological information stored in both expression and methylation datasets.

To this end, we unified both expression and methylation datasets into a single "bi-omic" dataset composed of 378 luminal-A samples for which both types of data are available. From each dataset we selected the top 1000 variable features (top 1000 variable genes from the RNA-Seq dataset, and top 1000 variable CpGs from the methylation dataset). We then clustered the samples using a variant of K-Means for which the distance metric is formulated as the average of the correlation based distances on the two data types, i.e., for samples s, t

$D_{st} = \frac{d_{st}^{Exp} + d_{st}^{Meth}}{2}$ where the distance d_{st} for each data type is 1 minus the correlation between the 1000-long vectors of samples s and t.

Interestingly, applying the method on the luminal-A samples did not produce an improved partition neither in terms of stability (repeated executions yielded significantly different results) nor in terms of survival prediction compared with the separate partitions. We assume this result can be attributed to fact that the two datasets impose very different partitions on the samples, making this dataset sub-optimal target for the described integrative clustering approach.

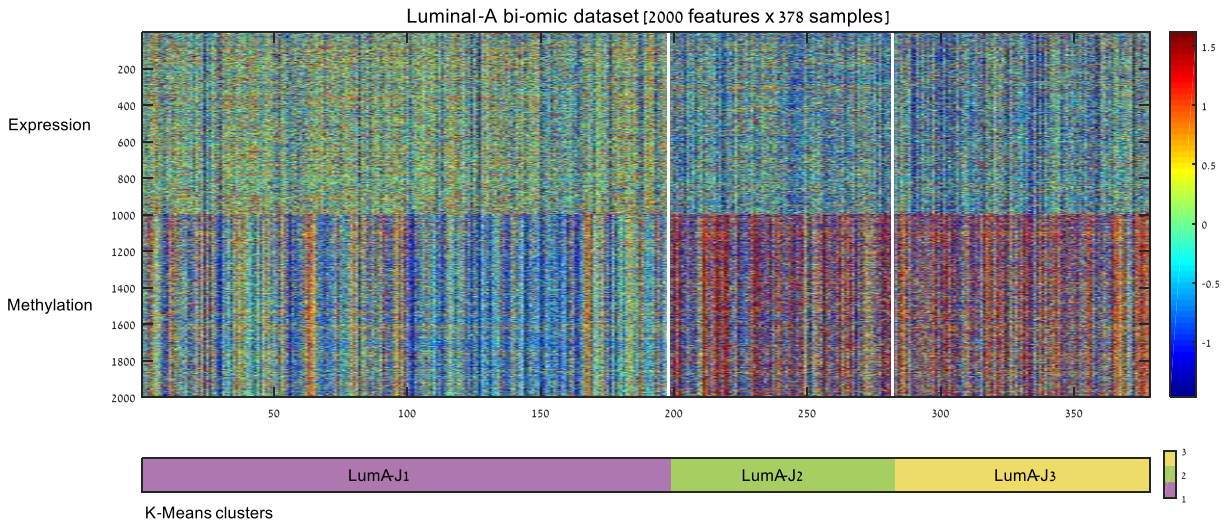


Figure S1.11A: Joint clustering of 378 luminal-A samples to 3 using the k-means algorithm based on top 1000 variably expressed genes and top 1000 variably methylated CpG islands.

7.1.12. LumA-R1/2 clusters are enriched for the ILC classes defined by TCGA

We compared our RNA-Seq based partition of the luminal-A samples to the three ILC (Invasive Lobular Carcinoma) classes recently defined by TCGA. A Chi-square test determined that the two partitions are related (p=1.2e-04, based on the 104 ILC samples appearing on both datasets). The hypergeometric test we used to evaluate the enrichment of specific ILC classes within each of our luminal clusters. LumA-R1 cluster was found to be significantly enriched for the proliferative ILC class (p=8.1e-04), whereas the LumA-R2 cluster was found to be significantly enriched for the Reactive-like ILC class (2.4-e04).

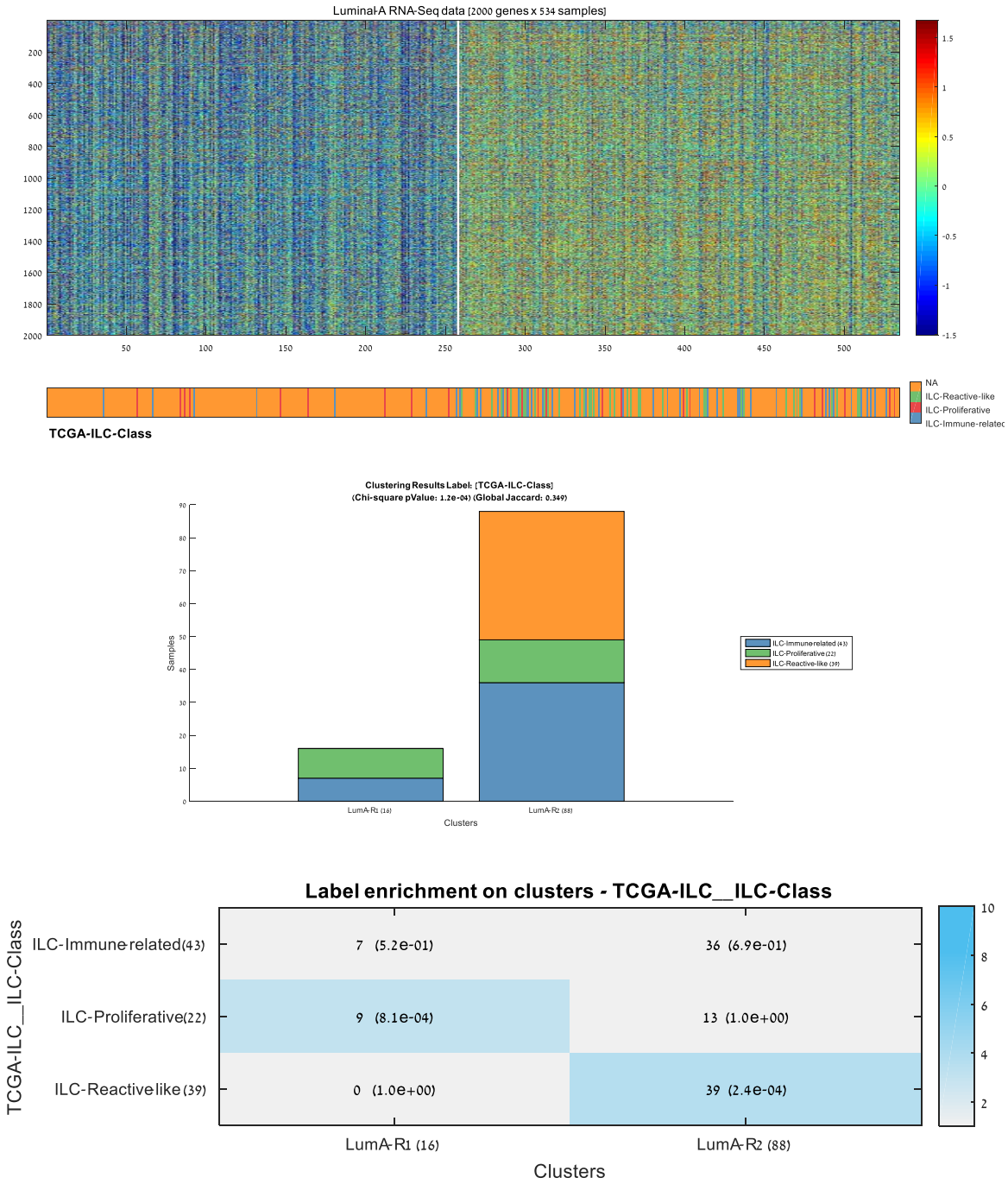


Figure 7.1.12A: Comparison of the two luminal-A subgroups we identified and TCGA's ILC (Invasive Lobular Carcinoma) classes.

7.1.13. LumA-M1 samples are enriched for the Epi-LumB group identified by Stefansson et al.

For comparing our methylation-based luminal-A clusters to the bad outcome luminal group described by Stefansson et al.[226] (named Epi-LumB as it was largely composed of luminal-B samples), we first kept only samples that appeared both in our partition and in Epi-LumB labels for TCGA's Meth450 dataset, and then we calculated enrichment for the Epi-LumB label in our clusters. Cluster LumA-M1 was found to be enriched with the Epi-LumB label (p=1.6e-07), enforcing our observation that this group is associated with a bad outcome (though labeled as luminal-A by PAM50).

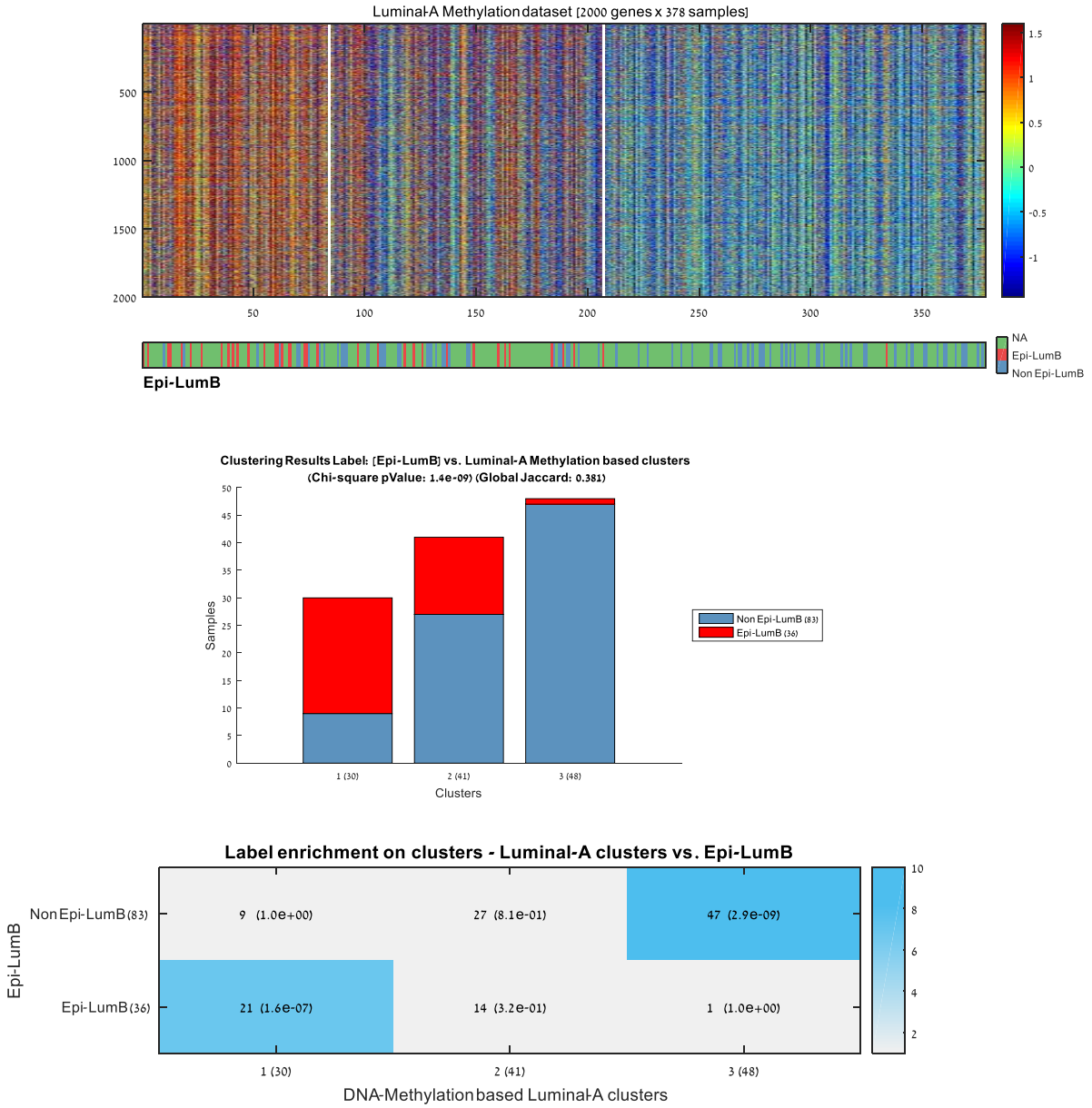
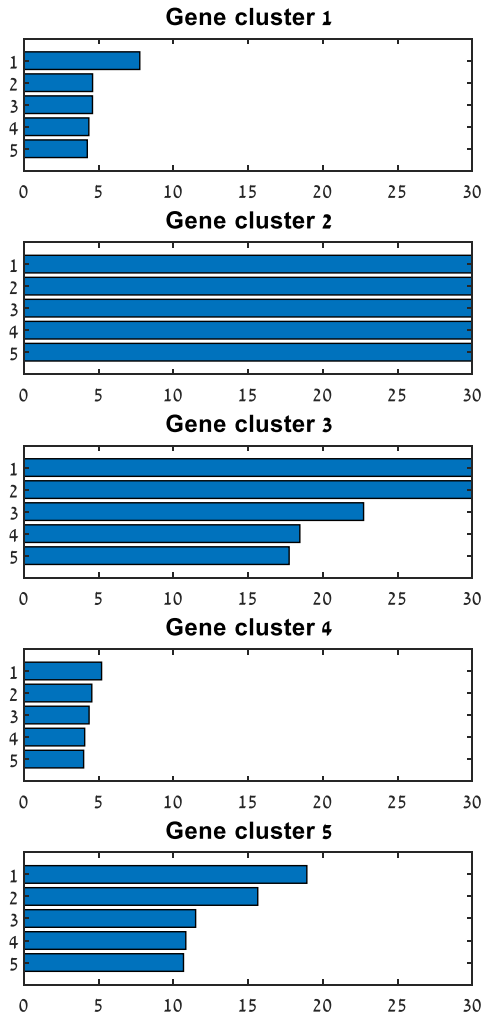


Figure S1.13A: Comparison of the methylation subgroups we identified and the Epi-LumB subgroups.

7.2. Supplement 2: Skin cancer subtypes



1.76e-08	**melanin biosynthetic process**
2.53e-05	negative regulation of neuron apoptotic process
2.56e-05	eye pigment biosynthetic process
4.49e-05	neuron projection development
5.69e-05	cell-cell adhesion mediated by cadherin

6.60e-69	cornification
1.74e-50	epidermis development
3.20e-36	keratinocyte differentiation
9.29e-35	peptide cross-linking
1.19e-33	keratinization

1.20e-44	immune response
1.50e-32	adaptive immune response
1.90e-23	inflammatory response
3.47e-19	T cell costimulation
1.81e-18	regulation of immune response

6.30e-06	ion transmembrane transport
2.85e-05	regulation of membrane potential
4.32e-05	chloride transport
8.52e-05	negative regulation of myoblast differentiation
1.00e-04	nervous system process

1.18e-19	nervous system development
2.27e-16	cell adhesion
3.24e-12	cell-cell signaling
1.47e-11	axon guidance
2.10e-11	chemical synaptic transmission

Figure S2.1: Gene ontology enrichments on the five gene clusters. The analysis was performed in PROMO, using FDR-corrected hypergeometric test p-values. The five most significant GO terms are listed for each gene cluster.

