A transcriptomic analysis of intratumor and stromal heterogeneity in breast cancer

by

Sadiq Mehdi Ismail Saleh

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Department of Biochemistry, McGill University Montreal QC, Canada

December 2016

ABSTRACT

The management of breast cancer is complicated by inter- and intra-tumour heterogeneity. In particular, triple negative breast cancer (TNBC) is a difficult to treat, molecularly heterogeneous cancer subtype that lacks actionable targets. The heterogeneity of the TNBC microenvironment (stroma) has not been well characterized despite the key role that it may play in tumor progression. Similarly, the impact of intra-tumoral heterogeneity on therapeutic response and patient outcome remains largely unknown.

To address these challenges I investigated the transcriptome of tumor-associated stroma isolated from TNBCs (n=57), as well as comprehensive single-cell gene expression profiling from a treatment-resistant breast patient-derived xenograft (PDX) displaying heterogeneity for the therapeutically targetable HER2 receptor (n=33 cells).

Analysis of the TNBC stroma identified four stromal properties associated with T cells (T), B cells (B), invasive epithelial cells (E), or a desmoplastic reaction (D) respectively. My method, entitled STROMA4, assigns each sample as either low, intermediate, or high for each property independently in stromal or bulk expression profiles. I provide evidence that TNBCType, a previously reported subtyping scheme for TNBC, underestimates the complexity of some tumors, and show how stratification by the STROMA4 method can predict patient benefit from therapy with increased sensitivity. Combining the STROMA4 property assignments generates a novel TNBC subtyping scheme, and analysis of this subtyping scheme revealed that only 15 of 81 possible subtypes had larger than expected populations. This combinatorial

approach revealed that the B, T and E properties are prognostic only when the D property is not high, providing a potential explanation for misprediction by existing classifiers.

Analysis of single-cell RNA-seq (scRNA-seq) data from a PDX heterogeneous for HER2 expression identified distinct cellular subpopulations and revealed a predominantly basal breast cancer subtype. Unsupervised hierarchical clustering distinguished two major cellular subpopulations with differential expression of EGFR, which was validated immunohistochemically in the resected tumour. Further investigation into differences between the EGFR-high and -low cells in the scRNA-seq data indicated that EGFR-high cells were more "stem-like", which was then validated experimentally. The presence of EGFR-high stem cells in this PDX model, as well as in other PDX models, is associated with sensitivity to EGFR inhibition.

Analysis of the TNBC stroma, using a multi-parameter classification model, produces a simple ontology that captures TNBC heterogeneity, and informs how tumor-associated properties and biologies interact to affect prognosis; while analysis of the scRNA-seq data identified two groups of cells with differential expression of EGFR and stem-like characteristics, which is associated with response to EGFR inhibition. Thus, this work adds to our understanding of the contribution of inter- and intra-tumoral heterogeneity to the complexity of the cancer ecosystem, and the effect it has on response to therapy.

RÉSUMÉ

L'hétérogénéité inter et intra-tumorale participe à la complexité de la biologie du cancer du sein. Plus particulièrement, le cancer du sein triple négatif (CSTN) est difficile à traiter en raison de son hétérogénéité au niveau moléculaire et manque de cibles thérapeutiques actionnables. L'hétérogénéité du microenvironnement tumoral des CSTN reste peu caractérisée malgré le rôle clé que ce dernier peut jouer dans la progression tumorale. De façon similaire, l'impact de l'hétérogénéité intra-tumorale sur la réponse thérapeutique et la survie des patients demeure inconnue.

Afin d'élucider ces mécanismes, j'ai analysé le transcriptome du stroma tumoral isolé à partir de CSTN (n=57) ainsi que le profil d'expression cellulaire (cellules individuelles ; n=33) d'une xénogreffe dérivée de tumeur de patient (XDP) résistante à la thérapie et arborant une hétérogénéité d'expression du récepteur HER2. Ce récepteur peut être ciblé de façon thérapeutique en clinique.

L'analyse du stroma des CSTN a permis d'identifier quatre propriétés stromales associées aux cellules T (T), B (B), aux cellules épithéliales invasives épithéliales (E), ou à une réaction desmoplasique (D). La méthode que j'ai développée, intitulée « STROMA4 », assigne un score faible, intermédiaire ou élevé pour chaque propriété à partir des profils d'expression génique du stroma (stroma tumoral) ou de la tumeur globale (tumeur en entier). J'ai pu montrer que le « TNBCType », une méthode de sous-typage des CSTN ayant préalablement été publiée,

sous-estime la complexité de certaines tumeurs. De plus, la stratification des patientes selon la méthode « STROMA4 » permet de prédire la réponse à la thérapie avec une meilleure sensibilité que la méthode « TNBCType ». La combinaison des différentes propriétés identifiées par la méthode « STROMA4 » génère une nouvelle stratification des CSTN. L'analyse de cette nouvelle classification a permis de montrer que seul 15 des 81 sous-types possibles (d'après les différentes combinaisons de scores des propriétés) sont représentés par une population plus grande qu'attendue. Cette approche combinatoire révèle que les propriétés B, T et E sont pronostiques seulement quand la propriété D est de faible score. Ceci pourrait expliquer, en partie du moins, la mauvaise prédiction des classificateurs existants.

L'analyse des données provenant du séquençage ARN de cellules isolées (scRNA-seq) d'une XPD hétérogène pour l'expression de HER2 identifie des sous-populations cellulaires distinctes et révéle, de façon dominante, un sous-type basal de cancer de sein. La classification hiérarchique (« hierarchical clustering » en anglais) non supervisée distingue deux sous-populations cellulaires majeures ayant des taux d'expression du récepteur EGFR différentes au niveau de l'expression génique. Cette différence est validée au niveau protéique par immunohistochimie sur un échantillon de tumeur humaine réséquée au moment de la chirurgie de la patiente. L'étude des différences entre les cellules à forte et faible expression de EGFR par scRNA-seq indique que les cellules ayant une expression élevée de EGFR arborent des propriétés de cellules « souches ». Ce résultat a été validée de façon expérimentale. La présence de cellules ayant un fort taux d'expression de EGFR dans ce modèle ainsi que dans d'autres modèles XPD, est associée à une sensibilité vis à vis de l'inhibition de EGFR.

L’analyse du stroma des CSTN, utilisant un modèle de classification multi-paramètrique, génère une classification simple qui récapitule l’hétérogénéité des CSTN. Cette classification permet de comprendre comment les différentes propriétés biologiques associées à la tumeur interagissent et affectent le pronostic. D’autre part, l’analyse des données du scRNA-seq identifie deux groupes de cellules avec des différences de (i) niveaux d’expression de EGFR, (ii) propriétés de cellules « souches ». Ces cellules sont également associées avec une réponse à l’inhibition de EGFR. Ainsi, ce travail permet une meilleure compréhension de la contribution de l’hétérogénéité aux niveaux inter- et intra-tumoral à la complexité de l’écosystème du cancer et son effet sur la réponse à la thérapie.

ACKNOWLEDGEMENTS

This work was supported by the CIHR Systems Biology Training program, and a McGill Faculty of Medicine Internal Studentship (awarded as Gershman Memorial and Hugh E. Burke Fellowships). This work would also not have been possible were it not for the patients who so generously donated their tissue and consented to be part of this work.

A big thank you to Kanwal Hayat, Karima Hayat, and Tina Gruosso for working on the French translation of my abstract.

I would like to thank my thesis supervisors Dr. Morag Park and Dr. Michael T. Hallett for all their guidance, training, and support over the last 7 years. I appreciate the time you took to groom me in matters both academic and professional. I will fondly remember our conversations that allowed me to grow my perspective of the world, and to expand beyond the confines I had previously been restricted to. In particular I would like to thank Dr. Hallett for taking me under his wing, and having the patience to work with me, despite my lack of bioinformatics knowledge.

I would also like to thank past and present members of my Research Advisory Committee: Dr. Josie Ursini-Siegel, Dr. Thomas Duchaine, Dr. Vanessa Dumeaux, Dr. Jason Young, and Dr. Derek Ruths – for your insight and thought provoking questions.

A special thank you to Dr. Vanessa Dumeaux, Daniel Del Balso, and other past and present members of the Hallett lab for their advice (both academic and non-academic) during my

tenure at the lab. I would also like to thank Paul Savage for allowing me to work with him on the single cell RNA-seq project. It was an honour to have collaborated with you on this work, despite the hard time I gave you about this just being a “small project”. Additionally, I would like to thank the past and present members of the Park lab for the memorable times I had with them, and to confirm that I will probably have to change my sleeping schedule, and wake up earlier now that I am all “grown up”. I will especially remember with fondness my times playing volleyball, ice hockey (Go Squids!), Scopa, and the trips to Barbados and Mont Ste. Hilaire with members of the Park and Hallett labs.

To the parents I was born to, and to the parents who came into my life a little later, I would like to thank you for your being the platform that allowed me to get to where I am today. I know that without your unconditional love, I would not have had the courage to push forward, and get to this point. To my siblings: Muizz, Zarlasht, Nazar, Kanwal, Rahim, and Zoya – I know that you are probably all for being there to laugh with me when there was sunshine, and to help me push through when things were gloomy. Thank you to my adorable nieces and nephew: Arissa, Atiyah, and Yazdan – whose laughter helped light up my days. To my Aunt/Grandmother/friend, Dr. Boustan Hirji, thank you for your pearls of wisdom. I know that Montreal would have been a lot lonelier without you in it. I would also like to remember my grandparents: those who are with me today, and those who have moved on – for your love during my formative years, that have helped make me the man I am today.

Finally, I would not be where I am today without my *qurut ul ain*, my wife Karima. You are the rock in my life, and without you to push me, I am sure that I would have not had the

courage to push through, and achieve this milestone. I think you have probably put as much blood, sweat, and tears into this thesis as I have! Thank you for being the yin to my yang.

"From the very beginnings of Islam, the search for knowledge has been central to our cultures. I think of the words of Hazrat Ali ibn Abi Talib, the first hereditary Imam of the Shia Muslims, and the last of the four rightly-guided Caliphs after the passing away of the Prophet (may peace be upon Him). In his teachings, Hazrat Ali emphasized that 'No honour is like knowledge.' And then he added that 'No belief is like modesty and patience, no attainment is like humility, no power is like forbearance, and no support is more reliable than consultation.'

Notice that the virtues endorsed by Hazrat Ali are qualities which subordinate the self and emphasize others - modesty, patience, humility, forbearance and consultation. What he thus is telling us, is that we find knowledge best by admitting first what it is we do not know, and by opening our minds to what others can teach us."

- Address by His Highness the Aga Khan at the American University in Cairo

PREFACE

This thesis is written in the traditional format, and is divided into the following five chapters:

Chapter 1. **Literature review**

Chapter 2. **Results**

Chapter 3. **Discussion**

Chapter 4. **Experimental Procedures**

Chapter 5. **Bibliography**

Publications arising from this thesis

Chapter 2 contains material presented in the following research articles:

Saleh SMI, Bertos N, Gruosso T, Gigoux M, Souleimanova M, Zhao H, Omeroglu A, Hallett MT, Park M. Identification of interacting stromal properties in triple-negative breast cancer, In preparation for resubmission to *Cancer Research*.

Paul Savage, **Sadiq MI Saleh**, Yu-Chang Wang, Timothée Revil, Dunarel Badescu, Dongmei Zuo, Alexis Blanchet-Cohen, Leah Liu, Nicholas Bertos, Valentina Munoz-Ramos, Mark Basik, Jamil Asselah, Sarkis Meterissian, Marie-Christine Guiot, Atilla Omeroglu, Claudia Kleinman, Morag Park, Jiannis Ragoussis. A targetable EGFR-associated tumor-initiating program in breast cancer, In preparation for Cancer Discovery.

***Authors contributed equally**

Contribution of authors

TNBC stroma project: M.P. & M.H. conceived the project and supervised the research; N.B., M.S., and H.Z. generated the gene expression dataset; M.G. & T.G. generated the KI67 data and reviewed the genelists; S.M.I.S. analyzed the data and generated the figures; A.O. evaluated the H&E-stained sections from each sample and selected representative regions for isolation by LCM; S.M.I.S., M.H., N.B., and M.P. wrote the manuscript.

Single cell RNA-Seq project: PS conceived the study, participated in its design, developed PDX, performed functional experiments and drafted the manuscript. SMIS participated in study design, performed gene expression informatics analysis and helped draft the manuscript. TR performed exome sequencing analysis. YCW processed single-cells for genomic profiling. DB performed sequencing quality control and statistics. EI processed sequencing data to generate read counts. DZ assisted with immunofluorescence microscopy. NB coordinated directed biobanking. KS, IH,RR and CB helped optimize single-cell profiling and single cell protein detection assays. VMR, JA and SM coordinated clinical sample procurement. AO performed pathological assessment. SH and CK assisted with RNA-seq informatics analyses. MP and JR conceived the study, participated in its design and helped draft the manuscript. All authors read and approved the final manuscript.

Additional publications:

Liu X, Nugoli M, Laferrière J, **Saleh SM**, Rodrigue-Gervais IG, Saleh M, et al. Stromal retinoic acid receptor beta promotes mammary gland tumorigenesis. Vdi VDI, editor. Proceedings of the National Academy of Sciences of the United States of America. 2011;108:774–9.

Knight JF, Lesurf R, Zhao H, Pinnaduwage D, Davis RR, **Saleh SMI**, et al. Met synergizes with p53 loss to induce mammary tumors that possess features of claudin-low breast cancer. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2013; Available from: http://www.ncbi.nlm.nih.gov/pubmed/23509284

Tofigh A, Suderman M, Paquet ER, Livingstone J, Bertos N, **Saleh SM**, et al. The Prognostic Ease and Difficulty of Invasive Breast Carcinoma. Cell Reports [Internet]. 2014; Available from: http://www.cell.com/cell-reports/abstract/S2211-1247(14)00765-7

Oh E-Y, Christensen SM, Ghanta S, Jeong JC, Bucur O, Glass B, Montaser-Kouhsari L, Knoblauch NW, Bertos N, **Saleh SMI**, et al. Extensive rewiring of epithelial-stromal co-expression networks in breast cancer. Genome Biology. 2015;16:128.

Table of Contents

1. LITERATURE REVIEW

1.1 An overview of the normal breast

The mammary gland is present in female mammals that functions to feed offspring. The functions of the adult mammary gland are mediated through several distinct cell types, as is observed in other glandular tissues. During a female's lifetime the mammary gland can change quite dramatically; the mammary gland is only partially developed during embryogenesis, and final development occurs postnatally (Figure 1) [1].

1.1.1 Epithelial cells

Several types of epithelial cells can be identified in the mammary gland. Apically facing luminal epithelial cells make up the secreting cells that line the milk duct. The basal surface is lined by a layer of myoepithelial cells that contract to facilitate milk secretion. In addition to the myoepithelial cells, cell sorting experiments have identified additional cell types with putative stem-like functions in the basal layer [2].

During puberty the mammary epithelium invades into the mammary fat pad in response to hormonal cues [3,4], and undergoes rounds of proliferation and apoptosis during the menstrual cycle [5]. However, it is only during pregnancy that the epithelial cells develop the capability to secrete milk. When the stimuli for milk production are lost during weaning, the mammary epithelium loses its milk producing capability and returns to its pre-pregnancy state in a process known as involution [1].

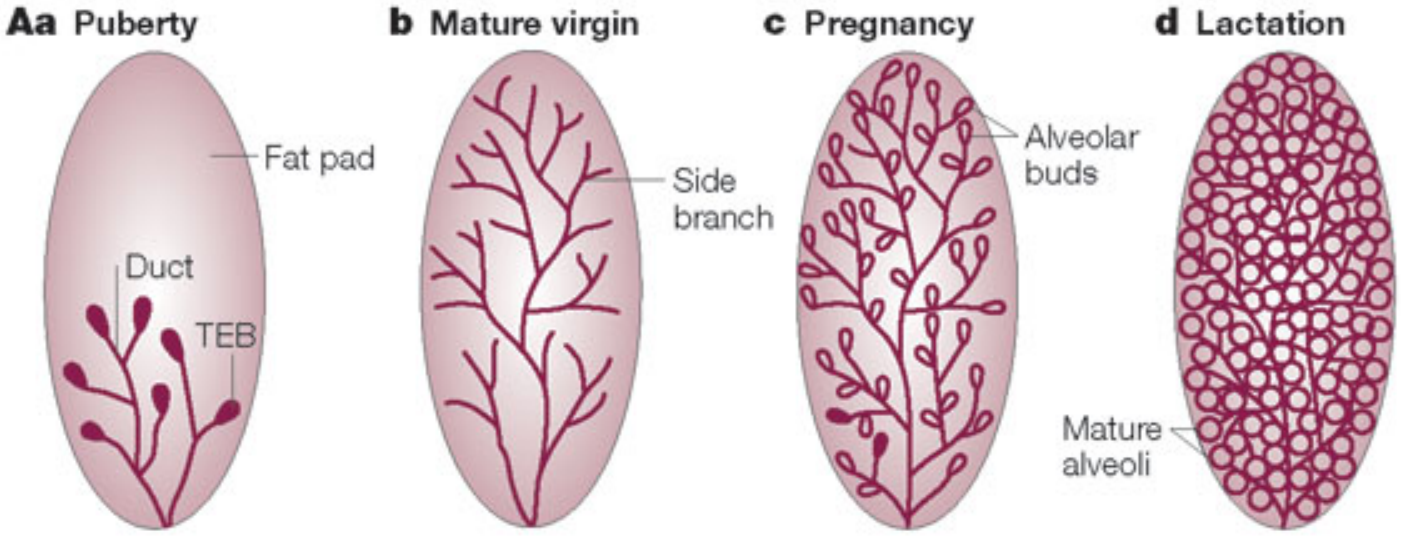


Figure 1: Schematic of mouse mammary gland development during puberty, pregnancy and lactation (Aa–d). Adapted from Figure 1, Hennighausen and Robinson, 2005

1.1.2 Stromal cells

Mammary epithelial cells are embedded within a microenvironment, or stroma, that is made up of a diverse mix of cell types. Fat-filled adipocytes comprise a large portion of the adult and non-lactating gland, are involved in milk production, and have endocrine signaling functions [6].

Fibroblasts are often found in close proximity to the basal epithelial layer, and function to support the mammary epithelial cells [7]. Fibroblasts have the ability to produce and remodel the extracellular matrix (ECM) [8], which results in the entrapment and release of growth factors that influence neighboring cellular functions [9].

Blood and lymphatic vessels are also present in normal breast stroma and play an important role, particularly during lactation where they serve to deliver nutrients and drain the breast of waste metabolites [10]. Distinct populations of immune cells are also present in the breast stroma. In addition to providing protection against infection, immune cells such as eosinophils, mast cells, and macrophages also perform additional functions during the branching, and involution phases of the mammary gland [11]. The immune system is also able to identify and kill transformed epithelial cells. This process is termed immune surveillance and is mediated by B-cells, T-cells, and natural killer cells, among other cells [12].

Sakura and colleagues demonstrated the substantial effect that stroma has on mammary epithelium development. They observed that mixing mammary epithelial cells with salivary

stromal cells resulted in the formation of structures more reminiscent of the salivary gland rather than a mammary gland [13].

1.2 Breast tumorigenesis

Normal cells go through a regulated cycle of growth and death to maintain tissue architecture and function. The abnormal, uncontrolled growth of cells can lead to either non-cancerous (benign) tumors or a malignant lesion. Tumors that originate from the epithelial cells lining the ducts are the most commonly occurring breast cancers, though tumors can also form in the lobules and other breast tissues. Tumors arising from the ductal epithelial tissue can be classified as ductal carcinoma in-situ (DCIS) or invasive ductal carcinoma (IDC). DCIS tumors are characterized by an abnormal proliferation of cells that fill the local duct but do not invade the surrounding tissue. In contrast, IDC tumors invade through the membrane surrounding the ducts into the local microenvironment [14].

1.3 Tumor epithelial-stromal interactions

Normal cells have checkpoint mechanisms present to prevent their progression to a malignant phenotype. In response to this tumor cells develop several complementary functions to circumvent these checkpoints. The deregulation of signals promoting progression through the cell cycle, enabling cell growth, promoting cell survival, and deregulating energy metabolism are some of the functions gained by tumor cells as they become malignant. These ‘hallmarks of

cancer' [15] are achieved through gain of function mutations in oncogenes, or through loss of function mutations in tumor suppressor genes. Epithelial cells also adapt to avoid destruction by immune cells, either by mitigating the effect of the immune cells or by escaping detection by immune cells [12].

As with the normal mammary gland, breast tumors are not made up solely of epithelial cells. The tumor interacts with its microenvironment and the tumor microenvironment provides complementary functions for tumor growth. In addition to contributing to functions such as tumor vascularization, stromal cells can also contribute to achieving the hallmarks of cancer [16].

The strong interactions between the stroma and epithelial compartments in breast tumors are not limited to the breast. In particular it has been observed that tumor cells can mobilize stromal cells from the bone marrow to allow a tumor located in the adjacent breast to change from being indolent to metastatic [17]. A few examples of how stromal cells aid in tumor progression are described below.

1.3.1 Tumor cell proliferation and growth

Although driving mutations can lead to chronic proliferation, it has also been observed that stromal cells have the capacity to support the hyperproliferation of adjacent epithelial cells. For example cancer associated fibroblasts can be induced to express and secrete growth factors, or to

degrade the ECM to release growth factors. This in turn stimulates proliferation in the neighboring epithelial cells [18].

1.3.2 Cell death resistance

Stromal cells also enable the tumor cell to resist cell death. For example tumor associated macrophages have been observed to adhere to tumor cells and mimic the interactions usually present between two epithelial cells. This allows the tumor cell to circumvent cell death pathways usually triggered by anoikis, or detachment from epithelial cells [19]. Cancer associated fibroblasts have also been observed to mediate tumor cell survival [20].

1.3.3 Preventing destruction by immune cells

The tumor stroma contains distinct subsets of mononuclear immune cells commonly referred to as "tumor infiltrating leukocytes" (TILs). Higher proportions of TILs in the tumor has been associated with better patient prognosis [21]. TILs consist of distinct cell types such as B-cells, CD4 T-cells, dendritic cells, and other immune cell types that have been observed to have both pro-tumorigenic and anti-tumorigenic roles. The presence of one cell type in particular in the tumor stroma, CD8+ cytotoxic T-cells, has been associated with good prognosis in several different cancers. This is likely due to the ability of cytotoxic T-cells to clear damaged cells by targeting them specifically and programming them to undergo apoptosis [22]. For malignant cells to grow into a tumor these epithelial cells must be able to evade or suppress these cytotoxic

T-cells, and other immune cells that would otherwise induce epithelial cell death. Infiltrating immune cells have been observed to develop phenotypes normally associated with wound healing and inflammation [23]. This aberrant activation of the immune cells can lead to the generation of an immunosuppressive microenvironment thus allowing the tumor to evade immune cell mediated death.

1.4 Breast cancer epidemiology

Breast cancer is a major health concern among women and, with the exception of non-melanoma skin cancers, is the most frequent cancer among women worldwide [24]. In Canada it is estimated that in 2015 alone 25,000 women will be diagnosed with breast cancer and 5,000 will die from it [25]. Historically, breast cancer incidence was lower in developing countries, but increases in life expectancy and changes in lifestyle has seen a rise of breast cancer incidence in these countries [24]. Despite the stability of breast cancer incidence rates over the last 20 years, the mortality rates of breast cancer have seen a dramatic decrease due to advances in screening and treatment of the disease. Current rates estimate that 88% of Canadian women diagnosed with breast cancer will survive past five years [26], which is a marked improvement when compared to the average survival rate of 76% observed between 1985-1987 [27].

The majority of breast cancer research focuses on female breast cancer. While male breast cancer does occur, its incidence is much lower. While 1 in 9 Canadian women are estimated to

develop breast cancer in their lifetime, the incidence rate for Canadian males is a much lower 1 in 220 [25].

Despite not knowing the etiology of most breast cancer cases, numerous risk factors have been linked to breast cancer incidence. In addition to the increased risk present for females, increased age, a family history of breast cancer, and mutations in genes such as BRCA1/2, TP53, and ATM [28], represent some additional risk factors associated with an increased risk to develop breast cancer [29].

1.5 Breast cancer in the clinic

In developed countries, most breast cancer patients are diagnosed by mammography. However, many breast cancers are diagnosed with masses that are not detected by mammograms, or during the interval between mammograms (15% and 30% respectively) [29]. The benefit of regular screening by mammograms has only been observed among women aged 50-74 [30]. For women above the age of 75 there is insufficient evidence to evaluate the benefits and harms. Among younger women, the sensitivity of the mammographic screens has been observed to be significantly lower [31]. This led to the recommendation for regular mammographic screening to be only applied to older women.

The decrease in mammogram sensitivity is associated with increased breast density, and is present even among older women [32]. Despite this decrease in mammogram sensitivity, the risk of breast cancer incidence is significantly higher among women with dense breasts. Thus there is

a large body of women at high risk for breast cancer who do not undergo regular mammographic screening, and for whom diagnosis of the disease may be delayed.

Following the identification of a suspicious growth, a biopsy is performed to confirm if the growth is malignant. Often the biopsy is performed by fine needle aspiration (FNA) which is a minimally invasive technique involving a small bore needle. FNAs are typically performed when the mass is palpable, but can also be performed on non-palpable lesions by leveraging the use of imaging technologies (e.g. ultrasound). Core needle biopsies (CNBs) and excisional biopsies are two alternate methods that provide additional accuracy in the evaluation of non-palpable lesions at the expense of being more invasive than the FNA method. CNBs use a larger hollow needle to remove suspicious tissue samples from the breast. CNBs can provide additional information over FNAs with regards to whether cells have breached the basement membrane, and thus are more informative when the malignancy of the lesion is unclear. In contrast to the non-surgical FNA and CNB methods, excisional biopsies are the surgical removal of either the entire breast mass or a suspicious section of the breast. Biopsies allow a clinician to determine whether a suspicious mass is malignant, and in the case of a malignant mass, can provide additional information for patient care [29].

1.6 Breast cancer subtypes

Histopathological examination of tumor biopsies is a common classification scheme used for breast cancer. While invasive ductal carcinoma (IDC) and invasive lobular carcinoma (ILC)

make up the majority of breast cancers, (85-90%), there are many additional histological subtypes with much lower rates of incidence. For the remainder of this thesis, the term 'breast cancer' will refer to IDC cases unless otherwise specified.

In addition to the histological subtypes, the tumor can also be classified by several additional features. Tumor grade is a prognostic score calculated based on the architecture of the tumor. By assessing nuclear atypia, mitotic activity, and tubule formation a grade between 1 and 3 can be assigned to the tumor, with higher grades being associated with progressively poorer outcome.

Tumors are also stratified into subtypes defined by differential protein expression of the estrogen receptor (ER) and progesterone receptor (PR), as well as expression and/or genomic amplification of human epidermal growth factor receptor 2 (HER2) (Figure 2). These proteins have been selected as subtype markers as they have targeted treatments that led to an improved patient prognosis within the associated subtype. These subtypes are typically determined from biopsied tissue. ER and PR status is typically determined by immunohistochemistry (IHC) testing, while HER2 can either be assessed by IHC or fluorescence in situ hybridisation (FISH). Tumor proliferation can also be assessed independently by IHC staining with Ki67, with higher staining intensity and number of positive cells associated with increased proliferation.

The main treatment modality implemented for breast cancer patients is surgery to remove the primary tumor. In addition to surgery ER/PR positive patients receive tamoxifen or aromatase inhibitors to inhibit the estrogen receptor or decrease estrogen levels respectively. HER2 positive patients are treated with Trastuzumab, a monoclonal antibody targeting the HER2 receptor.

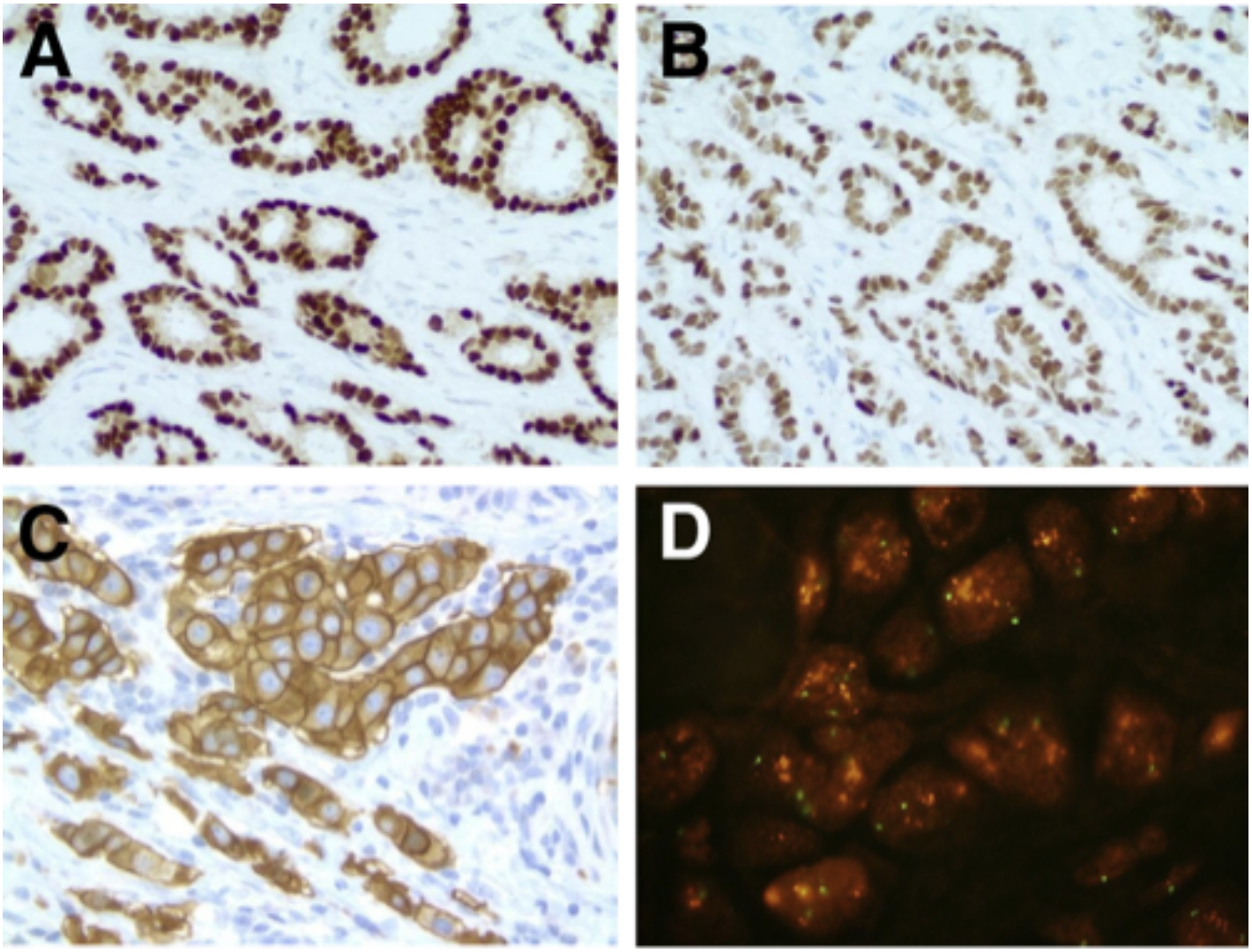


Figure 2: IHC (A–C) and FISH (D) analyses of breast cancer. A: ER+ strong intensity. B: PR+ moderate intensity. C: HER-2 overexpression, score 3+. D: HER-2 gene amplification. Original magnification: ×20 (A and B).

Additionally radiation therapy and chemotherapy may also be administered, especially among patients where the tumors have metastasized to the lymph nodes. Neo-adjuvant therapies can also be administered preoperatively to reduce tumor burden and limit the amount of breast tissue that is removed [29].

1.7 Triple-negative breast cancer

The 10-15% of breast carcinomas that lack expression/amplification of ER, PR or HER2 form a subtype that is termed triple-negative breast cancer (TNBC) [33]. Tumors of the TNBC subtype are associated with earlier age of onset, and higher grade at presentation [34]. Since tumors of this subtype are negative for ER, PR, and HER2, the only adjuvant treatments that patients diagnosed with TNBC receive are a combination of chemotherapy and radiation. The type and dose of chemotherapy that TNBC patients receive depends on the stage of the tumor, with stage I tumors receiving less toxic regimens than stage IV tumors [33]. Despite the general improvement in breast cancer patient outcome as a result of chemotherapy, TNBC patients display an overall poorer patient prognosis when compared to other subtypes of the disease. While some TNBC patients respond well to chemotherapy, other tumors show resistance to the same treatment indicating that the TNBC subtype is heterogeneous. It has been observed that the presence of infiltrating lymphocytes in TNBC tumors has been associated with better overall prognosis, possibly due to an increased sensitivity to chemotherapeutics [33]. Further

stratification may be able to further differentiate patients who may benefit from chemotherapy from those who don't.

1.8 Gene expression microarrays

Gene expression profiling is a class of methods aimed at capturing a quantitative snapshot of RNA levels in cell or tissues [35]. Gene expression microarrays have traditionally been utilized for gene expression analysis, and represent a low cost method to assess the transcriptional profile of a sample. These microarrays consist of DNA oligonucleotide probes adhered to a solid surface. These probes have been generated to be complementary to RNA sequences found within the species of interest and can vary in number from several thousand to several millions to more accurately quantitate the RNA levels in the sample. RNA from the sample of interest is first labeled with fluorescent nucleotides and upon exposure to the microarray the fluorescent RNA molecules bind to their complementary probes. A snapshot of the microarray is taken with a highly sensitive camera and the fluorescence of each probe is quantitated. The resultant fluorescent intensity gives an estimation of the abundance of each RNA within the sample.

1.9 Bioinformatic approaches

Analysis of large gene expression datasets would not have been possible if not for the development of bioinformatics approaches and tools. This thesis makes use of many such tools to ensure that the data is of good quality, and to analyze the data to address specific hypotheses

regarding the heterogeneity of breast cancer. An overview of standard methods used in the analysis of gene expression microarray datasets is presented below. These methods have been implemented in the projects presented as well as in previous work published by this lab [36–40].

1.9.1 Proper study design

The proper design of a dataset is of key importance as a poorly designed dataset may prevent proper analysis. Due to the highly sensitive nature of gene expression technologies, small differences during sample preparation may greatly impact the dataset. Examples of such differences include the ambient temperature of the room in which the samples are prepared, the batch of chemicals used to extract the RNA, or even the person who extracts the RNA. These differences are commonly referred to as 'batch effects' and they can be prevented by proper planning.

While some technical aspects can be controlled and planned for: e.g. ensuring that only one person handles all the samples or that only one batch of chemicals is used – other aspects are not as easily controlled. Environmental factors are an example of one such uncontrollable factor, and thus proper study design must be implemented to prevent such factors from ruining the dataset. Study design involved the planning of the experiment such that groups of interest are not handled in the same batch and are instead spread over different batches. For example, it would be improper to handle all the control samples one day and the treated samples from an experiment on another day. Doing so will likely result in differences between control and treated samples

being confused with batch effects resulting from handling on two different days and thus confound identification of relevant biological signals.

Proper study design becomes more complicated with larger datasets, especially when groups of interest are not known *a priori*, and batch effects may be unavoidable in some circumstances. In such scenarios tools to correct for batch effects may be utilized. One such method implemented by the LIMMA package [41] first fits a linear model to estimate the association of each feature with the batch effect, and then takes the residuals of this linear model to remove the association. However caution must be exercised when attempting to remove a batch effect so as not to hide relevant biological signals or introduce unintentional noise into the dataset.

1.9.2 Data normalization

After the dataset has been generated, the raw data needs to be harmonized to permits the comparison of distinct samples within the dataset. In the case of microarrays, the fluorescent value that the image extraction software captures is assumed to be a combination of the signal from the hybridized RNA and some background noise that results from technical variability. For transcripts with low signal intensities the background noise may mask the signal and lead to underestimation of the signal. Thus methods for determining and subtracting the background noise from the signal were developed. It has been observed that background subtraction introduces additional noise for some two-color Agilent arrays, and that the signal is most reliably estimated when no background subtraction is performed [42].

In order to assess the quality of individual arrays of 2-channel microarrays, MA plots can be utilized. These plots compare the log2-ratio of the two channels (M) and the average of the two channels (A). Gene expression datasets make use of two assumptions: (1) the first that the majority of genes do not significantly vary in expression across samples, and (2) the second that the variance of a probe is not correlated with overall intensity. Thus it can be assumed that the MA values should be centered around 0 and that samples with distributions that deviate strongly from 0 may be of poor quality. Methods such as LOESS [43] can correct for within-array bias by fitting a regression to the MA-values, and attempt to center the distribution around 0 by taking the residuals of the fit.

Lastly, "between-array" correction can be performed to permit comparison between different samples. This correction step is based on the assumption that there are no gross changes between the overall expression profiles between samples. Quantile normalization makes use of this assumption and transforms the samples such that they have the same empirical distribution across the dataset.

1.9.3 Class discovery

In a typical breast cancer dataset, pre-existing information such as patient survival and receptor status may be available to separate samples into distinct classes. However, class discovery can be used in cases where this information is not available, or when trying to identify novel classes to distinguish samples. Class discovery attempts to partition the samples in a gene

expression datasets into groups of similar samples in an unbiased manner based solely on the data presented.

Class discovery involves three steps, the first being the identification of appropriate features. One method of selecting features is to select transcripts with high variance as they are more likely to provide a signal for identifying subtypes. The second step for class discovery involves converting the features into a distance metric to determine the similarity of samples. Two distance measures often used are Euclidean distance, which represents the shortest distance between two points, and a correlation-based distance, which uses one minus the standard Pearson correlation coefficient, to measure similarity. Following feature selection and identification of an appropriate distance metric, clustering is used to group similar samples together [44].

Hierarchal clustering is a commonly used method to summarize data in a 2-dimensional space, to be easily visualized as a heatmap. With hierarchical clustering each sample is initially assigned to its own cluster. Next, the two closest clusters are joined based on the chosen distance metric. This is repeated iteratively until all clusters have been combined [45]. This clustering approach can be summarized using a tree structure, known as a dendrogram, where leaves that are closer together represent clusters that are more similar to each other. Clustering can be performed on both the features (genes) and samples independently to identify groups of samples with similar patterns of gene expression, and groups of genes with similar patterns of expression across samples.

While this approach groups samples based on how similar they are, it does not determine the number of groups present. To determine the number of groups, the stability of clusters can be measured using permutation based approaches [46] and highly stable clusters can be used to determine the appropriate number of groups.

Subsequent to the identification of groups of samples, these groups can be analyzed to determine if there is association with known clinical variables (e.g. tumor size, grade, stage, outcome, etc.). Typically this is done through the use of a enrichment test such as the hypergeometric test. This identification of patient subtypes, and their association with clinical variables, allows for a better understanding of the disease.

1.9.4 Linear ordering

Gene expression signatures are often used as a surrogate to estimate the level of activation of a pathway in a sample. Pathway activation levels may not always be discrete (e.g. on or off) and may instead be represented as a continuous variable. Clustering based approaches attempt to partition samples or features into distinct groups, and are therefore not appropriate to order samples based on a continuous variable. A more appropriate method is to linearly order samples according to increasing levels of the signature. Therefore, while clustering can be used to identify gene signature, linear ordering may be more appropriate to estimate the activation of pathways associated with the gene signature.

Several methods have been utilized to estimate a linear ordering, including the estimation of a 'metagene' by simply averaging the expression of all the genes within the signature [47]. However, these methods make assumptions about the distributions of the individual genes of the signature. In contrast, the method used in this thesis avoids making such assumptions by rank-ordering samples based on the sum of ranks of the genes in the signature.

1.9.5 Class distinction

While class discovery focuses on identifying groups of samples, class distinction builds on this stratification by identifying features that define these groups. The groups used for class distinction can be defined based pre-existing information for the samples (e.g. patient survival, receptor status) or based on classes identified by class discovery. A variety of statistical methods have been developed to perform class distinction.

Student's t-test is a traditionally used parametric method that tests the alternate hypothesis if the means of two groups are significantly different. The t-test assumes that the two groups are derived from two normal distributions with approximately equal variance, and that the groups are of a reasonable size. These assumptions are not always valid when analyzing gene expression datasets.

Modified versions of the Student's t-test have been developed for microarrays to circumvent some of these assumptions. LIMMA [41] is one such method that is attempts to reduce the uncertainty associated with determining the standard deviation for each probe by shrinking the

estimated variances for each probe to a pooled variance. This generates a far more stable inference of the t-statistic and improves the power of the analysis, especially for experiments with fewer samples.

These class distinction methods test each feature independently. It is expected that a large number of features may be wrongly identified as being significantly different (false positive) by chance alone. This is the multiple testing problem. One technique to minimize the number of these false positives is to adjust the raw p-values obtained from the individual tests. The Benjamini–Hochberg (BH) method [48] is one such method for estimating the False Discovery Rate (FDR) and adjusting the raw p-value. The BH method first orders the determined p-values from lowest to highest, and then determines the highest value of (k) that satisfies the equation:

$$P(k) < \frac{k}{m}\alpha$$

where k is the rank of the feature as determined by the p-value, P(k) is the p-value for the feature, m is the total number of features, and α is the FDR value to be calculated.

All features of rank 1...k are determined to be significant even if they did not satisfy the equation individually. The result of this is that for an FDR of 0.05, 5% of the features are estimated to have been falsely identified as differentially expressed for an FDR adjusted.