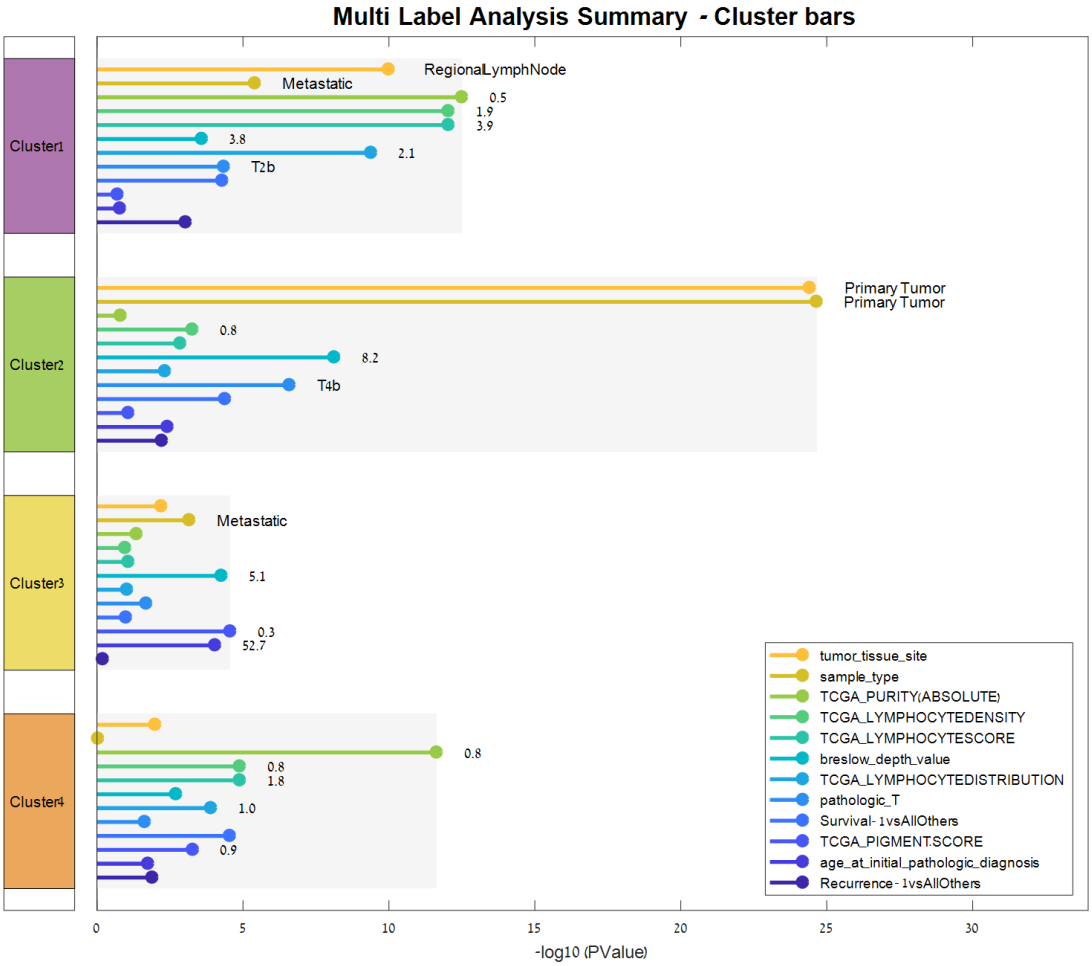


Figure S2.2: Enrichment for clinical labels on the four melanoma sample clusters. The four sample clusters were tested for enrichment for multiple clinical labels. The enrichments were tested using the hypergeometric test, and the top enriched labels on the clusters were plotted.

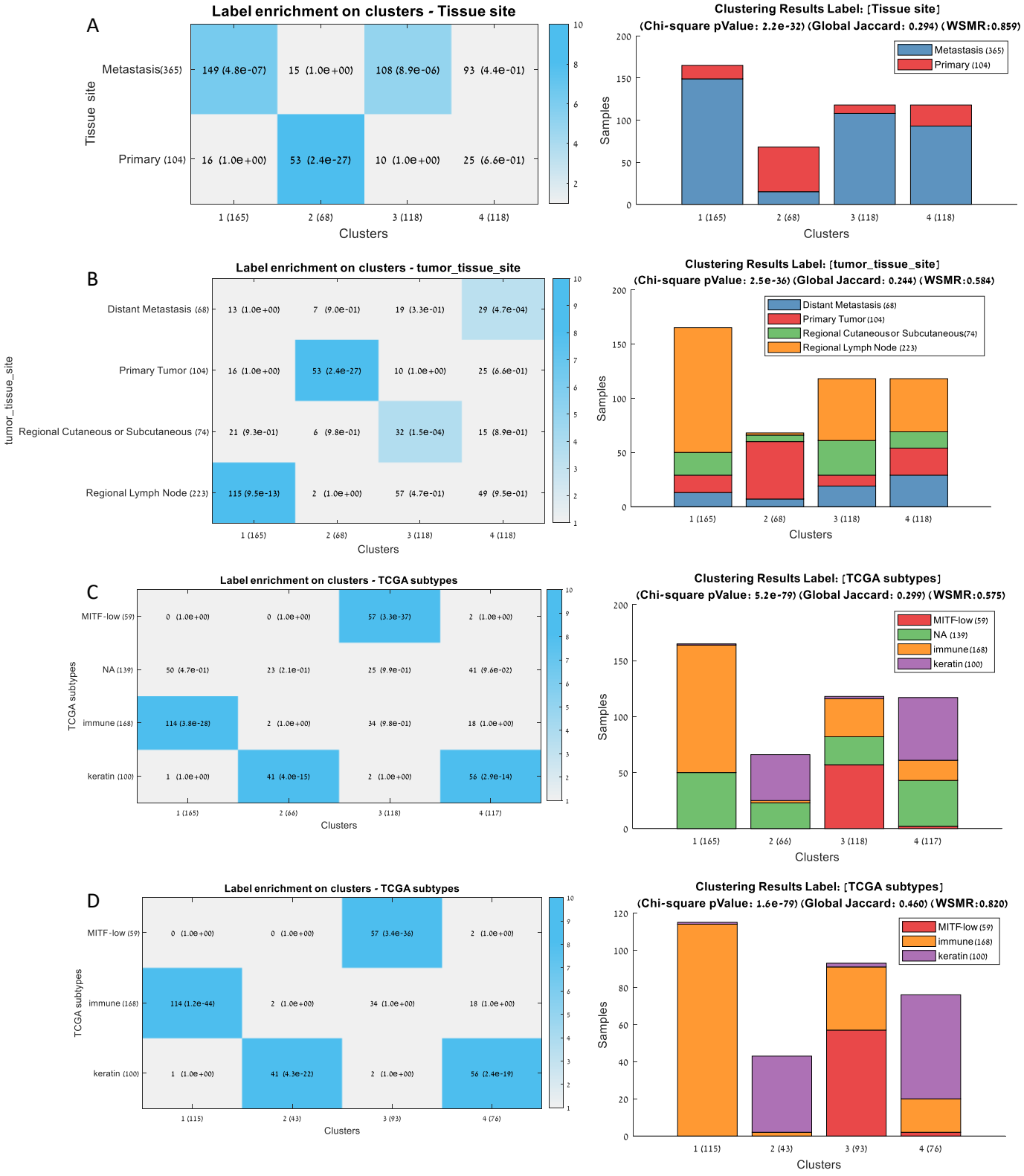


Figure S2.3: Characteristics of the four melanoma subtypes. Concordance between the four melanoma subgroups, tumor tissue sites, and TCGA's three transcriptomic subgroup labels. For each comparison, the histogram on the right shows the breakdown of samples in each subtype into categories, and the matrix on the left shows the confusion matrix. For each cell, the number of samples and p-value for enrichment based on the hypergeometric test is shown. **(A)** Primary vs. Metastasis **(B)** Detailed tissue site **(C)** TCGA's three transcriptomic subtypes, including NA value for new samples that were not included in TCGA's melanoma paper[48] **(D)** TCGA's three transcriptomic subtypes, omitting the NA samples.

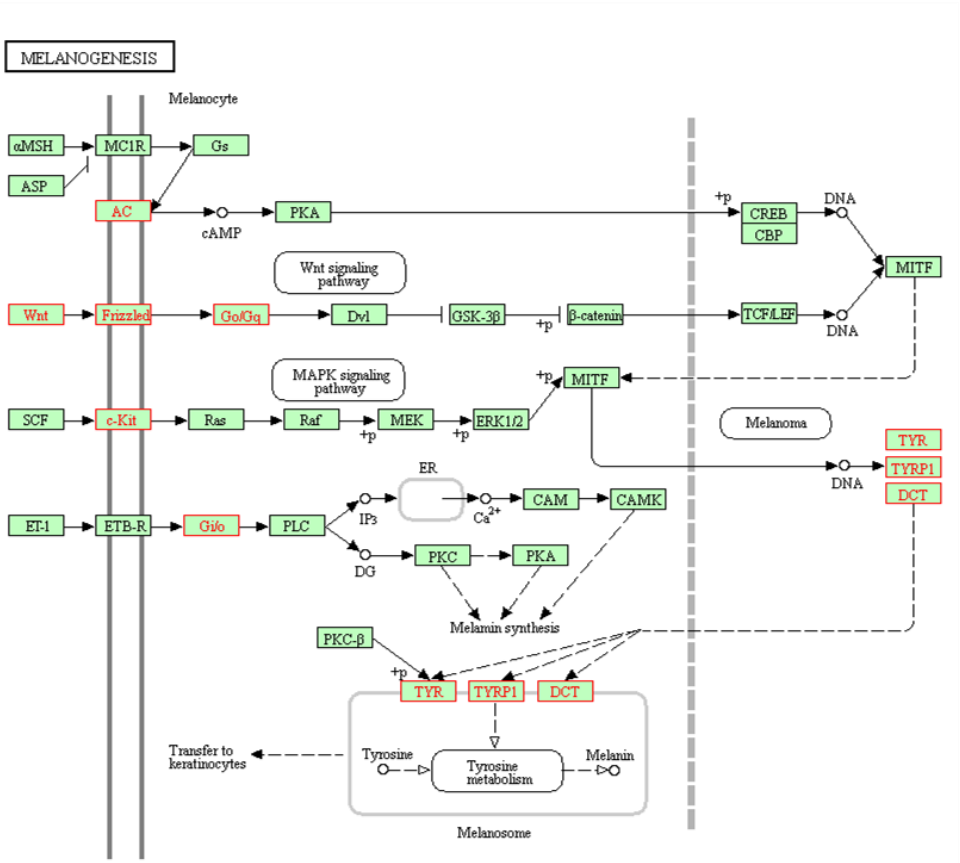


Figure S2.4: Sample-cluster 4 overexpressed genes that are enriched for the KEGG "Melanogenesis" pathway. Over-expressed genes (p<0.005) are marked in red.

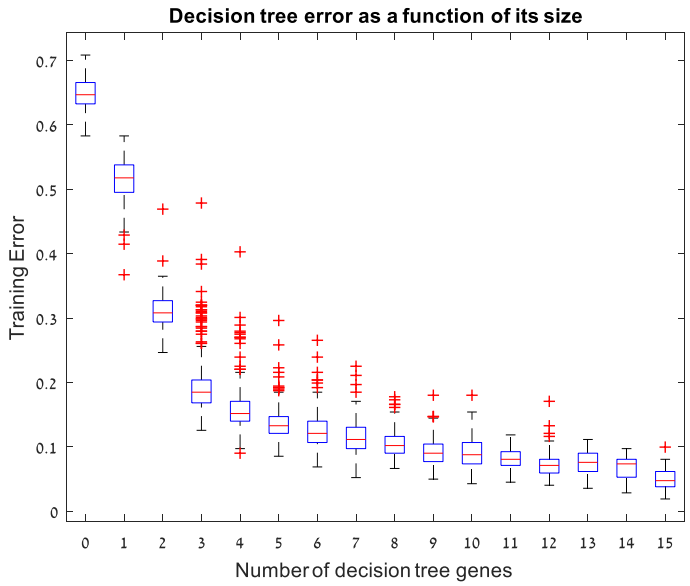


Figure S2.5: Error of decision tree classifiers as a function of the number of genes. For a varying number of genes (1-15) and for varying pruning levels (0-10), 30 decision trees were trained on resampled subsets of the dataset samples (resampling ratio of 0.9). The graph shows the average training error for each decision tree size. A three-gene classifier for predicting melanoma's molecular subtype gives a good balance between simplicity (avoiding over-fitting) and performance and reaches a training error that is close to that obtained by larger number of genes.

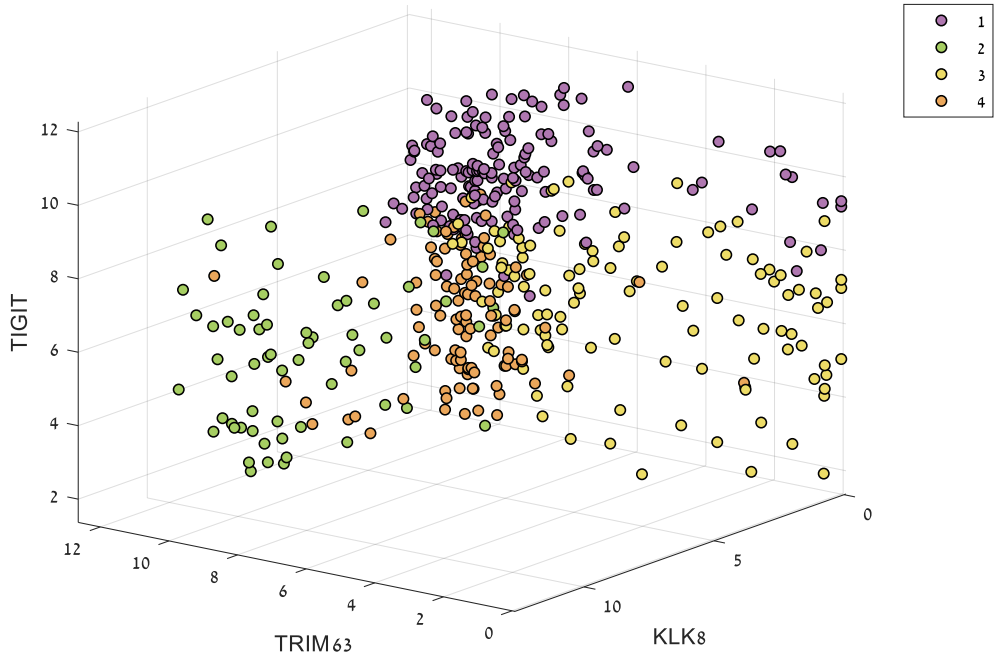


Figure S2.6: Dispersion of the 469 melanoma samples projected to the 3-dimensional space of the three selected classifier predictors: KLK8, TIGIT, and TRIM63. Samples are colored by the melanoma subgroup. The axis representing the expression level of the KLK gene distinguishes cluster 2 samples (green circles, "Keratin" subgroup) showing high levels of KLK8 expression, from all other clusters. The axis representing the expression of the TIGIT gene distinguishes cluster 1 samples (purple circles, "Immune" subgroup) showing high levels of TIGIT expression, from the other subgroups. Lastly, the axis representing the expression of the TRIM63 gene distinguished cluster 3 samples (yellow circles, "Melanogenesis-low" subtype) from cluster 4 samples (orange circles, "Melanogenesis-high" subtype).

		Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Sample number		**105**	**68**	**118**	**118**	**469**
Tissue site	**Primary Tumor**	16	53	10	25	**104**
	Regional Cutaneous or Subcutaneous	21	6	32	15	**74**
	Regional Lymph Node	115	2	57	49	**223**
	Distant Metastasis	13	7	19	29	**68**
TCGA's Transcriptomic subtypes	**Immune**	114	2	34	18	**168**
	Keratin	1	41	2	56	**100**
	MITF-Low	0	0	57	2	**59**
	NA	50	23	25	41	**139**

Table S2.1: Characterization of the four melanoma subgroups.

Gene Cluster	GO Term	#genes	Raw p-value	Empirical p-Value	Gene List
1	neurogenesis - GO:0022008	68	7.10E-14	2.00E-04	[GABRB3, ROBO2, ONECUT2, LAMC3, CTNND2, RASGRF1, PCSK9, KNDC1, NR2E1, DUSP15, CDH4, SOX1, DPYSL4, CDH1, CRTAC1, SALL4, EFHD1, NRTN, PLXNC1, PITX2, SOX6, SH3GL2, MNX1, EPHA5, NKX2-8, SERPINF1, ADRA2C, MAPK8IP2, POU3F1, ISL1, OLFM1, NRG3, DCT, RTN4RL1, KIT, LMX1B, MAPT, HAP1, BRSK2, ARX, PRDM13, PRELP, ADCY1, FSTL4, DLL3, TRPM1, VTN, GBX2, CTNNA2, PROM1, WNT4, SPTBN2, NKX2-2, MCOLN3, SYT3, GABRA5, MCF2, FZD9, DCDC2, L1CAM, POU4F1, GNAO1, CPEB1, NR4A3, NEURL1, RAB17, MDGA2, BMPR1B]
	somatodendritic compartment - GO:0036477	34	1.60E-08	4.00E-04	[PCSK2, CHRM1, CTTNBP2, CTNND2, CACNA1B, ADCY2, KNDC1, CPN1, KCNN1, KCNN2, PPARGC1A, ANKS1B, KCNH1, SPTBN2, SYT5, EPHA5, MME, GABRA5, SERPINF1, MAPK8IP2, L1CAM, GNAO1, OLFM1, CPEB1, MLPH, NOV, RTN4RL1, NEURL1, EEF1A2, RAB17, MAPT, BMPR1B, CRYAB, PDE9A]
	melanosome membrane - GO:0033162	6	2.34E-08	4.00E-04	[OCA2, SLC45A2, GPR143, DCT, TYRP1, TYR]
	cell body - GO:0044297	28	2.70E-08	6.00E-04	[PCSK2, CTNND2, CACNA1B, KNDC1, CPN1, KCNN1, KCNN2, PPARGC1A, KCNH1, SPTBN2, SYT5, EPHA5, GABRA5, SERPINF1, MAPK8IP2, L1CAM, GNAO1, OLFM1, CPEB1, NOV, RTN4RL1, NEURL1, EEF1A2, RAB17, MAPT, BMPR1B, CRYAB, PDE9A]
	secondary metabolic process - GO:0019748	9	4.43E-08	6.00E-04	[OCA2, SLC45A2, AS3MT, ABCC2, CITED1, DCT, TYRP1, CYP1A1, TYR]
	anatomical structure morphogenesis - GO:0009653	66	4.48E-08	6.00E-04	[ROBO2, RYR1, ONECUT2, CITED1, LAMC3, CTNND2, ONECUT1, LDB3, KNDC1, NR2E1, MYLK3, CDH4, SOX1, DPYSL4, KRT27, SALL4, RBPMS2, CAPN3, NRTN, PITX2, SOX6, SH3GL2, MNX1, KCNH1, EPHA5, NKX2-8, RIPK4, MAPK8IP2, MMP8, ISL1, OLFM1, CEACAM1, NOV, NRG3, SFRP5, ITGA7, CRYAB, MET, BRSK2, SLC24A4, HPGD, ARX, KLK4, GATA4, PRELP, ADCY1, CACNA1H, DLL3, TRPM1, VTN, GBX2, MYH14, TNNI3, CTNNA2, PROM1, WNT4, SPTBN2, CAP2, FOXF2, L1CAM, POU4F1, NR4A3, NOX5, RADIL, BMPR1B, RAMP1]
	neuron projection development - GO:0031175	31	4.80E-08	6.00E-04	[ROBO2, BRSK2, CTNND2, ARX, RASGRF1, PRELP, NR2E1, ADCY1, CDH4, DPYSL4, GBX2, CDH1, CRTAC1, EFHD1, NRTN, CTNNA2, SH3GL2, MNX1, SPTBN2, EPHA5, NKX2-8, MCF2, MAPK8IP2, L1CAM, POU4F1, ISL1, GNAO1, NR4A3, RTN4RL1, MAPT, BMPR1B]
	central nervous system development - GO:0007417	39	4.91E-08	6.00E-04	[ROBO2, CITED1, LAMC3, CTTNBP2, ARX, KNDC1, NR2E1, VTN, SOX1, GBX2, CDH1, CRTAC1, CTNNA2, SOX6, PITX2, PPARGC1A, SH3GL2, MNX1, NKX2-2, WNT4, SPTBN2, EPHA5, SLC6A17, S100A1, GABRA5, POU3F1, POU4F1, ISL1, GNAO1, BCAN, NR4A3, NRG3, DCT, RTN4RL1, NEURL1, MAL, HAP1, MAPT, MDGA2]
	melanin metabolic process - GO:0006582	6	7.19E-08	8.00E-04	[OCA2, SLC45A2, CITED1, DCT, TYRP1, TYR]
	developmental pigmentation - GO:0048066	8	2.97E-07	0.0034	[OCA2, SLC45A2, GPR143, CITED1, DCT, KIT, TYRP1, TYR]
2	skin development - GO:0043588	85	4.46E-72	2.00E-04	[FOXE1, ITGB4, KRT23, ABCA12, TGM1, LCE1B, CASP14, PRSS8, TGM5, KRT6C, TGM3, RPTN, TP63, KRT6B, KRT6A, DSP, KRT4, KRT2, KRT1, SPINK5, KRT79, KRT78, KRT7, KRT77, KRT5, OVOL1, KRT75, LOR, LCE1C, EREG, LCE2B, FLG2, LCE2C, CLDN4, LCE2A, DSG1, PKP1, PKP3, DSG3, IRF6, DSC1, IVL, DSC2, S100A7, DSC3, SPRR2E, FLG, SPRR3, SPRR2G, CSTA, KRT80, KLK5, ALOX12B, APCDD1, EVPL, KLK8, PPL, EGFR, LCE3D, SCEL, PERP, SFN, PI3, SPRR2A, SPRR2B, ALOXE3, SPRR2D, CDSN, C1orf68, KLK13, KRT13, GRHL3, KRT10, CNFN, ASPRV1, LCE3E, KRT19, KRT17, GJB3, KRT16, KRT15, KRT14, SPRR1A, FGFR2, SPRR1B]
	epidermal cell differentiation - GO:0009913	72	6.99E-62	2.00E-04	[KRT23, ABCA12, TGM1, LCE1B, CASP14, PRSS8, TGM5, KRT6C, TGM3, RPTN, TP63, KRT6B, KRT6A, DSP, KRT4, KRT2, KRT1, SPINK5, KRT79, KRT78, KRT7, KRT77, KRT5, KRT75, LOR, LCE1C, EREG, LCE2B, LCE2C, LCE2A, DSG1, PKP1, PKP3, DSG3, IRF6, DSC1, IVL, DSC2, S100A7, DSC3, SPRR2E, FLG, SPRR3, SPRR2G, CSTA, KRT80, KLK5, EVPL, KLK8, PPL, LCE3D, SCEL, PERP, SFN, PI3, SPRR2A, SPRR2B, SPRR2D, CDSN, C1orf68, KLK13, KRT13, KRT10, CNFN, LCE3E, KRT19, KRT17, KRT16, KRT15, KRT14, SPRR1A, SPRR1B]
	cornified envelope - GO:0001533	40	1.51E-55	2.00E-04	[SPRR2E, FLG, SPRR3, CSTA, SPRR2G, EVPL, CST6, PPL, TGM1, SCEL, LCE3D, LCE1B, PI3, SPRR2A, SPRR2B, RPTN, SPRR2D, DSP, CDSN, KRT2, C1orf68, KRT1, KRT10, CNFN, LOR, LCE3E, LCE1C, LCE2B, LCE2C, LCE2A, PKP1, DSG1, DSG3, PKP3, SPRR1A, DSC1, IVL, SPRR1B, DSC2, DSC3]
	epithelial cell differentiation - GO:0030855	87	3.37E-53	2.00E-04	[EHF, KRT23, TFCP2L1, ABCA12, TGM1, LCE1B, CASP14, PRSS8, TGM5, KRT6C, TGM3, RPTN, TP63, KRT6B, KRT6A, DSP, KRT4, KRT2, KRT1, SPINK5, KRT79, KRT78, KRT7, KRT77, FOXL2, KRT5, KRT75, LOR, LCE1C, EREG, LCE2B, LCE2C, LCE2A, ELF3, DSG1, PKP1, PKP3, DSG3, IRF6, RHCG, DSC1, IVL, DSC2, S100A7, DSC3, SPRR2E, FLG, SPRR3, SPRR2G, CSTA, KRT80, DLX3, KLK5, EVPL, KLK8, PPL, LCE3D, SCEL, RAB25, PERP, SFN, PI3, SPRR2A, SPRR2B, SPRR2D, CDSN, PSAPL1, WNT7B, AKR1C1, C1orf68, KLK13, KRT13, AKR1C2, PTK6, KRT10, CNFN, GRHL2, LCE3E, KRT19, KRT17, KRT16, KRT15, KRT14, CD24, SPRR1A, FGFR2, SPRR1B]
	peptide cross-linking - GO:0018149	28	3.95E-34	2.00E-04	[SPRR2E, FLG, SPRR3, CSTA, EVPL, TGM1, LCE3D, LCE1B, PI3, SPRR2A, TGM5, SPRR2B, TGM3, SPRR2D, DSP, KRT2, C1orf68, KRT1, KRT10, LOR, LCE3E, LCE1C, LCE2B, LCE2C, LCE2A, SPRR1A, IVL, SPRR1B]
	extracellular exosome - GO:0070062	118	5.88E-23	2.00E-04	[CALML5, CBLC, CALML3, DEFB1, AQP5, TGM1, CKMT1B, PRSS8, PRSS3, KRT6C, TGM3, KRT6B, CD177, KRT6A, GBP6, ENTPD2, KRT2, KRT1, SPINK5, SLC6A14, KRT79, KRT78, KRT7, KRT77, KRT5, NCCRP1, KRT75, SULT2B1, FLG2, SLPI, SPINT1, SCNN1B, SCNN1A, PKP1, RHCG, S100A9, DSC1, S100A8, DSC2, S100A7, KPRP, CRABP2, TACSTD2, SLC5A1, EVPL, PPL, FUT3, PKHD1, SFN, PI3, S100A14, CLCA4, PROM2, GGT6, CDSN, C1orf68, A2ML1, CNFN, CRNN, AKR1B10, LCN2, FAT2, SPRR1B, LAD1, CLIC3, PCDHGB5, ITGB4, DMKN, CASP14, CTSG, ITGB6, SERPINB3, DSP, SERPINB4, LYNX1, GPX2, MMP7, ARG1, ANXA3, SERPINB5, TMPRSS11D, EPN3, CEACAM5, DSG1, IRF6, DSG3, IVL, SPRR3, LRRC15, CSTA, SERPINB13, SBSN, CST6, SCEL, RAB25, RNASE7, MAL2, EPS8L2, LGALS7B, WNT7B, AKR1C1, CKMT1A, KLK13, KRT13, BBOX1, KRT10, PDZK1IP1, SLURP1, KLK11, KRT19, KRT17, KRT16, KRT15, KRT14, SAA1, S100P, NECTIN4, C1orf116]
	extracellular vesicle - GO:1903561	118	9.25E-23	2.00E-04	[CALML5, CBLC, CALML3, DEFB1, AQP5, TGM1, CKMT1B, PRSS8, PRSS3, KRT6C, TGM3, KRT6B, CD177, KRT6A, GBP6, ENTPD2, KRT2, KRT1, SPINK5, SLC6A14, KRT79, KRT78, KRT7, KRT77, KRT5, NCCRP1, KRT75, SULT2B1, FLG2, SLPI, SPINT1, SCNN1B, SCNN1A, PKP1, RHCG, S100A9, DSC1, S100A8, DSC2, S100A7, KPRP, CRABP2, TACSTD2, SLC5A1, EVPL, PPL, FUT3, PKHD1, SFN, PI3, S100A14, CLCA4, PROM2, GGT6, CDSN, C1orf68, A2ML1, CNFN, CRNN, AKR1B10, LCN2, FAT2, SPRR1B, LAD1, CLIC3, PCDHGB5, ITGB4,