1.9.6 Class prediction

The bioinformatics community also makes use of machine learning techniques to classify samples. This problem can be summarized as the ability to predict the class of a future sample based on a set of previously derived rules. To derive these rules the classifier is first trained on a dataset where the class labels of the samples are already known. Ideally, the classifier can then be tested on an independent dataset. If no such dataset is available, the testing can be performed using a leave *k* out cross validation. This approach repeatedly leaves random samples out of the original dataset and then uses these samples to optimize the classifier.

Naïve Bayes classifiers (NBCs) are a classification method utilized in our analyses. NBCs have been observed to perform well in the classification of microarrays [49]. NBCs are based on a model of conditional probability and assume that the genes used for classification act independently to classify samples. NBCs calculate a posterior probability for a class based on the following formula:

$$\text{Posterior} = \frac{Likelihood \times Prior}{Evidence}$$

Therefore, a simple NBC built to predict the outcome of a patient based on ER and HER2 status would calculate the following:

$$\Pr(Outcome\ of\ this\ patient) = \frac{[\Pr(ER|Outcome) \times \Pr(HER2|Outcome)] \times \Pr(Outcome)}{\Pr(ER, HER2)}$$

Given that the Prior and Evidence probabilities are constant when the values of the features are known, the likelihood probabilities are what influence the posterior probability. When there are multiple possible classes, a posterior probability is determined for each class, and the class with the highest posterior probability is determined to be the most likely classification.

1.9.7 Pathway analysis

Class distinction analyses can generate lists differentially expressed genes that vary in size from fewer than ten to several hundreds, or thousands, of differentially expressed genes. Since the size of these differentially expressed gene lists are often too large for manual analysis, bioinformatics tools are required to identify relationships between groups of genes that may be interacting, performing similar biological processes, or be part of the same pathway. These tools are termed 'pathway analysis tools' and can be classified into several distinct categories depending on the pathway genelists (signatures) used and the type of statistic used. In the analyses described in this thesis we use two pathway analysis methods: over-representation analysis and Gene Set Enrichment Analysis [50].

Over-representation approaches compare the differentially expressed genes against other publically available signatures. The overlap between the list of genes determined to be differentially expressed genes by class distinction, and publically available genelists is determined. Statistics such as the Fisher's Exact Test ask whether there are a surprisingly number of genes in common between the target pathway and the list of differentially expressed

genes. The FET compares two categorical groups and determines if the observed overlap between two groups is higher than would be expected by random chance alone.

In contrast to over-representation approaches, GSEA does not only utilize the subset of genes determined to be differentially expressed. Instead all genes are ranked based on the results from class distinction, and the enrichment of gene signatures at the low and high tails of these ranked lists is determined. The significance of the this enrichment is determined using permutations of the gene or sample labels.

While pathway analysis can offer a better understanding of large lists of differentially expressed genes, these methods are not always informative. Despite advancements of statistical approaches for pathway analysis, the lack of similar advancements in the gene signature databases means that often no significant pathways are identified. This can be especially pertinent when analyzing datasets with novel cell types for which signatures have not previously been identified.

1.10 Breast cancer informatics in the clinic

In the early 2000s, the introduction of a gene expression microarray that could simultaneously detect the level of thousands of transcripts promised to revolutionize the field of breast cancer research. It was hypothesized that generating gene expression profiles from breast cancer patients would enable further stratification of breast cancer in addition to what was observed by traditional IHC approaches, and that this in turn would allow for the stratification of

patients into distinct groups that are responsive to standard of care and thus required no additional intervention, and a distinct group where patients were non-responsive and thus require additional therapeutics. Two distinct approaches were implemented to stratify patients.

1.10.1 Stratifying patients by a clinical end-point: prognosis

The first type of approach made use of patient outcome information and attempted to build classifiers that could distinguish patients based on their clinical outcome. These initiatives built classifiers based on lists of genes associated with patient outcome. These genelists were derived using class distinction to identify genes that differed in expression based patient outcome [51], or were based on prospectively selected genes previously associated with patient outcome [52]. Therefore this approach built classifiers based on a set of genes that have differential expression between patients with poor and good clinical outcome. One successful classifier that resulted from this approach is a 70-gene panel [53] that was redeveloped as a clinical assay named 'Mammaprint'. Despite being identified as a prognostic geneset that stratifies patients based on clinical outcome, it was observed that the Mammaprint geneset also had the capacity to predict which patients would have good outcome without chemotherapeutic intervention. It has since been repurposed to distinguish high-risk patients that would benefit from chemotherapy from low-risk patients who will likely have good outcome despite not receiving chemotherapy among early-stage breast cancer patients [54].

1.10.2 Subtyping breast cancer

The second approach used class discovery to identify patient subtypes. As previously described, this approach labels samples with new class labels solely based on the data presented. Therefore these subtypes were identified independent of patient outcome information and other clinical characteristics of the tumor. These 'molecular subtypes' of breast cancer grouped patients that were most similar by their gene expression profiles (Figure 3) [55,56].

Despite being identified independently of clinical information, the molecular subtypes were differentially associated with ER/PR and HER2 status and with patient prognosis. In particular two subtypes were strongly associated with ER/PR positive patients (Luminal A and B), another subtype enriched for HER2 positive patients (HER2-enriched), and a subtype enriched for TNBC patients (Basal-like). A fifth controversial subtype (Normal-like) clustered with the samples taken from the normal breast and thus it is debated whether this subtype represents samples contaminated by normal tissue.

Following the identification of this subtyping scheme alternate schemes have also been proposed. These schemes have been developed using gene expression and other genomic technologies. An example of an alternative scheme are the IntClust subtypes that used the combined analysis of DNA and mRNA samples from ~2,000 samples to identify 10 distinct subtypes [57]. As was observed with the intrinsic subtypes, these IntClust subtypes also varied in their prognostic association. However the majority of the IntClust subtypes further stratified the Luminal subtypes, whereas there was little advance in the stratification of the HER2 or Basal-like subtypes.

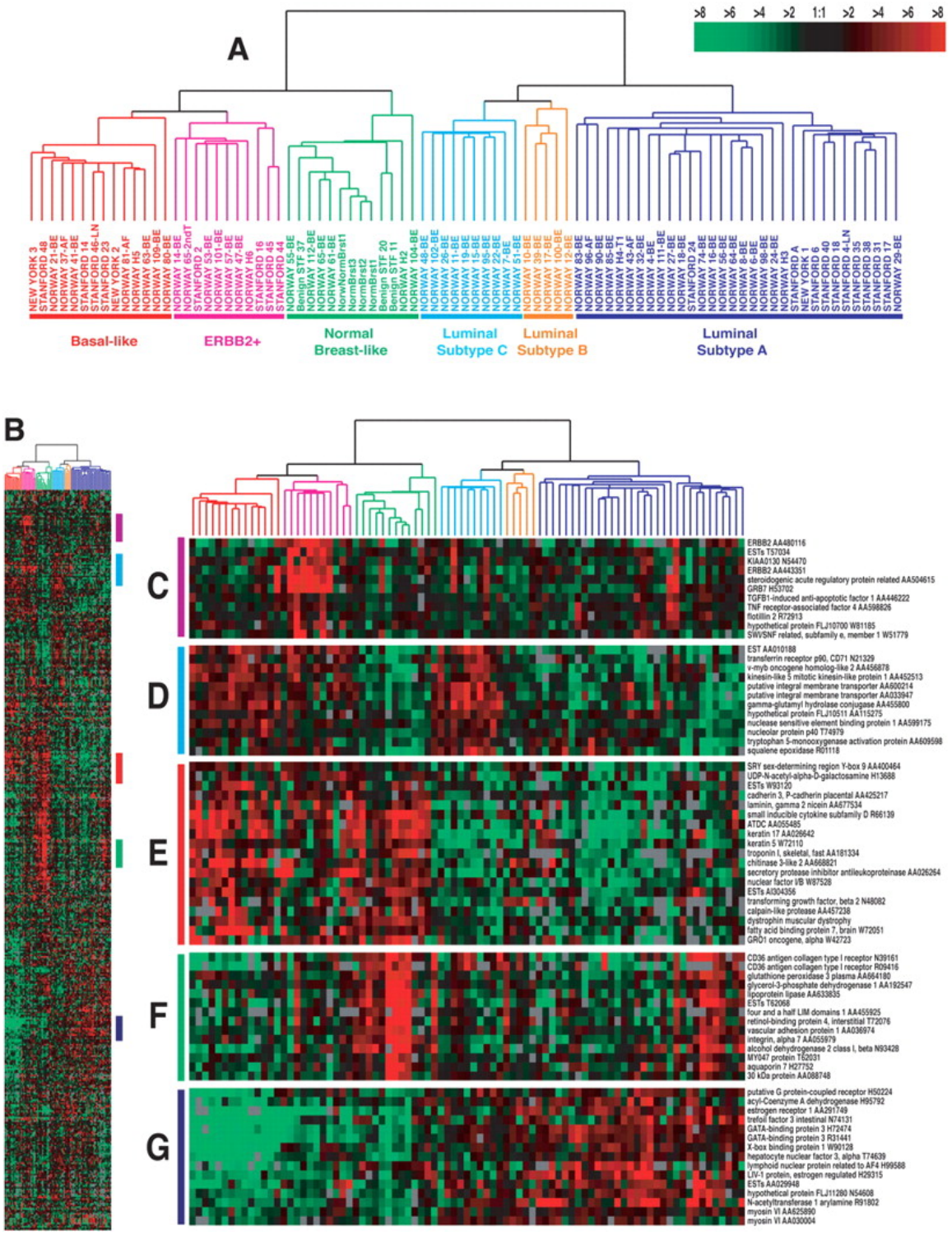


Figure 3: Gene expression patterns of 85 experimental samples representing 78 carcinomas, three benign tumors, and four normal tissues, analyzed by hierarchical clustering using the 476 cDNA intrinsic clone set. (A) The tumor specimens were divided into five (or six) subtypes based on differences in gene expression. The cluster dendrogram showing the five (six) subtypes of tumors are colored as: luminal subtype A, dark blue; luminal subtype B, yellow; luminal subtype C, light blue; normal breast-like, green; basal-like, red; and ERBB2+, pink. (B) The full cluster diagram scaled down (the complete 456-clone cluster diagram is available as Fig. 4). The colored bars on the right represent the inserts presented in C–G. (C) ERBB2 amplicon cluster. (D) Novel unknown cluster. (E) Basal epithelial cell-enriched cluster. (F) Normal breast-like cluster. (G) Luminal epithelial gene cluster containing ER.

Figure 1, Therese Sørlie et al. PNAS 2001;98:10869-10874, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications
Retrieved December 12 2016

These subtyping schemes for breast cancer propose varying numbers of patient partitions that range from just four subtypes via classic approaches based on ER, PR and HER2 status to 10 subtypes identified by IntClust through joint DNA and mRNA analysis.

However these subtyping schemes often make assumptions about the nature of the dataset. One such assumption is that the distribution of subtypes in the dataset used for classification mimics the distribution of subtypes in the training dataset. The AIMS approach was developed to circumvent this assumption [58]. AIMS uses a set of binary rules to classify samples individually and independent of a dataset and thus negates this assumption. By not requiring samples to be normalized before being assigned to subtypes, AIMS is able to classify samples independently of the dataset composition.

1.11 The importance of looking within subtypes

The efficacy of subtyping schemes is measured by the prognostic association of the subtypes. An investigation by Venet and colleagues [59] speculated that nearly all genes and processes are prognostic in breast cancer. In contrast, Tofigh and colleagues [40] observed that the majority of gene signatures are in fact associated with tumor subtype. This is due in part to the large transcriptional signature of the estrogen signaling pathway which is differential amongst the subtypes, and is associated with patient survival. Therefore the ability of most genes to predict patient subtype is confounded with their ability to predict patient prognosis. Further investigation revealed that the prognostic capacity of gene signatures within subtypes was significantly

reduced when investigated within individual subtypes [40]. This stressed the importance of identifying signatures that could stratify patients within a subtype, and to identify novel schemes to further sub-stratify patients within subtypes.

Additionally a subset of patients were identified whose observed outcome was consistently mispredicted by almost all reported prognostic gene signatures. The consistent misprediction of patients by signatures generated a novel definition of 'inherent difficulty'. Patients defined as inherently difficult have a clinical prognosis that can not correctly be predicted by gene signatures.

1.12 TNBC subtypes

Previous studies, including several high-throughput profiling efforts, have indicated that the TNBC subtype has higher levels of inter-tumoral (patient-to-patient) heterogeneity when compared to other subtypes with respect to both gene expression [40], and somatic genomic aberrations [60,61]. This heterogeneity may at least partially underlie why TNBC is a poor outcome subtype [62,63]. Several efforts have investigated whether there are subtypes within TNBC with distinct cellular processes and responses [47,64,65]. These studies however have used gene expression profiling of bulk samples enriched for epithelial cells of the tumor proper.

One scheme in particular, proposed by Lehmann and colleagues [64], has received much focus from the community and has since been translated into an assay for clinical implementation (INSIGHT TNBCTYPE™) *[66]*. This scheme identified six subtypes within

TNBC patients and demonstrated that these "TNBCType" subtypes are associated with differential responses to neoadjuvant chemotherapy [67]. The six subtypes were given distinct titles based on the elevated expression of specific genes and the association with different pathways. BL1 (basal-like 1), BL2 (basal-like 2), M (mesenchymal), MSL (mesenchymal stem-like), IM (immunomodulatory), and LAR (luminal androgen receptor). A second scheme [65] identified 4 subtypes that were strongly associated with the subtypes identified by Lehmann and colleagues, which lends credence to their validity.

1.13 Investigating tumor stromal heterogeneity

The previous studies mentioned have used gene expression profiles of whole tumor tissue, Finak and colleagues [36,37] used gene expression profiles derived from laser capture microdissected (LCM) stromal tissue to investigate stromal heterogeneity in a pan-breast cancer cohort. LCM presents a powerful tool to isolate tissue in a compartment specific manner that is amenable to gene expression profiling. This investigation revealed that tumor stroma was heterogeneous between tumors, and that there are distinct stromal subtypes. Additionally, they observed that these stromal subtypes are associated with patient prognosis, and can contribute to identifying patients that may require additional therapeutic intervention [37].

The importance of stromal heterogeneity within the TNBC subtype in particular was observed by Tofigh and colleagues [40] who observed that stromal gene expression signatures, and immune-related signatures in particular, were particular efficate within the TNBC subtype.

This was further strengthened by the identification of an immune-related signature in the TNBCType subtyping scheme. These observations warranted an unbiased investigation of the stroma of TNBC patients to identify additional sources of stromal heterogeneity. Additionally the current interest in the field of immunotherapeutics, which seeks to reactivate the immune cells in a tumor, would likely benefit from such an investigation.

1.14 Intratumoral heterogeneity

As mentioned previously, breast cancers display various sources of inter-tumoral heterogeneity, that is heterogeneity between distinct tumors, with respect to histopathological categories and the measurement of epithelial characteristics. While this heterogeneity can be exploited in the treatment of the disease by stratifying patients into subtypes that are associated with distinct prognosis, or by identifying patients that may respond to a specific treatment regimen, it is insufficient to completely explain the disparate response to treatment observed among patients within each of the subtypes.

In addition to inter-tumoral heterogeneity has also been observed within a tumor (intra-tumoral heterogeneity). A common example of this heterogeneity is observed when classifying a tumor based on its ER-positivity. Clinically, a tumor is defined as ER-positive when at least 1% of cells stain positive for the estrogen receptor [68], and it has been observed that tumors with higher number of ER-positive stained do respond better to endocrine therapy. Therefore these

tumors display substantial heterogeneity with respect to ER-positivity, and this heterogeneity is linked to therapeutic efficacy.

Intra-tumoral heterogeneity can be subclassified as spatial or temporal. Spatial heterogeneity refers to observable heterogeneity between different regions of a tumor. This heterogeneity can represent distinctions between different regions of the primary tumor, between the primary and metastatic lesions, or between metastatic sites. In addition to its observation at the histopathological level, proof of intra-tumoral heterogeneity has also been observed through genetic analysis of distinct regions of the tumor. It was observed that there were distinct genetic aberrations associated with distinct sections of the tumor [69]. This heterogeneity may be tied to therapeutic resistance as treatments that would be effective on one region of the tumor but not on the other. Thus individual treatment decisions must be necessary for distinct regions of the tumor.

In addition to spatial heterogeneity, tumors have been observed to evolve as they progress. This is similar to ecosystems that evolve under stress from external pressures. A commonly investigated form of temporal heterogeneity is related to the evolution of a tumor as it progresses from primary to metastatic disease [70]. A second form of temporal heterogeneity that has been studied are the changes present in the residual tumor after being treated [71]. These study of these sources of temporal heterogeneity can help identify potential treatment avenues for patients who do not respond to standard of care.

1.15 Patient-derived xenografts (PDXs)

The limited availability of patient derived breast cancer tissue coupled with the lack of accurate models for studying breast cancer has led to the development of PDX models [72]. PDXs are derived by growing of a human tumor in immune-compromised mouse. These PDXs have been observed to recapitulate the primary tumor using several distinct metrics. Historically tumors were implanted subcutaneously, but it was observed that when the tumors were transplanted orthotopically (i.e transplanted into the mammary fat pad of recipient mice) the resultant tumors more faithfully recapitulated the stromal environment of the primary tumor [73]. While the models do present with some limitations, the most prominent being the inability to investigate interactions between the immune system and the tumor, they do offer a renewable resource to study breast tumors in a biologically relevant setting [74].

PDXs have also been observed to mimic drug responses observed in the primary tumor. While the growth rate between different PDXs can be highly variable, they have been used to predict drug response for future patients, as well as to offer insight into what may cause drug resistance in these patients [74].

1.16 Single cell RNA sequencing (scRNA-seq) technology

While traditional transcriptomic analysis experiments average the gene expression profiles over thousands of cells, recent technological development have enabled the isolation and sequencing of single cells, thus providing a method for estimating the gene expression profile of

single cells. While the data from single cell experiments resemble data derived from bulk expression profiles (i.e. a matrix of m samples x n transcripts) the methods for isolating single cells and the limited availability of starting material in single cell experiments gives rise to additional sources of noise that need to be adjusted for [75]. These sources of noise include a large set of transcripts for which no reads are detected, as well as a strong relationship between the mean expression of a transcript and its variance.

1.17 Novel challenges presented with analyzing scRNA-seq data

The recent development of massively parallel sequencing technologies have allowed for the measurement of RNA frequencies by sequencing of sample cDNA. This development dramatically overcame some of the challenges faced by microarray based gene expression profiling, including the limited range of detection. Thus it has also allowed for novel implementations of gene expression profiling. Single cell RNA-sequencing (scRNA-seq) is one such implementation that allows us to profile the transcriptomes of single cells, a feat that was not possible with traditional microarrays. However this technology comes with some distinct challenges. These include the higher cost associated with sequencing to achieve sufficient coverage to detect low or rarely expressed transcripts, as well as the need to develop novel methodologies to normalize and analyze the data.

As a result of the challenges associated with analyzing this data, namely with the unbiased identification of distinct cell populations, scRNA-seq has seen limited use in the investigation of

intratumoral heterogeneity. Studies that have previously investigated intratumoral heterogeneity using single cells have either used markers from normal breast tissue to stratify single cells into distinct populations [76] or known protein markers to stratify cells prior to sequencing [77]. However neither of these approaches have allowed for a completely unbiased investigation into intratumoral heterogeneity.

Bioinformatic tools have traditionally been developed to investigate whole tumor expression profiles or profiles generated from a large number of cells. The analysis of single cell RNA-sequencing (scRNA-seq) data poses novel challenges that need to be addressed, and new workflows developed, to allow the proper utilization of this resource. Some of these challenges can be more easily addressed (e.g. through the inclusion of RNA standards to estimate technical variability) while others, such as the identification and removal of confounding factors, require further development [78]. The work in this thesis presents a novel workflow that addresses some of these challenges to allow the unbiased analysis of a scRNA-seq dataset.

1.18 Rationale

While the prognosis of breast cancer patients has seen a vast improvement, and the identification of patient subtypes has allowed for the tailoring of therapy, the response to treatment within patient subtypes still remains heterogeneous. While some efforts have been undertaken to identify drivers of this heterogeneity, the unbiased investigation of this heterogeneity and its functional consequences have not been investigated. This unbiased

investigation will allow for the identification of additional sources of heterogeneity not observed in normal tissues, or previously investigated tumor models.

Through the work presented in this thesis I will investigate two hypotheses regarding sources of tumoral heterogeneity. The first hypothesis is that the heterogeneity of the stroma of triple negative breast tumors is associated with patient prognosis. The second is that the heterogeneity exhibited within tumor epithelial cells of a single tumor has functional consequences, and that these are associated with response to therapy.

While the heterogeneity of triple negative breast cancer has been previously investigated [47,64,65], these studies utilized LCM isolated tumor epithelial profiles or whole tumor expression profiles biased toward epithelial content. The work presented here investigated LCM isolated tumor stromal profiles to identify novel sources of heterogeneity unidentified in these previous efforts.

Similarly, studies of intratumoral heterogeneity have focused on identifying clonal substructures and tumoral lineages through the DNA sequencing of single cells [79]. The analysis of intratumoral heterogeneity through the use of scRNA-seq prevents a novel approach as it allows for the identification of functional differences between clonal populations not easily determined by DNA sequencing.

2. RESULTS

2.1 Identification of interacting stromal properties in triple-negative breast cancer

2.1.1. Confirming tissue specificity of the LCM-derived material

To investigate stromal heterogeneity across TNBC tumors, 57 patient samples were selected based on negative ER, PR, and HER2 status according to clinical-pathological reports (Table 1). Tumor epithelial and non-epithelial (stromal) compartments were separately isolated by laser capture microdissection (LCM) and subjected to microarray-based gene expression profiling [36–38] (See methods). Matched histologically normal epithelial and stromal tissue were also isolated for a subset of cases (n=11).

Following quality control and data normalization, we investigated if there was a difference between the normal and tumor-associated stromal gene expression profiles. Under the hypothesis that the profiles should harbour two distinct patterns of expression corresponding to the normal and tumor-associated stromal components, we selected the most variable genes across all samples (IQR > 2, n=282 genes) as our features for subsequent analysis using the Partitioning Around Medoids (pam) function from the *cluster* package in R [version 2.0.1]. PAM requires a distance measure (correlation distance was chosen) and a specification as to the number of clusters k. We selected k=2 as we expected all samples to fall into two clusters along a single dimension (i.e., normal samples vs tumor-associated samples). These two clusters intuitively correspond to the subset of genes that are more highly expressed in normal (versus tumor-associated) stromal samples and the subset of genes that are more highly expressed in tumor-

No. Patients	57
Number of samples	
Tumor Stroma Samples	57
Normal Stroma Samples	11
Size	
<= 20mm	24
> 20mm	33
mean (mm)	25.17
Standard deviation	13.56
Grade	
Grade 1	0
Grade 2	5
Grade 3	51
Unknown	1
Lymph node status	
Positive	18
Negative	29
Unknown	10
Age	
<= 55	27
> 55	30
Age: mean in years (range)	58.63 (33 - 91)
Outcome	
Total relapse	16
Total relapse free	38
Unknown relapse information	3
Total follow up: mean in months (range)	67.30 (0.0 - 172.2)
Chemotherapy	
Patients receiving Chemotherpy	35
Patients NOT receiving Chemotherapy	8
Unknown	14

Table 1: Description of the patient and tumor characteristics of our cohort

associated (versus normal) samples. To linearly order the patient samples using this set of genes, we ranked samples based on the sum of expression of all the genes in the normal-enriched dataset, and the sum of expression of all genes in the tumor-associated-enriched dataset, after first negating these values. This method does not reweight genes and is thus an unbiased approach for ranking patients. We observed strong differences in expression of these highly variable genes (Wilcoxon test, $p < 0.01$), and that the normal stroma samples ordered separately from their tumor-associated counterparts (Figure 4). This confirmed the success of the LCM procedure in isolating distinct tissue compartments.

2.1.2. Expression profiling of microdissected tissue reveals four stromal properties in TNBC

To establish if differences are observed in TNBC stroma, the most variable genes in TNBC tumor stromal samples (IQR > 2, n=211 genes) were subjected to hierarchical clustering (Ward's algorithm, Pearson correlation distance). Four distinct clusters were observed that contained a significant number of genes with strong pairwise gene-gene correlations of expression (Figure 5A, colors along rows). These clusters, termed the characteristic gene sets, are statistically stable and reproducible (pvclust Approximately Unbiased p-value > 85%), and exhibit strong co-expression across the patient cohort.

To measure the level of expression of the stromal properties in TNBC tumors, patients were linearly ranked based on the overall amount of observed expression of the characteristic

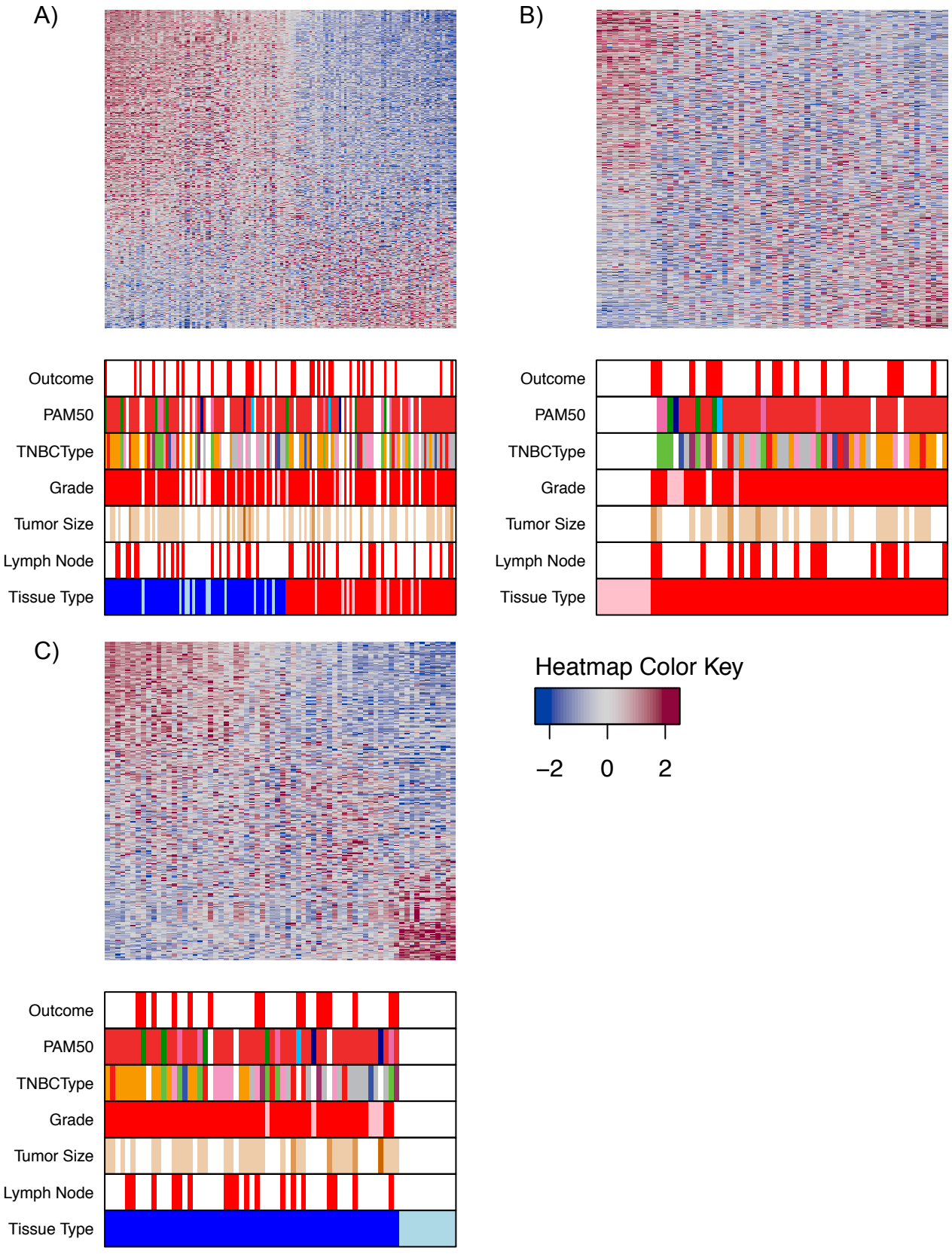


Figure 4: Laser capture microdissection (LCM) successfully isolates distinct compartments of the tumor. Separation of the most variable genes (IQR > 2) unbiasedly into two opposing directions using the Partitioning Around Medoids (pam) function and subsequent ranksum ordering of gene expression profiles distinguishes epithelial from stromal samples (A), and normal from tumor samples (B, C). Pink, light blue, dark blue, and red tissue types represent adjacent normal epithelium, adjacent normal stroma, tumor stroma, and tumor epithelial respectively. Rows represent transcripts and columns represent stromal samples. Values are centered and scaled per transcript across all samples and represented by the color key. Patients with the smallest sum of expression are ranked lowest (right) and those with the largest sum are ranked highest (left).

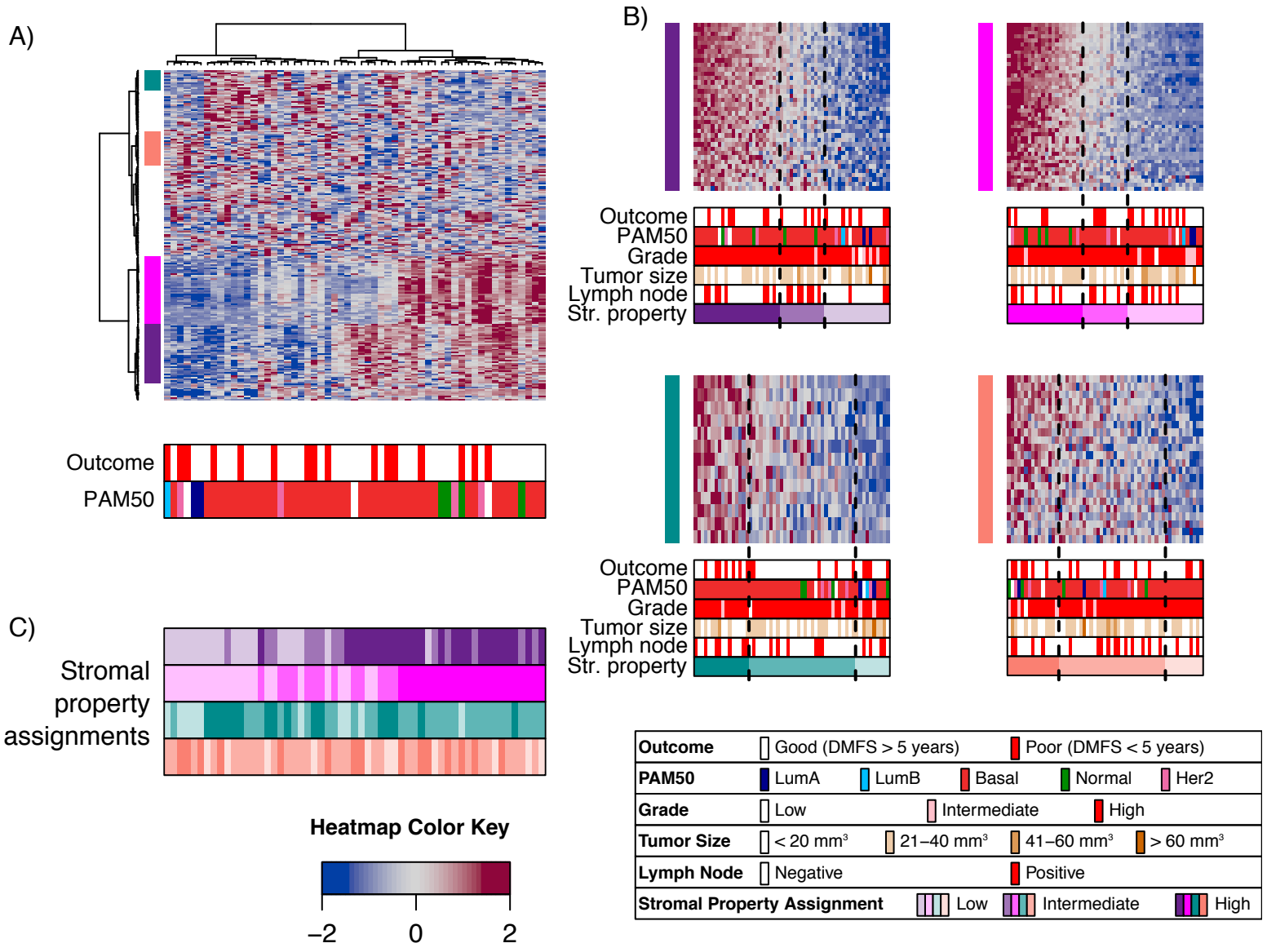


Figure 5: Hierarchical clustering identifies four stromal gene clusters in TNBC samples

A. Hierarchal clustering of tumor stromal gene expression profiles using genes with IQR > 2. Rows, transcripts; columns, samples. Stable clusters with A.U. > 0.85 and > 12 genes are indicated by colored bars at left. Values are centered and scaled per transcript across all samples before clustering. Scaled and centered values are represented by the color key.

B. Assignment of samples into 3 classes (high, intermediate (ROI), or low) for each property using ROI95 (classes demarcated by dashed lines in heatmaps). Patients with the smallest sum of expression are ranked lowest and depicted in lightest color (at right) and those with the largest sum are ranked highest and depicted in the darkest color (at left). Vertical colored bars at left of each heatmap correspond with the color assigned to samples high for that subtype. Rows, transcripts; columns, samples. Values are centered and scaled per transcript across all samples and represented by the color key.

C. Relationships between the assignments for each stromal property. Rows, transcripts; columns, patients as in panel A above. Patient rankings for each cluster are denoted by colors as in panel B. Note that samples can be high for multiple stromal properties.

genes for each stromal property independently (see Methods). A rank-based permutation test (ROI_{95}) [80], was applied to each linear ordering to estimate boundaries of regions that delineate samples that are low, intermediate or high for the characteristic gene set (Figure 5B, black bars). Hence, each patient sample is independently measured for each of the four ternary properties (low, medium, high). This approach differs from traditional subtyping approaches that partition the patient cohort into distinct, non-overlapping subtypes (Figure 5C).

2.1.3. Each stromal property is associated with markers of distinct cell types and processes

To characterize the molecular pathways and presence of specific cell types in each stromal property, we identified differentially expressed genes between patients deemed low versus those deemed high for each stromal property (LIMMA, FDR adjusted $p < 0.05$ after ROI_{95}, Table 2). For the first property, genes differentially expressed include both general (CD2, CD3D, IL-2Rα IL-2Rβ, IL-2Rγ), and lineage-specific (CD4, CD8A, CD8B) T cell-associated markers, as well as markers of a Th1-mediated anti-tumor response including IL-15 [81], granzymes (GZMA, GZMB, GZMK, GZMH) [82], markers of an interferon response (IFI30, IFIT5) [83], transcription factors involved in Th1 differentiation (STAT1, STAT4) [84,85], and TNFα-induced genes (TNFAIP2, TNFAIP8) [86,87]. These genes had greatest expression in patients deemed high for this property (purple, Figure 5B & C).

For the second property, genes differentially expressed between low and high classes (magenta, Figure 5B & C) include B-cell markers (CD19, CD79A, CD72), immunoglobulins (IGLL5, IGLL1, IGJ), and transcription factors associated with B-cell activation (POU2AF1, XBP1). These genes have greatest expression in high expressors of the property.

For the third property, differentially expressed genes (teal, Figure 5B & C) include keratins (KRT6B and KRT23), and metallothioneins. These genes are expressed by tumor epithelial cells [88] and thus may represent invasive tumor cells that have retained some of their epithelial characteristics due to tumor plasticity [89].

For the fourth property, differentially expressed genes (orange, Figure 5B & C) include multiple collagens (collagens 1A1/2, 3A1, 5A1/2, 8A1/2, 10A1, 12A1, 16A1), PDGFRB, FAP, in addition to collagen stabilizing and modifying enzymes (P4HA2, MMP2, LOXL1). All of these are factors associated with a desmoplastic reaction [90,91].

Additional pathway analysis for the first property identified signatures linked to the proliferation of T lymphocytes and activation of cytotoxic T cells, confirming the associations with T-cells (purple, Figure 5B & C). Similarly, pathway analysis for the second property identified signatures of B-cell proliferation (magenta, Figure 5B & C), and analysis of the fourth property identified a signature of desmoid-type fibromatosis [92] (hypergeometric test, $p < 0.05$; orange, Figure 5B & C; see also Methods and Table 3).

Stromal Property	Representative Significant Pathways from Ingenuity Pathway Analysis	Representative Genes	Property
	cell viability of B lymphocytes, quantity of B lymphocytes, differentiation of B lymphocytes, maturation of B lymphocytes	CD79A, POU2AF1, PDK1, PRDM1, TNFRSF13C, TNFRSF17, CD38, CD72, IGHM, IGLL1	B-cells (B)
		KRT6B, KRT23, Metallotheionins	Invasive Epithelial Cells (E)
	quantity of T lymphocytes, T cell development, activation of T lymphocytes, cytotoxicity of leukocytes	CD2, CD3D, IL-2Rα IL-2Rβ, IL-2Rγ, CD4, CD8A, CD8B, GZMBA, GZMB, GZMK, GZMH, STAT1, STAT4, TNFAIP2, TNFAIP8	T-cells (T)
	Hepatic Fibrosis / Hepatic Stellate Cell Activation, Adhesion of connective tissue cells	COL1A1, COL1A2, COL3A1, COL5A1, COL5A2, COL8A1, COL8A2, COL10A1, COL12A1, COL16A1, PDGFRB, FAP, P4HA2, MMP2, LOXL1	Desmoplastic stroma (D)

Table 3: Representative pathways and genes for the four stromal properties

Based on these observations, we labelled the four stromal properties as T (T-cell, purple), B (B-cell, magenta), E (invasive epithelial cells, teal), and D (desmoplastic reaction, orange) respectively.

2.1.4. Stromal properties can be accurately estimated in bulk expression profiles

The limited size of our dataset precludes an investigation into which, if any, of the stromal properties are associated with clinical variables (e.g., tumor size, grade, stage), or outcome information. We note also that the TNBC-restricted focus of our patient cohort greatly reduces the observed variability for many of these variables. For example, although low-grade TNBC tumors do occur, their epidemiologic frequency is low. In order to explore associations between the stromal properties and these variables in larger TNBC cohorts, we required methodology to "translate" signatures of the four properties to bulk expression tissue. The ability to explore bulk expression datasets would also allow us to explore associations between the stromal properties and patient subtyping schemes such as TNBCType.

The approach leverages our previous effort to generate gene expression profile bulk tumor samples for 54 of the 57 TNBC patients studied here [40]. Using the list of differentially expressed genes obtained from contrasting high versus low patient samples in our microdissected stroma-specific data (per property), we linearly ordered the matched patient samples in the bulk expression dataset. Then, the ROI_{95} procedure was used to delineate those regions in the order that correspond to patients with high, intermediate or low status for the property. We observed

statistically significant agreement between the two ternary classifications for all four properties (Kappa test, all $p < 0.01$; Figure 6 and Table 4). When the two classifications disagree, the vast majority estimate an intermediate stromal profile as either a high or low bulk profile, or vice versa (T property: 12/54, B property: 26/54, E property: 11/54, D property: 24/54). Only five disagreements estimated a high profile in stroma to be low in bulk, or vice versa (B property: 3/54, D property: 2/54). Together this suggests that the underlying signals from these stromal processes are conserved and detectable in bulk expression profiles despite their predominantly epithelial content, and that the stromal properties can be used to interrogate bulk expression datasets.

We then interrogated a large cohort of TNBC patient samples selected from a compendium of publicly available breast cancer datasets (n=1098) [40] to investigate potential associations between the stromal properties and clinical, patient, and outcome information. The compendium comprises 13 individual, non-overlapping microarray datasets generated from (non-microdissected) bulk tumor material. Estimation of the status for each stromal property was computed independently per dataset, and pooled across the constituent datasets (Table 5).