					DMKN, CASP14, CTSG, ITGB6, SERPINB3, DSP, SERPINB4, LYNX1, GPX2, MMP7, ARG1, ANXA3, SERPINB5, TMPRSS11D, EPN3, CEACAM5, DSG1, IRF6, DSG3, IVL, SPRR3, LRRC15, CSTA, SERPINB13, SBSN, CST6, SCEL, RAB25, RNASE7, MAL2, EPS8L2, LGALS7B, WNT7B, AKR1C1, CKMT1A, KLK13, KRT13, BBOX1, KRT10, PDZK1IP1, SLURP1, KLK11, KRT19, KRT17, KRT16, KRT15, KRT14, SAA1, S100P, NECTIN4, C1orf116]
	extracellular organelle - GO:0043230	118	1.00E-22	2.00E-04	[CALML5, CBLC, CALML3, DEFB1, AQP5, TGM1, CKMT1B, PRSS8, PRSS3, KRT6C, TGM3, KRT6B, CD177, KRT6A, GBP6, ENTPD2, KRT2, KRT1, SPINK5, SLC6A14, KRT79, KRT78, KRT7, KRT77, KRT5, NCCRP1, KRT75, SULT2B1, FLG2, SLPI, SPINT1, SCNN1B, SCNN1A, PKP1, RHCG, S100A9, DSC1, S100A8, DSC2, S100A7, KPRP, CRABP2, TACSTD2, SLC5A1, EVPL, PPL, FUT3, PKHD1, SFN, PI3, S100A14, CLCA4, PROM2, GGT6, CDSN, C1orf68, A2ML1, CNFN, CRNN, AKR1B10, LCN2, FAT2, SPRR1B, LAD1, CLIC3, PCDHGB5, ITGB4, DMKN, CASP14, CTSG, ITGB6, SERPINB3, DSP, SERPINB4, LYNX1, GPX2, MMP7, ARG1, ANXA3, SERPINB5, TMPRSS11D, EPN3, CEACAM5, DSG1, IRF6, DSG3, IVL, SPRR3, LRRC15, CSTA, SERPINB13, SBSN, CST6, SCEL, RAB25, RNASE7, MAL2, EPS8L2, LGALS7B, WNT7B, AKR1C1, CKMT1A, KLK13, KRT13, BBOX1, KRT10, PDZK1IP1, SLURP1, KLK11, KRT19, KRT17, KRT16, KRT15, KRT14, SAA1, S100P, NECTIN4, C1orf116]
	structural molecule activity - GO:0005198	54	5.09E-21	2.00E-04	[LAD1, WWC1, KRT23, LCE1B, EPB41L4B, KRT6C, KRT6B, KRT6A, DSP, KRT4, KRT2, KRT1, KRT79, KRT78, KRT7, KRT77, KRT5, KRT75, LOR, LCE1C, LCE2B, FLG2, LCE2C, CLDN4, LCE2A, PKP1, INA, IVL, SHANK2, SPRR2E, FLG, SPRR3, CSTA, KRT80, EVPL, PPL, LCE3D, MAL2, PI3, SPRR2A, SPRR2B, SPRR2D, C1orf68, KRT13, KRT10, LCE3E, MPP7, KRT19, KRT17, KRT16, KRT15, KRT14, SPRR1A, SPRR1B]
	intermediate filament - GO:0005882	28	4.63E-18	2.00E-04	[FLG, KRT80, KRT23, CASP14, KRT6C, KRT6B, KRT6A, DSP, KRT4, KRT2, KRT1, KRT13, KRT79, KRT78, KRT7, KRT77, KRT10, KRT5, KRT75, KRT19, KRT17, KRT16, KRT15, KRT14, PKP1, EPPK1, INA, SHANK2]
3	immune system process - GO:0002376	176	8.58E-85	2.00E-04	[FCN1, ADAMDEC1, FCMR, NCF1, ATP8A1, CLEC10A, AQP9, SIRPG, HP, LY75, PRF1, RORC, SLA2, CXCL13, IKZF3, CLU, IFI44L, VPREB3, GPR174, CYSLTR2, SIT1, RGS1, TBC1D10C, TNFSF11, HLA-DOA, HLA-DOB, ZNF683, ZBP1, GBP5, CD96, LAG3, PRKCB, THEMIS, HLA-G, LAX1, CHIT1, FCAMR, CD8B, IFI27, CD8A, PADI2, PRKCQ, CLEC4E, CARD11, SKAP1, IDO1, BLK, TNFRSF11B, GATA3, CD1C, LY9, PLAC8, SPTA1, C3, CD79B, CD79A, KLRK1, C7, UBD, CD19, BTLA, NLRP2, SLAMF7, TNFRSF17, SLAMF6, ICOS, HLA-DQA2, HLA-DQA1, SLAMF1, HLA-DRB5, SIGLEC14, KLRC2, BCL11B, TNFRSF9, CRTAM, SH2D1A, IFNLR1, LYZ, SELE, SELP, MARCO, CXCL10, CXCL11, PTPRC, SELL, CD27, KLRD1, IL7R, HAMP, PIGR, ITK, CIITA, CXCL9, TNFRSF13B, FASLG, CD3G, CTSW, PTPN22, LRMP, CD3E, ITGAL, CD3D, PIK3CG, JCHAIN, TNFSF13B, SPN, GNLY, KYNU, OLR1, CTLA4, CD38, CCR7, LBP, CCR5, CCR2, CR2, CR1, ITGA4, RHOH, PAX5, MMP9, ZAP70, HSH2D, AIM2, ITGAD, IFNG, LCK, BANK1, IL1B, XCL2, CHI3L1, TLR8, CD48, TLR10, LTB, SMPDL3B, MS4A1, HLA-DQB2, LTF, BIRC3, CCL14, CD5L, FGL2, CXCR5, CST7, IL2RG, LILRA3, LILRA4, CCL8, CCL5, CXCR3, TBX21, IL21R, CCL19, CCL18, IL12RB1, GBP1, IL33, PLA2G2D, CCL22, CCL21, ERAP2, TRAT1, CD70, GZMA, GZMB, LILRB1, GZMH, CD2, CD6, CD5, CAMK4, POU2AF1, CD7, CD247, PDCD1]
	regulation of immune system process - GO:0002682	116	3.02E-57	2.00E-04	[FCN1, CLEC10A, SIRPG, SLA2, CXCL13, IKZF3, CLU, SIT1, GPR171, TBC1D10C, TNFSF11, UBASH3A, HLA-DOA, HLA-DOB, ZNF683, GBP5, CD96, LAG3, PRKCB, THEMIS, HLA-G, LAX1, CD8B, CD8A, PADI2, PRKCQ, CLEC4E, CARD11, SKAP1, IDO1, BLK, KLRB1, GATA3, CD1C, KIR2DL4, SPTA1, C3, CD79B, CD79A, KLRK1, C7, CD19, BTLA, SLAMF7, STAP1, SLAMF6, ICOS, HLA-DQA2, HLA-DQA1, SLAMF1, HLA-DRB5, CRTAM, SH2D1A, IFNLR1, SELP, MARCO, CXCL10, CXCL11, PTPRC, SELL, CD27, KLRD1, IL7R, PIGR, ITK, CXCL9, TNFRSF13B, CD3G, PTPN22, CD3E, ITGAL, CD3D, TNFSF13B, SPN, CTLA4, CD38, CCR7, LBP, CCR2, CR2, CR1, ITGA4, MMP12, GREM1, FCER2, ZAP70, AIM2, IFNG, LCK, BANK1, IL1B, TLR8, CD48, TLR10, SMPDL3B, HLA-DQB2, LTF, BIRC3, CCL5, CXCR3, TBX21, CCL19, TIGIT, IL12RB1, GBP1, IL33, PLA2G2D, CCL21, TRAT1, LILRB1, CD2, CD6, CD5, CAMK4, CD247, PDCD1]
	regulation of cell activation - GO:0050865	61	1.40E-37	2.00E-04	[TNFRSF13B, SIRPG, PTPN22, CD3G, CD3E, SLA2, CD3D, IKZF3, TNFSF13B, SPN, SIT1, TBC1D10C, CD38, TNFSF11, CTLA4, CCR7, LBP, HLA-DOA, CCR2, ZNF683, LAG3, HLA-G, LAX1, ZAP70, IFNG, BANK1, LCK, IL1B, PRKCQ, HLA-DQB2, CARD11, IDO1, GATA3, SPTA1, KLRK1, CCL5, TBX21, BTLA, STAP1, CCL19, TIGIT, IL12RB1, ICOS, HLA-DQA2, HLA-DQA1, SLAMF1, IL33, PLA2G2D, HLA-DRB5, CCL21, LILRB1, CD2, SELP, PTPRC, CD6, CD5, CAMK4, CD27, CD247, PDCD1, IL7R]
	regulation of leukocyte cell-cell adhesion - GO:1903037	51	1.03E-36	2.00E-04	[SIRPG, PTPN22, CD3G, CD3E, CD3D, TNFSF13B, SPN, SIT1, TNFSF11, CTLA4, CCR7, HLA-DOA, CCR2, ZNF683, LAG3, ITGA4, HLA-G, LAX1, ZAP70, IFNG, LCK, IL1B, PRKCQ, HLA-DQB2, CARD11, IDO1, GATA3, SPTA1, KLRK1, CCL5, BTLA, CCL19, TIGIT, IL12RB1, ICOS, HLA-DQA2, HLA-DQA1, SLAMF1, PLA2G2D, HLA-DRB5, CCL21, LILRB1, CD2, PTPRC, CD6, CD5, CAMK4, CD27, CD247, PDCD1, IL7R]
	regulation of T cell activation - GO:0050863	50	1.36E-36	2.00E-04	[SIRPG, PTPN22, CD3G, CD3E, CD3D, TNFSF13B, SPN, SIT1, TNFSF11, CTLA4, CCR7, HLA-DOA, CCR2, ZNF683, LAG3, HLA-G, LAX1, ZAP70, IFNG, LCK, IL1B, PRKCQ, HLA-DQB2, CARD11, IDO1, GATA3, SPTA1, KLRK1, CCL5, BTLA, CCL19, TIGIT, IL12RB1, ICOS, HLA-DQA2, HLA-DQA1, SLAMF1, PLA2G2D, HLA-DRB5, CCL21, LILRB1, CD2, PTPRC, CD6, CD5, CAMK4, CD27, CD247, PDCD1, IL7R]
	positive regulation of leukocyte activation - GO:0002696	45	8.39E-32	2.00E-04	[SIRPG, CD3G, GATA3, CD3E, CD3D, TNFSF13B, SPN, SPTA1, KLRK1, CCL5, TBX21, BTLA, CD38, TNFSF11, CTLA4, STAP1, CCR7, CCL19, LBP, IL12RB1, ICOS, HLA-DQA2, HLA-DQA1, SLAMF1, CCR2, IL33, HLA-DRB5, CCL21, LILRB1, HLA-G, CD2, ZAP70, PTPRC, IFNG, CD6, LCK, CD5, IL1B, CD27, PRKCQ, CD247, PDCD1, IL7R, HLA-DQB2, CARD11]
	innate immune response - GO:0045087	60	1.28E-31	2.00E-04	[FCN1, ITK, CIITA, NCF1, CLEC10A, SLA2, CLU, PIK3CG, JCHAIN, KYNU, LBP, ZNF683, ZBP1, GBP5, CR2, CR1, HLA-G, ZAP70, AIM2, IFNG, IFI27, LCK, XCL2, TLR8, TLR10, SMPDL3B, CLEC4E, HLA-DQB2, LTF, BLK, CCL14, GATA3, LY9, C3, CCL8, KLRK1, C7, CCL5, UBD, NLRP2, SLAMF7, SLAMF6, CCL19, CCL18, IL12RB1, HLA-DQA2, GBP1, HLA-DQA1, SLAMF1, HLA-DRB5, SIGLEC14, CCL22, CCL21, KLRC2, SH2D1A, GZMB, IFNLR1, MARCO, CD6, KLRD1]
	external side of plasma membrane - GO:0009897	40	2.07E-31	2.00E-04	[FCN1, CXCL9, TNFRSF13B, CXCR5, FASLG, IL2RG, CD3E, SPN, CD79B, CD79A, KLRK1, CD19, CXCR3, CTLA4, CCR7, CCR5, IL12RB1, ICOS, SLAMF1, LAG3, TNFRSF9, LILRB1, GP1BA, SELP, CD2, FCER2, CXCL10, PTPRC, IFNG, SELL, CD8B, CD5, CD8A, CD27, TLR8, KLRD1, PDCD1, CD69, IL7R, MS4A1]
	regulation of cell adhesion - GO:0030155	59	7.14E-29	2.00E-04	[ADAMDEC1, SIRPG, PTPN22, CD3G, CD3E, CXCL13, CD3D, PIK3CG, TNFSF13B, SPN, SIT1, TNFSF11, CTLA4, CCR7, CYTIP, HLA-DOA, CCR2, ZNF683, LAG3, ITGA4, HLA-G, LAX1, GREM1, ZAP70, IFNG, LCK, IL1B, PRKCQ, HLA-DQB2, CARD11, SKAP1, IDO1, GATA3, SPTA1, KLRK1, CCL5, ABI3BP, BTLA, CCL19, TIGIT, IL12RB1, ICOS, HLA-DQA2, GBP1, HLA-DQA1, SLAMF1, PLA2G2D, HLA-DRB5, CCL21, LILRB1, CD2, PTPRC, CD6, CD5, CAMK4, CD27, CD247, PDCD1, IL7R]
	T cell activation - GO:0042110	36	3.54E-26	2.00E-04	[ITK, RORC, PTPN22, CD3G, GATA3, ITGAL, CD1C, CD3E, SLA2, CD3D, LY9, PIK3CG, SPN, TBX21, CCR7, SLAMF6, CCL19, PLA2G2D, CCL21, BCL11B, CRTAM, THEMIS, RHOH, LILRB1, CD2, ZAP70, HSH2D, PTPRC, IFNG, CD8B, LCK, CD8A, CD7, CLEC4E, IL7R, CARD11]

4	chloride channel activity - GO:0005254	10	3.01E-06	0.0254	[GABRP, GLRA2, CLCN4, TTYH1, FXYD3, GABRA3, FXYD1, SLC26A4, ANO5, GABRG2]
	neurogenesis - GO:0022008	53	5.00E-06	0.0392	[ALK, ATP8A2, PPP1R9A, RND2, KIF17, HOXC10, RIMS2, GRIP1, CHL1, SOX8, PHGDH, SOX9, NEFH, TRIM67, CHRNB2, MYOC, COL25A1, OLIG1, OLIG2, GFRA3, MAG, SFRP1, OLFM3, RARB, ASPA, CRB1, NLGN1, PLPPR5, LRP4, BHLHE22, UGT8, SLITRK2, PTPRZ1, MAP2, CNR1, FLRT1, ADGRG6, SPP1, APOD, LINGO2, BCHE, KCNJ10, EYA1, CNTN6, LGI4, S100B, SORL1, MT3, LIN28A, LHX2, FABP7, FGF13, HCN1]
5	cell adhesion - GO:0007155	81	1.34E-28	2.00E-04	[PCDHGB7, SPON1, TENM3, TNC, HBB, ICAM5, SLC7A11, ARHGAP6, HAPLN4, HAPLN1, COMP, CDH2, ITGB8, NRCAM, EDIL3, PCDHAC2, POSTN, KIRREL2, ACTN2, APLP1, OMD, EPDR1, PCDHA13, PCDHA11, PCDHA10, CLDN11, IL1RAPL1, ADGRB1, PKP2, ITGA8, COL8A1, EPHA3, FREM2, ASTN1, PCDH10, NRXN1, NTM, NRXN3, ADAM22, NRXN2, THBS2, THBS4, COL19A1, ADD2, ACAN, NT5E, SRPX2, EFS, RELN, FLRT3, PCDHA5, PCDHA4, SPOCK1, PCDHA3, NCAM1, NCAM2, PCDHA7, PCDHA6, NLGN4Y, NLGN4X, PCDH9, ANGPT1, NEGR1, PCDH20, BMP7, PTPRD, NFASC, KRT18, PCDHB2, ITGA10, PCDHB16, PCDHB6, CNTN1, ITGBL1, PCDHB5, CNTN3, PCDHB3, CNTN4, NECTIN3, SDK2, ADGRL3]
	biological adhesion - GO:0022610	81	2.22E-28	2.00E-04	[PCDHGB7, SPON1, TENM3, TNC, HBB, ICAM5, SLC7A11, ARHGAP6, HAPLN4, HAPLN1, COMP, CDH2, ITGB8, NRCAM, EDIL3, PCDHAC2, POSTN, KIRREL2, ACTN2, APLP1, OMD, EPDR1, PCDHA13, PCDHA11, PCDHA10, CLDN11, IL1RAPL1, ADGRB1, PKP2, ITGA8, COL8A1, EPHA3, FREM2, ASTN1, PCDH10, NRXN1, NTM, NRXN3, ADAM22, NRXN2, THBS2, THBS4, COL19A1, ADD2, ACAN, NT5E, SRPX2, EFS, RELN, FLRT3, PCDHA5, PCDHA4, SPOCK1, PCDHA3, NCAM1, NCAM2, PCDHA7, PCDHA6, NLGN4Y, NLGN4X, PCDH9, ANGPT1, NEGR1, PCDH20, BMP7, PTPRD, NFASC, KRT18, PCDHB2, ITGA10, PCDHB16, PCDHB6, CNTN1, ITGBL1, PCDHB5, CNTN3, PCDHB3, CNTN4, NECTIN3, SDK2, ADGRL3]
	extracellular matrix - GO:0031012	59	8.97E-25	2.00E-04	[VIT, SPON1, CPXM2, ELN, TNC, PCSK6, HAPLN4, HAPLN1, COMP, PODNL1, FGF9, EMILIN3, COL10A1, EDIL3, TIMP4, POSTN, APLP1, OMD, P3H2, WNT16, ASPN, SFRP2, MMP13, MMP16, COL8A1, COL4A5, ANGPTL4, EMID1, FREM2, COL11A1, PTN, THBS2, THBS4, COL19A1, COCH, ACAN, RELN, FLRT3, EPYC, GPC3, SPOCK1, CILP2, NDP, GPC4, WNT2, GPC6, LRRN3, TFPI2, BMP7, MFAP5, LRFN5, CILP, SMOC1, OGN, MFAP2, COL20A1, COL9A3, LRRN1, FMOD]
	neurogenesis - GO:0022008	99	1.28E-23	2.00E-04	[STMN2, TNC, SOX2, FGF5, SALL1, CDH2, DPYSL5, KIF5C, KIF5A, NRCAM, POSTN, OMD, SOX11, ANK3, POU3F2, DKK1, ISL2, SFRP2, DOK5, ADGRB3, ADGRB1, EPHA3, NGEF, ASTN1, DLX1, NDRG4, DLX2, SHC3, DLX5, NTM, LPAR1, EFNB3, FLRT3, NPTX1, NKX2-5, NKX6-1, WNT2, PLXNA4, RAP1GAP2, NGFR, LRRN3, SYT1, BDNF, LIF, INHBA, BMP7, PTPRD, BMP2, LRFN5, GDNF, TRPV4, OGN, CNTN1, ZNF536, LRRN1, CNTN4, SDK2, SNAP25, TENM3, AREG, UCHL1, NEFL, NEFM, ZNF521, FOXD1, EDN3, GFRA1, WNT16, GFRA2, GAP43, HAND2, ALDH1A2, IL1RAPL1, DCX, SHANK1, LTK, SEMA3A, NRXN1, SEMA3B, NRXN3, ADAM22, PTN, RELN, ERBB4, EPYC, SLITRK6, SPOCK1, SLITRK5, NCAM1, CSMD3, NCAM2, WASF3, NTRK2, NLGN4X, NEGR1, VAX1, NFASC, FMOD, ADGRL3]
	anatomical structure morphogenesis - GO:0009653	115	3.80E-23	2.00E-04	[TNC, SOX2, COMP, SALL1, CDH2, BMPER, FGF9, DPYSL5, KIF5C, KIF5A, NRCAM, MYOZ2, POSTN, IGFBP5, ACTN2, OSR1, APLP1, OMD, NPY1R, SOX11, ANK3, DKK1, SFRP4, RBP4, ISL2, SFRP2, DOK5, ADGRB3, ADGRB1, PKP2, COL8A1, FREM2, DLX1, NDRG4, DLX2, SHC3, DLX5, DLX6, EFNB3, FLRT3, NDP, NPTX1, NKX2-5, NKX6-1, WNT2, PLXNA4, STRA6, NGFR, TFAP2B, LRRN3, BDNF, LIF, INHBA, BMP7, BMP2, LRFN5, GDNF, TRPV4, OGN, LRRN1, CNTN4, SDK2, THRB, CXCL8, TENM3, SYCP2, ELN, MEOX2, AREG, TMEM100, UCHL1, NEFL, COL10A1, FOXD1, GFRA1, WNT16, GFRA2, ALDH1A3, GAP43, MMP13, MMP16, HAND2, ALDH1A2, ITGA8, ANGPTL4, MDFI, SHANK1, SEMA3A, NRXN1, COL11A1, SEMA3B, KCNA2, NRXN3, PTN, HOXD11, PTGS2, SRPX2, RELN, ERBB4, EPYC, GPC3, SLITRK6, SLITRK5, NCAM1, GPC4, NTRK2, ANGPT1, EYA4, VAX1, NFASC, KRT18, WT1, MFAP2, FMOD, NECTIN3]
	axon development - GO:0061564	46	4.61E-21	2.00E-04	[SHC3, DLX5, SEMA3A, NRXN1, SEMA3B, NRXN3, TNC, UCHL1, EFNB3, RELN, FLRT3, DPYSL5, EPYC, KIF5C, KIF5A, SLITRK6, NEFL, SLITRK5, NEFM, NCAM1, NRCAM, NPTX1, NCAM2, PLXNA4, NGFR, FOXD1, LRRN3, BDNF, OMD, GFRA1, ANK3, VAX1, BMP7, GFRA2, ISL2, NFASC, LRFN5, GAP43, GDNF, DOK5, OGN, ADGRB1, LRRN1, CNTN4, FMOD, EPHA3]
	synapse organization - GO:0050808	31	5.00E-20	2.00E-04	[NRXN1, NRXN3, TNC, NRXN2, RELN, CDH2, FLRT3, LRRTM2, SLITRK6, NRCAM, NLGN4Y, NTRK2, NLGN4X, BDNF, ANK3, DKK1, PTPRD, NFASC, GDNF, GLRB, PCDHB2, ADGRB3, IL1RAPL1, PCDHB16, PCDHB6, COL4A5, PCDHB5, PCDHB3, SDK2, ADGRL3, SHANK1]
	regulation of nervous system development - GO:0051960	60	2.03E-17	2.00E-04	[SNAP25, TENM3, STMN2, SOX2, CDH2, LRRTM2, NEFL, NRCAM, SOX11, DKK1, POU3F2, ISL2, SFRP2, ADGRB3, IL1RAPL1, ADGRB1, EPHA3, NGEF, SHANK1, DLX1, LTK, NDRG4, DLX2, SEMA3A, NRXN1, LPAR1, NRXN3, PTN, THBS2, SRPX2, RELN, FLRT3, ERBB4, SLITRK6, SPOCK1, SLITRK5, CSMD3, NKX2-5, NKX6-1, WNT2, WASF3, PLXNA4, RAP1GAP2, NGFR, NTRK2, LRRN3, NEGR1, SYT1, BDNF, LIF, VAX1, BMP7, PTPRD, BMP2, TRPV4, CNTN1, ZNF536, LRRN1, CNTN4, ADGRL3]
	cell morphogenesis involved in neuron differentiation - GO:0048667	42	3.78E-17	2.00E-04	[SHC3, DLX5, SEMA3A, NRXN1, SEMA3B, NRXN3, UCHL1, EFNB3, RELN, FLRT3, DPYSL5, EPYC, KIF5C, KIF5A, SLITRK6, SLITRK5, NCAM1, NRCAM, NPTX1, PLXNA4, NGFR, FOXD1, LRRN3, BDNF, OMD, GFRA1, ANK3, VAX1, BMP7, GFRA2, ISL2, NFASC, LRFN5, GAP43, GDNF, DOK5, OGN, ADGRB1, LRRN1, CNTN4, FMOD, SHANK1]
	neuron projection morphogenesis - GO:0048812	44	4.84E-17	2.00E-04	[SHC3, DLX5, SEMA3A, NRXN1, SEMA3B, NRXN3, UCHL1, EFNB3, RELN, FLRT3, DPYSL5, EPYC, KIF5C, KIF5A, SLITRK6, NEFL, SLITRK5, NCAM1, NRCAM, NPTX1, PLXNA4, NGFR, POSTN, FOXD1, LRRN3, BDNF, OMD, GFRA1, ANK3, VAX1, BMP7, GFRA2, ISL2, NFASC, LRFN5, GAP43, GDNF, DOK5, OGN, ADGRB1, LRRN1, CNTN4, FMOD, SHANK1]

Table S2.2: Gene ontology enrichments for the five gene clusters. Top significant enrichments for GO terms are listed for each gene cluster. Enrichments were calculated using TANGO.

Gene Cluster	KEGG Pathway	#genes	Raw p-value	Corrected p-value	Enrichment factor	Gene list
1	Melanogenesis	9	7.25E-05	0.00414	5.07	[GNAO1, DCT, KIT, TYRP1, FZD9, ADCY2, ADCY1, TYR, WNT4]
	Calcium signaling pathway	10	0.00119	0.0442	3.22	[RYR1, CHRM1, GNAL, CACNA1B, ATP2B3, CACNA1D, ADCY2, ADCY1, MYLK3, CACNA1H]
	Maturity onset diabetes of the young	4	8.83E-04	0.0357	9.11	[PKLR, ONECUT1, MNX1, NKX2-2]
3	Natural killer cell mediated cytotoxicity	17	3.07E-11	3.52E-09	8.19	[KLRC2, PRKCB, SH2D1A, PRF1, GZMB, FASLG, ITGAL, HLA-G, KIR2DL4, PIK3CG, ZAP70, KLRK1, IFNG, LCK, KLRD1, CD48, CD247]
	Graft-versus-host disease	12	6.87E-13	9.53E-11	19.2	[HLA-DRB5, IFNG, IL1B, PRF1, GZMB, FASLG, KLRD1, HLA-DOA, HLA-DQA2, HLA-G, HLA-DOB, HLA-DQA1]
	B cell receptor signaling pathway	8	1.87E-05	0.00129	6.99	[CD79B, CD79A, CR2, PRKCB, CD19, CARD11, CD22, PIK3CG]
	Allograft rejection	10	1.44E-10	1.39E-08	17.7	[HLA-DRB5, IFNG, PRF1, GZMB, FASLG, HLA-DOA, HLA-DQA2, HLA-G, HLA-DOB, HLA-DQA1]
	Primary immunodeficiency	14	4.83E-17	1.56E-14	26.2	[CIITA, TNFRSF13B, IL2RG, CD3E, CD3D, CD79A, ZAP70, PTPRC, CD8B, LCK, CD8A, CD19, IL7R, ICOS]
	Leukocyte transendothelial migration	8	3.87E-04	0.0171	4.56	[ITK, NCF1, ITGA4, PRKCB, RHOH, ITGAL, MMP9, PIK3CG]
	Hematopoietic cell lineage	21	1.12E-19	5.41E-17	15.8	[CR2, HLA-DRB5, CR1, ITGA4, CD3G, GP1BA, CD1C, CD3E, CD3D, CD2, FCER2, CD8B, CD5, CD8A, IL1B, CD19, CD7, CD38, IL7R, MS4A1, CD22]
	Autoimmune thyroid disease	10	5.32E-09	4.70E-07	12.6	[HLA-DRB5, PRF1, GZMB, CTLA4, FASLG, HLA-DOA, HLA-DQA2, HLA-G, HLA-DOB, HLA-DQA1]
	Type I diabetes mellitus	11	3.27E-11	3.52E-09	16.8	[HLA-DRB5, IFNG, IL1B, PRF1, GZMB, FASLG, HLA-DOA, HLA-DQA2, HLA-G, HLA-DOB, HLA-DQA1]
	Chemokine signaling pathway	22	1.59E-13	3.08E-11	7.63	[CCL14, ITK, CXCL9, CCL22, CCL21, NCF1, PRKCB, CXCR5, CXCR6, CXCL13, PIK3CG, CXCL10, CXCL11, CCL8, CCL5, CXCR3, XCL2, CCR7, CCL19, CCL18, CCR5, CCR2]
	Cytokine-cytokine receptor interaction	35	6.30E-23	6.11E-20	8.79	[CCL14, CXCL9, TNFRSF13B, CXCR5, FASLG, TNFRSF11B, CXCR6, IL2RG, CXCL13, TNFSF13B, CCL8, CCL5, CXCR3, IL21R, TNFRSF17, TNFSF11, CCR7, CCL19, CCL18, CCR5, IL12RB1, CCR2, CCL22, CCL21, CD70, TNFRSF9, IFNLR1, CXCL10, CXCL11, IFNG, IL1B, XCL2, CD27, LTB, IL7R]
	Asthma	5	8.35E-05	0.0045	10.9	[HLA-DRB5, HLA-DOA, HLA-DQA2, HLA-DOB, HLA-DQA1]
	Toll-like receptor signaling pathway	8	1.59E-04	0.00771	5.19	[CXCL10, CXCL11, CXCL9, CCL5, IL1B, TLR8, LBP, PIK3CG]
	Systemic lupus erythematosus	8	0.00117	0.0442	3.85	[C3, HLA-DRB5, IFNG, C7, HLA-DOA, HLA-DQA2, HLA-DOB, HLA-DQA1]
	Cell adhesion molecules (CAMs)	22	7.20E-17	1.74E-14	10.9	[CADM3, HLA-DRB5, ITGA4, ITGAL, SELE, HLA-G, CD2, SELP, SPN, PTPRC, CD6, CD8B, SELL, CD8A, CTLA4, PDCD1, HLA-DOA, ICOS, HLA-DQA2, HLA-DOB, HLA-DQA1, CD22]
	Antigen processing and presentation	12	9.42E-09	7.62E-07	8.94	[CIITA, HLA-DRB5, CD8B, KLRC2, CD8A, KLRD1, HLA-DOA, HLA-DQA2, HLA-G, HLA-DOB, HLA-DQA1, KIR2DL4]
	T cell receptor signaling pathway	17	6.73E-13	9.53E-11	10.3	[ITK, CD3G, CD3E, CD3D, PIK3CG, ZAP70, PTPRC, CD8B, IFNG, CD8A, LCK, CTLA4, PRKCQ, CD247, PDCD1, ICOS, CARD11]
5	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	9	4.28E-05	0.00259	5.38	[DES, CDH2, ACTN2, CACNA2D1, ITGA10, PKP2, ITGA8, ITGB8, CACNG4]
	Neuroactive ligand-receptor interaction	16	2.54E-04	0.0117	2.77	[GABRA2, GRIA1, CHRM3, GRIA2, THRB, LPAR1, NPY1R, ADRB1, GRIK2, PRLR, MCHR1, GHR, GABRR1, GLRB, F2RL2, NTSR1]
	ECM-receptor interaction	9	1.17E-04	0.00596	4.74	[COMP, RELN, COL11A1, ITGA10, TNC, ITGA8, ITGB8, THBS2, THBS4]
	Dilated cardiomyopathy	10	3.59E-05	0.00232	4.92	[PLN, DES, CACNA2D1, ITGA10, ITGA8, ITGB8, ADRB1, ADCY8, CACNG4, ADCY5]
	Cell adhesion molecules (CAMs)	15	3.14E-07	2.34E-05	5.03	[NLGN4X, NEGR1, NRXN1, NRXN3, NRXN2, CLDN11, NFASC, CDH2, CNTN1, ITGA8, ITGB8, NRCAM, NCAM1, NCAM2, NECTIN3]
	TG F-beta signaling pathway	8	7.28E-04	0.0307	4.12	[COMP, BMP2, FST, BMP8B, INHBA, THBS2, BMP7, THBS4]

Table S2.3: Enrichment analysis for KEGG pathways performed using PROMO on the five gene clusters. Top significant KEGG pathways are displayed for each gene cluster. Enrichments were calculated using TANGO.

Id	Gene Symbol		p-value	Mean diff.	Id	Gene Symbol	p-value	Mean diff.
1	**OCA2**		8.28E-11	5.70	**51**	DUSP9	1.33E-09	1.70
2	**TYRP1**		6.71E-08	5.66	**52**	WNK2	4.53E-06	1.68
3	ITGB1BP3		1.10E-12	3.29	**53**	LAMA1	6.48E-06	1.68
4	SLC7A4		1.18E-06	3.02	**54**	SEPT3	4.46E-11	1.68
5	IP6K3		8.43E-06	2.94	**55**	CDK2	1.92E-14	1.65
6	C14orf34		5.34E-08	2.78	**56**	KCNAB2	3.11E-13	1.62
7	ABCB5		5.83E-12	2.68	**57**	MGC16025	6.70E-13	1.62
8	GABRA5		5.62E-10	2.67	**58**	PRR5-ARHGAP8	4.87E-05	1.61
9	KRTAP19-1		1.06E-11	2.65	**59**	SNCB	7.12E-06	1.58
10	FAM69C		5.10E-16	2.65	**60**	GNAO1	1.15E-07	1.58
11	**KIT**		4.68E-07	2.42	**61**	MAST1	4.27E-07	1.57
12	VGF		2.36E-05	2.40	**62**	HRK	2.01E-07	1.56
13	SLC6A17		3.69E-12	2.30	**63**	LOC148145	2.18E-05	1.55
14	MGAT5B		5.03E-14	2.30	**64**	PLAC2	3.51E-05	1.55
15	ACCSL		7.56E-11	2.23	**65**	C6orf176	6.14E-09	1.53
16	GNAL		2.53E-09	2.22	**66**	SFTPC	1.36E-07	1.52
17	ABCC2		4.07E-09	2.20	**67**	RIMS4	2.97E-06	1.49
18	ONECUT1		2.53E-09	2.17	**68**	ONECUT2	2.75E-05	1.48
19	NECAB2		8.68E-09	2.16	**69**	FZD9	6.86E-09	1.48
20	PRODH		1.73E-07	2.08	**70**	ARHGAP8	2.72E-05	1.48
21	PNMA6A		9.15E-17	2.08	**71**	LOC100127888	2.55E-09	1.46
22	CNTFR		2.96E-05	2.07	**72**	**TRIM63**	5.22E-16	1.44
23	POU4F1		2.19E-07	2.03	**73**	EPHA5	2.27E-07	1.44
24	TRPM1		6.21E-12	2.03	**74**	DGCR5	8.67E-06	1.44
25	SLC5A10		9.89E-07	2.00	**75**	TMEM151A	1.05E-05	1.43
26	**SILV**		1.95E-15	1.96	**76**	C1QL4	2.51E-08	1.42
27	FOXF2		2.92E-09	1.94	**77**	CPNE7	2.05E-06	1.41
28	CDK15		2.41E-07	1.90	**78**	GBX2	8.10E-05	1.40
29	SLC16A6		4.76E-05	1.90	**79**	FSTL4	1.84E-06	1.40
30	NKX2-8		5.75E-05	1.89	**80**	NRTN	1.99E-05	1.40
31	L1CAM		2.05E-06	1.88	**81**	TFAP2A	7.61E-22	1.39
32	CDH3		1.18E-10	1.87	**82**	DUSP8	2.01E-08	1.38
33	BRSK2		3.30E-09	1.86	**83**	C6orf218	4.38E-12	1.35
34	PITX2		6.15E-05	1.85	**84**	ZNF703	4.75E-14	1.32
35	DPYSL4		1.59E-10	1.84	**85**	HES6	6.04E-08	1.32
36	KIF1A		5.41E-06	1.84	**86**	LGI3	1.30E-05	1.31
37	PRRT4		8.44E-07	1.83	**87**	NCRNA00052	3.20E-07	1.30
38	RTN4R		1.19E-14	1.81	**88**	C15orf59	8.46E-05	1.28
39	ADAM11		2.45E-11	1.81	**89**	LOC390595	9.69E-06	1.28
40	CA14		4.30E-14	1.80	**90**	TPCN2	1.79E-10	1.28
41	NR4A3		8.93E-10	1.80	**91**	ADAMTSL5	1.29E-07	1.26
42	MCF2L		6.71E-09	1.80	**92**	GPRC5A	7.47E-05	1.26
43	TSPAN10		9.07E-12	1.78	**93**	**DCT**	6.70E-05	1.26
44	TPPP		1.21E-11	1.77	**94**	LRRC39	7.68E-07	1.25
45	KCNH1		9.86E-07	1.77	**95**	ITPKB	6.75E-11	1.25
46	GMPR		1.59E-12	1.76	**96**	CELF5	2.07E-05	1.25
47	KREMEN2		2.72E-08	1.74	**97**	MANEAL	4.15E-07	1.25
48	DLL3		1.70E-09	1.74	**98**	TTYH2	1.84E-13	1.25
49	SULT4A1		4.74E-05	1.72	**99**	ANKRD9	7.75E-13	1.24
50	SEMA6A		6.74E-20	1.72	**100**	HES4	1.58E-07	1.24

Table S2.4: List of the 100 most differentially expressed genes distinguishing cluster 4 samples from and all other clusters. Genes are sorted by descending fold-change. P-value was calculated using the rank-sum test applied on [Melanogenesis-high] samples (n=118) vs. [Immune,Keratin,Melanogenesis-low] samples (n=350). p-value cutoff: p<0.0001.