2.1.5. Stromal properties are associated with outcome in bulk expression profiles

To test associations between the stromal properties and clinical variables (e.g. tumor size, grade, stage, outcome, etc.) requires a larger cohort of TNBC patient samples. Due to the unavailability of TNBC stromal datasets, we developed and tested a statistical method to

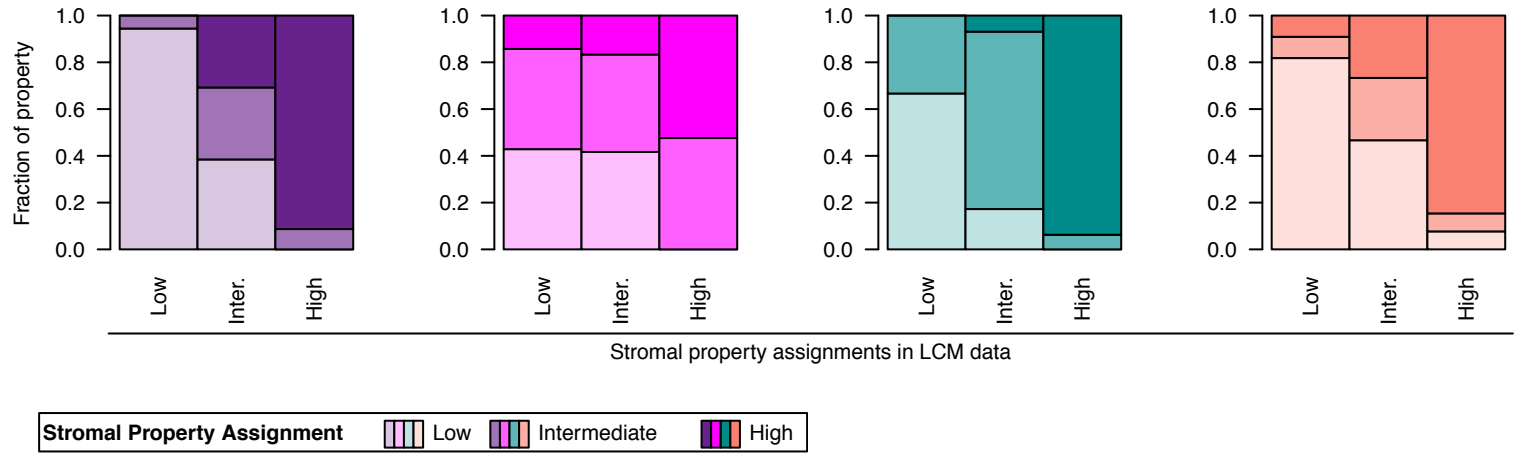


Figure 6: Concordance between assignments of matched patients in LCM stroma and bulk expression datasets. Columns, patient assignments from LCM stroma profiles; bar color density, assignments to low (light), intermediate (medium), or high (dark) from bulk expression profiles. Bar plots represent T (T-cell, purple), B (B-cell, magenta), E (invasive epithelial cells, teal), and D (desmoplastic reaction, orange) stromal properties respectively,

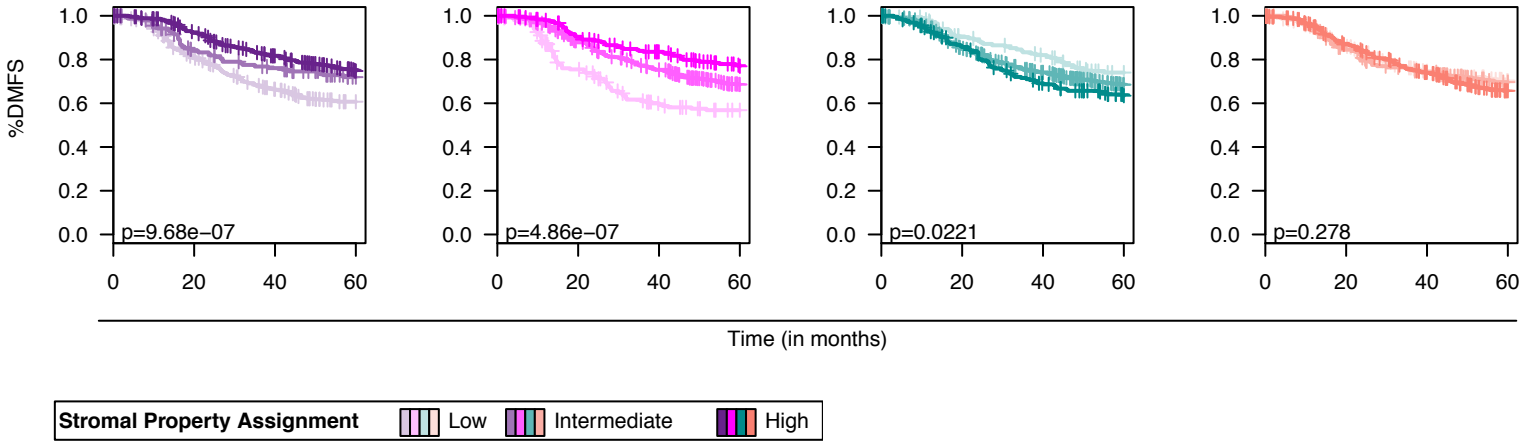


Figure 7: Kaplan-Meier survival analysis of the stromal properties for distant metastasis free survival of TNBC patients in external TNBC bulk expression datasets (n=1,098). Log-rank test p-values are indicated at bottom left for each graph.

T stromal property	Whole tumor low	Whole tumor Intermediate	Whole tumor high	Kappa test p-value
LCM low	17	1	0	1.34E-10
LCM intermediate	5	4	4	
LCM high	0	2	21	
B stromal property	Whole tumor low	Whole tumor Intermediate	Whole tumor high	Kappa test p-value
LCM low	9	9	3	9.54E-03
LCM intermediate	5	5	2	
LCM high	0	10	11	
E stromal property	Whole tumor low	Whole tumor Intermediate	Whole tumor high	Kappa test p-value
LCM low	6	3	0	2.88E-10
LCM intermediate	5	22	2	
LCM high	0	1	15	
D stromal property	Whole tumor low	Whole tumor Intermediate	Whole tumor high	Kappa test p-value
LCM low	9	1	1	5.89E-05
LCM intermediate	14	8	8	
LCM high	1	1	11	

Table 4: Agreement between the ternary classifications in stroma and whole tumor samples for all four properties

estimate the status of each stromal property in bulk expression data. Briefly this method uses the list of differentially expressed genes for each property and the ROI_{95} method to assign samples as low, intermediate, or high. This method, entitled STROMA4, was applied to a large cohort of TNBC patient samples (n=1098) selected from 13 individual non-overlapping publicly available breast cancer datasets [40] (See Methods). Stromal property assignments were computed independently per dataset, and pooled across the constituent datasets (Table 5). This enabled us to test if the low, intermediate and high partitions of each property stratified patients by clinical outcome. While the D property (orange) did not demonstrate significant association with outcome, the T, B, and E properties (purple, magenta, teal) were significantly correlated with outcome (log-rank test, distant metastasis free survival (DMFS) at 5 years all $p < 0.05$, Figure 7). This demonstrates that that the T, B and E properties of the stroma inform on clinical outcome for TNBC patients.

2.1.6. Stromal properties are associated with clinical variables

Stromal property assignments for the compendium were assessed for association with clinical variables. We observed that patients low for the T property tend to have intermediate or low grade tumors (FET, $p < 0.01$); however of the 369 T-low tumors only 24% are of intermediate or low grade. The ternary partitions of the D property are associated with grades I-III, while the partitions of the E property are strongly associated with lymph node status (both Kappa test, $p < 0.01$). Again, although there are significant associations here for the D and E properties with

Stromal Property Assignment	Number of patients	Fraction
T high	460	0.42
T intermediate	168	0.15
T low	470	0.43
B high	260	0.24
B intermediate	620	0.56
B low	218	0.20
E high	239	0.22
E intermediate	675	0.61
E low	184	0.17
D high	444	0.40
D intermediate	206	0.19
D low	448	0.41

Table 5: Summary of assignments of stromal proeprties across our TNBC compendium

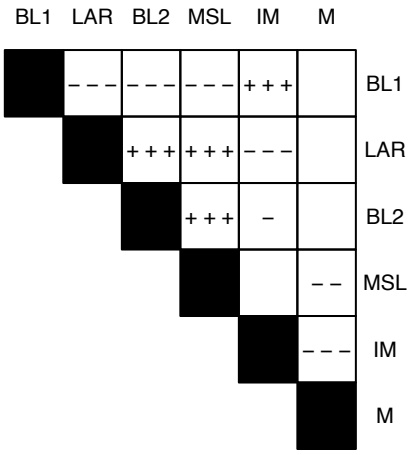
respect to grade and lymph node status, these are not one-to-one relationships. For example, 92% of the 368 D-low tumors are grade III.

2.1.7. Stromal properties succinctly summarize TNBC heterogeneity

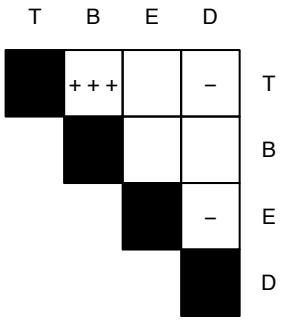
To establish whether the stromal properties are associated with subtypes of TNBC derived from bulk tumor gene expression profiles (TNBCType) [64] we first applied our approach to the TNBCType subtypes. Using the data from Lehmann and colleagues [64], the gene sets that underlie each of the six TNBC subtypes were subjected to our methodology, estimating their activation as either low, intermediate or high across the TNBC compendium (Methods). This procedure places the six 'Lehmann properties' in a format that allows comparison with the four stromal properties.

To examine if any of the Lehmann properties interact (eg whether samples positive for one subtype, are also positive for the second), associations between all possible states of all possible pairs of Lehmann properties were determined. Subtyping schemes partition patient samples into disjoint groups with the assumption that if a patient belongs to one subtype (for which they have the appropriate molecular profile), they do not belong to another subtypes (their profile is sufficiently distinct). However, we observe strong statistical correlations between 11 out of 15 pairs of TNBCType properties (Figure 8A, Kappa test, all p values < 0.01). While the mesenchymal (M) and immunomodulatory (IM) properties display a near perfect ($p < 1e\text{-}10$) anti-correlation that is consistent with distinct subtypes, we also observe an equally strong (anti-

A)



B)



Association

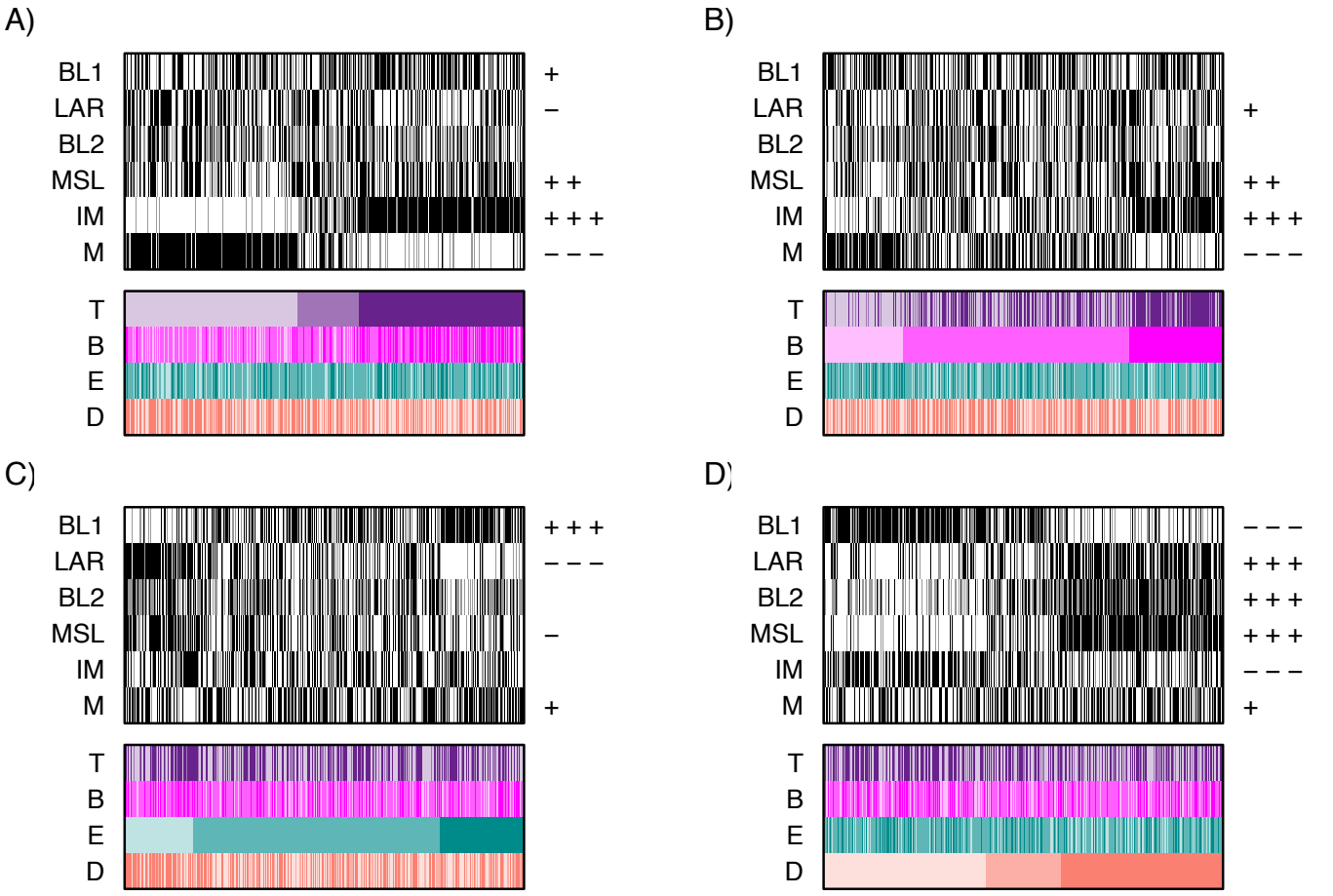
	Positive	Negative
p < 0.01	+	–
p < 1e–05	+ +	– –
p < 1e–10	+ + +	– – –

Figure 8: Agreement within theTNBCType properties and our stromal properties
A. Agreement between the Lehmann properties as determined by the Kappa test
B. Agreement between the stromal properties as determined by the Kappa test

)correlation between the basal-like-2 (BL2), luminal androgen receptor (LAR), and mesenchymal stem-like (MSL) properties. Hence, for example, even though by our method many patients are high for both BL2 and LAR, each will be assigned to only one of these subtypes under the TNBCType approach, This correlation does not support their identification as distinct subtypes.

To establish if the four stromal properties show similar associations we repeated this analysis with the stromal properties (Figure 8B). Of the 6 pairs tested, only the T and B stromal properties were strongly correlated ($p < 1e-10$), while the D property showed a weak anti-correlation ($p < 0.01$) with the T and E properties.

To investigate associations between the Lehmann properties and our TNBC stromal properties we performed a similar analysis comparing the Lehmann and stromal properties (Kappa test, Figure 9A-D). Notably the T stromal property (Figure 9A), and to a lesser extent the B property (Figure 9B), captures the inversely-correlated M and IM properties ($p < 1e-10$). The stromal E property exhibits strong correlation with the BL1 property and anti-correlation with the LAR property ($p < 1e-10$; Figure 9C). Patient samples estimated high for the D property are almost always estimated high for the BL2, LAR, and MSL properties and low for the BL1 property ($p < 1e-10$, Figure 9D). These observations highlight that the 'Lehmann properties' are strongly associated with the stromal properties, and suggests that TNBC heterogeneity can be succinctly summarized by three distinct properties related to immune infiltration (B and T), androgen receptor signaling (E), and a desmoplastic stroma (D).



Association

	Positive	Negative
p < 0.01	+	–
p < 1e–05	+ +	– –
p < 1e–10	+ + +	– – –

Figure 9: Relationships between the properties for TNBCType and our stromal properties Heatmaps depicted summarize ROI95 assignments for each TNBCType property across the TNBC compendium. Samples are colored white, grey and black to represent low, intermediate, and high subtype assignments respectively. Stromal properties are colored as before. Associations are determined by the Kappa test.
A. Patients are ordered by the T property
B. Patients are ordered by the B property
C. Patients are ordered by the E property
D. Patients are ordered by the D property

2.1.8. The D property is the stromal image of tumor proliferation

High expression of the MSL and BL1 Lehmann properties was observed to be negatively and positively correlated with the proliferative index of the tumor respectively [64], and in the preceding subsection we show that our D stromal property is also strongly positively and negatively associated with the MSL and BL1 Lehmann properties respectively, suggesting that the D property reflects the stroma-derived image of tumor proliferation.

We investigated this relationship in three ways. First, samples with expression of the D property had a significantly lower percentage of Ki67-positive tumor cells than samples with low expression of D (two sided t-test $p < 0.05$, Figure 10). Second, using a gene signature of proliferation [93] and the ROI_{95} to estimate proliferative states, we observed a strong statistical association between expression of the D property and expression of the proliferative signature (Figure 11A, B, Kappa test, $p < 0.01$). Third, we observed that low expressors of the D property are associated with higher grade (Figure 12). High-grade tumors tend to have a higher mitotic index [94]. Together these findings indicate that the D property is the stromal image of tumor proliferation from the neighbouring epithelial cells.

2.1.9. Stromal property interactions induce 15 enriched subtypes with larger than expected populations

The many alternative subtyping schemes for breast cancer propose varying numbers of patient partitions that range from just four subtypes via classic approaches based ER, PR and

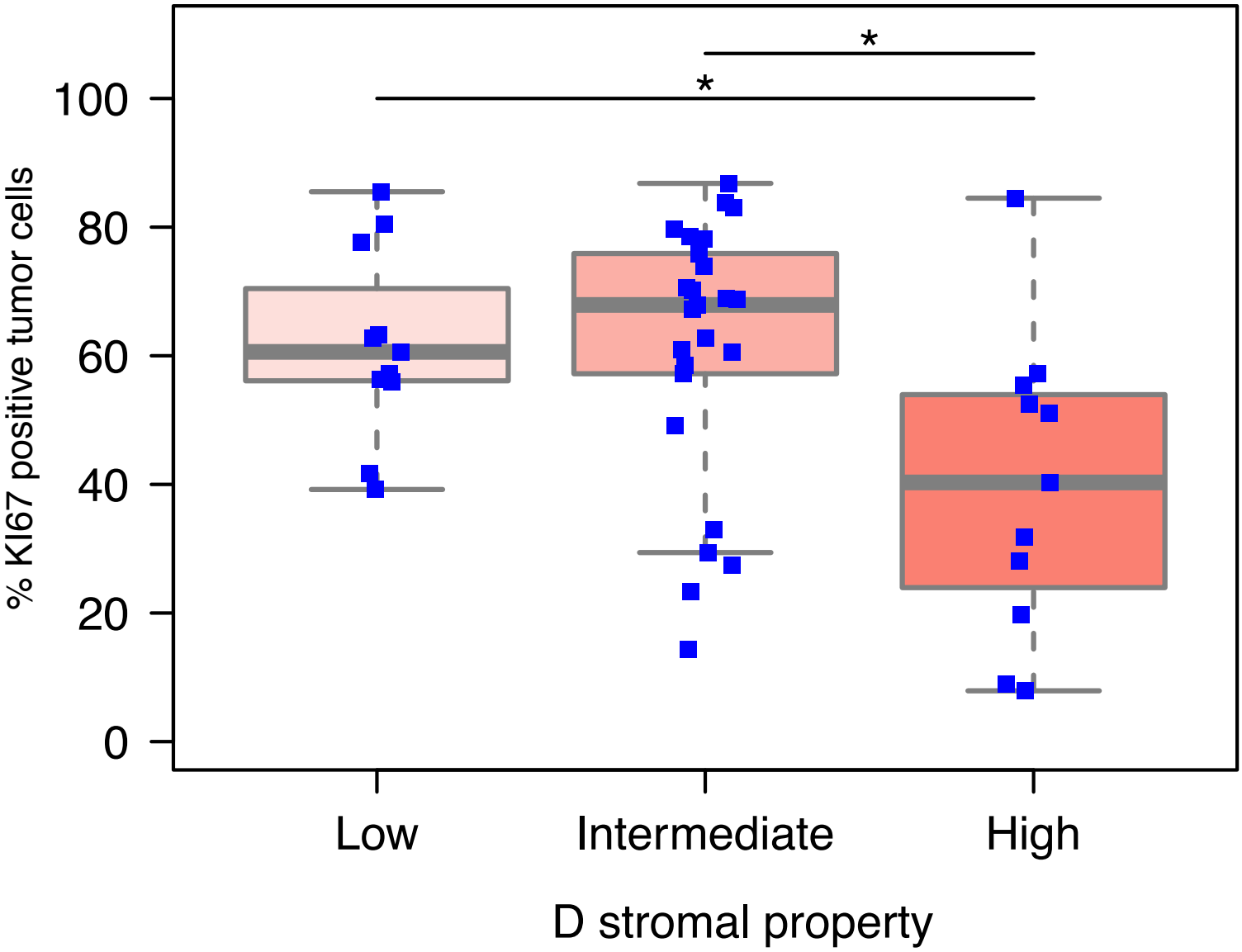


Figure 10: Boxplot showing the association of the D property with Ki-67 staining as a marker for proliferation. * indicates comparisons which are significantly different (two-sided two-sample t-test, $p < 0.05$)

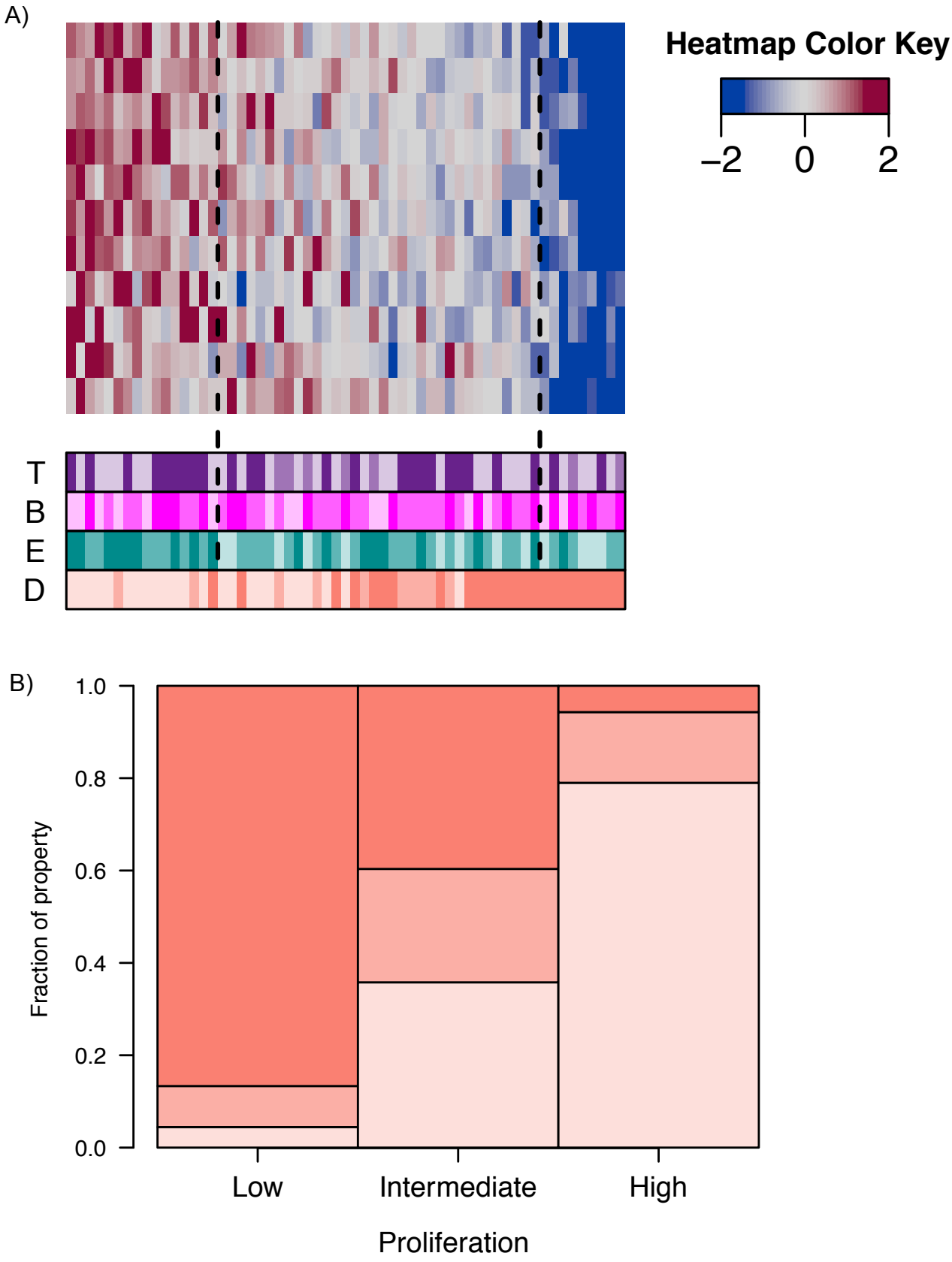


Figure 11: Increased levels of the D property are associated with decreased proliferation
A. Ordering of patient-matched bulk expression dataset using a proliferation signature indicates an association with D status (via ROI95).
B. Barplots indicating the fraction of high, intermediate, and low proliferation patients that are high, intermediate, or low for the D property. Patients were assigned to classes by ROI95 using the proliferation signature (horizontal axis; panel B) and the characteristic gene set for the D property (vertical axis).

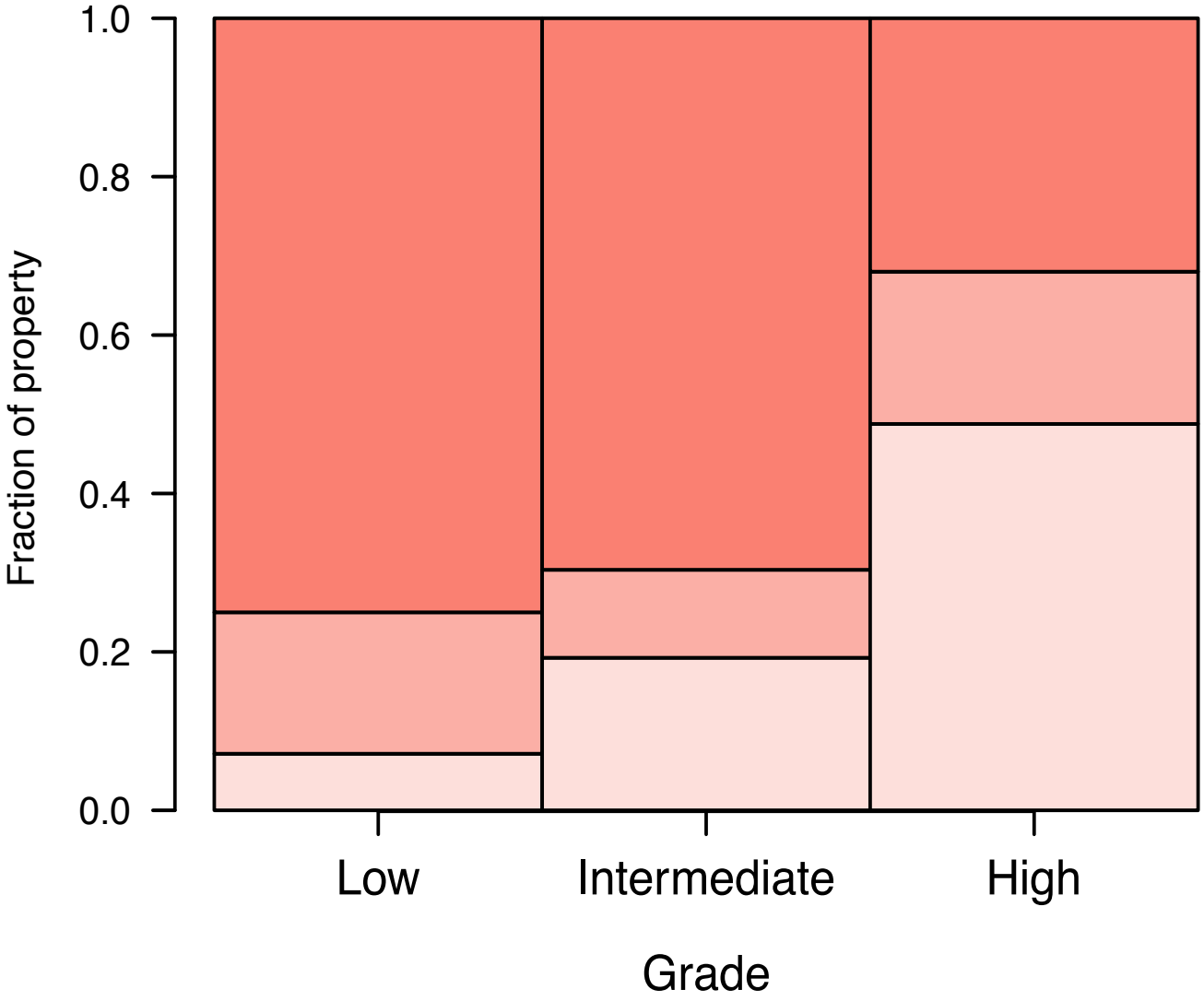


Figure 12: Barplots indicating the fractions of patients that are low, intermediate or high for the D property with respect to grade.

HER2 status to 10 subtypes identified by IntClust through joint DNA and mRNA analysis [57]. Similarly, there have been different numbers of "sub-sub" types proposed for TNBC [64,65]. Our four ternary stromal properties suggests that as many as 3^4(=81) such sub-subtypes could exist; each such subtype is described as a combination of the four properties (eg T-high, B-high, E-low, D-int). However, using our compendium of ~1000 TNBC profiles, we observed that the number of patients assigned to each of the 81 subtypes varied significantly, with some subtypes populated by many samples and others with very few.

To measure this statistically, the enrichment or depletion of subtype populations were assessed relative to a background model (binomial-based test, see Methods). Only 15 of the 81 subtypes are significantly enriched beyond levels that would be observed solely by chance across our compendium (Table 6). Many of the 15 enriched subtypes were either of the form T-high, B-high/B-intermediate, or of the form T-low, B-low/B-intermediate, suggesting a strong interaction between the T and B properties. Conversely, 17 of the 81 subtypes were significantly depleted, suggesting selection against some specific combinations of stromal properties (for example, the T-high, B-low/intermediate, D-high, E-high combination), in turn suggesting a complex interaction between these properties. The remaining 49 subtypes were populated as would be expected under a null binomial model.

Subtype	Number of successes	Hypothesized probability of success	Hypothesized number of successes (rounded)	p-value (enrichment)	p-value (depletion)	Enriched/Depleted
T-low,B-Intermediate,D-high,E-low	44	0.016378759	18	1.16E-07	0.999999956	Enriched
T-low,B-low,D-low,E-Intermediate	44	0.021317034	23	7.65E-05	0.999961767	Enriched
T-low,B-low,D-Intermediate,E-high	13	0.003470644	4	0.000166983	0.999955949	Enriched
T-high,B-high,D-Intermediate,E-Intermediate	28	0.011441759	13	0.000106353	0.999955615	Enriched
T-high,B-high,D-low,E-Intermediate	49	0.02488305	27	9.45E-05	0.999950454	Enriched
T-low,B-low,D-low,E-high	21	0.007547809	8	0.000140421	0.999948949	Enriched
T-low,B-low,D-high,E-low	15	0.005758983	6	0.002220861	0.99915791	Enriched
T-high,B-high,D-high,E-low	16	0.006722373	7	0.003848074	0.998401694	Enriched
T-high,B-high,D-low,E-low	16	0.006782935	7	0.004183773	0.998247347	Enriched
T-high,B-high,D-high,E-Intermediate	42	0.02466088	27	0.004215117	0.997465284	Enriched
T-low,B-low,D-Intermediate,E-Intermediate	20	0.009802029	11	0.007179468	0.996493291	Enriched
T-low,B-low,D-high,E-high	16	0.007480418	8	0.010072303	0.995370615	Enriched
T-Intermediate,B-Intermediate,D-high,E-Intermediate	36	0.021477234	24	0.009602033	0.994171835	Enriched
T-high,B-high,D-low,E-high	17	0.008810443	10	0.020139737	0.989756745	Enriched
T-high,B-Intermediate,D-low,E-high	33	0.021009517	23	0.028549504	0.981661295	Enriched
T-high,B-high,D-Intermediate,E-high	8	0.00405123	4	0.081971238	0.962412399	
T-Intermediate,B-low,D-Intermediate,E-Intermediate	7	0.003503704	4	0.094992225	0.95774673	
T-low,B-low,D-high,E-Intermediate	31	0.021126703	23	0.067492574	0.954032293	
T-low,B-Intermediate,D-high,E-Intermediate	78	0.06008512	66	0.07428032	0.94121958	
T-high,B-high,D-Intermediate,E-low	6	0.003118939	3	0.132268612	0.940662506	
T-Intermediate,B-Intermediate,D-Intermediate,E-low	5	0.002716293	3	0.181571748	0.918110553	
T-Intermediate,B-Intermediate,D-low,E-high	12	0.007673041	8	0.144183169	0.914350637	
T-high,B-Intermediate,D-low,E-Intermediate	75	0.059336503	65	0.117569128	0.904801867	
T-high,B-high,D-high,E-high	13	0.008731778	10	0.170127314	0.89360361	
T-Intermediate,B-Intermediate,D-high,E-low	9	0.005854535	6	0.199437007	0.8840836	
T-Intermediate,B-low,D-low,E-Intermediate	11	0.007619706	8	0.221398293	0.860654927	
T-high,B-Intermediate,D-Intermediate,E-high	13	0.009660626	11	0.268493073	0.817277715	
T-high,B-low,D-low,E-high	10	0.007387217	8	0.2968454	0.805415659	
T-Intermediate,B-Intermediate,D-Intermediate,E-high	5	0.003528229	4	0.34651877	0.804802026	
T-Intermediate,B-high,D-high,E-high	4	0.003188997	4	0.463867823	0.725308773	
T-Intermediate,B-high,D-high,E-Intermediate	11	0.009006582	10	0.403040645	0.709869544	
T-Intermediate,B-high,D-low,E-low	3	0.002477246	3	0.5114835	0.709751411	
T-low,B-Intermediate,D-Intermediate,E-high	12	0.00987064	11	0.40130022	0.706913803	
T-Intermediate,B-high,D-Intermediate,E-Intermediate	5	0.00417873	5	0.484791226	0.687935497	
T-Intermediate,B-low,D-Intermediate,E-high	1	0.001240571	1	0.744106081	0.604891928	
T-Intermediate,B-Intermediate,D-Intermediate,E-Intermediate	11	0.009964663	11	0.533639267	0.586322097	
T-low,B-low,D-low,E-low	6	0.005810866	6	0.613926481	0.545274277	
T-high,B-Intermediate,D-high,E-low	17	0.016030274	18	0.590148909	0.505870875	
T-Intermediate,B-Intermediate,D-low,E-Intermediate	23	0.021670723	24	0.59374123	0.488884686	
T-low,B-low,D-Intermediate,E-low	2	0.002671961	3	0.791148954	0.437895132	
T-Intermediate,B-low,D-low,E-high	2	0.00269794	3	0.795561259	0.431385696	
T-Intermediate,B-Intermediate,D-high,E-high	7	0.007604532	8	0.728455604	0.404607381	
T-low,B-Intermediate,D-Intermediate,E-Intermediate	28	0.02787733	31	0.708935294	0.358755305	
T-Intermediate,B-low,D-Intermediate,E-low	0	0.000955084	1	1	0.35022374	
T-Intermediate,B-low,D-high,E-low	1	0.00205853	2	0.895920015	0.339813699	
T-Intermediate,B-low,D-low,E-low	1	0.002077075	2	0.898022241	0.335035244	
T-high,B-Intermediate,D-Intermediate,E-Intermediate	27	0.027284196	30	0.733435418	0.332922905	
T-Intermediate,B-high,D-Intermediate,E-low	0	0.001139091	1	1	0.286094103	
T-low,B-Intermediate,D-Intermediate,E-low	6	0.007599154	8	0.839295464	0.272203682	
T-Intermediate,B-high,D-high,E-low	1	0.002455128	3	0.932730253	0.249057136	
T-Intermediate,B-Intermediate,D-low,E-low	4	0.005907279	6	0.887911417	0.224392326	
T-high,B-low,D-Intermediate,E-low	1	0.00261511	3	0.943592444	0.218800387	
T-Intermediate,B-low,D-high,E-high	1	0.002673852	3	0.947124809	0.208527069	
T-Intermediate,B-high,D-Intermediate,E-high	0	0.00147958	2	1	0.196757821	
T-low,B-Intermediate,D-low,E-Intermediate	59	0.060626427	67	0.846508841	0.18664233	
T-high,B-Intermediate,D-low,E-low	13	0.016174691	18	0.900997575	0.153136231	
T-low,B-high,D-high,E-high	6	0.008921599	10	0.925684333	0.142383713	
T-low,B-high,D-Intermediate,E-low	1	0.003186742	3	0.969942099	0.135568008	
T-Intermediate,B-high,D-low,E-high	1	0.003217727	4	0.970950692	0.13201369	
T-Intermediate,B-high,D-low,E-Intermediate	6	0.009087722	10	0.93295501	0.130361597	
T-high,B-low,D-Intermediate,E-high	1	0.003396801	4	0.976151554	0.113098859	
T-low,B-Intermediate,D-low,E-high	17	0.021466246	24	0.935686472	0.098781269	
T-high,B-Intermediate,D-Intermediate,E-low	4	0.00743747	8	0.962636362	0.089680295	
T-low,B-Intermediate,D-low,E-low	11	0.016526315	18	0.97269202	0.050042459	
T-low,B-high,D-high,E-Intermediate	17	0.025196986	28	0.988954008	0.01950195	Depleted

T-low,B-high,D-high,E-low	2	0.006868512	8	0.995557801	0.01933253	Depleted
T-high,B-Intermediate,D-high,E-Intermediate	48	0.058806713	65	0.988457744	0.01652559	Depleted
T-low,B-Intermediate,D-high,E-high	13	0.021274583	23	0.992898091	0.013891054	Depleted
T-low,B-high,D-Intermediate,E-high	0	0.004139301	5	1	0.010520973	Depleted
T-low,B-high,D-low,E-high	2	0.009001974	10	0.999465084	0.002957237	Depleted
T-high,B-low,D-high,E-high	1	0.00732126	8	0.999686689	0.002850512	Depleted
T-Intermediate,B-low,D-high,E-Intermediate	1	0.007551673	8	0.999757183	0.00227151	Depleted
T-high,B-low,D-high,E-low	0	0.005636451	6	1	0.002016623	Depleted
T-high,B-Intermediate,D-high,E-high	10	0.020821932	23	0.999198514	0.001979431	Depleted
T-high,B-low,D-low,E-low	0	0.00568723	6	1	0.001906658	Depleted
T-low,B-high,D-low,E-low	0	0.00693039	8	1	0.000482727	Depleted
T-high,B-low,D-Intermediate,E-Intermediate	1	0.009593475	11	0.999974694	0.000294447	Depleted
T-high,B-low,D-low,E-Intermediate	7	0.02086348	23	0.999973741	9.22E-05	Depleted
T-low,B-high,D-Intermediate,E-Intermediate	1	0.011690493	13	0.999997531	3.45E-05	Depleted
T-high,B-low,D-high,E-Intermediate	4	0.020677199	23	0.999999743	1.56E-06	Depleted
T-low,B-high,D-low,E-Intermediate	5	0.025423986	28	0.999999983	1.00E-07	Depleted

Table 6: Enrichment of the 81 subtypes across our TNBC compendium. Subtypes in pink, green, and white are over-represented, under-represented, or represented at levels expected by chance alone respectively.

2.1.10. The D property is a master controller of the prognostic role of the T, B and E properties

Above we have investigated the prognostic capacity (DMFS at 5 yrs) of the stromal properties individually, of which the B, T and E properties were statistically significant. To establish if the interactions between the stromal properties have prognostic capacity in the TNBC compendium we focused on the 15 stromal subtypes that had larger than expected populations. We observe that all subtypes where D is high, were not significantly different in terms of the prognostic capacity (Figure 13A, log-rank test, $p > 0.05$). However, when D is low or intermediate, the spectrum of subtypes (varying states of B, T and E) do have significantly different survival characteristics (Figure 13B; log-rank, $p < 0.05$).

To verify our observation that the D property controls the prognostic capacity of the B, T, and E properties we performed univariate survival analyses of these three properties in D-high and D-low(/intermediate) patient cohorts. Patients were first stratified into a D-high or D-low cohort based on their expression of the D stromal property and then univariate survival analysis for the B, T, and E properties was performed as before.

2.1.11. The D property and the inherent prognostic difficulty of some patients

The vast majority of previously reported prognostic gene signatures for breast cancer are derived from bulk expression data from the tumor proper. These signatures measure a broad range of tumoral hallmarks and cancer-related processes (e.g., proliferation, genomic instability,

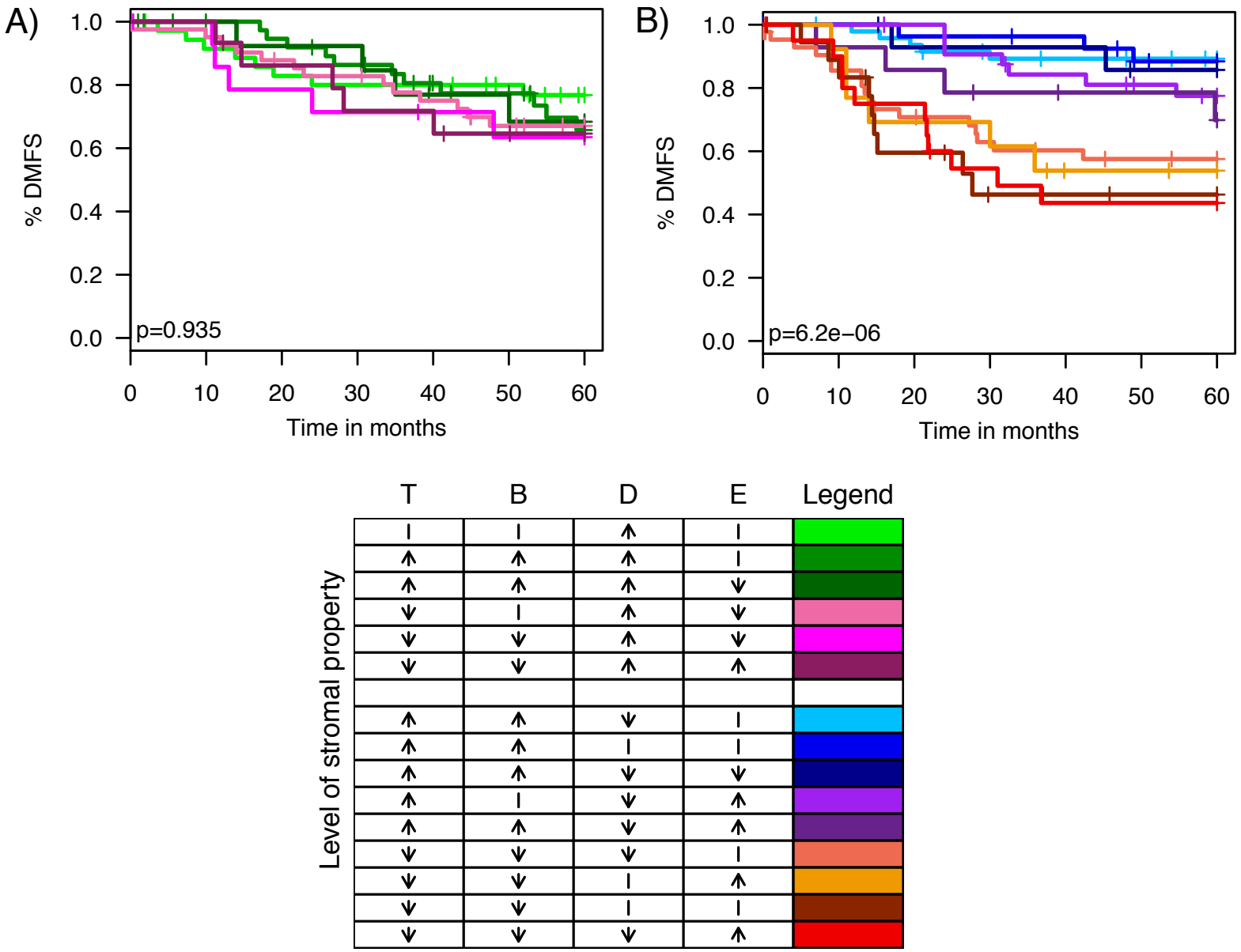


Figure 13: The D property controls the prognostic effect of the B, T, and E properties
A. Kaplan-Meier curves showing lack of prognostic value of B, T or E properties in D-high patients. Only the D-high subset of the 15 overrepresented combined classes is shown. ↑, |, and ↓ represent high, intermediate, and low assignments for stromal properties respectively.
B. Kaplan-Meier curves showing prognostic value of B, T and E properties in D-intermediate and D-low patients. Only the D-intermediate and D-low subsets of the 15 overrepresented combined classes are shown. ↑, |, and ↓ represent high, intermediate, and low assignments for stromal properties respectively.

immune response). In a previous effort [40], we identified a subset of patients whose observed outcome was consistently mispredicted by almost all reported prognostic gene signatures (inherently difficult patients). As our stromal D property appears to be a master controller of the prognostic capacity of the T, B and E properties, we hypothesized that patient samples estimated high for the D property might have higher inherent prognostic difficulty, i.e., gene signatures that predict prognosis will almost always incorrectly predict D-high patients. If this is true, then knowledge of the state of the D property might provide a significant breakthrough that would increase the accuracy of prognostic classifiers.

Using the inherent difficulty score from Tofigh and colleagues [40], we observed a significant difference in inherent difficulty of observed poor-outcome patients contingent on D-high versus D-low or D-intermediate status (two sample, two-sided t-test; $p < 0.05$; Figure 14, DMFS at 5yrs with blue and red representing good and poor outcome respectively). Thus, those poor outcome patients that are high for the D property are systematically mispredicted as good-outcome.

2.1.12. The TNBC stromal properties are generalizable to other patient cohorts

Although the stromal properties were identified within a TNBC cohort, a natural question is to ask if they have prognostic capacity in other breast cancer subtypes. It is well-established that signatures related to cell cycle and tumor proliferation provide prognostic information especially within ER-positive related cohorts [40,95,96], suggesting potential efficacy for the D stromal

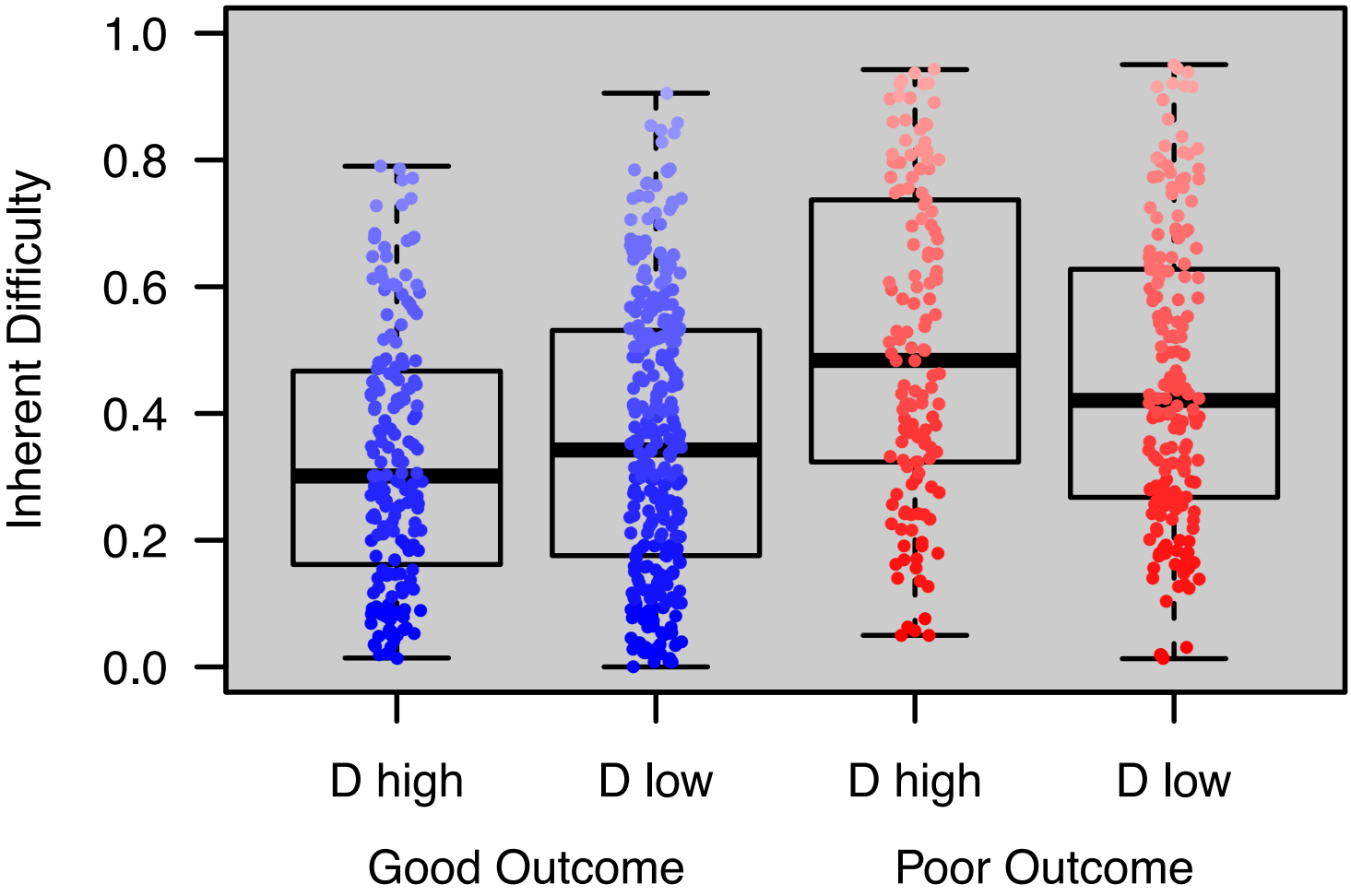


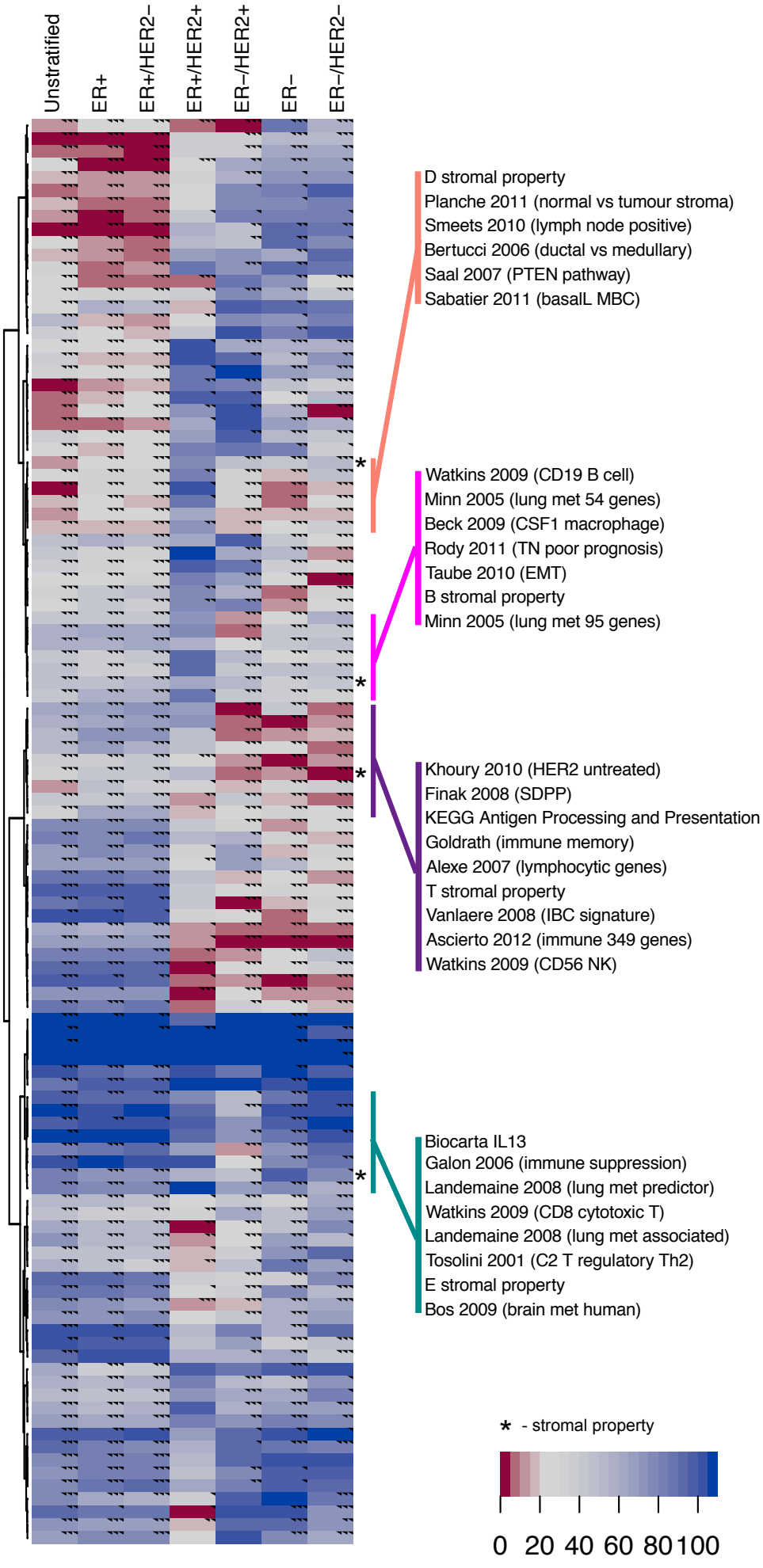
Figure 14: Boxplot showing the association of the D property with difficulty to predict poor prognosis but not good prognosis.

property. The prognostic capacity of immune-related signatures are also well established, particularly within ER-negative cohorts, suggesting efficacy for the T and B properties [40,95,97,98]. We asked if there is evidence that the microenvironmental states captured by our TNBC stromal properties can be used universally across the disease to predict disease progression.

Following the methodology of Tofigh and colleagues [40], we trained prognostic predictors for each of the stromal properties and compared these predictors with previously reported predictors (n ~ 120), trained in the same manner for each breast cancer subtype, across a large collection (n ~ 5000) of invasive breast cancer profiles (Figure 15). More specifically, for the gene set of each stromal property and each prognostic signature, a Naive Bayes' Classifier was generated under statistical cross-validation within datasets while reserving several complete datasets for additional independent validation. The use of such classifiers allows, for example, the learning procedure to "weight" specific genes within a gene signature as more or less important in predicting patient outcome. DMFS at five years was used as the clinical end point, and performance was measured by the product of accuracy, via standard survival analyses (log-rank test) and via a random sampling based approach.

As perhaps expected, the T predictor was one of the best predictors in the TNBC cohort. However, it was also significantly associated with prognosis in all subtypes except the ER positive/HER2 positive cohort. While the D predictor was also significant in these cohorts, it was among the top predictors of prognosis in unstratified pan-breast cancer analysis. Although the D

Figure 15: Heatmap depicting the subtype-specific performance of prognostic classifiers. Colors are proportional to the rank of the classifier within the specific patient cohort, with red representing the highest- performing classifiers relative to the remaining classifiers. Ticks represent the level of significance of the classifier (log-rank test, $p < 0.05, 0.01, 0.001$, respectively). Stromal property classifiers and the closest adjacent signatures have been highlighted.



predictor is significant in the TNBC subtype ($p < 0.001$), it was not among the highest-performing classifiers. The B predictor is significantly associated with prognosis in the same cohorts as the T predictor, although it never appears amongst the highest performing signatures. Interestingly the E predictor showed a stronger association with prognosis in unstratified, and in ER-positive cohorts, than within the TNBC subtype. Together these results suggest that the stromal properties are nearly universal to breast cancer, and have different prognostic capacities in different subtypes.

2.2 Functional consequences of intra-tumoral heterogeneity in a TNBC tumor

2.2.1. Identification of an index case to study intra-tumoral heterogeneity

To explore the possibility of functional heterogeneity existing in breast cancer and response to treatment, an index case, diagnosed as a hormone receptor-negative, high-grade invasive ductal carcinoma, displaying heterogeneity in histopathology as well as HER2 positivity by immunohistochemistry (IHC) and fluorescence *in situ* hybridization (FISH) , was selected for in-depth analyses. As a renewable source of tissue for functional studies, a patient derived xenograft (PDX) was established at the time of surgery, following a lack of response to neoadjuvant treatment. The PDX preserved histomorphology of the primary tumour, and was concordant for hormone receptor negativity and HER2 status, including HER2 IHC and FISH heterogeneity.

The similarity of the PDX with the primary tumour was confirmed by gene expression profiling, followed by unsupervised hierarchical clustering of the most variable genes (IQR > 2). This revealed that the PDX and patient primary tumour cluster adjacently within a dataset comprising of 428 breast tumours [40] (Figure 16). The tumour and PDX clustered with tumours of the basal-like intrinsic subtype, but subclustered within this subtype with tumours positive for HER2 by immunohistochemistry. Utilising gene expression tests for intrinsic subtype prediction, one of these, PAM50 [99], classified the patient's primary tumor and PDX as basal-like, whereas a more recent single-sample predictor of subtype, the absolute intrinsic molecular subtype (AIMS) [58], classified the primary tumour as HER2E and the PDX as basal-like, supporting the histological heterogeneity of this case.

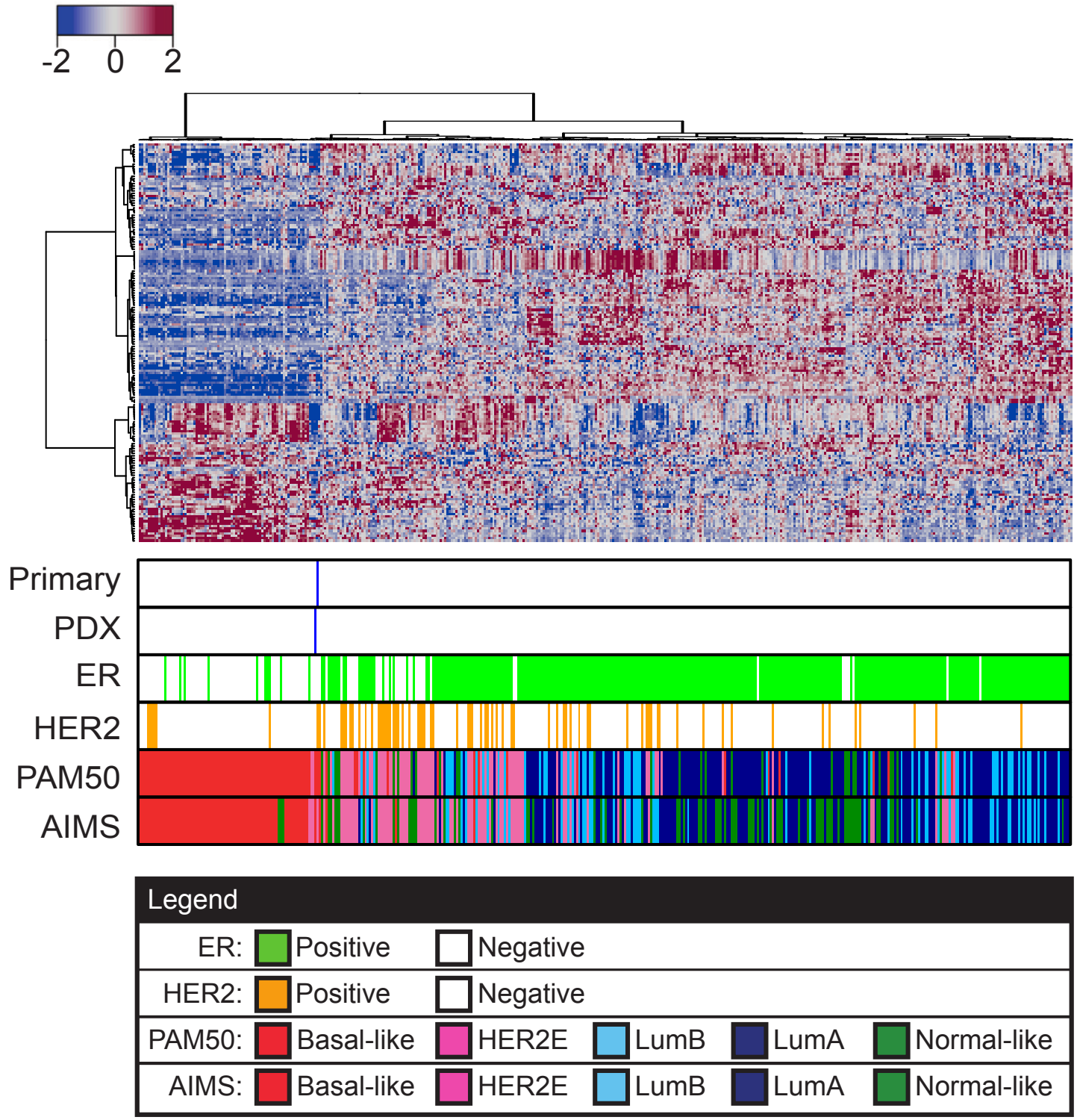


Figure 16: PDX and primary tumor display similar gene expression profiles.
Unsupervised hierarchical clustering of gene expression data from 429 breast tumours using the most variable genes (IQR > 2). Colored bars below indicate specific samples (primary or PDX), ER and HER2 status, and PAM50 and AIMS subtype. Heatmap color intensity represents row z-score.

2.2.2. Single-cell RNA-seq reveals intra-tumour heterogeneity

To uncover the extent of cellular diversity within this tumour and to begin to understand the functional consequences, viable single cells isolated from the PDX were genomically analysed by RNA sequencing (RNA-seq). RNA-seq was performed from 33 viable single cells isolated from the PDX and captured using the Fluidigm C1 autoprep system [100]. Quality control checks verified that the cells were viable at isolation, and that they resembled the bulk tumor. An average of 6,675,705 reads were generated for the single cell RNA-seq (scRNA-seq) dataset, 92.8% of which mapped to the human genome (hg19).

Due to the novelty of the sc-RNASeq technology standard normalization methodology has not yet been developed. A significant challenge present when normalizing sc-RNASeq data is to estimate and remove latent sources of variation to enable for the proper identification of biological signals. Different methods have been suggested to estimate this latent variation, and these methods make use of various factors including using cell cycle stage, library size, and exogenous spike-ins, among others [78].

To determine which of these factors could contribute to the latent sources of variation within our data, we calculated summary values for the library size, number of features detected, and for the controls for each sample within our dataset. These factors were compared to the 1st and 2nd principal components for our data. We observed that the strongest (anti-)correlation was observed between the number of features detected and the1st principal component confirming that this was a large source of latent variation. Additionally technical variability (estimated by

the ERCC spike in controls) and library size also contributed as sources of latent variation (Figure 17).

To mitigate the effects of these sources of variation, a novel method of normalization and analysis was developed based on the LIMMA/voom framework [41]. In addition to log2-scaling the read count values, this framework normalizes for differences in library size when samples are loaded. The effects of the remaining two sources of latent variation were estimated and removed using the removeBatchEffect function. This function fits the summary values for the two latent variables into a linear model for each gene and the residuals of the model returned as the corrected values. These corrected values were used for clustering and for class discovery. For class distinction, the summary values for latent variation were included in the model built to identify differentially expressed genes.

2.2.3. Identification of PDX single-cell subgroups

An exploration of transcriptome heterogeneity was performed by investigating subtype diversity at the cellular level. Two intrinsic subtype predictors, the single-sample AIMS and canonical, PAM50, were applied to the single-cell transcriptome data. Using AIMS, the majority of individual cells were classified as HER2E (59%) or basal-like (31%), with a minor population being classified as the normal-like subtype (10%) (Figure 18). This is in accordance with the classification of whole tumour samples derived from the patient's primary tumour and from the PDX, which were classified by AIMS as HER2E and basal-like respectively (Figure 16).

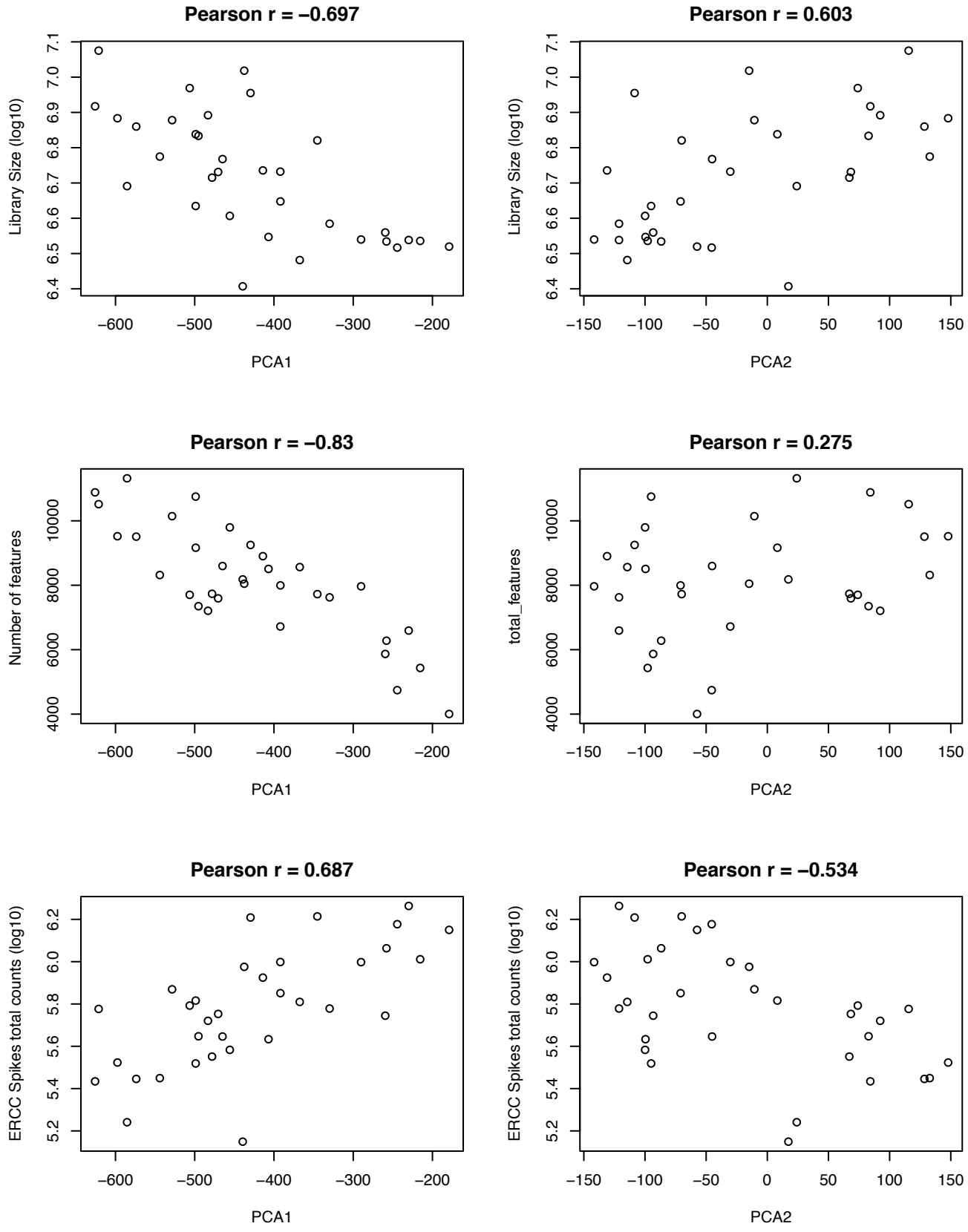


Figure 17: Principal component 1 is strongly (anti-)correlated with library size, number of features detected, and total ERCC spike in counts.

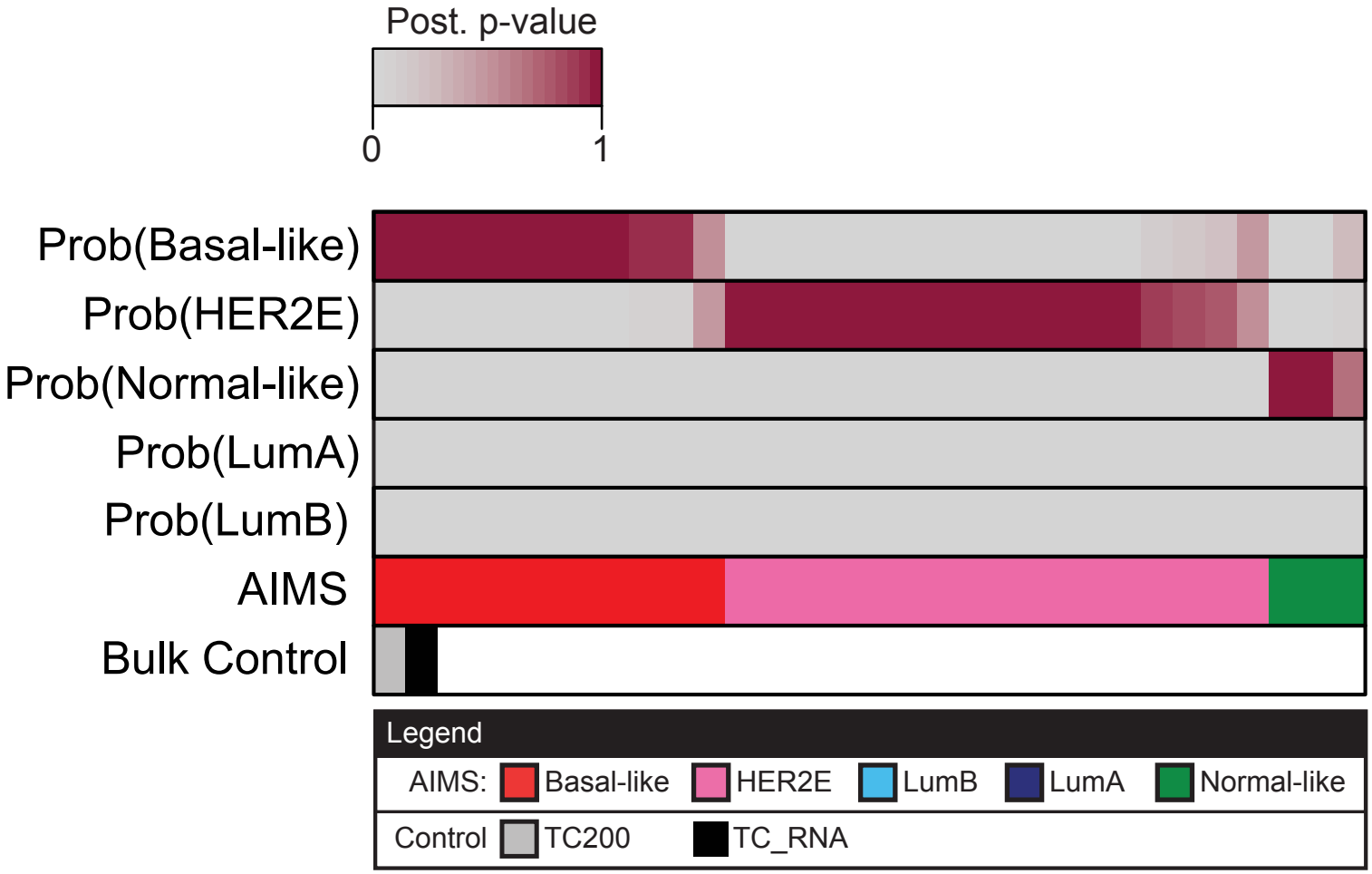


Figure 18: Absolute subtype assignment by AIMS and associated posterior p-values. Higher p-values (darker red) indicates increased likelihood to be of the associated subtype. TC200 and TC_RNA controls were determine to be of Basal-like subtype similar to the whole tumor sample. The majority of individual cells were classified as HER2E (59%) or basal-like (31%), with a minor population being classified as the normal-like subtype (10%)

However, when PAM50 was applied to single-cell transcriptomic data, cells classifications were distributed across all five subtypes (Figure 19). This result was unexpected as cells lacked expression of the hormone receptors (ER or PR) typically associated with the luminal phenotype. It has been observed that the PAM50 method is susceptible to misclassification in unbalanced datasets, where non-luminal samples become classified as a luminal subtype as the composition of the dataset is changed [58]. AIMS does not require the dataset scaling step used by PAM50 [58], supporting the suitability of AIMS as a single-sample predictor for breast cancer single-cell transcriptome data.

To identify if distinct subsets of tumour cells are also identified using unbiased approaches, class discovery using the 200 most variable genes was performed on the single-cell RNA-seq dataset (Figure 20A). This identified that tumour cells isolated from the PDX, segregated into one of two major clusters (Figure 20A, blue and orange).

To investigate differences between these clusters, class distinction was performed and identified a set of 575 genes differentially expressed between the clusters (edgeR FDR adjusted p-value < 0.01) (Figure 21A, Table 7). The epidermal growth factor receptor (EGFR) was among these genes and was a marker for the orange cluster (Fig 21B, Table 7). The heterogeneity of EGFR expression was validated by EGFR immunostaining which confirmed that EGFR was variable and identified distinct cell populations in the primary tumour as well as the PDX.

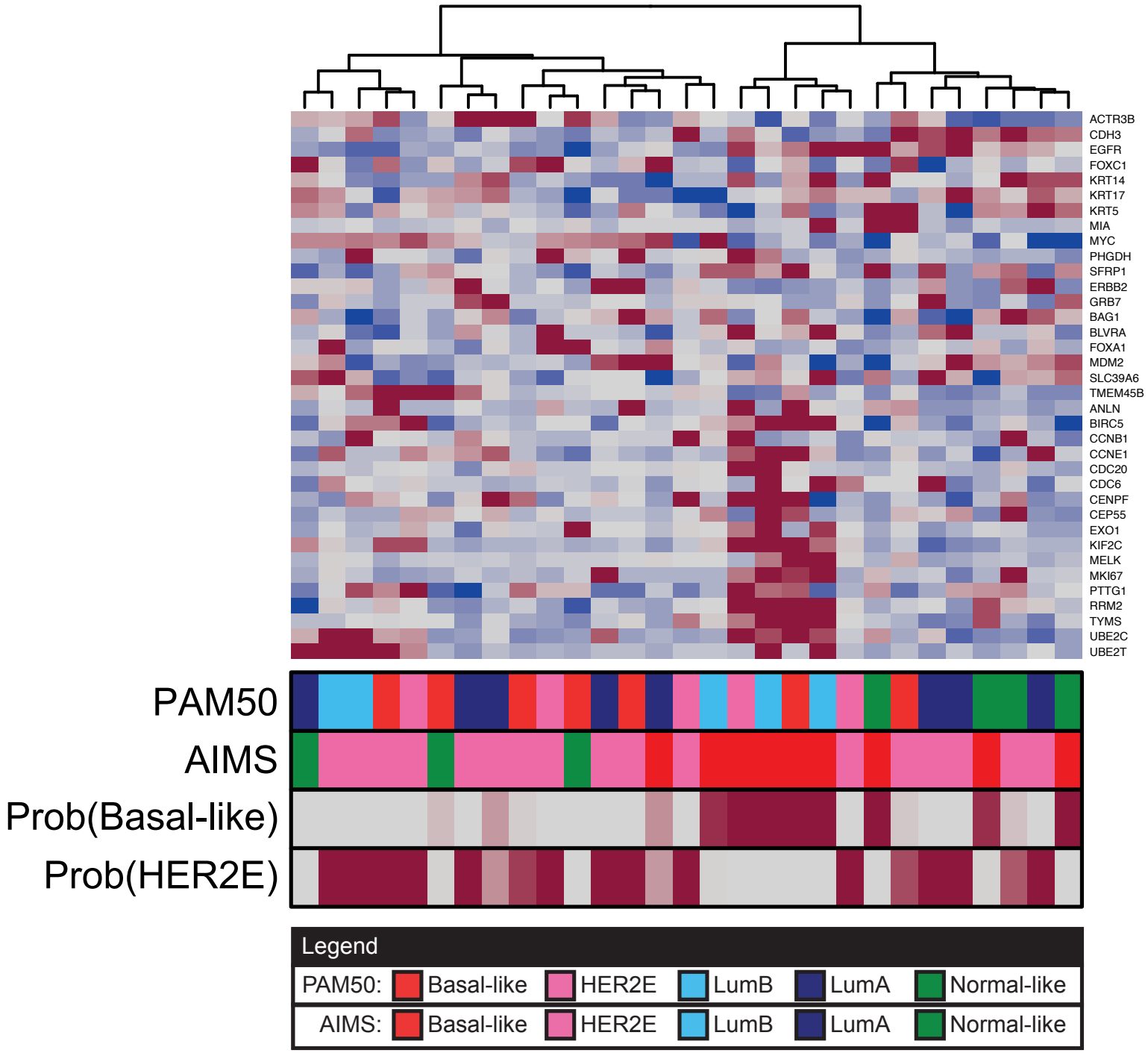


FIgure 19: PAM50 centroid-based assignment of molecular subtypes classifies cells across all 5 subtypes.

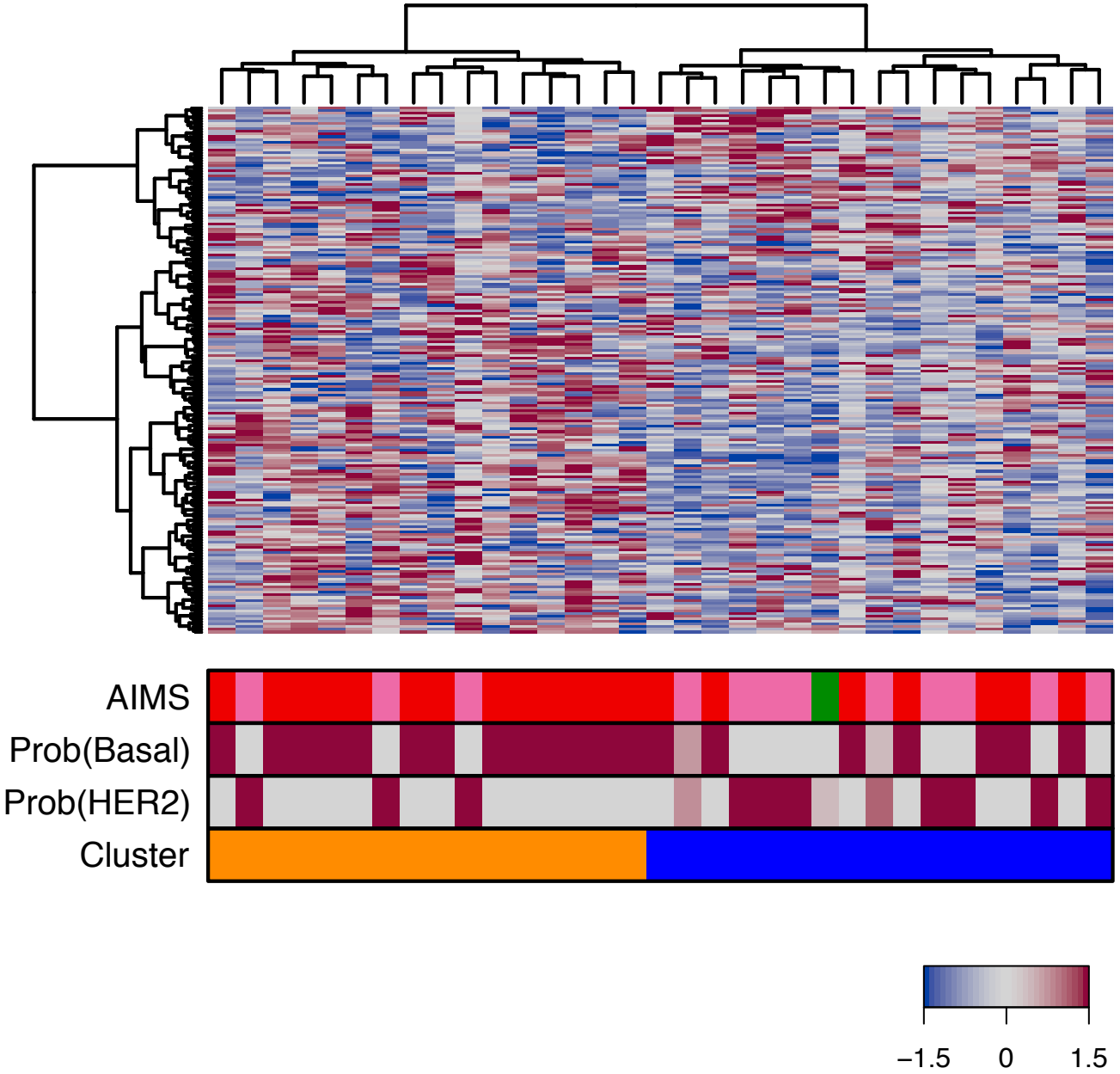


FIgure 20: Class discovery of single-cell RNA-seq data from 33 PDX-derived cells using 200 most variable genes (inter-quartile range). Colored bar below represents the cluster calls for the two subpopulations identified. Heatmap color intensity represents row z-score.

In addition to EGFR, the orange cluster is also enriched in basal-like markers (KRT14, KRT6A) (Table 7). Consistent with this, the majority (13/16) of single cells in the orange cluster were classified as basal-like by AIMS (Figures 20, 21A; Fisher's exact test p-value < 0.05).

To confirm that these observations were not associated with distinct cell cycle phases, enrichment analysis of cell cycle phase signatures were performed. While some associations with these signatures was observed, the significant signatures were all associated with the orange cluster (Figure 22). This indicates that these functional differences are not associated with distinct cell cycle phases, but that they may be affecting cell cycle processes.

2.2.4. Increase in stem cell characteristics among EGFR-high cells

To determine if functional differences exist between the blue and orange clusters an enrichment test was performed on the differentially expressed genes. In addition to genes associated with EGFR trafficking, we observe genes associated with a hypoxic response, and with oxidative phosphorylation (Table 8). These pathways have previously been associated with stem cell populations. Grün and colleagues [101] have proposed a method for identifying putative stem cell populations from single cell transcriptomic data. This method makes use of linkages between clusters and transcriptomic entropy to determine cells with more stem-like characteristics.

Given that we are comparing just two clusters, we do not expect distinct number of links between the clusters. We therefore calculated the transcriptomic entropy for each cell as

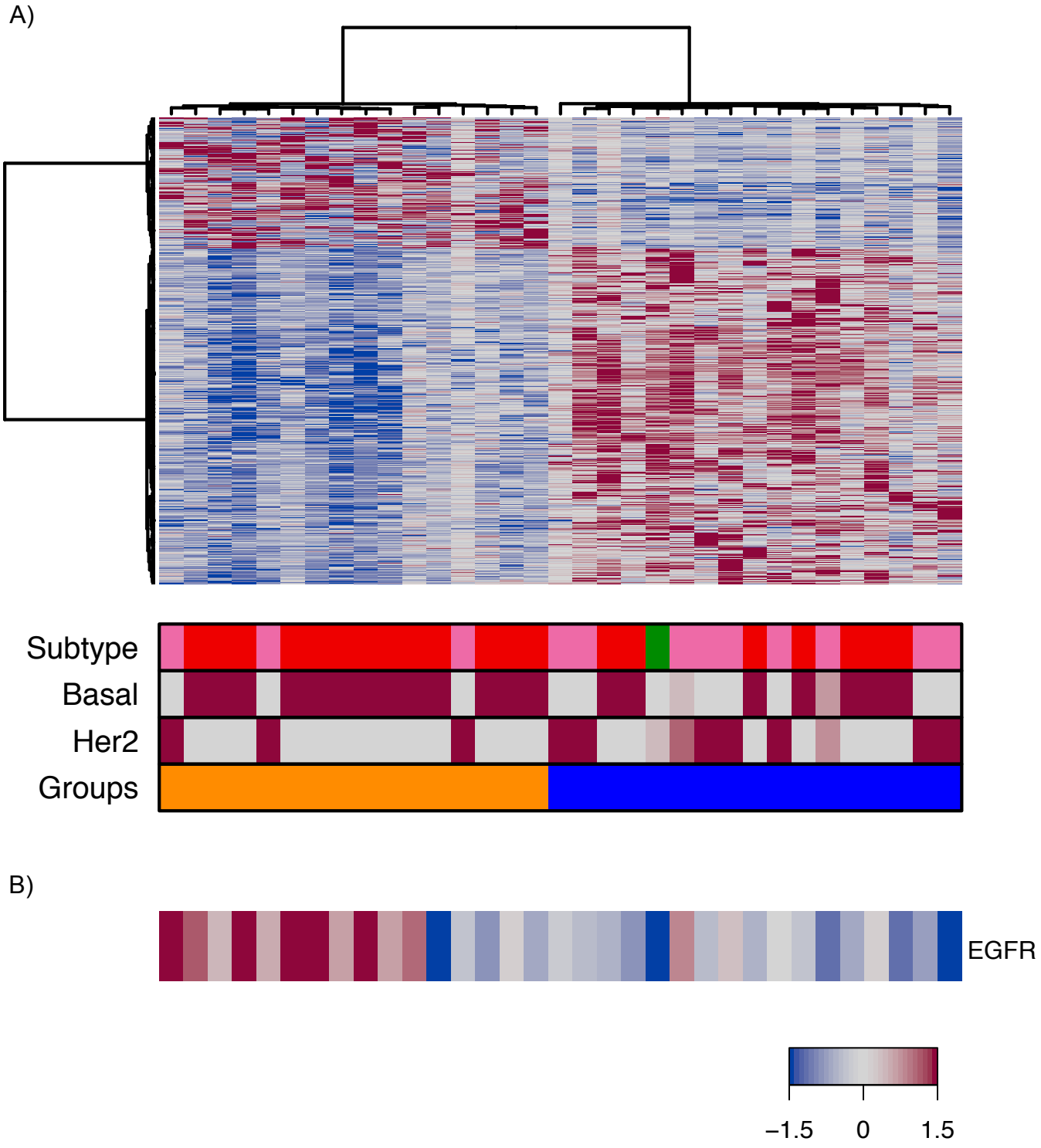


Figure 21: EGFR is significantly different between cell populations.
A. Hierarchical clustering of single-cell RNA-seq dataset using 575 genes differentially expressed between clusters identified in Figure 20 (edgeR FDR adjusted p-value < 0.01). Heatmap color intensity represents row z-score.
B. Subset of heatmap form Figure 21A depicting EGFR expression across clusters.