Patient number	Age at time of primary tumor diagnosis	Survival from T diagnosis (months)	Survival from regional lymph node (N) diagnosis (months)	Survival from distant metastasis (M) diagnosis (months)	Current status
1	81	over 60 months			Alive
2	67	over 60 months			Alive
3	66	over 60 months			Alive
4	88	24	4.15		Dead
5	74	19.75	6.77		Dead
6	67	19.48	17.38	0.85	Dead

Table S2.5: Clinical details for the six patients selected for Immunohistochemical staining. Patients 1-3 survived for more than 60 months after diagnosis and were therefore labeled as "Good Prognosis", whereas patients 4-6 survived less than 20 months and were therefore labeled as "Poor Prognosis".

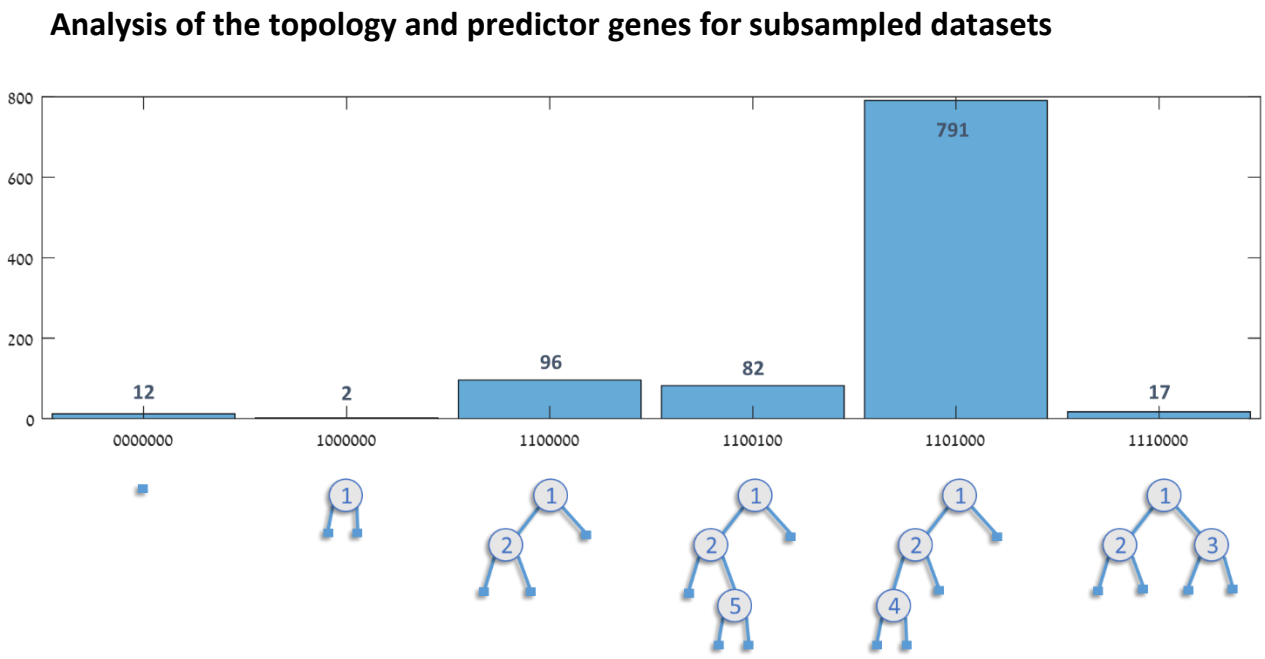


Figure S2.6: Distribution of the topology of 1000 decision trees. For analyzing the topology and biological function of the tree predictors, we trained 1000 3-gene decision trees by resampling the dataset samples (resample factor = 0.8). The most frequent topology was '1101000', identical to the topology of the final decision tree presented in Figure 4, which was trained on the entire dataset (Note that 1101000 and 1100100 are considered different since the left child of every node always corresponds to the "less than" subgroup.)

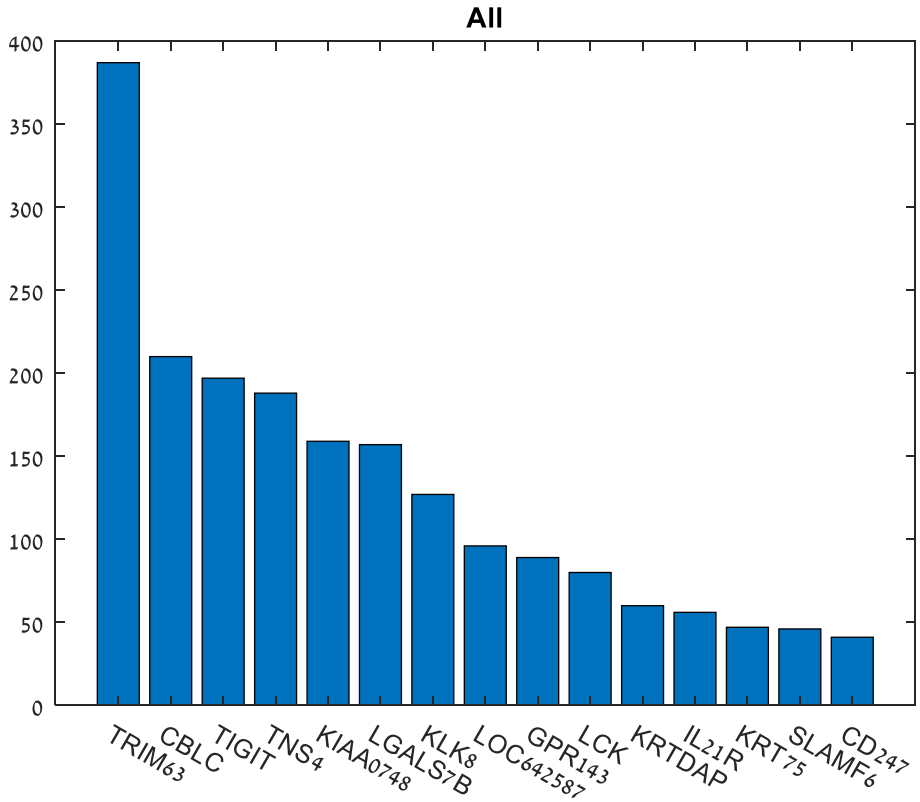


Figure S2.7: Most frequent genes in the 1000 decision trees (in all tree positions).

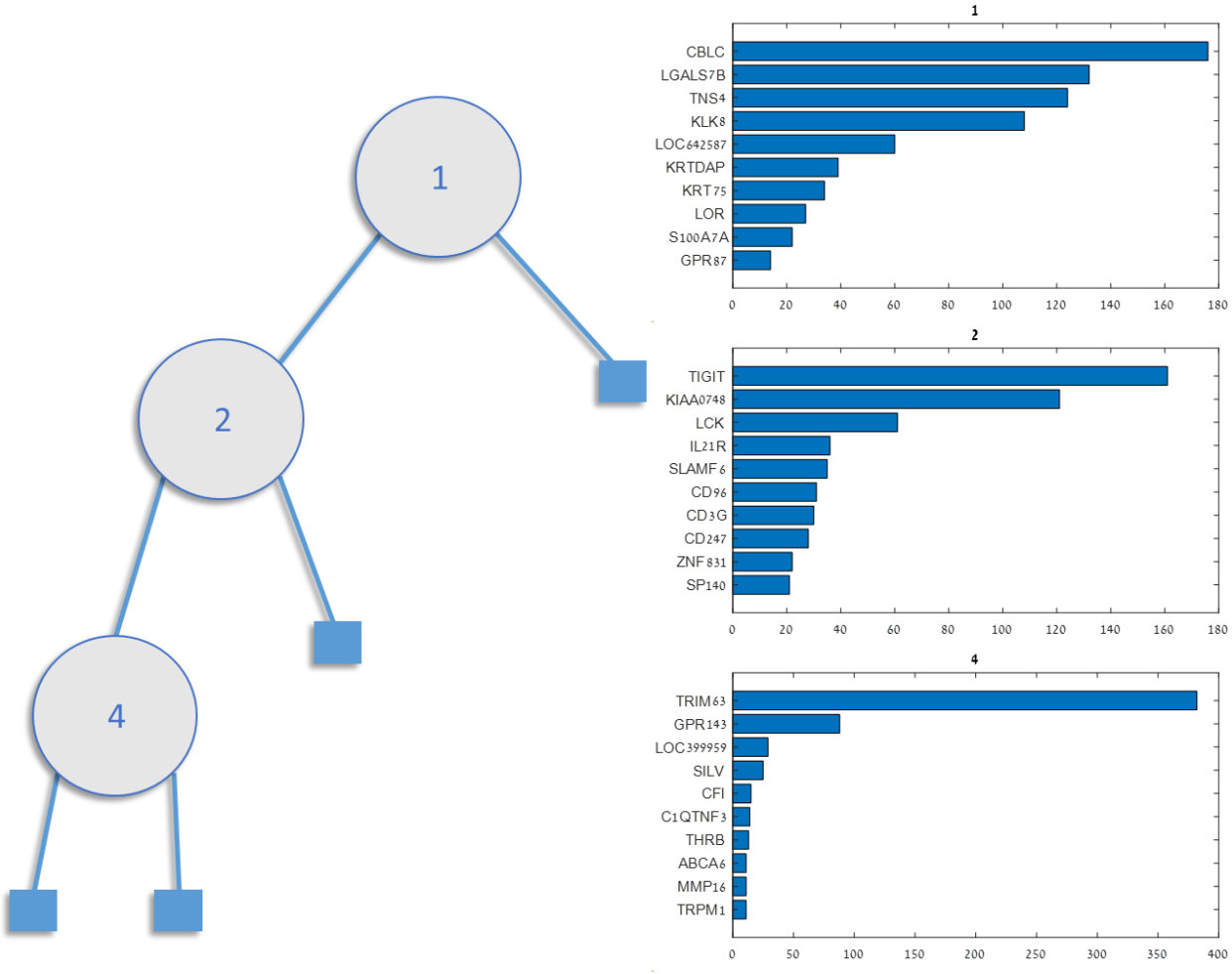


Figure S2.8: Most frequent predictor genes for each position in the tree and their biological function. For each position in the 1101000 topology, we generated a list of the 10 most frequent predictor genes as appearing on the 791 random tree variants. The analysis showed that the most frequent genes in each position are characterized by a specific biological function. Position 1, which forms the tree's root, was typically assigned with keratin and other skin related biomarkers, such as LGALS7B, TNS4 and KLK8. Position 2 was typically assigned with well-known immune markers such as TIGIT, KIAA0748, LCK, and IL21R. Position 4 was preferentially assigned with TRIM63, but also with typical melanogenesis genes such as GPR143, SILV, and TRPM1. Interestingly, the lists also included genes that are less familiar in their context here, such as LOC399959.

The results demonstrate the hierarchy of the biological functions by which melanoma samples can be partitioned into distinct subgroups, and also show that the final tree presentd in Figure 4 is a representative of a stable tree topology and is using predictor genes that are biomarkers of the above three biological functions.

7.3. Supplement 3: PROMO

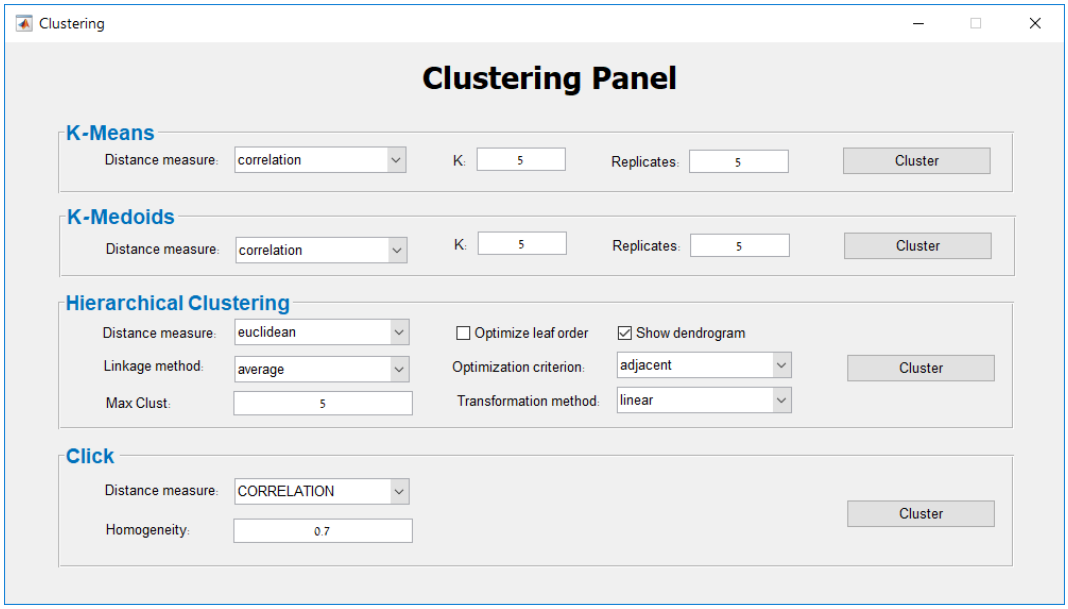


Figure S3.1: Clustering Panel The clustering panel allows the selection of a clustering algorithm and its relevant parameters. Clustering can be applied both on samples and on genes. The resulting clusters are added as a new sample label and can be explored on PROMO's main screen with respect to any other clinical label (See Figure 3).