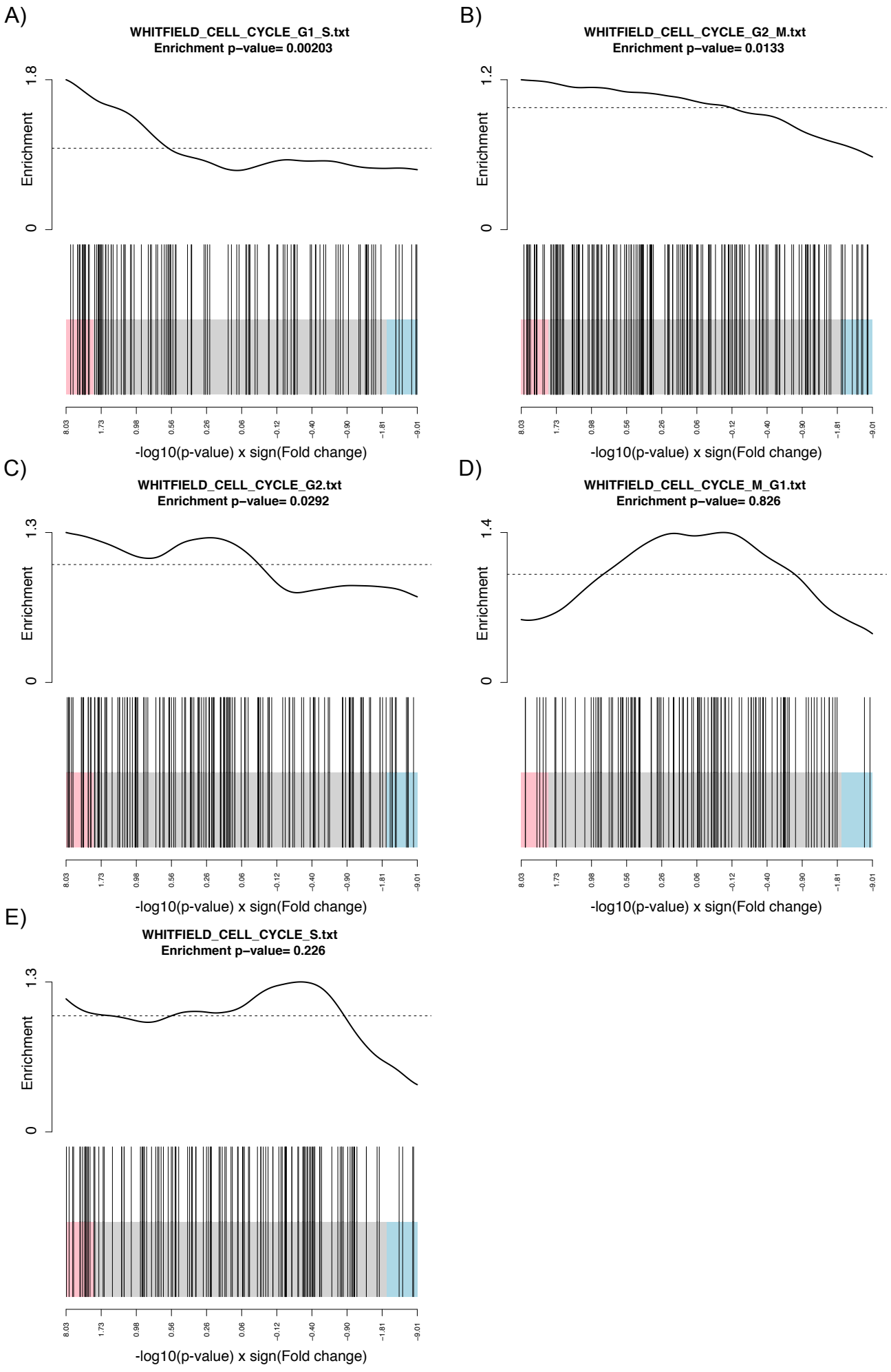


Figure 22: Enrichment of cell cycle stage signatures (A-E) in orange vs blue clusters. Top plot indicates enrichment score. Botttom plot indicates ranks of individual genes from the signature (lines) and association with orange cluster (left) or blue cluster (right)

Gene Set Name	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
HALLMARK_PROTEIN_SECRETION	96	Genes involved in protein secretion pathway.	10	0.1042	5.30E-08	2.65E-06
HALLMARK_ADIPOGENESIS	200	Genes up-regulated during adipocyte differentiation (adipogenesis).	13	0.065	1.57E-07	3.93E-06
HALLMARK_HYPOXIA	200	Genes up-regulated in response to low oxygen levels (hypoxia).	12	0.06	1.10E-06	1.10E-05
HALLMARK_TNFA_SIGNALING_VIA_NFKB	200	Genes regulated by NF-kB in response to TNF [GeneID=7124].	12	0.06	1.10E-06	1.10E-05
HALLMARK_XENOBIOTIC_METABOLISM	200	Genes encoding proteins involved in processing of drugs and other xenobiotics.	12	0.06	1.10E-06	1.10E-05
HALLMARK_INTERFERON_GAMMA_RESPONSE	200	Genes up-regulated in response to IFNG [GeneID=3458].	11	0.055	7.08E-06	5.90E-05
HALLMARK_FATTY_ACID_METABOLISM	158	Genes encoding proteins involved in metabolism of fatty acids.	9	0.057	3.65E-05	2.08E-04
HALLMARK_INFLAMMATORY_RESPONSE	200	Genes defining inflammatory response.	10	0.05	4.16E-05	2.08E-04
HALLMARK_KRAS_SIGNALING_UP	200	Genes up-regulated by KRAS activation.	10	0.05	4.16E-05	2.08E-04
HALLMARK_OXIDATIVE_PHOSPHORYLATION	200	Genes encoding proteins involved in oxidative phosphorylation.	10	0.05	4.16E-05	2.08E-04
HALLMARK_ALLOGRAFT_REJECTION	200	Genes up-regulated during transplant rejection.	9	0.045	2.21E-04	9.20E-04
HALLMARK_HEME_METABOLISM	200	Genes involved in metabolism of heme (a cofactor consisting of iron and porphyrin) and erythroblast differentiation.	9	0.045	2.21E-04	9.20E-04
HALLMARK_PEROXISOME	104	Genes encoding components of peroxisome.	6	0.0577	6.84E-04	2.63E-03
HALLMARK_BILE_ACID_METABOLISM	112	Genes involve in metabolism of bile acids and salts.	6	0.0536	1.01E-03	2.93E-03
HALLMARK_APICAL_JUNCTION	200	Genes encoding components of apical junction complex.	8	0.04	1.05E-03	2.93E-03
HALLMARK_COMPLEMENT	200	Genes encoding components of the complement system, which is part of the innate immune system.	8	0.04	1.05E-03	2.93E-03
HALLMARK_ESTROGEN_RESPONSE_EARLY	200	Genes defining early response to estrogen.	8	0.04	1.05E-03	2.93E-03
HALLMARK_IL2_STAT5_SIGNALING	200	Genes up-regulated by STAT5 in response to IL2 stimulation.	8	0.04	1.05E-03	2.93E-03
HALLMARK_APOPTOSIS	161	Genes mediating programmed cell death (apoptosis) by activation of caspases.	7	0.0435	1.33E-03	3.50E-03
HALLMARK_UV_RESPONSE_DN	144	Genes down-regulated in response to ultraviolet (UV) radiation.	6	0.0417	3.59E-03	8.28E-03
HALLMARK_E2F_TARGETS	200	Genes encoding cell cycle related targets of E2F transcription factors.	7	0.035	4.47E-03	8.28E-03
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION	200	Genes defining epithelial-mesenchymal transition, as in wound healing, fibrosis and metastasis.	7	0.035	4.47E-03	8.28E-03
HALLMARK_ESTROGEN_RESPONSE_LATE	200	Genes defining late response to estrogen.	7	0.035	4.47E-03	8.28E-03
HALLMARK_G2M_CHECKPOINT	200	Genes involved in the G2/M checkpoint, as in progression through the cell division cycle.	7	0.035	4.47E-03	8.28E-03
HALLMARK_GLYCOLYSIS	200	Genes encoding proteins involved in glycolysis and gluconeogenesis.	7	0.035	4.47E-03	8.28E-03
HALLMARK_MITOTIC_SPINDLE	200	Genes important for mitotic spindle assembly.	7	0.035	4.47E-03	8.28E-03
HALLMARK_MTORC1_SIGNALING	200	Genes up-regulated through activation of mTORC1 complex.	7	0.035	4.47E-03	8.28E-03
HALLMARK_UV_RESPONSE_UP	158	Genes up-regulated in response to ultraviolet (UV) radiation.	6	0.038	5.62E-03	1.00E-02
HALLMARK_APICAL_SURFACE	44	Genes encoding proteins over-represented on the apical surface of epithelial cells, e.g., important for cell polarity (apical area).	3	0.0682	1.00E-02	1.73E-02
HALLMARK_SPERMATOGENESIS	135	Genes up-regulated during production of male gametes (sperm), as in spermatogenesis.	5	0.037	1.23E-02	2.05E-02
HALLMARK_ANDROGEN_RESPONSE	101	Genes defining response to androgens.	4	0.0396	1.96E-02	3.16E-02

Table 8: Gene signatures from MSigDB (Hallmark gene sets) that have a significant overlap with the genes differentially expressed between blue and orange clusters

described, and compared the entropy between the two clusters. The transcriptomic entropy was observed to be higher among cells in the blue cluster compared to cells in the orange cluster (two sided t-test $p < 0.05$, Figure 23) leading to the hypothesis that cells in the blue cluster would possess more stem-like characteristics, and that the PDX would respond to EGFR inhibition. The increased stem-like properties of EGFR-high cells and sensitivity of the PDX to EGFR inhibition was subsequently confirmed by several independent assays.

2.2.5. Identification of tumors that respond to EGFR Inhibition

To validate our that our observation was applicable to other tumors as well, we identified four additional PDX models that displayed a similar mosaic pattern of EGFR expression as determined by immunofluorescence staining. Notably, three of these four tumors were also responsive to EGFR inhibition similar to the index tumor. To identify commonalities between these tumors and our index case we performed scRNA-seq on three of these additional tumors.

As was observed with our index case, the library size, number of features, and technical variability were strongly correlated with the first or second principal component for these tumors. Thus these sources of latent variation were assessed and removed as done with our index case. Similar to our index tumor, we identified a subset of cells with high EGFR expression and high transcriptomic entropy (> 0.7). In addition, these cells were enriched for the gene signature identified in the index tumor (Figure 24). We therefore hypothesized that these EGFR-high cells would have stem-like characteristics similar to what was observed in our index case. EGFR-high

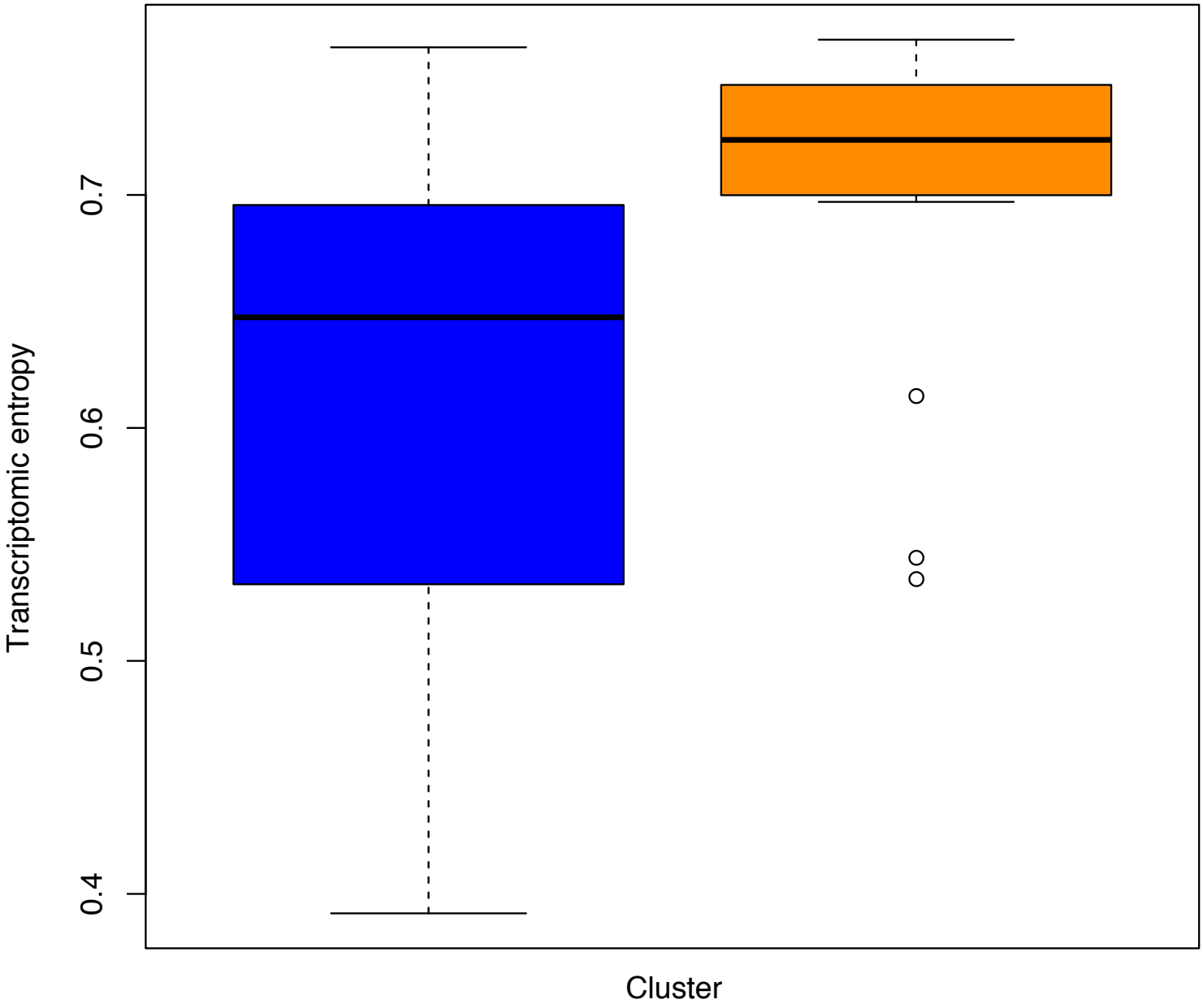


Figure 22: Transcriptomic entropy is significantly different between the clusters (two sided t-test p-value = 0.00757)

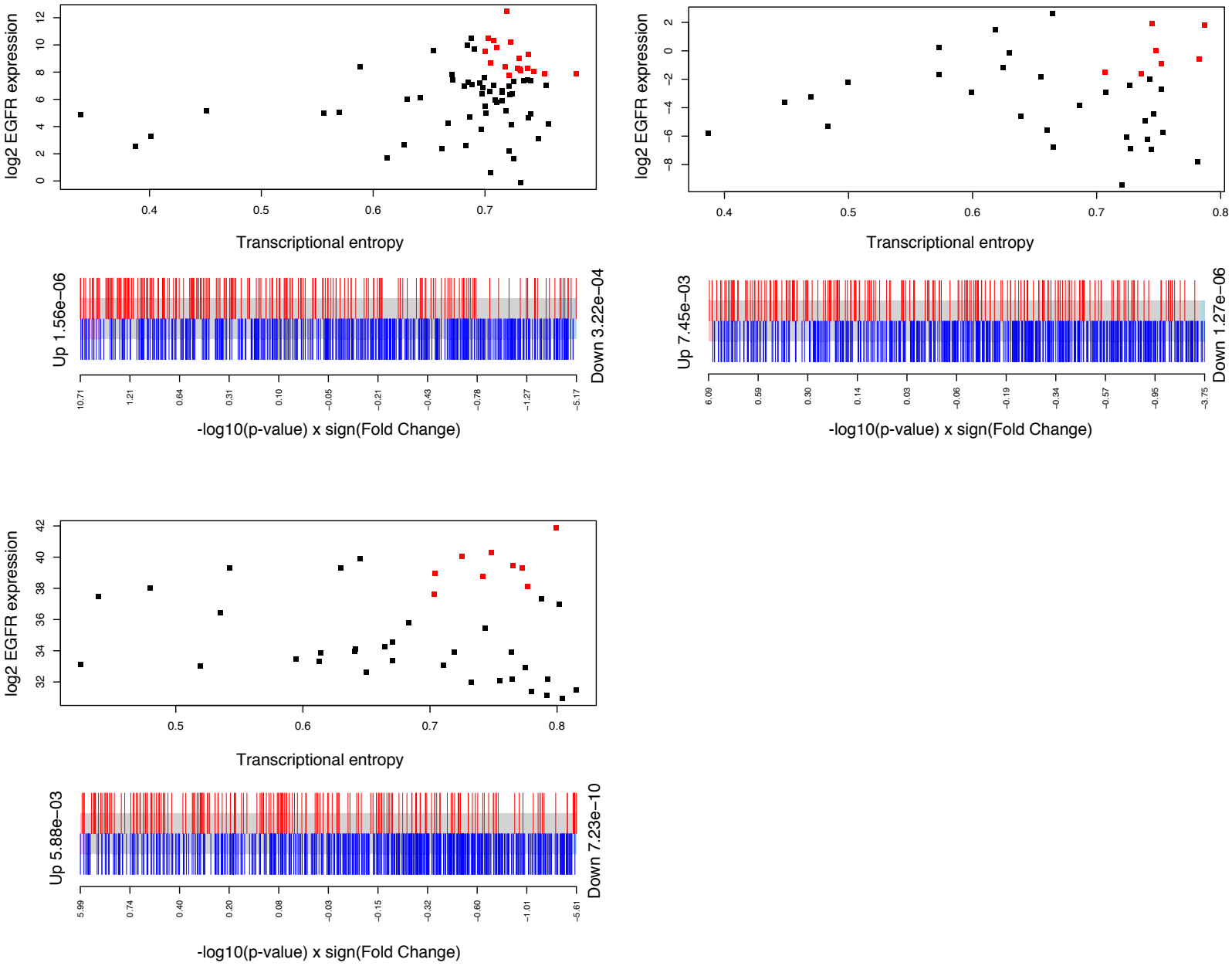


FIgure 24: GCRC1876 (A), HCI001 (B), and BCM2665 (C) PDXs have EGFR high "stem-like" cells that are enriched for the gene signature defined in the index tumor
Top: Identification of EGFR high "stem-like cells" (red)
Bottom: Enrichment of gene signature from index tumor. Up genes (higher expression in orange cluster) are and down genes (higher expression in blue cluster) are denoted by red and blue lines respectively. Enrichment of up and down genes aare indicated next to the graph.

cells isolated from these tumors by FACS were determined to have more stem-like characteristics than EGFR-low cells, similar to the index tumor.

3. DISCUSSION

The poor outcome TNBC subtype remains the focus of much research as we still lack effective classification markers, prognostic signatures, and targeted therapies. This study represents the first large-scale effort to investigate the tumor microenvironment in TNBC patients. With respect to the TNBC subtype, previous studies have focused on gene expression profiling [47,64,65] or DNA sequencing [102] of bulk material enriched for tumor cells. Efforts to study the tumor microenvironment, including our own [36–38,103], have used LCM to isolate stromal elements in a pan-BC fashion, not restricted to TNBC.

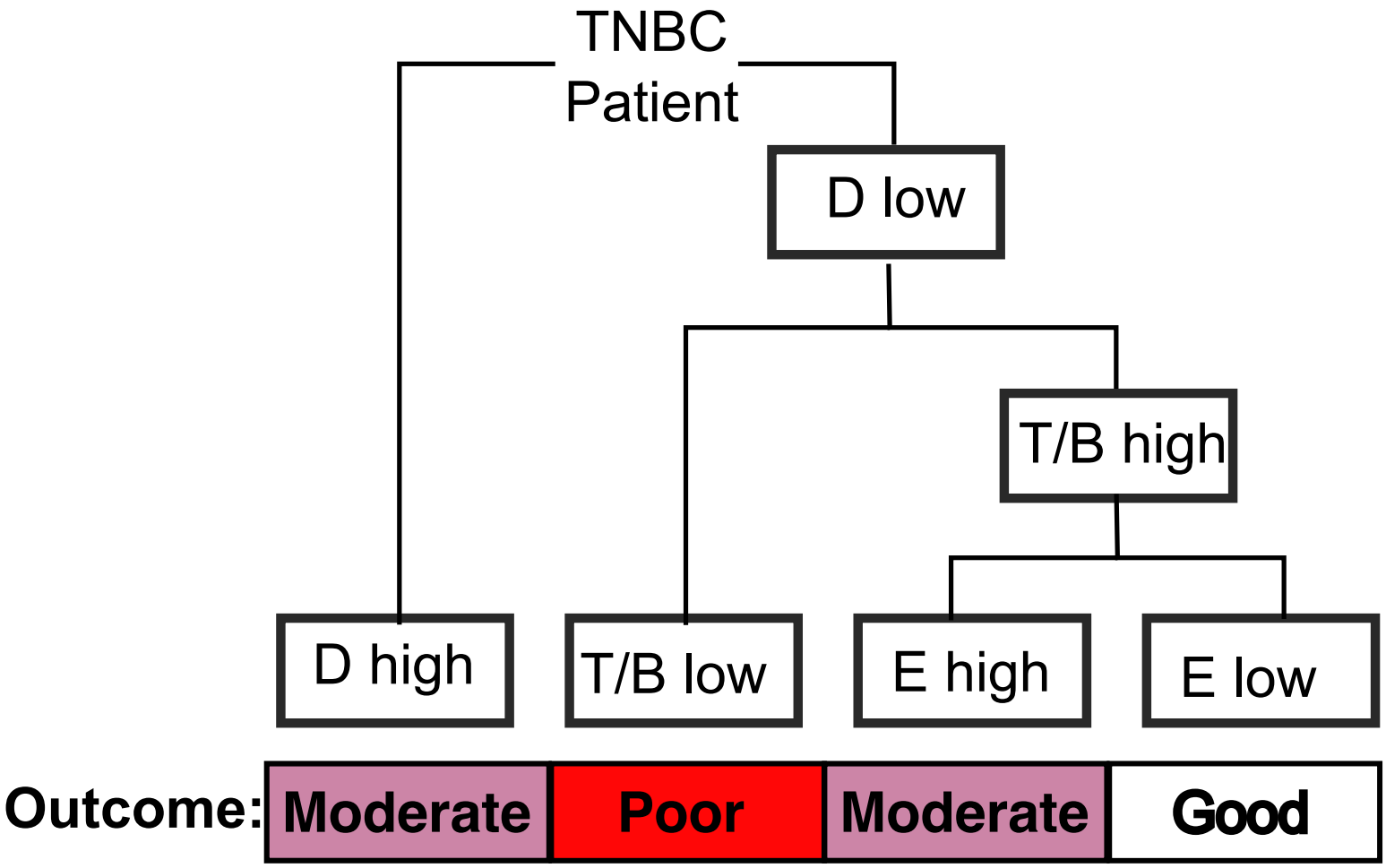
Here, we identify four stromal properties in TNBC patients. A key distinction between this study and previous work is that we allowed patients to express multiple properties, rather than assigning samples to individual distinct subtypes. We observe that this method better captures the heterogeneity of the TNBC stroma. Despite being discovered in LCM-derived material, these stromal properties are shown to hold true even when applied to matching bulk expression sample profiles. This allows us to assign levels of the four properties to a large bulk-derived compendium of TNBC patients, revealing that the activation state of the B, T, and E properties were associated with patient outcome, and that the D property is associated with tumor proliferation.

Using the infrastructure and concepts from Tofigh et al. [40], we show evidence that the vast majority of existing gene signatures for predicting patient prognosis fail for many D-high patient samples. In particular, these signatures often incorrectly predict D-high poor outcome patients to have good outcome. D-high patients are the least proliferative amongst the TNBC, although all

of these tumors are highly proliferative in comparison to non-TNBC tumors. This suggest that the D-high, least proliferative TNBC tumors have been problematic for existing prognostic signatures. Therefore, when our novel D signature identifies a sample as high, current prognostic predictors should not be utilized as they will likely fail. Conversely, it is only when a patient sample is deemed low (or intermediate) for D that the T, B, and E properties provide additional prognostic information. These observations allow us to generate a decision tree that ablates the complexity of having many (81) potential subtypes (Figure 25).

When the D property is high, the immune response within the microenvironment (estimated via the T and B properties), and the E property are insufficient to predict outcome. The desmoplastic stroma as encompassed by the D signature may suppress the tumor-antagonistic effects of a stimulated immune response. This is consistent with the observation that the cohort of samples deemed high for D have moderate to poor prognosis when compared to the entire TNBC cohort. Hence, the prior or concurrent targeting of desmoplastic stroma may enhance the therapeutic benefit of immunomodulatory therapy in this patient cohort.

Using bulk expression profiles, the TNBCType scheme [64] estimates that six subtypes capture the heterogeneity of TNBC patients. By applying our methodology to the genes that define their subtypes, we presented evidence that there is strong evidence that essentially every patient sample belongs to multiple Lehmann et al. subtypes. However, since their methodology assigns each patient to exactly one subtype (e.g., MSL and not LAR), these patients would be treated according to the standard of care for the MSL subtype, and potential anti-androgen



F**igure 25**: Decision tree to subtype TNBC patients based on observed associations with patient prognosis.

therapies suitable for the LAR subtype would be ignored despite the evidence that the patient is also LAR positive. Therefore application of our approach may improve the already positive findings from Lehmann and colleagues [64,67].

Ng et al. [95] suggest that TNBC may be characterized by three or possibly four properties related to androgen signaling, immune infiltration, and a desmoplastic stroma, all of which are captured by our four stromal properties presented here. This effort is the first to offer a quantitative estimate of the number of distinct, populated TNBC subtypes. If there are four core properties of TNBC and each property is measured as high, intermediate or low, then there are a total of 3^4(=81) possible subtypes. We found that 15 of these 81 subtypes had more patients than expected by chance, and 17 subtypes that had significantly fewer subtypes than expected by chance. The remaining 49 subtypes had populations within our datasets that are not significantly different than what we would expect by chance, and therefore larger TNBC cohorts are necessary to investigate their prognostic and predictive values.

We observed that the stromal properties, despite being discovered amongst TNBC patients, were associated with patient prognosis in other patient cohorts. The predictors derived from the stromal properties are observed to cluster closely with other predictors polling similar processes. Specifically the D predictor clusters with a predictor derived from comparing normal and tumor stroma [104], the T predictor clusters with other signatures linked to the presence of immune cells in the tumor [105], and the B predictor clusters amongst other B cell related signatures [47,106]. Interestingly the E property clusters amongst predictors of metastatic potential

(including metastasis to brain [107] and lung *[108]*) and amongst signatures related to immune suppression/pro tumor immune responses [109,110]. This suggests that the E property may represent an early signature in the microenvironment of invasion or metastases. Future work could determine if the molecular events underlying the E property could be therapeutically targeted to prevent metastases. It has been previously observed that signatures measuring proliferation have their best performance in unstratified analyses, likely because there are large concomitant differences in the proliferative indices and rate of poor outcome between ER+ and ER- patients [40]. Given that the D property correlates strongly with tumor proliferation, it is perhaps unsurprising that the D property predictor is one of the best predictors in unstratified analysis.

Our approach has therefore identified four interacting stromal properties that can be combined to subtype TNBC patients. These properties have also displayed broader applicability among other cohorts of breast cancer patients indicating that they play a key role in determining breast cancer outcome.

Although breast tumours are now considered to consist of a heterogeneous mixture of individual cells [102,111,112], the presence of multiple subtypes within breast tumours are still poorly understood. Thus, although whole tumour level data detect dominant transcriptional programs, they do not capture the true diversity of transcriptional subtypes within the tumour. Single cell RNA-sequencing enables us to investigate this heterogeneity. However the novelty of

the technology means that novel methods of normalization and analysis need to be developed to handle the data.

Although most of the single cells in our index tumor were discretely classified by AIMS as a single subtype, several cells have expression patterns consistent with two subtypes, most commonly basal-like and HER2E (Figure 18). This "dual" state may reflect an altered differentiation program or an interconversion between these phenotypic subtypes [113], possibly reflecting a subtype transition. Single cells classified with a dual subtype expression phenotype have been observed in single-cell transcriptomic studies of glioblastoma [114] establishing that transition subtypes may exist in multiple cancers and contribute to intra-tumoural heterogeneity and/or plasticity with distinct biological and clinical responses.

The expression of EGFR has been shown to be elevated in ~50% of basal-like breast cancers [115], which has lead to several efforts to target it. The EGFR monoclonal antibody, cetuximab, demonstrated underwhelming response rates in a TNBC phase II trial [116], despite strong pre-clinical data. Due to the bulk expression profiling used previously to assess EGFR pathway activation, it was unclear at the time whether EGFR heterogeneity was responsible for the lack of clinical efficacy observed in these trials.

This study identified a subset of tumors that exhibited a mosaic pattern of EGFR expression by immunostaining and single cell RNA-seq profiling. Due to its expression on the cell surface, EGFR could also be used as a marker to distinguish cells by FACS and enabled the isolation of live cell populations for use in subsequent functional assays. Investigation of these tumors

indicate that this variation is derived from two distinct cell populations present within these tumours, and that EGFR-high cells possess stem-like characteristics. The EGFR-mosaic subset of tumors also appears to be enriched for tumors that respond to EGFR-inhibition, and presents a putative diagnostic test for identifying patients who would respond to EGFR inhibition.

The heterogeneity of breast tumors and the implications it has for determining treatment and how patients responds to treatment are poorly understood. The goal set out at the beginning of this thesis was to identify additional sources of heterogeneity in tumors, and to determine if they affect clinical outcome in breast cancer patients. To this end we identified two sources of variation amongst ER-negative tumors: one involving heterogeneity amongst tumor epithelial cells within a tumor, a second involving heterogeneity in stromal cell composition between different tumors.

It is recognized that tumor composition is not static, and that it evolves as a tumor progresses. Previous studies have investigated breast cancer heterogeneity using whole genome sequencing of tumor DNA or through assessment of previously identified RNA markers. While the investigation of tumor DNA provides a snapshot of tumor epithelial heterogeneity, it does not indicate the functional consequences of this heterogeneity. The use of previously identified RNA markers from normal breast tissue [76] assumes that the heterogeneity in tumors is also observed among normal breast cells. Similarly studies investigating stromal heterogeneity used known stromal markers to differences in stromal cells.

Our approaches are the first to investigate the functional consequences of intratumoral or stromal heterogeneity in an unbiased manner. The profiling of tumor stromal samples and of single cells allowed us to assess the sources of variation in these samples without making any prior assumptions. This enabled us to identify previously unobserved sources of variation and to link this with response to therapy or patient outcome.

4. EXPERIMENTAL PROCEDURES

4.1 Methods for TNBC stroma analysis

Sample selection

Samples were collected from patients undergoing breast surgeries at the McGill University Health Centre (MUHC) between 1999 and 2012 who provided written, informed consent (MUHC REB protocols SDR-99-780 and SDR-00-966). All tissues were snap-frozen in O.C.T. Tissue-Tek Compound within 30 minutes of removal. Information regarding clinical variables was obtained through review of medical records. Samples used for this cohort (n=57) were reported as negative for ER and HER2 via immunohistochemistry (IHC) (ER and HER2) and/or fluorescence in situ hybridization (HER2). All patients were PR-negative (IHC), with the exception of one case with weak expression. Haematoxylin and eosin (H&E)-stained sections from each sample were evaluated by an attending clinical pathologist with expertise in breast tissue to identify representative areas of tumor and tumor-associated stroma, as well as histologically normal breast epithelium and stroma.

Microarray dataset normalization

R/Bioconductor (vers 3.20; Bioconductor 3.1) [117] was used for most analyses. Normalization was performed using the limma package [41] where loess was applied for dye bias correction, and quantile normalization was used across arrays. Replicates of non-control probes were aggregated by taking their mean value. To investigate technical error introduced in

the LCM/microarray procedure, the stromal and matched adjacent normal sample from a single patient were repeated and found to be highly concordant (>0.8). Replicate expression profiles were then averaged for the remainder of the analysis. The most variable probe was chosen when there were multiple probes for the same transcript.

Discovery of stromal properties

Probes with an interquartile range > 2.0 across all samples were used as features in hierarchical clustering (Ward's algorithm, Pearson correlation distance). Samples were mean-centered and scaled for each transcript across all patients prior to clustering. Pvclust (version 1.3-2) was used to measure cluster stability with 100,000 iterations and we selected clusters containing at least 12 genes with an Approximately Unbiased (AU) value of >85%.

Linear order and assignment of signatures using ROI_{95}

This unbiased approach, which is described in [80], ranks samples based on their expression of a specific gene set. In our case, we use the characteristic genes for each stromal and Lehmann et al. property. This method estimates each patient sample as either low, intermediate, or high using a random resampling technique with 1,000,000 iterations.

Differential Gene Expression and Pathway Analysis

To identify differentially expressed (DE) genes for each stromal property, we fitted a linear model comparing the levels for each stromal property using the R package limma [41] and corrected with Benjamini-Hochberg ($p < 0.05$). DE genes lists were examined using QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) and compared against the Molecular Signatures Database (MSigDB) for pathway analysis.

Assignment of other subtyping schemes to patients across large patient cohort

We used our compendium of 5,901 bulk expression profiles of invasive breast cancer samples from 13 non-overlapping datasets generated on different technologies *(5)*. We define poor outcome as an observed distant metastasis within 5 years of diagnosis (where available) and used ER and HER2 status as reported for each dataset where available. Since many datasets lacked information on PR status, we used ER and HER2 negativity to define TNBC patients. All patients within the compendium were also labeled with intrinsic subtype values via PAM50 [99], and TNBC patients were labeled by TNBCType via the web-based tool [118].

Assignment of stromal properties and Lehmann properties to TNBC whole tumor samples

To assign the stromal properties to TNBC patients across the compendium of bulk expression profiles, we performed ROI_{95} using the lists of differentially expressed genes for each stromal property. Our software to estimate the level of stromal properties in a sample is available as a Bioconductor package entitled STROMA4. STROMA4 was applied to each (of the 13) datasets of our TNBC compendium independently and the three classes (low, intermediate, high) were combined across all of these datasets. Lehmann properties were similarly assigned using the characteristic genelist for each Lehmann subtype from the original publication. Assignments by ROI_{95} were compared to assignments by the TNBCType web-based tool [118]. Only four disagreements for "high" classification by both tools were observed (Table 9) confirming the accuracy of the ROI_{95} assignments.

Statistical Analysis

Cohen's kappa statistic was used to measure agreement between two ROI_{95}-based categorizations into low, intermediate and high (fmsb package vers. 0.5.1). Enrichment analyses were performed via a one-sided Fisher's exact test in R. The minimum p-value between each one-sided test was used to determine if there was significant enrichment or depletion. For variables with more than two levels, multiple tests were performed to determine enrichment for each level individually against all other levels.

Subtype assigned by webtool	ROI_95 Assignment for matching subtype		
	Low	Intermediate	High
BL1	0	0	161
BL2	0	3	84
IM	0	0	184
LAR	0	0	76
MSL	1	0	61
M	0	0	172

Table 9: Concordance between TNBCType and ROI95 Assignments

For comparisons between a ternary variable and a binary variable (for example, lymph node status versus high, intermediate low stromal property), we removed the intermediate category and then used Cohen's kappa for the two binary variables. To determine association with distant metastasis free-survival we used a Cox proportional hazards regression model via the coxph function in R (vers. 2.38) using the three (ordinal) levels estimated by the ROI_{95}.

Testing for enrichment of combinations of stromal properties

To determine if the fraction of observed combinations of stromal properties was higher than expected we used a one-sided binomial test (stats package in R) [117]. The observed number of patients with the combination being tested was used for the number of successes, the total number of patients was used for the number of trials, and the hypothesized probability of success was determined as the product of the fractions for the individual property levels being tested as observed in the TNBC patient compendium. A p-value less than 0.05 was deemed to be significant.

Laser capture microdissection (LCM) and gene expression profiling.

Frozen sections embedded in Tissue-Tek O.C.T. compound (Sakura Finetek) were sectioned at 10 μm thickness, stained using the HistoGene kit (ThermoFisher), assessed by a practicing clinical pathologist with expertise in breast cancer (A.O.) and subjected to laser

capture microdissection on an Arcturus PixCell IIe LCM system to isolate non-epithelial (stromal) compartments of the tumor bed as identified above. All microdissections were performed within three hours of tissue staining. Patient-matched adjacent histologically normal stromal tissue that was at least 2mm outside of the tumor margin was isolated for a subset of patient samples (n=11). Total RNA was extracted from each population of microdissected cells using the Arcturus PicoPure RNA Isolation Kit (ThermoFisher) Following extraction, total RNA yield and quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies). For samples exhibiting distinct 28S and 18S peaks, 100 pg to 5 ng of total RNA were then subjected to two rounds of T7 linear amplification using the Arcturus® RiboAmp® HS PLUS Kit and labeled with Cy3 dye (using the Arcturus® Cy3 Turbo Labeling™ Kit) according to the manufacturer's protocol. Hybridizations were performed using a common reference design. The reference used for all arrays was Universal Human Reference RNA (Stratagene), subjected to two rounds of T7 linear amplification using the Arcturus® RiboAmp® HS PLUS Kit and labeled with Cy5 dye (Arcturus® Cy3 Turbo Labeling™ Kit) according to the manufacturer's protocol. Prior to microarray hybridizations, amplified products were quantified using a spectrophotometer (NanoDrop) and subjected to BioAnalyzer assays for quality control. Agilent Technologies SurePrint G3 Human GE 8x60K Microarrays (Cat#G4851A) were used for all experiments. Amplified RNA (300 ng) was subjected to fragmentation followed by 17 h of hybridization, washing, and scanning on an Agilent G2505C scanner according to the manufacturer's protocol (manual ID #G4140-90050). Cy3-labeled samples were hybridized against Cy5-labeled reference for all arrays. Microarray data were feature extracted using Agilent Feature Extraction Software

(v. 10.7.3.1) with the default parameters. A full description of the patient and tumor characteristics of our cohort is presented in Table 1.

Microarray dataset normalization

R/Bioconductor (vers 3.20; Bioconductor 3.1) [117] was used for most analyses. Normalization was performed using the limma package [41] where loess was applied for dye bias correction, and quantile normalization was used across arrays. Replicates of non-control probes were aggregated by taking their mean value. To investigate technical error introduced in the LCM/microarray procedure, the stromal and matched adjacent normal sample from a single patient were repeated and found to be highly concordant (>0.8). Replicate expression profiles were then averaged for the remainder of the analysis. The most variable probe was chosen when there were multiple probes for the same transcript.

Differential Gene Expression and Pathway Analysis

To identify differentially expressed (DE) genes for each stromal property, we fitted a linear model comparing the levels for each stromal property using the R package limma [41] and corrected with Benjamini-Hochberg ($p < 0.05$). DE genes lists were examined using QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) and compared against the Molecular Signatures Database (MSigDB) for pathway analysis.

Assignment of other subtyping schemes to patients across large patient cohort

We used our compendium of 5,901 bulk expression profiles of invasive breast cancer samples from 13 non-overlapping datasets generated on different technologies [40]. We define poor outcome as an observed distant metastasis within 5 years of diagnosis (where available) and used ER and HER2 status as reported for each dataset where available. Since many datasets lacked information on PR status, we used ER and HER2 negativity to define TNBC patients. All patients within the compendium were also labeled with intrinsic subtype values via PAM50 [99], and TNBC patients were labeled by TNBCType via the web-based tool [118].

Building prognostic predictors from stromal properties

We have previously used Naive Bayes Classifiers (NBCs) to investigate the prognostic capacity of 122 signatures and showed that a subset of these were prognostic in TNBC patients. NBCs may have an advantage over the linear ordering by ROI_{95} as they allow weighting of genes to better reflect an association with prognosis. To determine if the stromal properties could perform as prognostic predictors in addition to being classification predictors we trained an NBC for each stromal property to predict prognosis. The NBCs were trained under leave-one-out cross-validation for the four stromal properties in the TNBC patient cohort within four individual datasets of the compendium, for which there were sufficient numbers of event (distant metastasis within 5 years; poor-outcome) and event-free (good-outcome) individuals.

4.2 Methods for scRNA-seq analysis

Confirming similarity of PDX with primary tumor

The similarity of the PDX with the primary tumour was confirmed by gene expression profiling, followed by unsupervised hierarchical clustering of the most variable genes (IQR > 2).

Normalization and analysis of scRNA-seq datasets

The R statistical framework with Bioconductor (R version 3.20; Bioconductor version 3.1) was used to load the raw read counts, and to normalize and analyze the data for AIMS and class discovery and distinction. AIMS [58] was applied to the raw read count values to identify subtypes for each sample prior to normalization.

To normalize the data, lowly expressed transcripts (reads detected in 3 or fewer cells) were first removed. The number of features, library size, and total counts from the ERCC spike-ins were calculated using the scater package.

The data was log2 transformed and the first and second principal components were estimated using the prcomp function and compared to potential latent variables. The logCPM (log2) values for each sample was calculated using the voom function from the limma package [41]. Differences due to the number of features detected or technical variability (ERCC spike-ins) were modeled out for visualization and class discovery using the removeBatchEffect function from the limma package, and by adding them to the model for class distinction.

Class discovery was performed by clustering the 200 genes with the highest interquartile range. Class distinction was performed on the two resultant clusters using the edgeR package [119]. Enrichment of gene signatures was assessed by a hypergeometric test using signatures from MSigDB (Molecular Signatures Database).

Estimation of transcriptomic entropy

Transcriptomic entropy was estimated as previously described [101]. Briefly, for each transcript the fraction of reads is calculated by dividing the number of reads by the total number of reads in that cell. The entropy for each transcript is calculated by multiplying the fraction of reads by log2 of the fraction of reads and dividing this by log2 of the total number of transcripts. Lastly the sum of all the individual transcript entropies are calculated.

Identification of EGFR-high "stem-like" cells

The additional PDX scRNA-seq datasets were normalized as was the index dataset and the transcriptomic entropy was estimated. A subset of EGFR high "stem-like" cells were identified as cells with the top ⅓ EGFR expression and a transcriptomic entropy > 0.7. Class distinction was performed as before to the EGFR-high stem-like cells and the remaining cells.

Enrichment of the index tumor and cell cycle gene signatures

Genes were ranked from low to high by multiplying the negative of the -log10 p-value by the sign of the fold change. Enrichment of the signatures were then assessed using the barcodeplot and geneSetTest functions from the LIMMA package [41].

5. BIBLIOGRAPHY

1. Inman JL, Robertson C, Mott JD, Bissell MJ. Mammary gland development: cell fate specification, stem cells and the microenvironment. Development. 2015;142:1028–42.

2. Visvader JE, Stingl J. Mammary stem cells and the differentiation hierarchy: current status and perspectives. Genes Dev. 2014;28:1143–58.

3. Lyons WR. Hormonal Synergism in Mammary Growth. Proc. R. Soc. Lond. B Biol. Sci. 1958;149:303–25.

4. Nandi S. Endocrine Control of Mammary-Gland Development and Function in the C3H/He Crgl Mouse. J. Natl. Cancer Inst. 1958;21:1039–63.

5. Fata JE, Chaudhary V, Khokha R. Cellular Turnover in the Mammary Gland Is Correlated with Systemic Levels of Progesterone and Not 17β-Estradiol During the Estrous Cycle. Biol. Reprod. 2001;65:680–8.

6. Hovey RC, Aimo L. Diverse and Active Roles for Adipocytes During Mammary Gland Growth and Function. J. Mammary Gland Biol. Neoplasia. 2010;15:279–90.

7. Makarem M, Kannan N, Nguyen LV, Knapp DJHF, Balani S, Prater MD, et al. Developmental Changes in the in Vitro Activated Regenerative Activity of Primitive Mammary Epithelial Cells. PLOS Biol. 2013;11:e1001630.

8. Dumont N, Liu B, DeFilippis RA, Chang H, Rabban JT, Karnezis AN, et al. Breast Fibroblasts Modulate Early Dissemination, Tumorigenesis, and Metastasis through Alteration of Extracellular Matrix Characteristics. Neoplasia. 2013;15:249–IN7.

9. Simian M, Hirai Y, Navre M, Werb Z, Lochter A, Bissell MJ. The interplay of matrix metalloproteinases, morphogens and growth factors is necessary for branching of mammary epithelial cells. Development. 2001;128:3117–31.

10. Schedin P, Hovey RC. Editorial: The Mammary Stroma in Normal Development and Function. J. Mammary Gland Biol. Neoplasia. 2010;15:275–7.

11. Reed JR, Schwertfeger KL. Immune Cell Location and Function During Post-Natal Mammary Gland Development. J. Mammary Gland Biol. Neoplasia. 2010;15:329–39.

12. Kim R, Emi M, Tanabe K. Cancer immunoediting from immune surveillance to immune escape. Immunology. 2007;121:1–14.

13. Sakakura T, Nishizuka Y, Dawe CJ. Mesenchyme-dependent morphogenesis and epithelium-specific cytodifferentiation in mouse mammary gland. Science. 1976;194:1439–41.

14. Virnig BA, Tuttle TM, Shamliyan T, Kane RL. Ductal Carcinoma In Situ of the Breast: A Systematic Review of Incidence, Treatment, and Outcomes. J. Natl. Cancer Inst. 2010;102:170–8.

15. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011;144:646–674.

16. Hanahan D, Coussens LM. Accessories to the crime: functions of cells recruited to the tumor microenvironment. Cancer Cell. 2012;21:309–22.

17. McAllister SS, Gifford AM, Greiner AL, Kelleher SP, Saelzler MP, Ince TA, et al. Systemic endocrine instigation of indolent tumor growth requires osteopontin. Cell. 2008;133:994–1005.

18. Bhowmick NA, Neilson EG, Moses HL. Stromal fibroblasts in cancer initiation and progression. Nature. 2004;432:332–7.

19. Chen Q, Zhang XH-F, Massagué J. Macrophage binding to receptor VCAM-1 transmits survival signals in breast cancer cells that invade the lungs. Cancer Cell. 2011;20:538–49.

20. Castells M, Thibault B, Delord J-P, Couderc B. Implication of Tumor Microenvironment in Chemoresistance: Tumor-Associated Stromal Cells Protect Tumor Cells from Cell Death. Int. J. Mol. Sci. 2012;13:9545–71.

21. Sugie T, Toi M. Antitumor immunity and advances in cancer immunotherapy. Breast Cancer. 2017;24:1–2.

22. Charles A Janeway J, Travers P, Walport M, Shlomchik MJ. T cell-mediated cytotoxicity. 2001 [cited 2017 Apr 9]; Available from: https://www.ncbi.nlm.nih.gov/books/NBK27101/

23. Kanterman J, Sade-Feldman M, Baniyash M. New insights into chronic inflammation-induced immunosuppression. Semin. Cancer Biol. 2012;22:307–18.

24. WHO | Breast cancer: prevention and control [Internet]. WHO. [cited 2016 Dec 12]. Available from: http://www.who.int/cancer/detection/breastcancer/en/

25. Breast cancer statistics - Canadian Cancer Society [Internet]. www.cancer.ca. [cited 2016 Dec 12]. Available from: http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=on

26. Ellison LF, Gibbons L. Survival from cancer-up-to-date predictions using period analysis. Health Rep. 2006;17:19.

27. Gibbons L, Ellison LF. Leading cancers-changes in five-year relative survival. Health Rep. 2004;15:19.

28. Easton DF, Pharoah PDP, Antoniou AC, Tischkowitz M, Tavtigian SV, Nathanson KL, et al. Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. N. Engl. J. Med. 2015;372:2243–57.

29. Stephens FO, Aigner KR. Cancer of the Breast: An Overview. Basics Oncol. [Internet]. Cham: Springer International Publishing; 2016 [cited 2016 Jul 4]. p. 147–209. Available from: http://link.springer.com/10.1007/978-3-319-23368-0_12

30. U.S. Preventive Services Task Force*. Screening for breast cancer: U.s. preventive services task force recommendation statement. Ann. Intern. Med. 2009;151:716–26.

31. Kolb TM, Lichy J, Newhouse JH. Comparison of the Performance of Screening Mammography, Physical Examination, and Breast US and Evaluation of Factors that Influence Them: An Analysis of 27,825 Patient Evaluations. Radiology. 2002;225:165–75.

32. Carney PA, Miglioretti DL, Yankaskas BC, Kerlikowske K, Rosenberg R, Rutter CM, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann. Intern. Med. 2003;138:168–175.

33. Esposito A, Criscitiello C, Curigliano G. Highlights from the 14(th) St Gallen International Breast Cancer Conference 2015 in Vienna: Dealing with classification, prognostication, and prediction refinement to personalize the treatment of patients with early breast cancer. Ecancermedicalscience. 2015;9:518.

34. Boyle P. Triple-negative breast cancer: epidemiological considerations and recommendations. Ann. Oncol. 2012;23:vi7-vi12.

35. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting Global Gene Expression Analysis. Cell. 2012;151:476–82.

36. Finak G, Sadekova S, Pepin F, Hallett M, Meterissian S, Halwani F, et al. Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. Breast Cancer Res. 2006;8:R58.

37. Finak G, Bertos N, Pepin F, Sadekova S, Souleimanova M, Zhao H, et al. Stromal gene expression predicts clinical outcome in breast cancer. Nat. Med. 2008;14:518–27.

38. Pepin F, Bertos N, Laferrière J, Sadekova S, Souleimanova M, Zhao H, et al. Gene-expression profiling of microdissected breast cancer microvasculature identifies distinct tumor vascular subtypes. Breast Cancer Res. 2012;14:R120.

39. Knight JF, Lesurf R, Zhao H, Pinnaduwage D, Davis RR, Saleh SMI, et al. Met synergizes with p53 loss to induce mammary tumors that possess features of claudin-low breast cancer. Proc. Natl. Acad. Sci. U. S. A. [Internet]. 2013; Available from: http://www.ncbi.nlm.nih.gov/pubmed/23509284

40. Tofigh A, Suderman M, Paquet ER, Livingstone J, Bertos N, Saleh SM, et al. The Prognostic Ease and Difficulty of Invasive Breast Carcinoma. Cell Rep. [Internet]. 2014; Available from: http://www.cell.com/cell-reports/abstract/S2211-1247(14)00765-7

41. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.

42. Zahurak M, Parmigiani G, Yu W, Scharpf RB, Berman D, Schaeffer E, et al. Pre-processing Agilent microarray data. BMC Bioinformatics. 2007;8:142.

43. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 2002;30:e15–e15.

44. Quackenbush J. Computational analysis of microarray data. Nat. Rev. Genet. 2001;2:418–27.

45. Ward Jr JH. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. 1963;58:236–44.

46. Suzuki R, Shimodaira H. pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling [Internet]. 2014. Available from: http://CRAN.R-project.org/package=pvclust

47. Rody A, Karn T, Liedtke C, Pusztai L, Ruckhaeberle E, Hanker L, et al. A clinically relevant gene signature in triple negative and basal-like breast cancer. Breast Cancer Res. 2011;13:R97.

48. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 1995;289–300.

49. Consortium M. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat. Biotechnol. 2010;28:827–38.

50. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 2005;102:15545–50.

51. van de Vijver MJ, He YD, van 't Veer LJ, Dai H, Hart AAM, Voskuil DW, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. N. Engl. J. Med. 2002;347:1999–2009.

52. Sparano JA, Paik S. Development of the 21-Gene Assay and Its Application in Clinical Practice and Clinical Trials. J. Clin. Oncol. 2008;26:721–8.

53. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6.

54. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer. N. Engl. J. Med. 2016;375:717–29.

55. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406:747–52.

56. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. 2001;98:10869–74.

57. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.

58. Paquet ER, Hallett MT. Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype. JNCI J. Natl. Cancer Inst. 2014;107:dju357–dju357.

59. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput. Biol. 2011;7:e1002240.

60. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70.

61. Bergamaschi A, Kim YH, Wang P, Sørlie T, Hernandez-Boussard T, Lonning PE, et al. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes. Chromosomes Cancer. 2006;45:1033–40.

62. Andreopoulou E, Schweber SJ, Sparano JA, McDaid HM. Therapies for triple negative breast cancer. Expert Opin. Pharmacother. 2015;16:983–98.

63. Lehmann BD, Pietenpol JA. Clinical implications of molecular heterogeneity in triple negative breast cancer. Breast Edinb. Scotl. 2015;24 Suppl 2:S36-40.

64. Lehmann BDB, Bauer J a J, Chen X, Sanders ME, Chakravarthy a B, Shyr Y, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. 2011;121:2750–2767.

65. Burstein MD, Tsimelzon A, Poage GM, Covington KR, Contreras A, Fuqua S, et al. Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-negative Breast Cancer. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. 2014;21:1688–1699.

66. Ring BZ, Hout DR, Morris SW, Lawrence K, Schweitzer BL, Bailey DB, et al. Generation of an algorithm based on minimal gene sets to clinically subtype triple negative breast cancer patients. BMC Cancer. 2016;16:143.

67. Masuda H, Baggerly KA, Wang Y, Zhang Y, Gonzalez-Angulo AM, Meric-Bernstam F, et al. Differential Response to Neoadjuvant Chemotherapy Among 7 Triple-Negative Breast Cancer Molecular Subtypes. Clin. Cancer Res. 2013;19:5533–40.

68. Viale G. The current state of breast cancer classification. Ann. Oncol. 2012;23:x207–10.

69. Zardavas D, Irrthum A, Swanton C, Piccart M. Clinical management of breast cancer heterogeneity. Nat. Rev. Clin. Oncol. 2015;12:381–94.

70. Zardavas D, Maetens M, Irrthum A, Goulioti T, Engelen K, Fumagalli D, et al. The AURORA initiative for metastatic breast cancer. Br. J. Cancer. 2014;111:1881–7.

71. Balko JM, Giltnane JM, Wang K, Schwarz LJ, Young CD, Cook RS, et al. Molecular Profiling of the Residual Disease of Triple-Negative Breast Cancers after Neoadjuvant Chemotherapy Identifies Actionable Therapeutic Targets. Cancer Discov. 2014;4:232–45.

72. Zhang X, Claerhout S, Prat A, Dobrolecki LE, Petrovic I, Lai Q, et al. A Renewable Tissue Resource of Phenotypically Stable, Biologically and Ethnically Diverse, Patient-Derived Human Breast Cancer Xenograft Models. Cancer Res. 2013;73:4885–97.

73. Whittle JR, Lewis MT, Lindeman GJ, Visvader JE. Patient-derived xenograft models of breast cancer and their predictive power. Breast Cancer Res. 2015;17:17.

74. Aparicio S, Hidalgo M, Kung AL. Examining the utility of patient-derived xenograft mouse models. Nat. Rev. Cancer. 2015;15:311–6.

75. Bacher R, Kendziorski C. Design and computational analysis of single-cell RNA-sequencing experiments. Genome Biol. [Internet]. 2016 [cited 2016 Sep 16];17. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4823857/

76. Lawson DA, Bhakta NR, Kessenbrock K, Prummel KD, Yu Y, Takai K, et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. Nature. 2015;526:131–5.

77. Jordan NV, Bardia A, Wittner BS, Benes C, Ligorio M, Zheng Y, et al. HER2 expression identifies dynamic functional states within circulating breast cancer cells. Nature. 2016;537:102–6.

78. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. 2015;16:133–45.

79. Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Res. 2015;25:1499–507.

80. Paquet ER, Lesurf R, Tofigh A, Dumeaux V, Hallett MT. Detecting gene signature activation in breast cancer in an absolute, single-patient manner. Breast Cancer Res. BCR [Internet]. 2017 [cited 2017 Apr 8];19. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5361722/

81. Jabri B, Abadie V. IL-15 functions as a danger signal to regulate tissue-resident T cells and tissue destruction. Nat. Rev. Immunol. 2015;15:771–83.

82. Cullen SP, Brunet M, Martin SJ. Granzymes in cancer and immunity. Cell Death Differ. 2010;17:616–23.

83. Zitvogel L, Galluzzi L, Kepp O, Smyth MJ, Kroemer G. Type I interferons in anticancer immunity. Nat. Rev. Immunol. 2015;15:405–14.

84. Takeda K, Akira S. STAT family of transcription factors in cytokine-mediated biological responses. Cytokine Growth Factor Rev. 2000;11:199–207.

85. Chang H-C, Han L, Goswami R, Nguyen ET, Pelloso D, Robertson MJ, et al. Impaired development of human Th1 cells in patients with deficient expression of STAT4. Blood. 2009;113:5887–90.

86. Kumar D, Whiteside TL, Kasid U. Identification of a Novel Tumor Necrosis Factor-α-inducible Gene, SCC-S2, Containing the Consensus Sequence of a Death Effector Domain of Fas-

associated Death Domain-like Interleukin- 1β-converting Enzyme-inhibitory Protein. J. Biol. Chem. 2000;275:2973–8.

87. Sarma V, Wolf FW, Marks RM, Shows TB, Dixit VM. Cloning of a novel tumor necrosis factor-alpha-inducible primary response gene that is differentially expressed in development and capillary tube-like formation in vitro. J. Immunol. 1992;148:3302–12.

88. Tai S-K, Tan OJ-K, Chow VT-K, Jin R, Jones JL, Tan P-H, et al. Differential Expression of Metallothionein 1 and 2 Isoforms in Breast Cancer Lines with Different Invasive Potential: Identification of a Novel Nonsilent Metallothionein-1H Mutant Variant. Am. J. Pathol. 2003;163:2009–19.

89. Cheung KJ, Ewald AJ. A collective route to metastasis: Seeding by tumor cell clusters. Science. 2016;352:167–9.

90. Gandellini P, Andriani F, Merlino G, D'Aiuto F, Roz L, Callari M. Complexity in the tumour microenvironment: Cancer associated fibroblast gene expression patterns identify both common and unique features of tumour-stroma crosstalk across cancer types. Semin. Cancer Biol. 2015;35:96–106.

91. Gilkes DM, Semenza GL, Wirtz D. Hypoxia and the extracellular matrix: drivers of tumour metastasis. Nat. Rev. Cancer. 2014;14:430–9.

92. Beck AH, Espinosa I, Gilks CB, van de Rijn M, West RB. The fibromatosis signature defines a robust stromal response in breast carcinoma. Lab. Invest. 2008;88:591–601.

93. Nielsen TO, Parker JS, Leung S, Voduc D, Ebbert M, Vickery T, et al. A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor–Positive Breast Cancer. Clin. Cancer Res. 2010;16:5222–32.

94. Barnard NJ, Hall PA, Lemoine NR, Kadar N. Proliferative index in breast carcinoma determined in situ by Ki67 immunostaining and its relationship to clinical and pathological variables. J. Pathol. 1987;152:287–95.

95. Ng CKY, Schultheis AM, Bidard F-C, Weigelt B, Reis-Filho JS. Breast cancer genomics from microarrays to massively parallel sequencing: paradigms and new insights. J. Natl. Cancer Inst. 2015;107:djv015–.

96. Dai H, Veer L van't, Lamb J, He YD, Mao M, Fine BM, et al. A Cell Proliferation Signature Is a Marker of Extremely Poor Outcome in a Subpopulation of Breast Cancer Patients. Cancer Res. 2005;65:4059–66.

97. Teschendorff AE, Miremadi A, Pinder SE, Ellis IO, Caldas C. An immune response gene expression module identifies a good prognosis subtype in estrogen receptor negative breast cancer. Genome Biol. 2007;8:R157.

98. West NR, Milne K, Truong PT, Macpherson N, Nelson BH, Watson PH. Tumor-infiltrating lymphocytes predict response to anthracycline-based chemotherapy in estrogen receptor-negative breast cancer. Breast Cancer Res. 2011;13:R126.

99. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J. Clin. Oncol. 2009;27:1160–7.

100. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods. 2014;11:41–6.

101. Grün D, Muraro MJ, Boisset J-C, Wiebrands K, Lyubimova A, Dharmadhikari G, et al. De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. Cell Stem Cell. 2016;19:266–77.

102. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, et al. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature. 2012;486:395–9.

103. Boersma BJ, Reimers M, Yi M, Ludwig JA, Luke BT, Stephens RM, et al. A stromal gene signature associated with inflammatory breast cancer. Int. J. Cancer. 2008;122:1324–32.

104. Planche A, Bacac M, Provero P, Fusco C, Delorenzi M, Stehle J-C, et al. Identification of Prognostic Molecular Features in the Reactive Stroma of Human Breast and Prostate Cancer. PLOS ONE. 2011;6:e18640.

105. Alexe G, Dalgin GS, Scanfeld D, Tamayo P, Mesirov JP, DeLisi C, et al. High Expression of Lymphocyte-Associated Genes in Node-Negative HER2+ Breast Cancers Correlates with Lower Recurrence Rates. Cancer Res. 2007;67:10669–76.

106. Watkins NA, Gusnanto A, Bono B de, De S, Miranda-Saavedra D, Hardie DL, et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. Blood. 2009;113:e1–9.

107. Bos PD, Zhang XH-F, Nadal C, Shu W, Gomis RR, Nguyen DX, et al. Genes that mediate breast cancer metastasis to the brain. Nature. 2009;459:1005–9.

108. Landemaine T, Jackson A, Bellahcène A, Rucci N, Sin S, Abad BM, et al. A Six-Gene Signature Predicting Breast Cancer Lung Metastasis. Cancer Res. 2008;68:6092–9.

109. Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pagès C, et al. Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. Science. 2006;313:1960–4.

110. Tosolini M, Kirilovsky A, Mlecnik B, Fredriksen T, Mauger S, Bindea G, et al. Clinical Impact of Different Classes of Infiltrating T Cytotoxic and Helper Cells (Th1, Th2, Treg, Th17) in Patients with Colorectal Cancer. Cancer Res. 2011;71:1263–71.

111. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. The Life History of 21 Breast Cancers. Cell. 2012;149:994–1007.

112. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014;512:155–60.

113. Gupta PB, Fillmore CM, Jiang G, Shapira SD, Tao K, Kuperwasser C, et al. Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. Cell. 2011;146:633–44.

114. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344:1396–401.

115. Nielsen TO, Hsu FD, Jensen K, Cheang M, Karaca G, Hu Z, et al. Immunohistochemical and Clinical Characterization of the Basal-Like Subtype of Invasive Breast Carcinoma. Clin. Cancer Res. 2004;10:5367–74.

116. Carey LA, Rugo HS, Marcom PK, Mayer EL, Esteva FJ, Ma CX, et al. TBCRC 001: Randomized Phase II Study of Cetuximab in Combination With Carboplatin in Stage IV Triple-Negative Breast Cancer. J. Clin. Oncol. 2012;30:2615–23.

117. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available from: http://www.R-project.org/

118. Chen X, Li J, Gray WH, Lehmann BD, Bauer JA, Shyr Y, et al. TNBCtype: A Subtyping Tool for Triple-Negative Breast Cancer. Cancer Inform. 2012;11:147–56.

119. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
IGLL5	3.109535291	4.85E-13
ENST00000390323	3.009176813	9.51E-12
S77011	2.681057587	2.23E-13
lincRNA:chr1:41956288-41971088_R	2.609181576	1.27E-13
ENST00000390237	2.500430757	2.73E-07
ENST00000390549	2.276508688	6.43E-07
BF175071	2.268506055	3.93E-11
ENST00000390559	2.262660833	1.71E-06
TNRC18	2.097825836	5.80E-12
ENST00000390551	1.996786106	6.45E-06
lincRNA:chr1:228258277-228268427_F	1.976227452	2.23E-13
ENST00000390547	1.931763591	0.00114498
lincRNA:chr2:91677073-91695823_R	1.920373401	3.33E-11
IGLL1	1.703982926	3.33E-11
lincRNA:chr10:6779694-6889494_R	1.635107215	4.85E-13
LOC100653210	1.601937439	5.19E-10
LOC96610	1.601120006	4.63E-08
POU2AF1	1.557077304	2.94E-11
ENST00000390305	1.535984698	6.54E-07
IGJ	1.519562001	7.32E-08
KIAA0125	1.512273254	1.18E-06
CD79A	1.491487715	1.83E-12
ENST00000390312	1.489437436	2.25E-08
ENST00000453166	1.46506248	3.26E-06
A_33_P3303394	1.462473414	3.28E-07
FCRL5	1.432141974	5.80E-12
ENST00000354689	1.414765434	1.66E-07
ENST00000426099	1.388823878	2.23E-08
DQ098707	1.350569372	6.67E-08
FAM46C	1.328974411	1.73E-10
lincRNA:chr1:1100287-1108062_F	1.313565297	3.33E-11
lincRNA:chr2:20783219-20795366_F	1.300186906	2.55E-11
NP113779	1.244767863	3.44E-06
ENST00000390252	1.225619765	3.21E-07
lincRNA:chr2:131595455-131614380_F	1.220298668	9.07E-11
NYX	1.204872706	2.55E-11
CHIT1	1.196545618	0.02088633
CPNE5	1.192924351	5.19E-06
CD19	1.189994588	0.00023945

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
FER1L4	1.168418661	0.000401661
ENST00000390291	1.163609121	1.71E-06
ENST00000390243	1.161235568	8.94E-05
TNFRSF17	1.139558474	8.94E-09
SMR3A	1.132574214	9.25E-11
lincRNA:chr7:32953000-32983675_R	1.114126286	4.19E-10
TMEM238	1.111853565	6.25E-06
ENST00000390317	1.111577852	4.63E-08
ENST00000390265	1.104813366	1.54E-07
lincRNA:chr1:201509727-201545677_F	1.087861315	1.82E-10
ENST00000390285	1.070121629	3.09E-05
lincRNA:chr1:181063077-181073127_F	1.060183547	2.36E-09
THBS4	1.053330593	0.041442545
RAPH1	1.041303291	2.68E-08
lincRNA:chr6:158663162-158717262_F	1.029053524	1.12E-10
AK127961	1.006098105	1.97E-08
TXNDC5	0.995479452	1.75E-05
lincRNA:chr19:42774860-42784435_R	0.987777668	2.61E-08
ENST00000390294	0.960953586	8.30E-05
PIM2	0.959877531	8.94E-09
lincRNA:chr7:25978475-25989975_R	0.953335039	0.000428236
lincRNA:chr3:171509574-171527714_R	0.952638335	6.45E-06
ENST00000390247	0.951943083	2.78E-05
lincRNA:chr17:38679761-38683252_R	0.947869945	3.07E-08
AF194718	0.93506067	0.000411711
ENST00000498435	0.918140163	3.12E-05
ENST00000492167	0.91083253	9.15E-06
AB363267	0.908887912	0.001667634
lincRNA:chr10:72833019-72847269_F	0.90406052	1.68E-08
lincRNA:chr1:41238013-41247638_R	0.902389708	2.25E-08
lincRNA:chrX:45598006-45710454_F	0.899615206	4.60E-09
lincRNA:chr9:32936300-32949200_R	0.890093052	8.42E-09
ENST00000390610	0.882014831	0.001667634
ENST00000390319	0.875196955	0.000178224
lincRNA:chr18:33161080-33207449_F	0.852890853	4.76E-11
ENST00000491977	0.846068134	0.000206121
FKBP11	0.840207508	6.62E-06
A_33_P3299047	0.837374433	7.47E-08
ENST00000390622	0.816450146	0.001840528

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
Q8VGA8	0.810063018	2.61E-08
CCDC129	0.808888844	1.81E-08
XBP1	0.808356304	0.015357699
lincRNA:chr2:77321142-77331767_R	0.804533858	8.14E-07
ENST00000536864	0.786942609	3.28E-07
MZB1	0.78656737	0.015350583
lincRNA:chr2:56190790-56194076_R	0.781794041	5.19E-06
ENST00000468879	0.780045447	0.000150303
OR1S1	0.777367585	1.37E-07
MGC16025	0.762482078	1.91E-06
lincRNA:chr7:226042-232442_R	0.760481307	3.88E-08
DERL3	0.754314594	0.01096468
lincRNA:chr7:32969506-32973492_R	0.752810604	2.39E-06
ENST00000390605	0.738352951	0.003185384
XLOC_l2_012374	0.737052166	1.02E-06
LOC338739	0.732578166	6.43E-07
lincRNA:chr17:73598283-73599500_F	0.728031609	3.24E-07
ENST00000390633	0.713254958	0.006010811
ENST00000474213	0.70974215	0.013004403
ENST00000390436	0.707871557	3.80E-07
lincRNA:chr18:59259607-59415414_F	0.703983718	1.10E-06
C11orf20	0.703816214	0.000197306
lincRNA:chr10:89908970-89919320_F	0.702320098	4.07E-06
lincRNA:chr2:218461580-218481005_R	0.699217599	6.26E-08
lincRNA:chr14:21670160-21677035_F	0.698447698	4.75E-06
lincRNA:chr10:99179035-99184435_R	0.697072841	2.78E-05
lincRNA:chrX:71300975-71319175_F	0.695900284	4.31E-06
lincRNA:chr14:32360774-32374086_F	0.693932877	3.81E-06
ENST00000479981	0.680220113	0.000893748
A_33_P3225572	0.679971447	2.04E-07
lincRNA:chr1:27845013-27856588_F	0.679133066	1.70E-07
ZNF714	0.675463734	2.81E-06
SSR4	0.672952734	0.000582131
lincRNA:chr8:134368768-134379793_F	0.669772277	3.91E-06
DUSP26	0.666040406	0.007220589
ENST00000390615	0.664578966	0.000227729
lincRNA:chr12:117562392-117569365_R	0.66233842	6.43E-07
ENST00000532821	0.658746968	1.71E-06
lincRNA:chr13:44760050-44767325_F	0.655628721	6.38E-06

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
ENST00000390600	0.648276287	0.000179662
ENST00000307548	0.647733598	0.000135618
lincRNA:chr18:33466152-33528357_F	0.645025161	1.32E-06
lincRNA:chr12:76015433-76261508_F	0.644812813	3.93E-06
ENST00000390301	0.638317394	0.000850064
KRT73	0.633156093	1.54E-07
MEI1	0.630101081	0.000172879
ENST00000398957	0.627700629	6.91E-06
lincRNA:chr12:68421607-68422081_F	0.619600787	0.018939101
PGA3	0.613320693	1.01E-06
LOC283710	0.61275943	3.00E-05
SPRNP1	0.612287806	2.15E-06
BCORP1	0.612048278	1.50E-06
RARG	0.608954029	4.17E-06
lincRNA:chr2:64412771-64525721_R	0.604857972	2.60E-07
PHOSPHO1	0.604800089	0.004426847
STARD5	0.600377318	0.001185239
HERPUD1	0.596022035	0.000205103
FAM21C	0.595378916	0.000345195
lincRNA:chr18:59259608-59416029_F	0.583194565	7.15E-05
SSX3	0.583176642	5.71E-07
LOC100129196	0.583136459	1.50E-06
lincRNA:chrX:17782154-17792979_F	0.576622852	7.32E-08
lincRNA:chr13:51103249-51263599_R	0.576516877	0.00016006
THC2688744	0.569148622	1.12E-05
CD38	0.565382492	0.001969194
lincRNA:chr6:6683526-6701601_R	0.560212861	4.52E-06
CD27	0.55877568	0.013959524
lincRNA:chr5:102544301-102549901_R	0.556823464	7.50E-05
lincRNA:chrX:13269404-13285729_F	0.556777494	1.91E-05
lincRNA:chr6:113672332-113690598_R	0.556143895	2.78E-05
AF336885	0.551821271	0.000705801
lincRNA:chr2:178145929-178216641_R	0.548727161	3.06E-05
A_33_P3268167	0.542098265	9.39E-05
ENST00000390594	0.538803465	0.007071496
lincRNA:chr12:122243692-122252067_R	0.531513567	3.53E-06
lincRNA:chrX:17782154-17795079_F	0.531073296	8.32E-06
BCL2L15	0.530677782	0.000195587
PTMS	0.527920612	8.40E-05

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
LOC151009	0.52398244	1.12E-05
LOC100130811	0.523921026	0.012563456
ACTR3BP5	0.522266113	0.000127539
FAM153A	0.519225287	3.02E-06
LOC388210	0.517457754	3.31E-06
SPCS3	0.516153597	0.003497784
LOC389602	0.510994837	9.58E-06
LOC100288292	0.510031527	6.54E-07
ANKRD36BP2	0.503416812	0.001291853
IQSEC3	0.50152265	6.26E-06
lincRNA:chr1:180929852-180942126_R	0.501117906	1.35E-05
lincRNA:chr11:2004771-2007592_F	0.50067029	5.34E-05
lincRNA:chr14:104345747-104366672_F	0.49764252	2.78E-05
lincRNA:chr9:14587837-14588176_F	0.497599237	0.00019992
SLC12A5	0.496365034	0.000387407
IP6K2	0.492131383	0.002189116
ENST00000390468	0.491198279	3.44E-06
lincRNA:chr9:33011575-33024725_R	0.490511732	5.60E-06
ENST00000390297	0.490447773	0.002316305
SLAMF7	0.487665846	0.002244982
ENST00000424969	0.487049396	0.000205103
lincRNA:chr1:234780127-234798577_R	0.486808033	0.001153558
lincRNA:chr1:153768876-153777076_F	0.4862076	9.48E-06
lincRNA:chr2:232250156-232258431_R	0.483718573	1.28E-05
CORT	0.480339686	3.06E-05
HSH2D	0.478590017	0.002323409
KCNA3	0.477931715	8.20E-05
SLC43A1	0.476865853	0.000362842
OR1S2	0.476353705	6.43E-07
A_33_P3401084	0.475968087	5.90E-05
lincRNA:chr4:68294580-68337980_F	0.474363863	0.000184439
lincRNA:chr6:43988722-43993947_R	0.472065215	2.88E-05
AK130802	0.470029988	0.00154833
lincRNA:chr7:131549760-131565160_R	0.46768592	7.64E-05
DUSP15	0.464814936	9.00E-05
SEC14L1	0.463575564	0.000194945
PDK1	0.463254598	0.00019992
lincRNA:chr5:95919519-95995094_R	0.462898786	6.65E-06
PRDM1	0.461380658	0.015945645

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
lincRNA:chr16:11481046-11484311_F	0.460806487	4.32E-05
PSD4	0.457982027	1.03E-06
lincRNA:chr22:46451236-46516536_R	0.457616554	0.0002409
ENST00000390454	0.454507137	8.80E-07
ENST00000552290	0.452436721	9.83E-05
LRP3	0.449180369	1.81E-05
lincRNA:chr1:118374652-118400852_R	0.44836418	0.000287182
lincRNA:chr12:132832202-132856777_F	0.447953846	0.004840558
OR10J5	0.447838959	0.001252957
lincRNA:chr2:233444129-233451012_F	0.445748812	6.65E-06
SLC16A11	0.445155896	0.000123387
lincRNA:chr2:74212792-74262017_R	0.442658561	0.000158739
LOC652119	0.442526168	0.000181798
lincRNA:chr2:70212746-70263471_R	0.442494081	8.07E-05
lincRNA:chr7:93318289-93329939_R	0.440337356	5.97E-05
lincRNA:chr6:21666674-22221946_R	0.438442842	3.66E-05
lincRNA:chr2:74225470-74225673_R	0.437709059	4.02E-05
ADAMTS7	0.437188838	0.000428236
lincRNA:chr10:104838935-104844310_F	0.436333481	0.000737686
SEC11C	0.435284704	0.002879764
lincRNA:chrX:53685100-53711400_R	0.432541826	0.017283434
SEL1L3	0.431595621	0.01841489
GPR150	0.430293218	0.000282699
UBE2J1	0.42990845	0.035342804
ENST00000431767	0.429487613	0.004435321
ST6GAL1	0.429023043	0.021312603
lincRNA:chr5:1167850-1186275_R	0.428742767	0.000226064
lincRNA:chr18:3347700-3348205_F	0.428468405	2.01E-05
XLOC_l2_014959	0.427671078	0.00019992
ENST00000383417	0.426442303	1.61E-05
ZNF215	0.425087497	0.001182837
lincRNA:chr21:30565800-30660465_R	0.424838679	0.000856506
RP11-165H20.1	0.424336244	9.68E-05
lincRNA:chr10:5156300-5171275_F	0.422628211	0.0002409
lincRNA:chr5:149980882-149995557_F	0.42030632	0.000226064
A_33_P3407691	0.419947298	2.07E-05
DNAJC4	0.416550431	0.00018334
ENST00000390625	0.414089488	0.008716134
LOC150622	0.413802539	9.27E-06

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
A_33_P3227010	0.411552742	0.000879447
lincRNA:chr1:33869120-33896238_F	0.411407491	0.000205103
LOC100133286	0.404301087	0.000623021
lincRNA:chr2:241928152-241929111_R	0.403850803	0.000632364
lincRNA:chrX:117240572-117247172_R	0.403187366	0.000400152
A_33_P3329104	0.402501521	0.000591294
lincRNA:chr18:36967212-37153620_R	0.402462151	6.45E-06
lincRNA:chr12:54128308-54135158_R	0.401242792	0.013678758
ENST00000446887	0.400782334	0.000122476
lincRNA:chr4:110462151-110470201_F	0.400467049	0.000271495
lincRNA:chr4:2460833-2464216_R	0.400232402	6.76E-05
DA197111	0.399813587	0.000278821
CU691877	0.397475968	0.00049583
lincRNA:chr9:23671778-23672397_F	0.393752768	9.83E-05
CTSL1P2	0.392171061	0.00247629
lincRNA:chr5:92762769-92774694_R	0.391864225	0.001550957
P39188	0.391500045	0.000158739
A_33_P3273399	0.388376206	0.001215456
lincRNA:chr9:133230324-133233284_R	0.38448929	9.22E-05
lincRNA:chr2:37776096-37862046_R	0.384274131	1.45E-05
BM477328	0.383771191	0.000130193
LRRC16B	0.383529574	0.000893748
C1orf190	0.381010397	0.018990823
lincRNA:chr1:181063077-181073127_R	0.380696646	0.000705801
KCNN3	0.380356015	0.006542414
A_33_P3261074	0.380230567	1.80E-05
lincRNA:chr6:29983946-29992471_F	0.380096819	3.00E-05
NKX1-2	0.379448283	0.000176743
ITM2C	0.377599744	0.036318642
GPR119	0.376449225	0.000161857
lincRNA:chr13:50657699-50697649_R	0.375046675	0.000269623
C2CD4C	0.372990274	2.78E-05
lincRNA:chr13:67955599-68781749_F	0.371263127	7.86E-05
RBMY1B	0.370181182	1.69E-05
HOXB13-AS1	0.368619053	0.000521338
A_24_P110273	0.36771052	0.000518454
LOC100130345	0.364099127	0.00159856
lincRNA:chr5:74327969-74348269_F	0.364012393	0.007609433
lincRNA:chr10:72800569-72813581_F	0.361460259	0.004650449

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
NP511209	0.359610179	0.001291853
lincRNA:chr13:33888375-33923550_F	0.358513834	0.000108533
lincRNA:chr5:12625075-12747025_F	0.35804445	4.71E-05
GGT1	0.356556167	0.004647437
SPAG4	0.35490441	0.003952053
PARVG	0.354376366	0.000176743
SELK	0.353586004	0.017085212
lincRNA:chr14:59882547-59912047_F	0.352979489	0.000705801
MEX3D	0.351133407	0.019950065
TPP1	0.350430674	0.003109953
REEP1	0.35009443	0.001777006
lincRNA:chr13:103554249-103568024_R	0.349275477	1.23E-05
ENST00000478163	0.348506678	0.000124017
lincRNA:chr18:35246652-35296677_F	0.347733661	0.028641139
ENST00000485332	0.347138893	0.000428433
MOB1A	0.346005082	0.001455784
lincRNA:chr4:127694975-127709825_F	0.343951204	0.003769994
ERCC2	0.34364559	0.000818507
IQGAP2	0.34344513	0.013921771
LTB4R2	0.341801891	0.000656397
LOC497256	0.341013859	0.000634032
lincRNA:chr13:101223049-101236424_R	0.340294454	0.000204287
FAM90A10	0.34001973	0.00019992
MUC4	0.338958664	0.001036307
lincRNA:chr4:123620575-123629031_R	0.335257355	0.001247483
lincRNA:chr1:93796837-93806487_F	0.334237305	0.002755365
KRTAP10-1	0.33339494	0.000586091
UNC93B1	0.333065864	0.001200414
lincRNA:chr22:46476224-46493888_F	0.331092997	0.000256897
CC2D1A	0.330670584	0.00028513
ACAP1	0.329927783	0.000205587
ENST00000471857	0.32731059	0.008479071
TTC22	0.325385493	0.01076115
lincRNA:chr8:134368768-134379793_R	0.324578225	7.64E-05
lincRNA:chr12:101125994-101157894_F	0.323139502	0.001433066
lincRNA:chr16:11581190-11588611_F	0.322578928	0.001260818
TM6SF1	0.322108494	0.000705801
lincRNA:chr6:106858857-106926507_R	0.320452711	0.000818507
LOC100128675	0.320226154	0.001694532