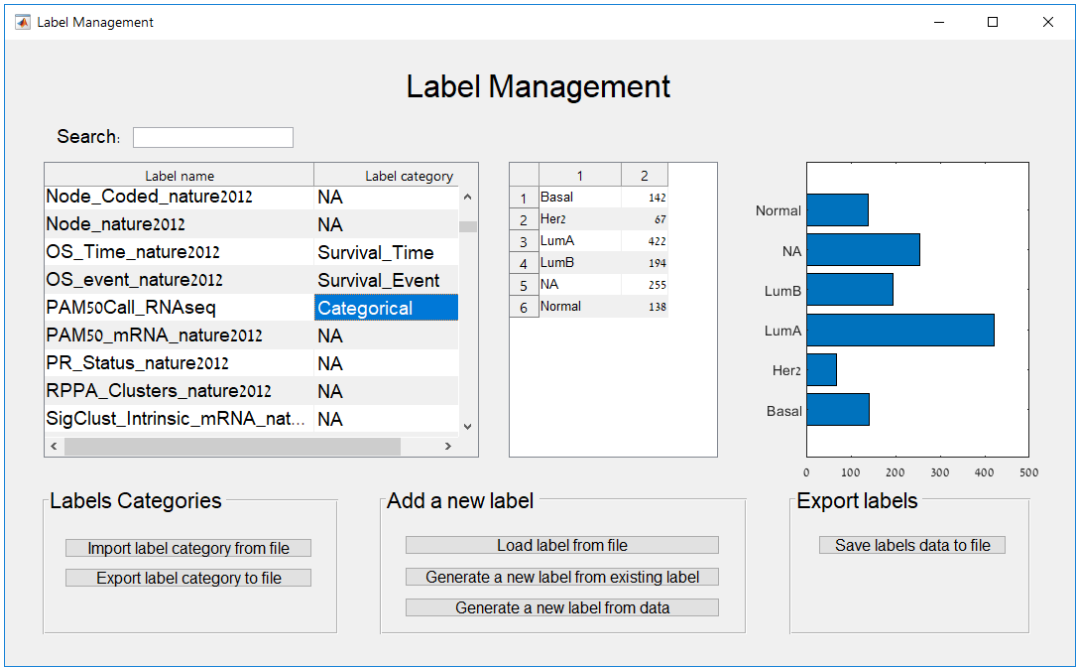
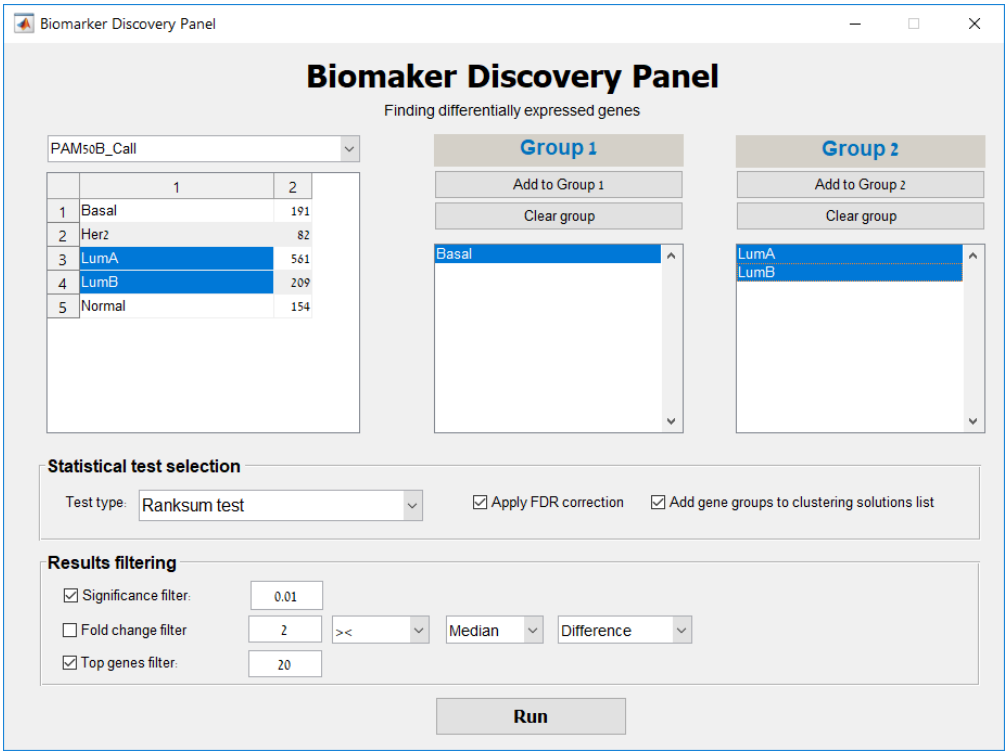


Figure S3.2: Label Management Panel. This panel allows the management of sample labels, including removing, renaming and viewing the distribution of values of a label. Labels can be assigned to category types, and those types determine the statistical test that can be used for calculating their enrichment on sample clusters. Both labels and their categories can be loaded and saved to files. New labels can be generated from existing labels (by uniting label values for instance), or from genomic data (e.g., translating the expression values of a selected gene to LOW/HIGH labels). Lastly, the distribution of values for the selected label is displayed as a histogram on the right.

A



B

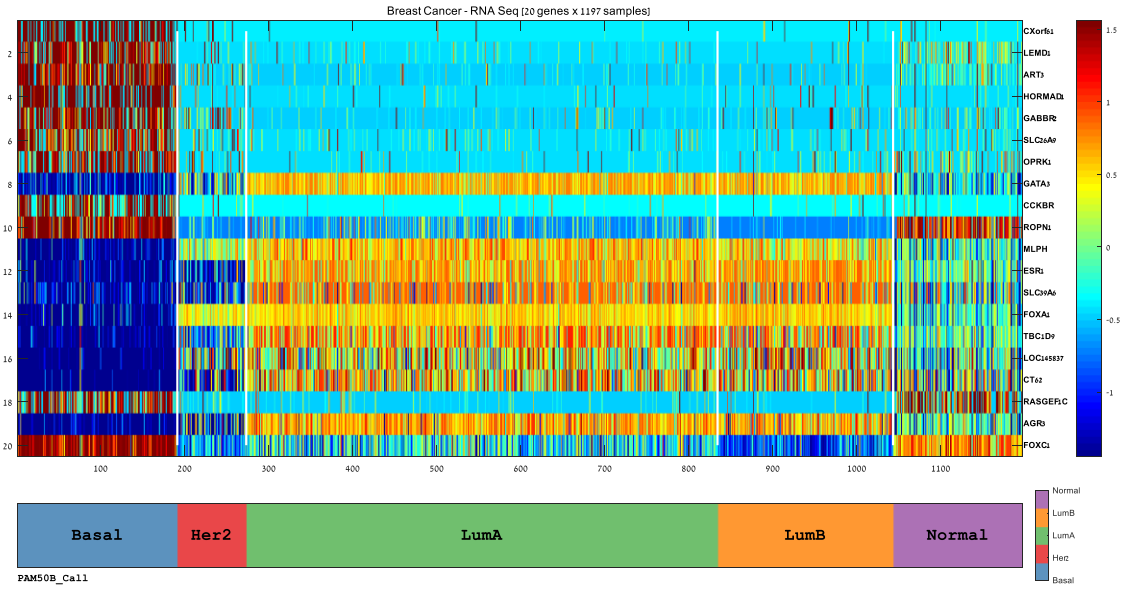


Figure S3.3: Biomarker Discovery (A) This panel is used for identifying genes that are differentially expressed between sample groups defined by any sample label. Statistical tests vary by label type, and include t-test, rank-sum test, ANOVA and Kruskal-Wallis. After optional filtering, the resulting list of genes is saved to a file sorted by p-value. Here two groups were defined, according to the PAM50 label. One group corresponds to the basal and the other to the LumA and Lum B categories. See Table S3.1 for the resulting set of differentially expressed genes. **(B)** The feature patterns of the identified genes are presented on PROMO's main screen together with any selected sample labels. Here we see the expression levels of the 20 genes that were identified by the test in A, after row normalization).

id	Gene Symbol	p-value (Test: Ranksum test on PAM50B_Call [Basal](n=191) vs. [LumA,LumB](n=770))	Fold Change
1	CXorf61	2.33E-123	4.5604
2	LEMD1	1.72E-122	3.1005
3	ART3	3.17E-118	5.294
4	HORMAD1	1.75E-113	5.8894
5	GABBR2	7.04E-111	4.2383
6	SLC26A9	4.21E-101	2.4335
7	OPRK1	2.22E-99	2.5534
8	GATA3	3.27E-99	-4.02715
9	CCKBR	5.86E-99	2.1373
10	ROPN1	8.44E-99	6.1879
11	MLPH	5.43E-98	-5.1038
12	ESR1	7.07E-98	-7.07625
13	SLC39A6	9.68E-98	-2.7197
14	FOXA1	3.21E-97	-6.7128
15	TBC1D9	6.71E-97	-4.06355
16	LOC145837	2.99E-96	-4.18775
17	CT62	3.12E-96	-3.65845
18	RASGEF1C	5.50E-96	2.1274
19	AGR3	1.12E-95	-9.2111
20	FOXC1	1.63E-95	4.26835

Table S3.1: List of differentially expressed genes. The 20 genes with the most significant differential expression between the groups defined in Figure S3.3A are shown. Genes are sorted by their rank-sum test p-values. Genes with positive fold change are over-expressed on the basal samples compared with the luminal samples. Here, for instance, we see that the Estrogen Receptor gene (ESR1) is ranked 12th and exhibits a significant under-expression on the basal tumors samples (the Triple-Negative subtype) compared to the luminal tumor samples.

נבחרו כאיזון ראוי בין דיוק לבין פשטות המסווג). אנו מקווים שעץ ההחלטה שיצרנו, כמו גם הגנים שזיהינו כבעלי ערך כסמנים, יתרמו בעתיד לשיפור הדיוק בסיווג של חולים לתתי קבוצות פרוגנוסטיות גם בקליניקה.

PROMO – כלי לניתוח מידע רחב היקף בתחום הסרטן

פרומו הינו כלי אינטראקטיבי בעל ממשק גרפי שפיתחנו על מנת לאפשר ניתוח מתקדם של אוספי מידע גדולים של סרטן בקלות ובמהירות ובכך להנגיש את אוספי מידע אלו לקהל חוקרים רחב יותר. פרומו מממש את מרבית שלבי הניתוח שביצענו על אוסף דגימות סרטן השד ועל אוסף דגימות המלנומה כפי שתוארו בפרקים הקודמים. הכלי מותאם לניתוח של אוספי מידע גדולים (הכוללים אלפי דגימות), ומתמחה בניתוח של מידע רחב היקף ביחד עם נתונים קליניים הזמינים עבור כל חולה, ומאפשר מגוון עשיר של סוגי ניתוחים.

סוג הניתוח הראשי בו תומך פרומו הוא זיהוי תתי סוגים של סרטן. השלבים העיקריים בניתוח כזה הם: (1) ייבוא של מידע רחב היקף ביחד עם מידע קליני ממגוון מקורות ופורמטי קבצים. (2) סינון ונירמול של אוסף הנתונים כדי לסלק מידע מיותר שאינו נחוץ להמשך הניתוח. (3) מגוון אפשרויות לחקר התפלגות הנתונים וויזואליזציה שלהם כדי לזהות את התכונות הבסיסיות של נתוני האוסף. (4) הפעלת מגוון אלגוריתמי קיבוץ (Clustering) על שורות או על עמודות מטריצת הביטוי מאפשרים חלוקה של תכונות האוסף (לרוב גנים) או של דגימות האוסף לתתי קבוצות. (5) מבחני העשרה על קבוצות הדגימות תוך שימוש במידע הקליני הזמין לכל דגימה מאפשר לאפיין קלינית את קבוצות הדגימות שהתקבלו. מבחני העשרה על תכונות האוסף (גנים) מאפשרים לאפיין את הפונקציות הביולוגיות של הגנים בקבוצות השונות. (6) איתור של סמנים על בסיס מבחנים סטטיסטיים לזיהוי גנים מפרידים או על בסיס ניתוח הישרדות. (7) לבסוף, פרומו מאפשר יצירה אוטומטית של עצי החלטה פשוטים עבור תווית דגימות נבחרת.

סיכום

בעבודה זו עשינו שימוש במגוון שיטות אלגוריתמיות וסטטיסטיות לצורך ניתוח נתונים רחבי היקף בתחום הסרטן. הניתוח נעשה על ידי שילוב של מידע ביולוגי רחב היקף עם מידע קליני מסוגים שונים. בניתוח של דגימות סרטן השד זיהינו דפוס מבוסס ביטוי גנים ודפוס מבוסס מתילציה אשר מאפשרים להעריך טוב יותר את סיכוי ההישרדות של דגימות מסוג Luminal-A. בניתוח של דגימות המלנומה, זיהינו קבוצה של דגימות בעלת שרידות נמוכה המאופיינת על ידי ביטוי יתר של גנים הקשורים למלנוגנזה, ויצרנו עץ החלטה פשוט שמאפשר לסווג דגימה חדשה לאחת מארבע תתי הסוגים שזיהינו. לבסוף, פיתחנו כלי בשם פרומו שמאפשר ניתוח מהיר של אוספי נתונים ביולוגיים גדולים בתחום הסרטן בשילוב עם המידע הקליני הזמין. אנו מקווים שהאבחנות שהעלינו והכלי שפיתחנו יסייעו בשיפור הסיווג בסרטן ויתרמו לקידום חזון הרפואה האישית.

קבוצה של דגימות Luminal-A אותה כינינו LumA-M1, שהתאפיינה בפרופיל היפרמתילציה ובסיכויי הישרדות נמוכים יותר באופן מובהק בהשוואה לשתי הקבוצות האחרות. ניתוח העשרה על אתרי ה-CpG שמתילציה גבוהה בהם מאפיינת את קבוצת הדגימות LumA-M1, הראה שמדובר בגנים הקשורים להתפתחות (Developmental genes).

ניתוח הישרדות מבוסס על Cox regression, העלה ששתי מערכות הסיווג שמצאנו הינן בעלות ערך פרוגנוסטי. שיוך של דגימה לתת הקבוצה LumA-R1 על בסיס נתוני ביטוי גנים מעלה את סיכויי הישנות הגידול, ואילו שיוך של דגימה לתת הקבוצה LumA-M1 על בסיס נתוני מתילצית דנ"א מוריד את סיכויי ההישרדות של המטופל באופן מובהק.

לסיכום, הניתוח שביצענו על נתוני ביטוי גנים ונתוני מתילציית דנ"א בדגימות של סרטן השד מסוג Luminal-A זיהה שני דפוסים בעלי ערך פרוגנוסטי שעשויים לסייע בעתיד בחלוקה טובה יותר דגימות אלו לתת סוגים מדויקים יותר.

שיפור הסיווג של סרטן העור מלנומה

המטרה של פרויקט זה הייתה לשפר את הסיווג מבוסס ביטוי הגנים של גידולי מלנומה כפי שהוצע על ידי TCGA בשנת 2015. סיווג זה כלל חלוקה של הדגימות לשלוש קבוצות טרנסקריפטומיות (Immune-high, Keratin, MITF-low) שהראו הבדל מובהק בסיכויי ההישרדות שלהן [48] .

חלוקה בלתי מונחית של 469 דגימות המלנומה שהורדו ממאגר ה-TCGA לארבע, זיהתה קבוצות ברורות בעלות הבדל מובהק בסיכויי ההישרדות של החולים. החלוקה בוצעה על בסיס פרופילי ביטוי של 2000 הגנים בעלי השונות הגבוהה ביותר באמצעות אלגוריתם k-means. אפיון של ארבע תתי הקבוצות שהתקבלו בחלוקה שביצענו על בסיס מידע קליני ועל ידי השוואה לשלוש תתי הקבוצות שהוגדרו על ידי TCGA העלתה שתת הקבוצות 1 ו-3 בחלוקה שלנו מקבילות לתת הקבוצות Immune-high ו-MITF-low אצל TCGA (בהתאמה). עם זאת, תת הקבוצה Keratin של TCGA פוצלה בחלוקה שלנו לשתי תתי קבוצות ברורות: תת קבוצה 2 הכילה בעיקר גידולים ראשוניים, הייתה בעלת סיכויי ההישרדות הנמוכים ביותר, והתאפיינה בביטוי יתר של גנים מסוג קרטין. תת קבוצה 4 הכילה בעיקר גרורות, הייתה בעלת סיכויי הישרדות נמוכים למדי, והתאפיינה בביטוי יתר של גנים מסוג מלנוגנזה (Melanogenesis).