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
lincRNA:chr21:45573097-45578347_R	0.319639117	0.010183009
lincRNA:chr4:83812701-83821724_F	0.318501981	0.001667634
lincRNA:chr6:27663246-27683586_R	0.318001782	0.000591294
lincRNA:chr14:61762697-61774172_R	0.316915035	0.000395418
lincRNA:chr20:31148894-31161690_F	0.315891026	0.014976857
A_33_P3263747	0.315674731	8.49E-05
lincRNA:chr2:120443455-120452280_F	0.315640247	8.51E-05
FLJ34208	0.314780833	3.38E-05
lincRNA:chr2:201638597-201642676_R	0.313156967	0.000743941
A_33_P3247210	0.311566008	0.002904411
lincRNA:chr6:6683526-6811251_R	0.308893741	0.000226064
LILRA1	0.308445379	0.000240103
lincRNA:chr21:15443204-15457462_R	0.307111128	0.000271444
USP48	0.305275175	0.021333814
LOC439951	0.305026421	0.001029597
lincRNA:chr7:30216750-30313150_F	0.302398355	0.006870822
HAP1	0.301324039	0.002187596
THC2573121	0.299742802	0.000705801
A_33_P3316691	0.299469574	0.002677158
XLOC_l2_001196	0.297994624	0.003581242
SERPINA6	0.297329985	0.048728048
lincRNA:chr13:113549599-113574924_R	0.295828901	0.00067025
lincRNA:chr2:43148721-43166396_F	0.29570783	0.000271759
XM_003118986	0.295640324	0.016104413
lincRNA:chr8:22847955-22856530_F	0.29559879	0.000723355
lincRNA:chr12:9480083-9506383_R	0.294986649	0.002774937
lincRNA:chr13:74245699-74258424_R	0.294391911	0.003371184
LY96	0.294383512	0.046848849
lincRNA:chr2:216473880-216713730_R	0.293473225	0.035231271
lincRNA:chr1:223185802-223197477_F	0.291823673	0.001247483
lincRNA:chr8:105607224-105688174_R	0.291592237	0.000234701
Q29HP5	0.288905846	0.001886293
SMCR2	0.28812959	0.000788884
lincRNA:chr21:36128280-36139780_F	0.286819545	0.000705801
SLC25A45	0.2865845	0.004426847
lincRNA:chr17:70417969-70419213_R	0.2863442	0.000898036
FLJ40434	0.284957994	0.004070697
A_33_P3333677	0.284810304	0.00019992
lincRNA:chr8:90615284-90628709_F	0.284523882	5.21E-05

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
THC2606816	0.284379822	0.001432392
TNFRSF13C	0.282020037	0.002959014
MRPS31	0.281696741	0.049654454
lincRNA:chrX:57001900-57013275_R	0.281337694	0.000205103
ENST00000433071	0.280686019	0.001432392
lincRNA:chr22:27067640-27070585_R	0.279790657	0.000705801
lincRNA:chr8:143274368-143286859_F	0.279080972	0.000518454
lincRNA:chr11:65068424-65079274_R	0.279076405	0.002093836
NEGR1	0.278803166	0.020532626
LOC100129648	0.27874931	1.60E-06
lincRNA:chr6:44006347-44033347_R	0.27692197	0.031462948
AK022213	0.275639012	0.000500739
XLOC_009868	0.274831736	0.017533391
AMPD1	0.274736475	0.042195933
RAB11FIP4	0.272838217	0.008749138
lincRNA:chr8:142273943-142310368_F	0.272783564	0.01099177
lincRNA:chr20:25213650-25221750_R	0.272431229	0.000529355
DNAL4	0.272222796	0.005303536
THC2485300	0.271966072	0.000666522
ARHGEF7	0.271637693	0.001061816
A_33_P3414228	0.271575062	0.001550957
lincRNA:chr1:145689868-145694918_R	0.271079777	0.003710585
lincRNA:chr6:138961057-139020207_F	0.270123705	0.018025858
MOP-1	0.269955253	0.022972249
A_33_P3291329	0.269913138	0.012682645
A_33_P3388883	0.268020454	0.001905163
TRIM78P	0.267711996	0.001252957
TBC1D3	0.266750468	0.014967474
HPX-2	0.265323273	0.03016607
lincRNA:chr3:136738660-136787685_F	0.264117781	0.000466732
lincRNA:chr15:69951596-69988646_R	0.263360921	0.001668282
lincRNA:chr9:36419075-36431725_F	0.262493823	0.02916401
lincRNA:chr10:80112419-80123044_F	0.262203138	0.025558052
PRDM4	0.261018226	0.002609476
lincRNA:chr5:61931044-61948469_F	0.260638246	0.001855437
lincRNA:chr15:38364787-38365169_F	0.260251322	0.001724242
lincRNA:chr5:134465126-134478076_R	0.258971144	0.004160236
ENST00000390606	0.258247961	0.000177631
A_33_P3274080	0.258018065	0.00227751

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
LOC100130463	0.257899321	0.004994878
FNDC3A	0.257355248	0.005354422
CD72	0.255986407	0.034483146
lincRNA:chr20:47715043-47723493_R	0.255558458	0.025244864
CEACAM19	0.255333587	0.003460138
KRTAP4-12	0.255275303	0.005303536
EIF4G1	0.254638083	0.037170383
lincRNA:chr12:54128308-54135158_F	0.252744313	0.002153181
EVX1	0.251421819	0.00129941
BX648392	0.24981138	0.003159261
lincRNA:chr18:60249120-60256770_R	0.24972558	0.031191338
NP1243929	0.249317492	0.007419368
C1orf229	0.248572217	0.005062546
lincRNA:chr5:133837051-133848051_R	0.247870031	0.003175298
PLA2G4E	0.247490834	0.018587671
lincRNA:chr6:3999676-4018612_R	0.24656788	0.027599651
SUV39H1	0.246552762	0.01930593
lincRNA:chr1:30142313-30152488_F	0.245602659	0.047265486
lincRNA:chr12:57775433-57820783_F	0.244793107	0.002585852
ADARB2-AS1	0.24456143	0.002612787
BX096650	0.244558415	0.018241896
lincRNA:chr7:33664500-33886275_R	0.244538504	0.001555645
lincRNA:chr10:47041819-47063144_R	0.243575391	0.000126815
lincRNA:chr5:149866432-149882132_R	0.243326818	0.019454423
lincRNA:chr8:27416133-27435295_R	0.24286332	0.003801005
lincRNA:chr5:141143016-141169841_R	0.242062987	0.006870822
MYOG	0.241278469	0.001665356
lincRNA:chr2:72150317-72161342_F	0.240949731	0.010628919
ENST00000462693	0.240801916	0.02154164
AK124642	0.240258868	0.003532018
SPACA3	0.239448003	0.010375418
AK123110	0.239084628	0.048614162
GHSR	0.238340042	0.000163812
C9orf100	0.23812664	0.023377118
lincRNA:chr18:59236195-59282370_F	0.237146175	0.022972249
A_33_P3321682	0.236337924	0.007419368
lincRNA:chr13:33868250-33874375_F	0.235275271	0.015945645
BE561442	0.235158929	0.015055522
LOC729305	0.234757699	0.020785605

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
UTS2R	0.234490132	0.002448219
ADAM17	0.234434509	0.011595529
ARSF	0.234006942	0.020900314
LOC646743	0.232022219	0.014967474
lincRNA:chr6:3999676-4018493_R	0.230825072	0.020785605
LINC00277	0.229675584	0.012872201
AA593742	0.228736097	0.000205818
lincRNA:chr5:131804581-131808735_F	0.2284958	0.002031285
lincRNA:chr3:175987356-176039381_F	0.22778946	0.002765515
DNAJB13	0.226922006	0.016019224
XLOC_005849	0.225843353	0.004912274
BC036215	0.224665387	0.002933591
XLOC_003462	0.223781252	0.00468934
NPY2R	0.22308496	1.67E-05
lincRNA:chr11:122009090-122080940_R	0.222377402	0.029783627
SH3GL3	0.221507126	0.012629946
KCNQ3	0.221286277	0.017454076
A_33_P3271885	0.220072826	0.019815855
lincRNA:chr7:54850800-54872648_R	0.218183577	0.009848631
OR4X2	0.217252116	0.003105988
LENG8	0.21536278	0.047241794
LINC00494	0.215111835	0.047241794
CDCP2	0.213034809	0.005026028
lincRNA:chr2:105421343-105428293_F	0.212378553	0.015065214
lincRNA:chr2:8001249-8036724_F	0.209446485	0.001217874
A_33_P3388527	0.209350733	0.01997125
Q39472	0.207966054	0.002095219
lincRNA:chr2:152627254-152653328_R	0.206228063	0.00227751
lincRNA:chr11:38083599-38261799_R	0.206013897	0.005303536
ENST00000434007	0.204836116	0.036368994
ENST00000359888	0.203399792	0.002360926
A_33_P3257479	0.203170105	0.004721845
OR4F15	0.202833837	0.011865028
lincRNA:chr10:114590927-114592210_F	0.202266671	0.017234411
lincRNA:chr2:8710460-8717085_F	0.20224809	0.003718991
lincRNA:chr3:15928546-15939996_F	0.201825264	0.02255201
A_33_P3210069	0.201312679	0.009859255
FCRLA	0.200172308	0.019663597
lincRNA:chr12:127650247-127665172_F	0.200137839	0.006454838

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
CALML3	0.198907175	0.001579301
LOC145845	0.198841061	0.020532626
TRIM53P	0.197642762	0.005589169
XLOC_009483	0.196608748	0.002723375
lincRNA:chr6:125229601-125277876_R	0.194292964	0.030733033
lincRNA:chr18:33016877-33028102_R	0.193386491	0.040293652
BJ995728	0.193032192	0.005132739
lincRNA:chr4:6683199-6683665_R	0.192462707	0.03016607
lincRNA:chr5:131806223-131811702_R	0.191553303	0.005832283
DLG4	0.189384536	0.02469916
lincRNA:chr1:156851776-156860276_R	0.189284947	0.017542597
PSPN	0.188841031	0.018929415
FRG2C	0.187488814	0.015998528
lincRNA:chr11:70606952-70667252_F	0.186729364	0.018929415
C10orf53	0.186265517	0.004647437
lincRNA:chr8:103749649-103767499_R	0.186129128	0.01610056
LOC100128591	0.185794539	0.03827278
TUBB4A	0.185539198	0.016683429
LOC284757	0.185356041	0.019933885
ATP13A2	0.184930287	0.036283
lincRNA:chr8:99929224-99951124_F	0.184639367	0.043504949
lincRNA:chr8:127836493-128531243_R	0.184354483	0.006858791
lincRNA:chrX:133675909-133683184_F	0.184284335	0.009933196
ENST00000359838	0.183890555	0.015861367
lincRNA:chr11:75122727-75130002_R	0.181920424	0.049654454
lincRNA:chr7:55030831-55040930_R	0.180952962	0.023879794
lincRNA:chr14:69273097-69292047_F	0.180488746	0.019383236
lincRNA:chr4:104263901-104301551_F	0.180424571	0.028904018
lincRNA:chr2:232366756-232372381_F	0.177584386	0.005690804
lincRNA:chr3:50266021-50271246_R	0.176617486	0.012069387
LOC100131551	0.176616084	0.02088633
SEZ6	0.176325635	0.003458127
lincRNA:chr4:16239177-16275952_R	0.175717166	0.001016427
LOC642947	0.175445099	0.048817388
DCTN1	0.174180067	0.037911253
lincRNA:chr2:218595605-218618480_F	0.171981686	0.039261003
lincRNA:chr11:61685224-61691249_F	0.171564866	0.000674491
XLOC_002473	0.169437429	0.043949052
SNORA71C	0.168473298	0.010628919

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
DCAF6	0.16829148	0.020770084
lincRNA:chr14:32398493-32399352_F	0.167524745	0.029067316
lincRNA:chr9:123606154-123611279_F	0.167491857	0.008479071
lincRNA:chrX:138994509-138999734_F	0.167100869	0.009960159
lincRNA:chr2:218834305-218867070_R	0.16562507	0.012371427
lincRNA:chr5:77285769-77295769_F	0.164498333	0.03416495
LOC148696	0.164336877	0.044946119
ENST00000425471	0.160535106	0.024598534
A_33_P3245133	0.159765485	0.011825369
CENPC1	0.159046549	0.04096689
lincRNA:chr5:60464943-60477543_F	0.158069114	0.022837557
OR4C15	0.157501829	0.001471181
SCAMP4	0.157186255	0.005019841
lincRNA:chr21:44919572-44938621_R	0.157022047	0.025906564
lincRNA:chr3:128268891-128271202_F	0.15684402	0.046766266
lincRNA:chr8:130695693-130741043_F	0.156758	0.001629653
A_33_P3381762	0.156545389	0.046766266
lincRNA:chr8:58126971-58142821_R	0.155254082	0.045465917
lincRNA:chr16:50304099-50310474_F	0.153578536	0.04096689
AK001094	0.153506567	0.011366551
TTC16	0.153004701	0.04774716
lincRNA:chr4:90365352-90484827_F	0.151283236	0.019208771
THC2588995	0.145663788	0.043022672
lincRNA:chr8:107055124-107072278_F	0.144604743	0.011825369
PIP5K1B	0.143200673	0.042769078
A_33_P3292126	0.1372898	0.030733033
TTTY13	0.133554639	0.020473123
lincRNA:chr11:29331349-29342299_R	0.132774208	0.014976857
SLC34A1	0.132743358	0.007419368
lincRNA:chr9:33011575-33024725_F	0.131111996	0.042939093
lincRNA:chr7:105551628-105553084_F	0.129052582	0.047228006
XPC	0.128145436	0.028331048
lincRNA:chr18:48872618-49058541_F	0.126108776	0.020698683
lincRNA:chr8:128976068-129037395_R	0.126080904	0.016749195
lincRNA:chr1:101518537-101561912_R	0.125502036	0.02463236
lincRNA:chr8:124456094-124469294_F	0.124071864	0.001455784
lincRNA:chr3:136790035-136808485_R	0.119521247	0.027119106
KRTAP23-1	0.117511559	0.034130375
SLC13A2	0.115929932	0.037371278

B-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
lincRNA:chr15:79641195-79678020_F	0.114356699	0.046766266
H2AFB2	0.110543394	0.045154445
HEPACAM	0.105052649	0.011825369
TTLL11	0.10178753	0.046766266
CHST5	0.09810802	0.034130375
SLC38A8	0.095763956	0.019208771
PDCD1LG2	0.081657648	0.033068172
X56665	0.078917432	0.045154445
HPX	0.076574674	0.039376045
CRKL	-0.070871075	0.034376844
A_33_P3253857	-0.090196193	0.02916401
SLC22A4	-0.110054629	0.034658694
LOC100130887	-0.121437032	0.041547541
CD46	-0.164916787	0.017785031
C1orf112	-0.243277025	0.009447456
ATP2B4	-0.287486592	0.016599421
SNHG6	-0.322935502	0.035357178
CATSPER1	-0.32358229	0.042231468
lincRNA:chr8:67834170-67838228_F	-0.419228117	0.002051418
BBOX1	-0.423295305	0.044382792
ORM2	-0.754406777	0.045964782

Invasive Epithelial Cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
MIA	2.320053033	5.20E-05
SERPINA3	2.287009902	0.020306368
KRT6B	2.15149619	0.047729093
KRT23	2.104182337	0.025104492
SAA1	1.681427942	0.004984948
LAMC2	1.488050696	0.039911072
ROPN1	1.483906812	0.016303728
MT1G	1.359709082	0.039911072
CBS	1.295060015	5.20E-05
lincRNA:chr1:205404014-205407007_R	1.2535815	6.07E-06
IGF2BP3	1.080272296	0.034734973
MT1E	1.040984277	0.021804873
MT1F	0.990579063	0.004984948
MT1M	0.903658952	0.025870536
MT1L	0.833900771	0.021804873
MT1H	0.805295842	0.010868213
MT1B	0.783949899	0.020306368
MT1A	0.696983104	0.021804873
FRMD3	0.615819012	0.010868213
PTX3	0.550470419	0.025870536
C11orf9	0.517274815	0.021804873
NDUFA5	-0.253579484	0.039911072
WAPAL	-0.285653293	0.004984948
CTNND1	-0.309457244	0.021804873
LRBA	-0.362928096	0.004984948
NDUFAF1	-0.363760916	0.022344852
C12orf29	-0.460321175	0.03746099

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
CXCL9	2.139516119	5.87E-05
UBD	1.969430566	5.61E-08
NKG7	1.892015021	5.77E-09
CCL18	1.865147886	6.83E-05
GNLY	1.860953813	1.33E-07
ZNF683	1.78696826	8.38E-07
HLA-DQA1	1.757642541	0.005475434
IDO1	1.679267451	2.22E-07
GBP5	1.664691492	7.37E-08
GZMA	1.65474284	8.77E-10
CD2	1.574712393	1.13E-09
GZMH	1.540330199	1.25E-05
CXCR3	1.434825469	2.72E-10
ICOS	1.423626962	6.33E-06
CD3D	1.421210032	1.05E-10
CCL19	1.39803513	0.013990341
SIRPG	1.381595311	1.60E-09
KLHDC7B	1.355178098	0.000539038
CCR5	1.322297237	2.17E-11
CD52	1.314644954	4.57E-09
GZMK	1.308057017	5.61E-08
GPR171	1.306662947	2.62E-07
PVRIG	1.291280942	2.00E-09
MARCO	1.28932417	0.013011425
LAMP3	1.274108276	1.19E-05
CCR7	1.269844026	1.73E-05
IL18RAP	1.244349199	5.83E-07
GZMB	1.243544848	1.25E-06
CD8A	1.236542633	7.46E-10
IL2RG	1.219271249	1.08E-07
LGALS2	1.217765871	0.007390867
TIFAB	1.196571533	0.001323503
S1PR4	1.190953134	2.93E-07
AMICA1	1.184174622	0.000275842
LOC100508196	1.15950281	6.83E-05
ENST00000466254	1.154838041	9.37E-10
TBC1D10C	1.127571442	4.31E-07
SELL	1.116622454	2.35E-07
CD8B	1.101355138	5.95E-05

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
ENST00000390622	1.094093785	3.71E-05
RARRES1	1.093194915	0.033941338
RASGRP1	1.066627634	1.76E-07
A23747	1.065606935	2.72E-09
RAC2	1.050374031	1.47E-11
BATF2	1.046762584	0.001401365
CXCR2P1	1.030663069	0.000601885
GBP4	1.026767555	0.000356911
HLA-DPA1	1.004452887	0.002909403
STAT1	1.003030116	1.12E-05
BIRC3	0.972402234	2.46E-05
CXCL13	0.967951725	4.85E-05
CD40LG	0.966844772	2.40E-06
CCL5	0.958595387	8.38E-07
BATF	0.953148814	2.33E-05
FAM26F	0.950036441	8.89E-07
ENST00000390477	0.947147926	0.00015533
TRAF3IP3	0.947116072	1.22E-06
SKAP1	0.946162523	1.83E-05
C15orf48	0.92357029	0.007390867
PSTPIP1	0.921279323	0.000105715
ENST00000425189	0.912404498	0.002011439
IL7R	0.910971194	0.017064534
PATL2	0.906063016	5.29E-06
DENND1C	0.901574147	5.29E-06
RASAL3	0.897593522	7.25E-06
lincRNA:chr14:23018310-23025460_R	0.897435254	1.18E-07
LCK	0.892628187	2.48E-09
LOC100131733	0.88642743	7.23E-06
MIAT	0.886114285	0.0019702
C16orf54	0.879467837	1.21E-07
C1QA	0.878195877	0.000120226
HCST	0.87748847	1.05E-10
NCR3	0.870007925	0.000682857
IL21R	0.869353074	0.00070423
SAMD9L	0.86429737	3.24E-06
HLA-F	0.852137968	1.85E-05
IFI30	0.847781772	0.00116269
TAP1	0.838799882	9.70E-06

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
HLA-C	0.838055873	0.004381237
LCP1	0.836314668	1.73E-05
THC2502506	0.828717222	2.13E-05
IL10RA	0.825330284	2.62E-07
CD3G	0.823456995	1.08E-06
C1QB	0.814102547	2.60E-05
SH2D2A	0.808541952	2.37E-07
ORM2	0.794761465	0.043103307
ZBED2	0.793000569	1.04E-05
CTLA4	0.79256965	0.000857656
RARRES3	0.791585018	0.023267731
TMIGD2	0.790295137	0.00108626
OR10H2	0.789433901	5.33E-05
FPR3	0.782863403	0.001254777
RGL4	0.781581596	4.31E-07
CXorf65	0.780773687	0.005254664
ATP8A1	0.780761546	1.73E-05
SLAMF1	0.776052677	0.000134476
HLA-DOB	0.775966242	0.021977685
FYB	0.771643395	1.60E-09
FASLG	0.770054477	6.07E-06
PTPRC	0.767180035	0.001401365
HLA-DRB1	0.765132932	0.000827624
RTP4	0.76105925	0.014496118
CYTIP	0.760402605	5.23E-06
STAT4	0.753403487	1.08E-06
LAG3	0.748775608	0.000841256
PDCD1	0.748491814	0.001231425
TFEC	0.742423543	0.001470338
SERPINA1	0.740653307	0.024069081
NLRC3	0.735109231	2.68E-07
SPOCK2	0.733570678	0.000455915
IGFLR1	0.731859402	0.032773456
PPP1R16B	0.730991912	2.70E-07
WARS	0.723387976	5.56E-05
TMC8	0.723041239	0.000279929
JAK3	0.719559135	1.69E-05
WIPF1	0.715732418	0.001301367
lincRNA:chr14:23018310-23025460_F	0.710408611	0.001898614

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
INSL3	0.707480993	0.002409961
CD97	0.706850974	1.61E-07
DOCK8	0.703874338	3.59E-08
TNFAIP8	0.703095887	2.63E-06
CD4	0.696084473	0.001612455
SAMD3	0.695917291	4.17E-07
ARHGAP25	0.694219481	4.01E-06
SOD2	0.693798375	0.0081295
ISG20	0.693522036	0.00057949
ITGB7	0.6931622	0.003674371
BST2	0.689295665	0.023267731
GIMAP7	0.68895296	0.001908074
TIGIT	0.688362707	5.95E-05
LY9	0.686722141	5.56E-05
WDFY4	0.684987888	0.000739878
FGD3	0.683298298	0.000279929
HMHA1	0.681830465	9.70E-06
GIMAP4	0.678653513	0.009893057
ARHGAP9	0.671981931	2.59E-06
CYBB	0.670442505	0.001449207
SNX10	0.668633462	0.000221447
PSMB10	0.668550913	2.48E-09
SOCS1	0.667837378	0.000201876
CASP1	0.664602207	3.62E-05
LILRB4	0.664128055	0.001518847
CD27	0.65523303	0.003886183
RHOH	0.649557395	0.000765615
VPREB3	0.644199025	0.032265961
HLA-G	0.643251582	0.023079676
OAS2	0.639119661	0.046056202
ACSL5	0.638300271	0.000702692
TYMP	0.638028334	0.00286531
DENND2D	0.637774706	2.49E-05
ASB2	0.63275113	0.04445213
CD38	0.62705977	0.001004586
RGS18	0.623657816	0.000136011
SLC9A9	0.62356833	0.000740523
IRF7	0.621863169	0.016389196
SUSD3	0.620804523	0.025433964

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
HLA-DOA	0.618837023	0.017582852
LAP3	0.617680443	0.000449567
FCER1G	0.617373183	0.000783881
APOC1	0.613897759	0.04972821
GMFG	0.611176967	9.88E-06
IL32	0.609751081	0.000810805
CMPK2	0.609305925	0.022933684
TOX2	0.608436171	0.031636754
CD6	0.600495079	5.32E-05
LGALS9C	0.597679178	0.000513193
HCLS1	0.59590339	5.64E-05
SLAMF8	0.595117389	0.012885809
IRF8	0.592018932	0.001401365
ENST00000517927	0.591317282	0.003100875
LCP2	0.588063266	0.000608546
ARHGAP4	0.588041295	0.000682286
LILRB3	0.585731869	0.047838175
IL15	0.584059267	0.033024765
KLRB1	0.583577498	0.004868771
TNFRSF8	0.583287446	0.003053036
P2RY6	0.581616779	0.048736493
CXCR6	0.580179698	0.000476548
SLC2A6	0.578733926	0.04805169
FAM78A	0.578313796	0.000199006
FAM113B	0.57664248	0.002728343
NUAK2	0.57654064	0.002813175
ITGB2	0.576519666	0.01253306
SLAMF7	0.573111054	0.00051213
PRKCB	0.571194428	0.000735708
LOC439949	0.5668703	0.000198211
LAIR1	0.565784698	0.001703539
EMB	0.563586387	0.000419414
MYO1F	0.562500076	0.000151047
TXNDC3	0.561378882	0.011102966
CD28	0.561296222	0.009511801
GIMAP1	0.559139338	0.00030383
MNDA	0.559018488	0.025749671
HLA-DMA	0.555823462	0.007485655
TLR1	0.554900681	0.003416834

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
INPP5D	0.553745137	1.98E-05
ABI3	0.553655361	0.023821712
AIF1	0.552482822	0.000120226
LRMP	0.550375747	0.000682286
TAP2	0.548599086	0.026830772
PSME2	0.547852063	0.000181491
PYCARD	0.543134875	0.000268881
DAPP1	0.543106513	0.003613365
B2M	0.540259229	0.04394629
HLA-DMB	0.539485064	0.019067719
HLA-E	0.536561895	0.00199459
HLA-A	0.526958213	0.030326702
CD79A	0.526501899	0.021048921
FLJ32255	0.526122267	0.036368787
GIMAP6	0.525471062	0.040282341
PTK2B	0.523942182	0.000236913
FAM20A	0.523286695	0.030326702
DHX58	0.518742778	0.007123335
PRIC285	0.515366867	0.036539587
MCOLN2	0.512746905	0.02498066
UBASH3A	0.512053254	0.000400731
C12orf35	0.511460034	0.005003397
VAMP8	0.511153502	0.009873334
SCO2	0.511056299	0.005003397
NR1H3	0.509996682	0.02652472
ARHGDIB	0.508993547	0.000230203
LST1	0.5084503	0.041824674
CLIC2	0.505473483	4.12E-05
HLA-DRB3	0.501930891	0.001420266
ICAM1	0.501443746	0.013002138
SLA	0.497124746	0.013003062
TRIM7	0.496704012	0.04478332
RLTPR	0.490483539	0.000614709
CYTH4	0.490409614	0.00243601
KLRC1	0.490307972	0.024540903
GPSM3	0.489871416	3.61E-05
MS4A4A	0.486742013	0.035232256
LOC100132707	0.485606208	0.041165814
PPP2R2B	0.485079231	0.016408526

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
lincRNA:chr6:33091097-33099422_R	0.483041241	0.040030506
MMP25	0.4814842	0.006492563
DOCK10	0.481214522	0.003259531
GSTK1	0.480371947	0.019767583
PLAC8	0.480031364	0.001296974
JAK2	0.47773593	0.002450424
NFKBIE	0.476420296	0.005479488
MFNG	0.475659824	0.003118164
APBB1IP	0.474250763	0.017778975
TLR7	0.473839679	0.00301502
PARP10	0.469544053	0.020662461
SP140	0.467483563	0.015944644
VNN2	0.463548561	0.033133354
GPR155	0.460287511	0.00954799
CORO1A	0.459810607	0.046632115
NCEH1	0.459027543	0.001652497
PTPN7	0.458725173	5.56E-05
FCHO1	0.456974071	0.000532479
TYROBP	0.448277263	0.016555252
PRKCH	0.4449266	0.028863207
KLRC4	0.442354844	0.013696512
TNFRSF14	0.441149588	0.000689211
CIITA	0.44027343	0.038864409
RNASE6	0.439465119	0.005112366
VAMP5	0.437337726	0.043772581
NMI	0.434928072	0.002566889
IL2RB	0.433190515	1.83E-05
IDO2	0.432966626	0.02366632
MPEG1	0.432953212	0.012580665
TRIM22	0.429248702	0.003255135
CYFIP2	0.428341608	0.002902279
TNFAIP8L2	0.425972572	0.008993052
LINC00426	0.423850335	1.69E-05
CDC42SE2	0.422359598	7.91E-05
GBP2	0.419674561	0.000400731
IFIT5	0.419174994	0.021885502
TNFAIP2	0.414592743	0.03987753
SASH3	0.41422149	0.000185305
LAT	0.41391962	0.000449567

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
KIR2DS2	0.413392316	0.003053036
AV659465	0.411542111	0.03822468
NFKB1	0.410372555	0.029010967
SH2D1B	0.408640035	0.010095257
TXK	0.405687527	0.045611011
MLKL	0.402545677	0.011682649
GIMAP5	0.401137263	0.013990341
LTB	0.400639709	0.031494059
LOC219731	0.398718589	0.049794867
RIMBP3	0.397773445	0.018269025
PLCB2	0.396570467	0.040178773
PLEKHA2	0.395527251	0.000783881
RHEBL1	0.393654138	0.024340177
TNFRSF25	0.393626991	0.027539506
C17orf87	0.393593611	0.000539038
GLRX	0.393096457	0.028600846
NR3C1	0.391361982	0.022456811
VMA21	0.390774186	0.010538684
WDR67	0.389047353	0.040805369
EFHD2	0.387983578	0.003370072
RAB37	0.385435212	0.012139801
ZBP1	0.385185221	0.003497863
SARDH	0.384395106	0.034208639
FYN	0.38383282	0.040030506
THC2601170	0.382798524	0.000759866
PTPN22	0.382021796	0.000400731
OBFC2A	0.381037307	0.001449207
CHST12	0.378504381	0.00387389
lincRNA:chr1:89874262-89947412_R	0.376092342	0.006049508
TNFSF13B	0.373595944	0.030184844
RASSF4	0.356176553	0.034081117
ARAP2	0.352181828	0.002516393
CASP10	0.351378131	0.024340177
AIM2	0.350488863	0.001676112
CD69	0.348970495	0.002748052
CXCL11	0.348852708	0.002382932
FLJ35776	0.348408828	0.019067719
ST8SIA4	0.34565134	0.000682286
PREX1	0.337572696	0.005457181

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
ITM2B	0.336585989	0.029033716
MYO7A	0.334438281	0.033796483
RPS6KA3	0.331211585	0.004584005
USF2	0.331200731	0.047977164
LPXN	0.330972806	0.001259872
MPP1	0.326821964	0.014272003
STK4	0.326362066	0.00980755
LOC100507429	0.32355073	0.008695999
PPP3CC	0.320882998	0.008791236
TNFRSF10B	0.319883972	0.040805369
ZNF831	0.319177449	0.009772069
SGTB	0.316557425	0.039793192
CD53	0.31458335	0.023115671
FAM65B	0.312778378	0.000685636
FLI1	0.311673281	0.019778736
PTGER4	0.311472786	0.039657076
TREX1	0.30963184	0.011821306
TRAFD1	0.30844417	0.023896809
LEPROTL1	0.308079223	0.01292431
SERPINB9	0.305987162	0.009013992
OAZ1	0.301188468	0.004050529
GPR18	0.30037311	0.001259872
GBP3	0.299744644	0.005614489
RNASEH2B	0.297958429	0.006225539
CD5	0.297872461	0.001565606
RNF19A	0.291441775	0.023115671
UBR1	0.286440766	0.016408526
NECAP2	0.285619779	0.003975998
CCNDBP1	0.283892851	0.016810486
AHR	0.260724002	0.049179016
AK091525	0.255770209	0.018269025
BANK1	0.254593617	0.030184844
PPP1R18	0.254018461	0.047697992
ANKRD22	0.247387748	0.034253664
XRN1	0.247301873	0.031268657
MICB	0.246624997	0.023300626
CD96	0.235811101	0.022301871
TLR3	0.225860531	0.000706702
GAB3	0.220224996	0.000900197

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
SPPL2A	0.21941485	0.046056202
RFFL	0.218005063	0.023677283
GLCCI1	0.216553428	0.023107773
NUB1	0.212938529	0.044746285
BTLA	0.205993957	0.026174873
LINC00324	0.205991253	0.016027504
STAT5B	0.205537737	0.030411791
CXCL10	0.196510218	0.009328456
STYK1	0.193194343	0.036702374
RTKN2	0.192684516	0.015426372
SP110	0.186959673	0.01837062
ITGA4	0.185564238	0.003579361
HERC1	0.170710674	0.010678154
ITGAL	0.170157846	0.032282454
CTSW	0.165305309	0.029956494
RASSF5	0.164931217	0.014496118
CYLD	0.161554297	0.04746914
KAT2B	0.138947059	0.009491283
GBP1P1	0.117178049	0.043356842
CD48	0.080994861	0.023300626
lincRNA:chr8:107282463-107284434_F	-0.086812108	0.019767583
GPR32	-0.096540018	0.007485655
lincRNA:chr11:72855977-72913777_R	-0.112449972	0.047838175
TNKS	-0.115856548	0.022338145
TTLL11	-0.118684426	0.017778975
FAM114A1	-0.133261501	0.029674634
lincRNA:chr10:131977935-131987135_F	-0.13833648	0.029081773
lincRNA:chr1:85933672-85934049_R	-0.140916443	0.007564397
A_33_P3362548	-0.146057906	0.04592172
SPIN1	-0.150586053	0.040178773
lincRNA:chr7:105551628-105553084_F	-0.16004247	0.009293247
TTC3	-0.1728851	0.023169641
RAI14	-0.189181205	0.018269025
ZNF782	-0.208115197	0.017310536
lincRNA:chr6:3708951-3719951_R	-0.214787937	0.006134935
ARHGEF7	-0.221048736	0.025091853
C3orf75	-0.224834613	0.043341985
IL17RD	-0.224957775	0.022473303
GTF2IRD1	-0.230333664	0.043631101

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
MFAP2	-0.232331084	0.002343686
PTGER3	-0.235366324	0.007608077
WASL	-0.245495665	0.027356516
WASF1	-0.279085701	0.026479075
KAT8	-0.28246554	0.014071411
FHOD3	-0.294040308	0.047697992
lincRNA:chr12:54523008-54559358_R	-0.299039159	0.028863207
GLRX5	-0.300633705	0.002745854
ARMCX2	-0.306123563	0.042841914
ARHGEF37	-0.308212434	0.023113721
COQ7	-0.315739832	0.033278899
SORT1	-0.337950416	0.040030506
PPP1R13B	-0.339304269	0.046632115
RNF24	-0.342358403	0.047977164
OSBPL10	-0.346109213	0.024780211
THAP4	-0.363980357	0.001385874
CYTH3	-0.373500683	0.027734719
FZD1	-0.37467226	0.015728638
FGFR2	-0.375523113	0.033359772
FERMT2	-0.386832134	0.026023354
DPCD	-0.395027881	0.032773456
LGALSL	-0.396782486	0.037364604
PTGFRN	-0.404190336	0.032694573
TRIM45	-0.406166404	0.023821712
KIAA1217	-0.406756842	0.042067742
SIX5	-0.40931717	0.0124702
BPHL	-0.409480863	0.047838175
CTSF	-0.410505347	0.045748097
ITGB5	-0.41160257	0.031432571
FAM69B	-0.41530026	0.024962379
SMTN	-0.41673699	0.046632115
FNBP1L	-0.417243173	0.035232256
TRO	-0.420031976	0.0124702
ALDH7A1	-0.42013283	0.018269025
ZCCHC14	-0.42785925	0.007997362
GLI3	-0.434755055	0.046056202
FV367791	-0.44519922	0.023677283
GPR125	-0.44973623	0.009772069
CMTM4	-0.456979293	0.045435078

T-cells		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
MMP2	-0.465645293	0.035232256
CLEC11A	-0.466617823	0.019067719
FAP	-0.470845088	0.008695999
LEPREL4	-0.475567397	0.023079676
CHKA	-0.478590219	0.016810486
SNURF	-0.482419586	0.014655544
TMEM98	-0.495980526	0.013696512
CERCAM	-0.497456802	0.019767583
TGFB1I1	-0.535149824	0.024069081
AEBP1	-0.540743928	0.016741858
LONRF2	-0.541192638	0.028863207
PARD3	-0.544646335	0.022064564
SPON2	-0.572273316	0.047697992
TMEM25	-0.588995456	0.03360061
EPCAM	-0.595012315	0.033941338
THBS2	-0.59870053	0.027539506
PXDN	-0.607636254	0.002516393
KHDRBS3	-0.691829728	0.01021203
SRPX2	-0.692416535	0.000740523
ZNF541	-0.692890005	0.018464555
PNMAL1	-0.703233719	0.034806621
HILPDA	-0.717463453	0.000265993
SHANK2	-0.725387371	0.004083918
LEPREL2	-0.739410552	0.000532479
ENST00000372591	-0.772200789	0.025749671
CNIH3	-0.784126418	0.005832888
SYT7	-0.866317294	0.015558068
SPP1	-0.981227914	0.032939557
TET1	-1.070259359	0.001433068

Desmoplastic stroma (D)		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
CXCL14	1.964953007	0.001594263
COMP	1.827539439	7.25E-05
COL10A1	1.409766974	0.001283756
CILP	1.320262092	0.022913059
F13A1	1.306128699	0.007853826
MFAP4	1.284543691	0.000909989
FBLN1	1.248319033	0.001061442
DCN	1.222102956	0.000454777
COL12A1	1.19710349	0.000467466
WISP2	1.196053412	0.015606989
KANK4	1.163531347	0.046833258
COL8A2	1.139929998	7.65E-07
NKD2	1.136097134	0.021782282
ITGA11	1.12621455	0.000266049
COL8A1	1.11783849	0.000273848
ADRA2A	1.107365656	0.002628289
lincRNA:chr3:112308735-112318605_R	1.085394386	0.004664713
TIMP3	1.083855611	7.27E-05
SSC5D	1.07845779	0.003621717
HTRA1	1.076282758	0.000266049
CDH11	1.066934007	2.38E-05
lincRNA:chr3:112315643-112316945_R	1.057783056	0.003190571
LUM	1.025421985	0.000218907
C1orf151-NBL1	1.020964502	8.88E-05
MFAP5	1.006128534	0.014713187
OGN	0.989122398	0.002965255
DACT3	0.978606891	0.000190629
SYNDIG1	0.972903895	0.000895093
THBS2	0.968476624	0.000569621
SPARC	0.952382976	0.002628289
PALM	0.949857292	0.002777819
FNDC1	0.944748653	0.001016
COL1A2	0.939687209	0.00075133
CTSK	0.925107621	0.002041941
COL5A1	0.914967352	0.001037315
PTPRD	0.914024909	7.25E-05
COL16A1	0.912583718	0.000481831
TAGLN	0.908591067	0.011767443
FAM155A	0.897912723	0.011137795

Desmoplastic stroma (D)		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
CHRD	0.890964208	0.001085786
COL5A2	0.888841088	0.001752753
PPP1R3C	0.888517849	0.000944426
PDGFRL	0.874027095	0.00012541
COL1A1	0.869343357	0.001037315
AEBP1	0.85844977	0.000394681
MMP2	0.845625553	0.000218907
C1S	0.84555879	0.000394681
PPAPDC1A	0.83518839	0.002777819
CFH	0.828294693	0.000421061
MGC24103	0.811175238	0.011137795
COL3A1	0.806772181	0.002118081
ADAMTSL2	0.798715748	0.030854368
OSBPL5	0.78623637	0.010203199
ANTXR1	0.784796536	0.011137795
RASL11B	0.78279096	0.000827911
PCSK5	0.7775006	0.010782031
MXRA5	0.774002762	0.001547409
STMN2	0.773375388	0.021672786
FV367791	0.761135703	0.000245264
MRC2	0.761015725	0.000217452
SFRP4	0.758165976	0.025957206
SNED1	0.756108722	0.006843101
FMOD	0.75542576	0.002041941
LEPREL2	0.753787751	0.003190571
VCAN	0.74734068	0.000481831
CCDC80	0.742676413	0.013899723
FBXL7	0.739595233	0.002041941
EFEMP2	0.736199164	0.000116695
SRPX2	0.730181943	0.002804915
C1QTNF6	0.728190486	0.001283756
TIMP2	0.723279301	0.011524326
PRRX1	0.717851258	0.001005976
lincRNA:chr2:100851143-100863413_F	0.707525231	0.000394681
VSTM4	0.707263796	0.038023676
RAB31	0.700524948	0.001112703
THY1	0.700519439	0.00538503
lincRNA:chr2:216585154-216585719_F	0.690044555	0.013601144
CERCAM	0.689741455	0.00236183

Desmoplastic stroma (D)		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
HMCN1	0.683303011	0.000915054
CLEC11A	0.681966624	0.001112703
KIAA1217	0.679428936	0.000752925
NUAK1	0.678199164	0.038664568
C12orf70	0.677555846	0.011137795
SULF1	0.674476465	0.027326014
GPX8	0.672356794	0.000245264
ST6GAL2	0.672053302	0.002041941
SPON1	0.669813939	0.021672786
NOX4	0.668034894	0.010782031
MRVI1	0.667301017	0.012151582
ISM1	0.66408556	0.013588165
PLAT	0.658633932	0.002777819
FAM198B	0.653972215	0.008080131
ZFHX4	0.645246954	0.000342713
SYNC	0.639494092	0.000337148
SSPN	0.638835261	0.011911451
PRICKLE1	0.63151806	0.011524326
CTHRC1	0.630425267	0.015066331
MIR100HG	0.630170959	0.001005976
ZEB1	0.627411221	0.001061442
DPYSL3	0.622570146	0.014298774
GLI3	0.621226423	0.005495086
ITGBL1	0.620186004	0.003693311
CYS1	0.618599744	0.000782721
ZNF503	0.618171084	0.01380618
PDGFRB	0.616871735	0.041900257
FHL1	0.60956728	0.012286873
FIBIN	0.607013268	0.020248282
CYTH3	0.604873681	0.000569621
RECK	0.604379453	0.001061442
C1R	0.601426464	0.012151582
RARRES2	0.596442155	0.011767443
HHIPL1	0.591719915	0.043390524
DACT1	0.588906433	0.025957206
TMEM200A	0.588724591	0.001283756
MSRB3	0.585492801	0.013899723
PPIC	0.584846956	0.001406862
KANK2	0.582511038	0.008237649