התמקדות בגנים הקשורים למלנוגנזה ומבוטאים ביתר בתת קבוצה 4 בחלוקה שקיבלנו (וזכתה לכינוי Melanogenesis-high), העלתה שגנים אלו הינם פרוגנוסטיים (עשויים לשמש כסמנים לחיזוי הישרדות) ואף עשויים לרמז על קשר מנגנוני בין אברון המלנוזום שאליו גנים אלו קשורים לפי הספרות לבין הישרדות. מחקר נוסף נדרש כדי לאמת השערה זו.

לסיום, אימנו מסווג מסוג עץ-החלטה לחיזוי תת סוג של מלנומה על בסיס ביטוי גנים. המסווג מאפשר סיווג של דגימה חדשה לאחת מארבע תתי הקבוצות שזיהינו, על בסיס מספר קטן של גנים (3 גנים

תוצאות

שיפור הסיווג של סרטן השד

סכמת הסיווג המולקולרי המקובלת בסרטן השד מבוססת על ביטוי גנים, קרויה PAM50 ומכילה את תתי הסוגים Basal-like, Her2, Luminal-A, Luminal-B. בפרויקט זה עשינו שימוש בנתוני ביטוי גנים (Gene Expression) ובנתוני מתילציית דנ"א (DNA Methylation) הזמינים עבור מאות דגימות סרטן השד במאגר הנתונים של ה-TCGA לצורך שיפור הסיווג של גידולי סרטן השד לקבוצות בעלות משמעות קלינית.

התחלנו בהורדת פרופילי ביטוי הגנים של 1148 דגימות (1035 דגימות מגידולים של סרטן השד ו-113 דגימת שד נורמלי) ממאגר ה-TCGA וחלוקתם בצורה בלתי מונחית ל-5 קבוצות בעזרת אלגוריתם ה-k-means על בסיס 2000 הגנים בעלי השונות הגבוהה ביותר. השוואת החלוקה אותה קיבלנו לתווית ה-PAM50 הצביעה על דמיון מתון בין שתי צורות החלוקה, כאשר דגימות ה-Luminal ובמיוחד דגימות ה-Luminal-A הראו הטרוגניות משמעותית המרמזת על האפשרות לחלק קבוצה זו לתתי סוגים עדינים יותר.

בהמשך, חילקנו רק את 737 דגימות ה-Luminal (Luminal-A ו-Luminal-B) לשתי קבוצות, תוך שימוש באותה שיטת חלוקה בלתי מונחית בעזרת אלגוריתם k-means על בסיס נתוני הביטוי של 2000 הגנים השונים ביותר על דגימות אלו. חלוקת דגימות ה-Luminal אותה קיבלנו הפרידה טוב יותר את הדגימות מבחינת הישרדות וגם מבחינת הסיכוי להישנות הגידול, בהשוואה לחלוקת התווית PAM50 לקבוצות Luminal-A ו-Luminal-B. תוצאה זו מראה שהשיטה בה אנו מחלקים את הדגימות מזהה אפיון ביולוגי כלשהו של הדגימות שהינו בעל חשיבות קלינית. גם בחלוקה זו, השונות הגדולה ביותר שנצפתה הייתה בקרב דגימות Luminal-A, ועל כן החלטנו להתמקד בתת קבוצה זו.

חלוקה בלתי מונחית נוספת לשתי קבוצות, הפעם רק של 534 דגימות ה-Luminal-A, חילקה את הדגימות לשתי קבוצות ברורות אותן כינינו LumA-R1 ו-LumA-R2 (n=258 ו-n=276 בהתאמה). באופן מעניין, תת-קבוצות אלו פיצלו את קבוצת ה-Luminal-A של סיווג ה-PAM50, והדגימו שוני מובהק סטטיסטית בסיכוי להישנות סרטן. תת הקבוצה LumA-R2 התאפיינה בסיכוי נמוך יותר להישנות הגידול תוך 5 שנים ביחס לקבוצה LumA-R1, והתאפיינה גם בביטוי יתר של מספר גדול של גנים הקשורים למערכת החיסון , המועשרים בגנים הקשורים להפעלת תאי T.

בחלקו השני של הפרויקט ביצענו ניתוח דומה על פרופילי מתילציית דנ"א של דגימות סרטן השד שהורדו אף הם ממאגר ה-TCGA. חלוקה בלתי מונחית של 378 דגימות ה-Luminal-A ל-3 קבוצות בעזרת אלגוריתם k-means על בסיס 2000 אתרי ה-CpG בעלי השונות הגבוהה ביותר, חשפה תת-

סרטן העור מלנומה מתחיל בהתחלקות לא מבוקרת של תאים בעור בשם מלנוציטים [40] האחראים על הפקת הפיגמנט מלנין, והפצתו לתאי העור הסובבים אותם. הטיפול בגידולי מלנומה ראשוניים הוא קל יחסית, וכולל הסרה של הגידול בניתוח [41]. אולם, גידולי מלנומה נוטים להתפשט במהירות יחסית לאיברים מרוחקים בגוף ולייסד שם גרורות. במצב זה, הטיפול הוא מאתגר הרבה יותר [42]. גם במלנומה, החלטות טיפוליות נעשו בהתחלה על סמך פרמטרים קליניים ופתולוגיים שחילקו את הגידולים למספר תתי סוגים [45][44], אך מלבד סיוע באבחון, לא הייתה לתתי סוגים אלו משמעות קלינית ברורה [44]. עם התפתחותן של טכנולוגיות גנומיות רחבות היקף, זוהו מספר גנים, בהם BRAF, NRAS, NF1, אשר קיומה של מוטציה בהם משפיע על התפתחות הגידול ועל הנטייה שלו לשלוח גרורות לאתרים מרוחקים [47][46]. בשנת 2015, פורסם מאמר של קבוצת ה-TCGA שהגדיר עבור מלנומה סיווג לארבעה תתי-סוגים בהתבסס על מוטציות שכיחות (BRAF, NRAS, NF1, ו-WT), ובמקביל הגדיר גם סיווג לשלושה תתי סוגים בהתבסס על פרופיל ביטוי גנים (High-Immune, Keratin ו-MITF-Low). בין שני הסיווגים היתה תאימות נמוכה ורק הסיווג השני, שמבוסס על פרופיל ביטוי גנים, הראה קשר לשרידות.

טכנולוגיות רחבות היקף וחזון הרפואה המותאמת אישית

בשנים האחרונות פותחו מספר טכנולוגיות רחבות היקף המאפשרות מדידה של מספר גדול מאוד של תכונות ביולוגיות בדגימה [50]. טכנולוגיות רחבות היקף אלו מכונות בכללותן 'Omics' היות והן מספקות נתונים רחבי היקף בתחומי ה-Genomics (חקר הדנ"א), Transcriptomics (חקר רנ"א וביטוי גנים), Epigenomics(חקר שינויים אפיגנטיים על הדנ"א), Proteomics (חקר חלבונים) וכן הלאה.

פרויקט ה-TCGA [79] הינו דוגמא למאגר נתונים גנומי של סרטן הכולל 11,000 דגימות שמייצגות 33 סוגי סרטן. הדגימות נדגמו על ידי מספר טכנולוגיות רחבות היקף כולל DNA-Seq למיפוי שינויים ברצף הדנ"א, RNA-Seq למדידת רמות ביטוי mRNA ו-miRNA, Methylation arrays למדידת רמות מתילציית ה-DNA, SNP arrays למדידת שינויים במספר העותקים בגנום, ועוד. בנוסף, כולל מאגר הנתונים גם מידע קליני מפורט לגבי כל אחת מהדגימות הכולל גיל, מין, תת סוג היסטולוגי, משך הישרדות ועוד. מאגר ה-TCGA הפך למשאב יקר ערך בתחום חקר הסרטן היות והוא כולל מידע רחב היקף ורב ממדי על גידולים מסוגים שונים כפי שלא היה זמין בעבר [83][80] [81].

אחד השימושים המבטיחים ביותר לניצולו של המידע הביולוגי רחב ההיקף בתחום הסרטן, הוא בתחום הרפואה המותאמת אישית [73][72]. אם בעבר סווגו חולי סרטן לקבוצות גסות על פי מאפיינים בסיסיים של הגידול, והטיפול שניתן שלהם היה אחיד למדי, תחום הרפואה המותאמת אישית מבטיח לסווג כל חולה לקבוצה מדויקת יותר, ולספק לו טיפול מותאם אישית שמבוסס על הפרופיל הגנטי המדויק של הגידול שלו. לשם פיתוח גישה זו יש לזהות תתי סוגים מדויקים עבור כל סוג סרטן, לפתח טיפולים ספציפיים עבור כל תת סוג, ולזהות סמנים (Biomarkers) שיאפשרו את סיווגו של חולה לתת סוג מדויק [75].

תקציר

רקע כללי

סרטן

המונח "סרטן" מתייחס לקבוצה גדולה של מחלות המתאפיינות בחלוקה בלתי מבוקרת של תאים ולעיתים גם בהתפשטותם לרקמות סובבות ולאיברים מרוחקים. המחלה הינה גורם התמותה השני בעולם [1]. בשנת 2018, אחד מכל שישה מקרי מוות בעולם נגרם מסרטן וסה"כ נרשמו 9.6 מיליון מקרי מוות ו-18.1 מיליון מקרים חדשים של המחלה [2]. סרטן יכול להופיע בכל איבר בגוף, אך סוגי הסרטן השכיחים ביותר הם בריאות, בשד ובמעי.

גידולים סרטניים מתפתחים בתהליך רב-שלבי שבמסגרתו תאים בריאים מומרים לתאים ממאירים בעקבות רצף של שינויים גנטיים ואפיגנטיים. שינויים אלו מאפשרים לתאים המומרים להגביר את קצב החלוקה שלהם ולרכוש תכונות חדשות. בהתחלה, המסה ההולכת וגדלה של תאים ממאירים נשארת במקום ההיווצרות הראשוני של הגידול (סרטן ראשוני, Primary), אולם שינויים נוספים המתרחשים בתאי הגידול המתרבים עלולים לגרום לתאים להתנתק מהגידול הראשוני ולפלוש לרקמות בריאות או להיכנס לדם או ללימפה. תאים אלו עשויים לעבור דרך מחזור הדם או מערכת הלימפה אל אתרים חדשים בגוף בהם יוכלו להקים מושבות חדשות של תאי הגידול, המכונות גרורות (Metastases). רוב מקרי המוות במחלת הסרטן הם תוצאה של הגרורות שמתפשטות לרקמות בריאות ופוגעות בתפקוד האיברים [4].

הטיפולים המסורתיים בסרטן אינם ספציפיים למאפיינים הגנטיים של הגידול והם כוללים ניתוח להסרת הגידול, הקרנות, כימותרפיה או שילובים שלהם. לאחרונה החל שימוש בטיפולים מתקדמים וספציפיים יותר כגון טיפולים חיסוניים (המגבירים את יכולת מערכת החיסון של הגוף להילחם בגידול), טיפולים הורמונאליים (המעכבים את קצב החלוקה של גידולים תלויי הורמון) וטיפולים מכוונים (שתוקפים מולקולה ספציפית הנדרשת לחלוקת התאים באופן בלתי מבוקר) [11] .

בסרטן השד, החלטות טיפוליות התבססו בתחילה על פרמטרים כגון גודל הגידול, מיקומו, מצב בלוטות הלימפה ושלב היסטולוגי. בהמשך, נעשה שימוש גם במידע לגבי הסטטוס של שלושה קולטני הורמונים (אסטרוגן, פרוגסטרון ו-HER2) בקביעת סוג הטיפול [14]. עם התפתחותן של טכנולוגיות רחבות היקף שמאפשרות את מדידת רמת הביטוי של מספר גדול של גנים בדגימות סרטן שד שונות, הוגדרו מספר תתי סוגים מולקולריים לסרטן השד [21][20][19]: Basal-like, HER2-enriched, Luminal-A and Luminal-B. לתתי סוגים מולקולריים אלו היתה תאימות מסוימת לשיטות הסיווג הקודמות, והם הראו מספר מאפיינים קליניים כגון רמת סיכון ותגובה לתרופות, מה שהביא להתבססותם הנרחבת. בשנת 2009 פורסם מסווג בשם PAM50 שמאפשר סיווג של דגימת סרטן השד לאחת מארבע הקבוצות הנ"ל על פי חתימת הביטוי של 50 גנים [23].

תמצית

מחלת הסרטן הינה גורם המוות השני בשכיחותו בעולם. המחלה מאופיינת בחלוקה בלתי מבוקרת של תאים ולעיתים גם בהתפשטות לרקמות סובבות ולאיברים מרוחקים. הטיפול במחלת הסרטן הוא מאתגר בשל היותה מאוד הטרוגנית: גם גידולים שמוצאם באותו האיבר בגוף עשויים להיבדל מאוד זה מזה מבחינה ביולוגית, מבחינת סיכויי ההישרדות ומבחינת האופן בו הם מגיבים לתרופות.

בשנים האחרונות, מספר מאגרי נתונים גנומיים גדולים בתחום הסרטן נעשו זמינים. מאגרים אלו כוללים נתונים שמקורם במגוון טכנולוגיות רחבות-היקף בנוסף למידע קליני מקיף עבור אלפי דגימות סרטן מסוגים שונים. בהתאם לחזון הרפואה המותאמת אישית, שילוב של מידע גנומי רחב היקף עם מידע קליני באמצעות שיטות סטטיסטיות ואלגוריתמיות שונות מאפשרות לזהות באופן חישובי תתי סוגים של סרטן שהינם בעלי חשיבות קלינית ועשויים להשפיע על אבחון, פיתוח תרופות וטיפול בסרטן.

בעבודה זו, פיתחנו גישה לשיפור הסיווג של סרטן לתתי סוגים בהתבסס על נתונים רחבי היקף, ויישמנו אותה על סרטן השד ועל סרטן עור מסוג מלנומה. הניתוח שביצענו על נתוני סרטן השד חשף הטרוגניות משמעותית בתת הסוג Luminal-A וחילק את הדגימות הכלולות בתת סוג זה לשתי תת-קבוצות פרוגנוסטיות על בסיס דפוסי ביטוי גנים ורמות מתילציה. הניתוח שביצענו על נתוני סרטן העור זיהה קבוצה של דגימות מלנומה בעלות הישרדות נמוכה והמאופיינות על ידי ביטוי ביתר של גנים המעורבים ב-Melanogenesis. כמו כן, פיתחנו מסווג מולקולרי פשוט המבוסס על רמות הביטוי של שלושה גנים לחיזוי תת-סוג במלנומה. לבסוף, אנו מתארים את PROMO, שהינו כלי תוכנה אינטראקטיבי שפיתחנו לצורך ניתוח נתונים רחבי-היקף וזיהוי תתי סוגים של סרטן. הכלי מכליל את השיטה לזיהוי תתי סוגים אותה הפעלנו בפרויקטים לשיפור הסיווג של סרטן השד וסרטן העור.

אוניברסיטת תל-אביב TEL AVIV UNIVERSITY

הפקולטה למדעים מדויקים ע"ש ריימונד ובברלי סאקלר

בית הספר למדעי המחשב ע"ש בלבטניק

זיהוי תתי-סוגים של סרטן באמצעות ניתוח מידע גנומי רחב היקף

חיבור לשם קבלת תואר "דוקטור לפילוסופיה"

מאת **דביר נתנאלי**

בהנחייתו של **פרופ' רון שמיר**

הוגש לסנאט של אוניברסיטת תל אביב

דצמבר 2019