Desmoplastic stroma (D)		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
SGCD	0.581828282	0.002118081
GPR124	0.571822163	0.012286873
FHOD3	0.570182726	0.00013999
MFAP3L	0.567242349	0.002804915
EMILIN1	0.565840231	0.021400242
HOXC6	0.564377244	0.003888844
lincRNA:chr14:96548697-96561747_F	0.560492413	0.006763746
CTSO	0.558540298	0.002628289
GPC6	0.557988474	0.044187966
LRRC15	0.555532993	0.000394681
PPAPDC3	0.549995206	0.005588328
GREM2	0.549986218	0.04425057
ASPN	0.547549272	0.043123479
XG	0.538170458	0.001061442
PMP22	0.535647296	0.038865831
PDGFC	0.529065488	0.000782721
LOXL1	0.524965988	0.000116695
C14orf132	0.524102998	0.004013811
IFI27L2	0.521835491	0.008341379
GLT8D2	0.52150705	0.000394681
MARVELD1	0.520400355	0.004385737
CYBRD1	0.51884758	0.005970936
lincRNA:chr14:96507172-96661947_F	0.517510657	0.012789429
LOC642361	0.515327541	0.010782031
CTSF	0.5081671	0.022663317
FKBP9	0.506562589	0.010782031
FAP	0.505828995	0.016183318
C1QTNF5	0.502519327	0.036922225
AKAP13	0.493280647	0.001444762
HSPA12A	0.491580707	0.044027757
TSHZ3	0.486649798	0.015451329
CARD6	0.482963615	0.034633694
C5orf62	0.474217651	0.021672786
ITGB5	0.466077344	0.034267078
FZD1	0.462028908	0.007506771
GXYLT2	0.461222067	0.04207749
ALKBH7	0.460130026	0.026297168
ABCB4	0.458142849	0.009225625
RUNX1	0.456503406	0.010782031

Desmoplastic stroma (D)		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
ZCCHC14	0.450585835	0.019003733
A_33_P3413997	0.450569245	0.025957206
LOC400236	0.444531031	0.015606989
DAPK3	0.444233857	0.042017256
CRAT	0.441170722	0.022913059
TRO	0.440570086	0.028422003
PDLIM5	0.440514474	0.046833258
FERMT2	0.432583315	0.031793089
BNC2	0.419417922	0.004201866
P4HA2	0.41445027	0.043341122
LIMA1	0.411118237	0.002804915
CYP2U1	0.411110638	0.014254277
PTN	0.406057718	0.02480253
CLMP	0.398972358	0.03321986
SFRP2	0.398440932	0.03696446
CMTM3	0.396942522	0.040862309
15-09-11	0.396023957	0.009012745
A_19_P00325768	0.394051985	0.015606989
UBR1	0.390950088	0.002246154
XLOC_001952	0.38712735	0.032178083
JDP2	0.383940379	0.004521294
C9orf3	0.382586363	0.00591145
CRY2	0.381052938	0.015866857
ECM2	0.380593464	0.014254277
SPOCK1	0.380380376	0.049883397
MVP	0.373898397	0.030603087
MYOF	0.372522336	0.014227493
MAGI2-AS3	0.364839684	0.014164032
lincRNA:chr12:46826133-46974783_F	0.363150787	0.045364578
PRAF2	0.359869116	0.045431558
lincRNA:chr3:114031960-114042235_R	0.337537826	0.039162916
KDELC2	0.336342667	0.049864996
GALNT10	0.332409391	0.024179081
DAB2	0.312341171	0.033681743
A_33_P3218564	0.31141921	0.034010287
PCOLCE	0.309031949	0.030854368
TLN2	0.304007202	0.027324495
GBP3	0.301532187	0.022761455
CILP2	0.299068098	0.046833258

Desmoplastic stroma (D)		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
SORCS2	0.297293263	0.015294362
IGFL2	0.291279505	0.03032998
SNTB2	0.281470467	0.008866293
TRPC1	0.281075662	0.020146363
COPZ2	0.274767736	0.003362133
RNASEL	0.273102827	0.015323055
ARL2BP	0.267211925	0.008080131
ZFPM2	0.259766999	0.013899723
ZFHX3	0.252060581	0.040133476
LOC283867	0.250248854	0.029467846
LPAR1	0.238708633	0.044458153
SEC14L2	0.194083745	0.030854368
RTN2	0.183198242	0.022761455
PRR5L	0.18302381	0.032876283
C14orf56	0.178886022	0.033015661
PRRX2	0.177941544	0.006699127
lincRNA:chr1:85933672-85934049_R	0.156412163	0.011524326
LOC727993	0.124437114	0.013077003
FGF4	-0.090045389	0.046833258
IL17F	-0.094160377	0.044027757
AW593215	-0.102584455	0.03027083
RNU4ATAC	-0.110298475	0.014673179
AB305952	-0.121540596	0.005495086
A_33_P3248077	-0.130189258	0.040895176
CCAR1	-0.139924274	0.049275705
TRDMT1	-0.140129883	0.049350435
LOC401588	-0.156148932	0.013899723
MAPKAPK5	-0.18259776	0.022761455
RAVER2	-0.192651748	0.011524326
MAPKBP1	-0.196999559	0.0361124
MELK	-0.21027299	0.005867391
BRCA1	-0.212101557	0.017503303
CBLL1	-0.218636987	0.035087692
AURKA	-0.229102799	0.015323055
YY1AP1	-0.234773719	0.010434893
EEF1E1	-0.236798751	0.020146363
DHX9	-0.241223626	0.019285026
WDR3	-0.241595384	0.015589714
CWC27	-0.243006039	0.044991597

Desmoplastic stroma (D)		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
GPT2	-0.253473709	0.006843101
ELAC1	-0.256056912	0.044907848
FAM83D	-0.282610203	0.035087692
ENST00000440540	-0.283524671	0.025645253
LRR1	-0.291530162	0.022901507
CDK1	-0.305734401	0.01298707
TDG	-0.313227353	0.039162916
CEP250	-0.313391953	0.015606989
LTV1	-0.325359955	0.044027757
NCL	-0.326807968	0.018351011
FBXO5	-0.328030858	0.019285026
UTP3	-0.332313481	0.006843101
SMC4	-0.338448788	0.011524326
SLC7A11	-0.33873362	0.015451329
CHAF1A	-0.339563936	0.013899723
DUSP12	-0.340010047	0.002162944
NUP107	-0.379819152	0.012286873
FDPS	-0.381649085	0.030854368
DNAJC2	-0.382433797	0.015981712
HJURP	-0.39274995	0.015606989
PDCD2	-0.394617995	0.046896851
USP1	-0.405679545	0.023712516
LOC152217	-0.408174626	0.023357417
SLC25A3	-0.408287848	0.040405414
SF3A3	-0.412402833	0.04363653
TIMELESS	-0.414528536	0.01433629
SNRPF	-0.415504138	0.015606989
MAGOH	-0.422082561	0.046896851
NASP	-0.424600742	0.011137795
CDC7	-0.426127953	0.040895176
PFDN2	-0.437870713	0.007506771
HNRNPA3	-0.438122484	0.018165483
PAXIP1	-0.451106264	0.006240825
NUP205	-0.453051284	0.001507005
GMPS	-0.462769843	0.020115294
TXN	-0.465220551	0.04363653
RRM1	-0.467760592	0.017194183
TRIM24	-0.478323148	0.004013811
KIF2C	-0.487451153	0.035087692

Desmoplastic stroma (D)		
Gene	**log2 Fold Change**	**FDR Adjust p-value**
TFDP1	-0.487506755	0.021672786
DKC1	-0.494389901	0.010782031
PTTG1	-0.495748139	0.046833258
PDHA1	-0.496830255	0.025904805
EZH2	-0.508642116	0.006843101
HAUS8	-0.517230281	0.03716752
NUSAP1	-0.522965666	0.022913059
FAM189B	-0.526020435	0.003987493
LMNB1	-0.526202194	0.001098792
PPIF	-0.528833459	0.009226491
HSPA14	-0.540921317	0.002118081
UBE2T	-0.546132917	0.030854368
CDCA5	-0.54613389	0.03518804
PAICS	-0.546966857	0.000895093
EBNA1BP2	-0.551771184	0.035087692
TYMS	-0.607822544	0.027439312
UNG	-0.61036562	0.008470762
GTPBP4	-0.623201456	0.004431711
UBE2C	-0.645372322	0.01433629
MCM7	-0.698536661	0.030854368
C10orf35	-0.704736936	0.012286873
STMN1	-0.71700265	0.042725898
RAD54L	-0.750315331	0.001005976
CDCA8	-0.843093687	0.029467846

Table 2: Differentially expressed genes between patients deemed low versus those deemed high for each stromal property (LIMMA, FDR adjusted p < 0.05 after ROI95)

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
DDOST	1650	10.695509	7.077294456	32.99157344	9.26E-09	7.65E-05
WDR7	23335	10.650178	4.42970173	16.44545715	5.01E-05	0.003107059
ALDH2	217	10.282923	5.209429091	23.94108214	9.93E-07	0.000532617
MRPS18A	55168	10.014386	3.810064536	20.31108552	6.58E-06	0.001265698
KATNA1	11104	9.9954849	4.285648497	15.55863519	8.00E-05	0.00400272
ACOT7	11332	9.9918353	5.097069987	26.27153665	2.97E-07	0.000327043
HMCES	56941	9.8300971	4.620531094	20.49762775	5.97E-06	0.001204152
BPGM	669	9.7841638	4.555812814	19.7075178	9.02E-06	0.001387666
PPIH	10465	9.760565	4.706831089	18.96249849	1.33E-05	0.001582074
NT5DC1	221294	9.6371352	5.500055048	24.93024868	5.94E-07	0.000427417
ZNF511	118472	9.552955	4.522318221	14.73614695	0.000123653	0.005024493
NT5DC2	64943	9.4783187	3.466538204	21.43637805	3.66E-06	0.000975649
LGALSL	29094	9.4543945	6.025967366	20.21523007	6.92E-06	0.001285872
NBAS	51594	9.4445886	6.017064143	16.39792253	5.13E-05	0.00312437
YIPF1	54432	9.4170809	4.890424043	18.81946187	1.44E-05	0.001605676
KYNU	8942	9.4141009	6.934894853	29.34239639	6.07E-08	0.000170277
HIBCH	26275	9.3982454	4.799649121	19.6414347	9.34E-06	0.001391891
MAVS	57506	9.3455803	4.133417518	14.59536036	0.000133242	0.005233142
NRSN2	80023	9.3260042	4.119691624	13.59088622	0.000227286	0.006772722
ANKS6	203286	9.3098514	3.192355195	18.14990893	2.04E-05	0.001940641
TIMM22	29928	9.2836042	4.865709196	20.6450433	5.53E-06	0.001163955
SLC25A1	6576	9.2592543	3.581369988	16.47548396	4.93E-05	0.003075646
LXN	56925	9.2222074	4.285350292	13.09919503	0.000295423	0.007767409
BDH2	56898	9.1840593	3.260668192	17.86605769	2.37E-05	0.002084939
ARG2	384	9.1551527	4.806681114	13.48704157	0.000240217	0.006909048
RGS14	10636	9.0274379	3.977988717	17.1128106	3.52E-05	0.002611089
NRM	11270	8.9546785	3.762018486	13.97522288	0.000185236	0.00613016
ARMCX6	54470	8.9449807	4.341179663	16.17203892	5.78E-05	0.003356438
NINJ1	4814	8.9425707	4.250956335	16.17668955	5.77E-05	0.003356438
SLC38A10	124565	8.8603547	5.126208581	17.34642089	3.11E-05	0.002429872
MRM2	29960	8.8472691	5.931213422	21.71661498	3.16E-06	0.000915808
EDNRA	1909	8.7827804	2.816929666	14.164435	0.000167507	0.005888917
ZNF615	284370	8.7655749	4.460756933	12.3571335	0.000439305	0.009546939
CREG1	8804	8.7038153	4.184452307	18.35589885	1.83E-05	0.001803929
SLC23A2	9962	8.6643641	3.866747446	14.25637848	0.000159519	0.005760108
HDDC3	374659	8.6567356	3.307964844	12.93887887	0.000321828	0.008052022
DIAPH2	1730	8.6500952	3.535009152	17.88575605	2.35E-05	0.002080292
IDH1	3417	8.6484292	5.746365475	18.77442686	1.47E-05	0.001633008
SQRDL	58472	8.6346391	5.206002294	17.90323109	2.32E-05	0.002080292
AGPAT1	10554	8.6119874	3.947901796	14.49308812	0.000140675	0.005334681
SEC23B	10483	8.5932009	6.806819792	28.74955159	8.24E-08	0.000170277
DHRS4L2	317749	8.5621898	2.505768725	17.22468443	3.32E-05	0.002542666
CDC20	991	8.4919468	3.627446142	12.34620652	0.000441883	0.00957781
FAM19A3	284467	8.4371737	2.914429385	12.6610544	0.000373352	0.008721029
NEO1	4756	8.4325475	4.760423338	13.67386524	0.00021746	0.006637216

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
MYBPC1	4604	8.3710099	5.785875512	19.84624987	8.39E-06	0.001387666
HHLA3	11147	8.3358462	3.36129162	13.53115502	0.000234636	0.006831696
ARFRP1	10139	8.3065409	4.594101986	15.35540521	8.91E-05	0.004184572
C4orf46	201725	8.2966485	2.776566706	16.03697001	6.21E-05	0.003451273
ZFP90	146198	8.1583733	4.713895388	17.69547473	2.59E-05	0.002198643
CDCA7	83879	8.1487554	4.093690381	13.33608524	0.000260347	0.007260748
TRAF7	84231	8.1043299	2.689152336	15.37903344	8.80E-05	0.004168092
GBP2	2634	8.1001272	4.328756362	12.22984931	0.00047031	0.009971777
ZSCAN12	9753	8.0663942	5.020359961	13.90940598	0.000191836	0.006257563
ZMIZ1	57178	8.0427978	2.138215175	14.78467309	0.000120511	0.004982528
BNIP3L	665	8.0298146	6.388779987	22.29164613	2.34E-06	0.00088036
C18orf25	147339	7.9359115	4.533841666	14.54639138	0.00013675	0.005262231
C20orf24	55969	7.9069412	5.06700578	22.53969052	2.06E-06	0.000851071
TNFAIP1	7126	7.9044745	6.295749032	21.92198559	2.84E-06	0.000883454
ACOT8	10005	7.9003237	1.868422754	13.82639011	0.0002005	0.006436729
SORT1	6272	7.8611044	3.568105034	15.27833287	9.28E-05	0.004262059
MT1L	4500	7.8409738	2.980382486	12.30753864	0.000451132	0.009714621
ZNF236	7776	7.7203691	3.465985247	13.20026494	0.00027991	0.007649749
ALAS1	211	7.7140479	5.273724744	13.78856634	0.000204577	0.006518887
ASCC1	51008	7.6905704	5.339691769	16.07635813	6.08E-05	0.003422289
FBXO34	55030	7.6283281	4.586497419	12.35988269	0.000438658	0.009546939
FH	2271	7.5664088	3.451258709	13.44032513	0.000246273	0.006974078
OLFML2A	169611	7.5314584	3.887230048	13.06395645	0.000301033	0.007840129
COMMD10	51397	7.529175	4.612900737	12.27225663	0.000459742	0.009828777
CRYZ	1429	7.5169434	6.447027484	19.51853044	9.96E-06	0.001435744
COPG1	22820	7.501829	5.359857867	16.4245077	5.06E-05	0.003107059
KLF13	51621	7.4800816	5.0444239	16.68242224	4.42E-05	0.002877127
DHRS4	10901	7.473421	2.026063775	19.99572724	7.76E-06	0.001387666
INHBB	3625	7.4445967	4.010754015	17.91844912	2.31E-05	0.002080292
SASS6	163786	7.4038038	2.496373679	15.79402257	7.06E-05	0.003711019
MFSD5	84975	7.3432382	3.472276288	18.36492754	1.82E-05	0.001803929
RNASEH1	246243	7.3115213	4.481678005	18.85649978	1.41E-05	0.001596371
CLDN12	9069	7.223355	3.893543323	13.67342219	0.000217511	0.006637216
POGLUT1	56983	7.1872656	5.457975054	21.28499479	3.96E-06	0.001022818
C1orf53	388722	7.1788274	2.457152497	12.87703404	0.000332639	0.008222995
VAMP3	9341	7.1752011	5.548264393	20.67100766	5.45E-06	0.001163955
ZNF623	9831	7.1714018	5.514408816	12.42946008	0.000422614	0.009343839
RCC2	55920	7.1510278	4.772811023	18.75190371	1.49E-05	0.00164028
TMEM248	55069	7.0610501	6.428830481	13.07591272	0.000299117	0.00782182
DOCK1	1793	7.0538308	3.420009898	13.13467579	0.00028988	0.007707464
C1orf115	79762	7.0383719	1.62822834	12.77483983	0.000351313	0.008426237
SETD3	84193	7.0373712	5.405679477	19.87321467	8.28E-06	0.001387666
ARFIP2	23647	6.9684731	5.295510082	16.97692135	3.78E-05	0.002662756
SDHAF2	54949	6.8237268	4.595171965	12.99744398	0.000311916	0.007936115
CHD1L	9557	6.8178147	6.289398681	14.53696881	0.000137436	0.005262231

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
WDR54	84058	6.810171	4.977999642	12.89132676	0.000330108	0.008197195
FAAH2	158584	6.808084	4.376155353	14.07658266	0.000175515	0.005968147
EPRS	2058	6.7934027	7.458495149	19.71695384	8.98E-06	0.001387666
SDHAF4	135154	6.7918664	4.29305006	14.64053729	0.000130086	0.00518401
SMIM19	114926	6.7567787	3.219298022	13.57537875	0.000229172	0.006783202
GLI3	2737	6.7502244	2.250531001	13.10720119	0.000294163	0.007758955
ATP13A3	79572	6.7045699	5.337711103	12.87069663	0.000333767	0.008226296
ZDHHC13	54503	6.6795098	4.985056265	14.95373487	0.00011018	0.004684197
RWDD3	25950	6.650587	2.091039207	14.07258057	0.000175889	0.005968147
WLS	79971	6.6324449	0.729913871	15.82583291	6.94E-05	0.003704931
PNP	4860	6.6273936	6.499320607	21.97522185	2.76E-06	0.000883454
ATP6V1E1	529	6.6134959	5.762525849	27.66646433	1.44E-07	0.000179292
VEGFC	7424	6.6111029	0.905467622	13.14010693	0.000289041	0.007707464
SLC36A4	120103	6.6008775	4.41924514	12.47503187	0.000412428	0.00915534
COMT	1312	6.578476	5.826487064	25.60911469	4.18E-07	0.000384104
ECH1	1891	6.5559928	6.75907346	25.24467699	5.05E-07	0.000417574
VRK2	7444	6.4698052	3.60969523	15.44974352	8.47E-05	0.004105641
KRT14	3861	6.4464354	7.264603984	25.08993842	5.47E-07	0.000427417
ZNF426	79088	6.441064	3.427620944	13.59186582	0.000227168	0.006772722
RBBP8	5932	6.4365148	6.281491127	14.95575362	0.000110062	0.004684197
RBL2	5934	6.4333236	5.406396054	13.5525981	0.00023197	0.006783202
FDPS	2224	6.4226955	7.314359837	25.75614646	3.87E-07	0.000376863
PLIN2	123	6.4168123	4.92248438	12.68555636	0.000368491	0.008655762
FBXO7	25793	6.4149906	5.663346159	14.68433383	0.000127098	0.005093146
CTBS	1486	6.413642	2.711258868	14.27445983	0.000157994	0.005717521
TCF25	22980	6.3842708	3.7966207	16.64337336	4.51E-05	0.00288027
SNAP23	8773	6.3817022	3.513363322	15.45690395	8.44E-05	0.004105641
CCNF	899	6.3796285	3.261273135	12.62491484	0.000380639	0.008841308
CLDN10	9071	6.3718308	6.377486938	17.50276306	2.87E-05	0.002292074
MAF	4094	6.3679738	3.65354726	12.60085028	0.000385571	0.008892033
CENPE	1062	6.3611903	3.602441735	12.55434139	0.000395287	0.008992712
TBCB	1155	6.3174596	3.586779311	16.95550401	3.83E-05	0.002681554
CTSL	1514	6.3126598	5.569801438	13.22864528	0.000275703	0.007561486
SLC37A3	84255	6.2843896	3.772262271	16.52101971	4.81E-05	0.003025494
SCYL1	57410	6.2134803	1.793961253	13.17826564	0.000283215	0.007666863
LAMTOR3	8649	6.2019673	5.975742906	16.78138803	4.19E-05	0.002819707
OSTF1	26578	6.0301287	4.768433109	13.11772726	0.000292515	0.007752576
CCM2	83605	5.9362952	5.21885616	12.3578235	0.000439142	0.009546939
DERA	51071	5.8499731	6.218644541	21.9062579	2.86E-06	0.000883454
KCTD20	222658	5.7514149	5.210505945	18.60581001	1.61E-05	0.001703944
ZMYM4	9202	5.7499026	4.840768252	15.1757389	9.80E-05	0.004364474
ZFP91	80829	5.7478836	6.454566345	21.98983924	2.74E-06	0.000883454
FMR1	2332	5.6507117	3.974384887	14.35068876	0.000151724	0.005551359
PEBP1	5037	5.6031434	5.985973539	19.95460612	7.93E-06	0.001387666
LTBP1	4052	5.601331	6.729679032	12.88821201	0.000330658	0.008198538

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
IQGAP3	128239	5.5921064	1.763845213	14.36779848	0.000150352	0.005537897
RBM45	129831	5.5836373	4.619842553	13.32396954	0.000262035	0.00729551
NFKB1	4790	5.5681406	5.917471444	17.53230152	2.82E-05	0.002281333
RBM41	55285	5.5635991	2.595617845	16.56557224	4.70E-05	0.002977884
ERP44	23071	5.5454126	5.671343131	17.59792637	2.73E-05	0.002267831
PARL	55486	5.5306037	7.17425752	17.08947253	3.57E-05	0.002611089
LAMC2	3918	5.4946851	7.748621842	28.11276841	1.14E-07	0.000179292
OPTN	10133	5.4821531	4.210281201	16.73480763	4.30E-05	0.002866481
TIGAR	57103	5.4497357	6.429674666	12.58765748	0.000388303	0.008919093
EIF4E2	9470	5.4290053	5.869887834	18.87789734	1.39E-05	0.00158945
ADAM10	102	5.3670953	5.602747356	13.985345	0.000184241	0.006125575
ZNF740	283337	5.3657125	1.205473682	15.74243489	7.26E-05	0.003786399
SMS	6611	5.353874	5.574567117	12.4933563	0.000408402	0.009114911
MSN	4478	5.3488372	6.507102007	17.66049389	2.64E-05	0.002216716
EGFR	1956	5.2518251	6.730520939	15.25663118	9.38E-05	0.004263837
DHRS7	51635	5.2126749	6.333722879	13.0177227	0.000308557	0.007923777
CBL	867	5.1587369	5.454951351	14.83197649	0.000117526	0.004909276
ECHS1	1892	5.1563804	5.862218143	12.99853283	0.000311735	0.007936115
COPS7A	50813	5.1561476	6.139757618	12.77397427	0.000351475	0.008426237
PPP4C	5531	5.1544648	4.596033486	15.94755656	6.51E-05	0.003554403
TMOD3	29766	5.1150886	5.03926498	16.39539035	5.14E-05	0.00312437
PNPO	55163	5.1147398	5.267950873	15.30892881	9.13E-05	0.004240604
RBFOX2	23543	5.1103297	5.617658646	18.9481387	1.34E-05	0.001582074
FAM114A1	92689	5.0698323	6.022942282	18.91344023	1.37E-05	0.001582074
VCL	7414	4.997882	6.752791547	13.86707337	0.000196206	0.006337607
FAM136A	84908	4.8042283	6.324536755	13.90075423	0.000192721	0.006274059
C14orf166	51637	4.7914806	7.080350494	22.03182609	2.68E-06	0.000883454
TMEM60	85025	4.7858025	6.742119257	12.82634289	0.000341773	0.008309532
UCK1	83549	4.7603181	-0.101699022	12.29287884	0.00045469	0.009753127
PARP6	56965	4.7453342	6.905386303	20.91173456	4.81E-06	0.001089558
TMEM179B	374395	4.7300943	6.158211976	14.01081402	0.000181762	0.006069071
AKR1A1	10327	4.7290886	5.443656321	13.44632962	0.000245486	0.006963717
IGFBP4	3487	4.7131793	7.526845936	29.33409282	6.09E-08	0.000170277
TPX2	22974	4.6995335	5.975026073	12.40702995	0.000427721	0.009431524
DAP3	7818	4.6863186	7.296291518	20.2178715	6.91E-06	0.001285872
MRPL15	29088	4.6165994	5.62544345	14.21346175	0.000163199	0.005854625
LRRC16A	55604	4.6072967	4.858377939	19.43571362	1.04E-05	0.00147992
HACD3	51495	4.6048292	7.959186803	24.45460411	7.61E-07	0.000483937
NRDC	4898	4.5930233	7.110490688	15.19913234	9.67E-05	0.00434787
TCEA1	6917	4.5903478	5.780574617	14.20859542	0.000163621	0.00585708
AKIP1	56672	4.583956	5.108661038	13.5696263	0.000229875	0.006783202
LTA4H	4048	4.4833844	6.387875017	14.06256297	0.000176829	0.005968147
VPS41	27072	4.47753	5.668583826	15.07936438	0.000103084	0.004489268
RPN2	6185	4.45508	8.255927152	13.092995	0.000296402	0.00778079
GNG12	55970	4.3790378	6.798346558	15.29507277	9.20E-05	0.004259858

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
CBX5	23468	4.3684484	6.534932876	19.01781824	1.30E-05	0.001576209
PTPRF	5792	4.3093628	6.028800519	18.61622762	1.60E-05	0.001703944
MTPN	136319	4.2614373	4.867951977	13.60340576	0.000225776	0.006768624
PLAT	5327	4.2417356	6.254916714	13.81568842	0.000201645	0.00644768
TCF3	6929	4.206733	1.835534949	12.54224645	0.000397854	0.008992712
IARS2	55699	4.1270235	6.868574451	15.66101663	7.58E-05	0.003897078
HNRNPM	4670	4.1226035	5.507226212	14.07434334	0.000175725	0.005968147
KRT18	3875	4.1114841	7.49772582	12.99142259	0.000312921	0.007949447
MAP7D3	79649	4.0660394	5.349200836	13.63897123	0.000221539	0.006710272
COX7A2L	9167	4.0053919	7.228939481	12.62557784	0.000380504	0.008841308
PPT1	5538	3.9725498	6.644148277	13.71544856	0.000212698	0.006621864
TMED10	10972	3.9369979	8.823695186	18.22649937	1.96E-05	0.001885822
MALL	7851	3.828436	5.270219944	14.07950135	0.000175243	0.005968147
RAB5A	5868	3.7248507	6.409128066	12.95818661	0.000318526	0.008030146
HP1BP3	50809	3.6119046	6.750962477	15.48999451	8.29E-05	0.004058324
TMEM106B	54664	3.596183	4.971024923	12.2188623	0.000473088	0.009979496
TUFM	7284	3.5934374	6.640684262	12.36029766	0.000438561	0.009546939
TRIM29	23650	3.5792612	7.287663454	14.61087581	0.00013215	0.005233142
CCDC47	57003	3.5791083	7.210926638	15.13461565	0.000100111	0.004426822
UBR4	23352	3.5440205	7.6906679	14.53500228	0.000137579	0.005262231
COPB2	9276	3.5108161	7.902057511	12.99798606	0.000311826	0.007936115
TAPBP	6892	3.3761876	7.607900051	14.09720227	0.000173602	0.005968147
LONP2	83752	3.2974776	6.716924276	13.15451879	0.000286827	0.007688071
NEDD8	4738	3.264488	8.112575195	13.94682869	0.000188055	0.006170729
LAMP1	3916	3.1470017	2.726260773	12.85439631	0.000336687	0.00826109
KRT3	3850	3.1030634	-1.024895545	13.46854623	0.000242596	0.006920307
EMP2	2013	3.0668112	8.483185068	14.90644964	0.000112976	0.004765662
ZC3H11A	9877	3.0414494	7.497939274	17.44916549	2.95E-05	0.002346279
CALU	813	2.8308167	8.376476836	13.16769449	0.000284817	0.007666863
KRT6A	3853	2.3083226	11.48497681	13.712368	0.000213047	0.006621864
COMP	1311	-2.119547	0.156936351	12.94597237	0.000320611	0.008045913
SPATA4	132851	-2.150935	0.506759743	14.53230768	0.000137776	0.005262231
RBFOX3	146713	-2.161992	1.318433977	16.65831336	4.48E-05	0.00288027
NASP	4678	-2.169713	10.28133033	14.43706269	0.000144922	0.005424503
LMOD3	56203	-2.173807	1.465272796	13.69851056	0.000214625	0.006622131
LY9	4063	-2.177446	-0.449273701	12.52975695	0.000400522	0.009012015
ANKS1B	56899	-2.189074	0.385722065	15.98642784	6.38E-05	0.003516983
COL6A3	1293	-2.190445	2.734643048	12.5593068	0.000394238	0.008992712
SCUBE1	80274	-2.21472	1.060596581	12.5990831	0.000385936	0.008892033
C1orf158	93190	-2.226069	-0.151326003	13.73736768	0.00021023	0.006621864
PDE11A	50940	-2.246523	1.994210058	13.76269614	0.000207414	0.006559034
TIE1	7075	-2.296875	1.276569916	12.77081309	0.00035207	0.008426237
SERAC1	84947	-2.315348	1.982847273	12.83141761	0.000340847	0.008309532
GSDMB	55876	-2.323188	-0.024179421	13.03797838	0.000305237	0.00787522
MIXL1	83881	-2.323549	6.870287599	13.86184175	0.000196753	0.006342887

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
ZBTB8B	728116	-2.340974	3.700491592	13.16042802	0.000285924	0.007676305
FCRL2	79368	-2.347736	-0.299044129	15.51053148	8.20E-05	0.004050421
ZNF320	162967	-2.360122	1.271456612	14.12511903	0.000171044	0.005930257
LINC01134	100133612	-2.373711	0.260387181	13.01168127	0.000309554	0.007936115
PIWIL1	9271	-2.398244	2.815640386	13.83801031	0.000199264	0.006411342
TNFSF13B	10673	-2.398831	5.444639608	13.00575217	0.000310536	0.007936115
CACNA1D	776	-2.400582	1.994718517	13.24157549	0.000273808	0.007544984
IL10	3586	-2.403536	0.467209797	13.75540112	0.000208221	0.006571685
CACNG8	59283	-2.41529	-0.535103129	13.1349442	0.000289839	0.007707464
BTBD19	149478	-2.433386	0.238994616	18.54733772	1.66E-05	0.001734779
LINC00970	101978719	-2.434407	0.004183317	12.85224489	0.000337075	0.00826109
CSF2RA	1438	-2.447061	0.716369452	12.55913688	0.000394274	0.008992712
MS4A10	341116	-2.447544	1.103628905	15.49596054	8.27E-05	0.00405754
GBP1P1	400759	-2.456727	6.168240881	13.82328739	0.000200832	0.006436729
KCNMA1-AS1	101929328	-2.458039	2.514236232	13.70604614	0.000213765	0.006621864
RIPPLY3	53820	-2.470268	-0.22663217	14.8242626	0.000118007	0.004909276
MIB2	142678	-2.476678	0.343525525	12.70681003	0.000364326	0.008618624
SPN	6693	-2.481578	-0.154356294	13.35999902	0.000257048	0.007205198
FLG-AS1	339400	-2.486914	0.717192328	14.32631231	0.000153701	0.005598931
CATSPERD	257062	-2.48939	1.982598539	14.2798637	0.000157541	0.005713632
CRB1	23418	-2.49159	-0.628374054	12.91610187	0.000325768	0.008113771
CTRC	11330	-2.494119	2.216544506	13.46432664	0.000243142	0.00692098
L1TD1	54596	-2.503687	3.415472192	13.35155246	0.000258209	0.007217082
NCMAP	400746	-2.506217	0.6370653	12.65419564	0.000374724	0.008740736
NEGR1	257194	-2.532388	7.693341157	16.74167891	4.28E-05	0.002866481
DRAXIN	374946	-2.564747	1.434757872	13.21422425	0.000277832	0.007607274
HACD4	401494	-2.580546	-1.102714876	12.23536222	0.000468922	0.009967916
CD3G	917	-2.58408	0.62501907	14.40933426	0.000147071	0.005458554
CCDC180	100499483	-2.589708	2.116800078	14.70267497	0.000125868	0.005089291
HCK	3055	-2.595787	-0.42334637	13.13570701	0.000289721	0.007707464
CSPG4P12	440300	-2.61021	2.778750003	14.45732661	0.000143371	0.005413403
FLG2	388698	-2.616329	-0.481120982	13.76260931	0.000207424	0.006559034
PKN2-AS1	101927891	-2.621054	0.482522992	14.18930052	0.000165308	0.005886328
SLC30A2	7780	-2.621567	0.043237588	12.35290049	0.000440302	0.009556053
PKNOX2	63876	-2.623931	-0.112755575	12.28598568	0.000456372	0.009776535
TNR	7143	-2.643476	-0.031312046	12.32398789	0.000447174	0.009642769
MIRLET7BHG	400931	-2.65318	2.88743665	13.49605863	0.000239065	0.006902073
ZNF215	7762	-2.657537	-1.109364407	13.00232888	0.000311104	0.007936115
LILRA5	353514	-2.659056	0.211699619	16.25006093	5.55E-05	0.00329028
GBP6	163351	-2.661051	-0.50529544	13.55126197	0.000232135	0.006783202
TNFAIP8L1	126282	-2.661919	2.228636981	14.65866447	0.000128841	0.005146794
LRRC74B	400891	-2.663117	-0.087343374	14.0322423	0.000179703	0.006028243
FZD4	8322	-2.677242	-1.039431744	12.93050581	0.00032327	0.008075901
FAM221A	340277	-2.68425	0.786539386	12.84280737	0.000338779	0.008275814
GALNT16	57452	-2.688577	-1.066408212	13.39501351	0.000252294	0.007095987

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
SEMA4C	54910	-2.699877	1.087093335	12.69442903	0.000366747	0.008629663
KCNMB1	3779	-2.700426	2.027031216	13.11139827	0.000293505	0.007753959
TMEM229B	161145	-2.702188	-0.906535577	12.84365141	0.000338626	0.008275814
RGSL1	353299	-2.714202	3.959364844	13.97420914	0.000185336	0.00613016
SLC2A5	6518	-2.715839	0.397638984	16.12756272	5.92E-05	0.003392124
LRRC27	80313	-2.725933	1.601712454	12.43949229	0.00042035	0.009318698
RIMKLA	284716	-2.729171	1.095120938	14.14062307	0.000169641	0.005906348
MIR519A2	574500	-2.730205	1.786519642	21.03608506	4.51E-06	0.001060074
TRIM73	375593	-2.74957	1.117173651	16.10605068	5.99E-05	0.003392124
MRLN	100507027	-2.757373	-0.904277633	12.32936641	0.000445888	0.009642769
	729348	-2.768445	7.111712451	12.37587141	0.000434918	0.00953935
KLRD1	3824	-2.77147	0.184106034	13.174373	0.000283804	0.007666863
ITIH5	80760	-2.790119	0.940243255	14.7603828	0.000122073	0.005006425
	400748	-2.800446	0.796374978	17.52090145	2.84E-05	0.002281333
CYB5RL	606495	-2.804531	0.16533517	16.11011577	5.98E-05	0.003392124
CCDC158	339965	-2.810075	-0.698440675	14.23373323	0.00016145	0.005817134
RNF125	54941	-2.81328	0.645915833	19.357546	1.08E-05	0.001493796
ZNF718	255403	-2.819588	1.348409709	14.56271965	0.00013557	0.005262231
SIDT2	51092	-2.827722	-0.61405536	12.55301598	0.000395567	0.008992712
KLK10	5655	-2.833825	0.858873897	13.48244116	0.000240806	0.006913984
MAN1C1	57134	-2.843384	1.250421835	17.57302324	2.76E-05	0.002274862
NPFFR1	64106	-2.845468	0.78688442	18.61612774	1.60E-05	0.001703944
TPH1	7166	-2.859544	4.377800794	12.76618754	0.000352941	0.008434893
	101928733	-2.861167	-0.9860682	12.95030227	0.00031987	0.008045913
ABCA8	10351	-2.865248	0.020288671	14.82371965	0.000118041	0.004909276
PIGR	5284	-2.876352	1.923868384	18.03411959	2.17E-05	0.002015973
LINC00924	145820	-2.877185	0.416331062	16.27102972	5.49E-05	0.003265772
ABCC9	10060	-2.893959	5.831215523	13.02739935	0.000306967	0.007907496
	101927560	-2.895869	-0.361961645	15.0923328	0.000102378	0.004479185
FOXL2NB	401089	-2.901981	-0.877820179	14.76618795	0.000121698	0.005006425
BRDT	676	-2.905803	-0.435523902	21.85657463	2.94E-06	0.000883454
	101928514	-2.906527	-0.246664084	17.96824873	2.25E-05	0.002063764
MYO3B	140469	-2.912011	-0.268188642	13.98315136	0.000184456	0.006125575
TMEM116	89894	-2.917647	0.358171226	12.59853564	0.000386049	0.008892033
AGRN	375790	-2.920726	0.907365791	15.27829002	9.28E-05	0.004262059
CNR2	1269	-2.922768	1.787229433	20.96816043	4.67E-06	0.001072627
SCRG1	11341	-2.928214	1.399709151	13.95932034	0.000186809	0.006154284
SLC16A10	117247	-2.932039	-0.692060062	12.27121125	0.00046	0.009828777
ACTG1P17	283693	-2.934841	-0.298322524	13.10400319	0.000294665	0.007759836
PLA2G2C	391013	-2.940976	-0.456584125	14.75923786	0.000122148	0.005006425
SLC15A1	6564	-2.946767	0.563669445	16.03790832	6.21E-05	0.003451273
PLCE1	51196	-2.952876	3.242482819	15.40384017	8.68E-05	0.00413744
CROCCP3	114819	-2.967846	0.988617918	14.03387525	0.000179547	0.006028243
ACSL6	23305	-2.979587	2.461238638	17.40189438	3.03E-05	0.00239385
EGR1	1958	-2.981661	10.06963609	14.71774647	0.000124865	0.005061336

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
RBM20	282996	-2.982335	-0.336057837	22.69571112	1.90E-06	0.00082599
SLC7A14	57709	-2.987843	-0.628329901	12.7043885	0.000364798	0.008618624
SLFN12L	100506736	-2.994845	-0.814215155	17.52755772	2.83E-05	0.002281333
ATP8B2	57198	-3.000731	-0.00357993	23.93125938	9.98E-07	0.000532617
DFFB	1677	-3.029438	0.898541106	19.01611243	1.30E-05	0.001576209
CEP112	201134	-3.03682	-0.801723485	15.85358273	6.84E-05	0.003689795
PAXBP1-AS1	100506215	-3.039814	0.94862503	12.39287916	0.000430974	0.009490615
SV2B	9899	-3.053989	-0.854510995	12.24802771	0.00046575	0.009926001
	101929586	-3.075912	-0.454204097	17.08735722	3.57E-05	0.002611089
PCSK6-AS1	105371027	-3.08235	0.061750441	16.69315931	4.39E-05	0.002877127
ZSCAN2	54993	-3.092789	0.793251864	14.35857569	0.00015109	0.005551359
IFIT3	3437	-3.098434	0.578973551	18.26858134	1.92E-05	0.001862815
CFAP74	85452	-3.104792	1.22678192	24.47283345	7.54E-07	0.000483937
DNAJC5B	85479	-3.113288	-0.404810587	18.63895364	1.58E-05	0.001703944
LINC01482	101928104	-3.116282	-0.778719522	13.66987149	0.000217923	0.006637216
KCNQ1OT1	10984	-3.117184	8.745263517	18.32022255	1.87E-05	0.001827152
MPP4	58538	-3.154962	2.053723633	12.38132758	0.000433649	0.009536812
C1orf61	10485	-3.159018	0.962016782	17.84897913	2.39E-05	0.002092604
C1orf127	148345	-3.160194	-0.765377516	13.56744833	0.000230142	0.006783202
SPNS1	83985	-3.169162	0.606052915	14.16209351	0.000167716	0.005888917
FRRS1	391059	-3.18543	1.270474733	18.94181727	1.35E-05	0.001582074
CIITA	4261	-3.188319	-0.449683589	16.03479984	6.22E-05	0.003451273
RNF222	643904	-3.191933	-0.632186111	19.9064498	8.13E-06	0.001387666
HES1	3280	-3.200339	6.819022148	13.50036964	0.000238516	0.006902073
DUOX2	50506	-3.205662	4.647993492	14.98975374	0.000108097	0.00461938
SPATA6	54558	-3.221605	0.489696354	15.79969594	7.04E-05	0.003711019
LINC00683	400660	-3.236063	-0.856326062	13.46774114	0.0002427	0.006920307
FAM92A1P2	403315	-3.260861	-0.598875377	15.56147977	7.99E-05	0.00400272
BTBD8	284697	-3.265013	-0.053240421	18.39194531	1.80E-05	0.001791449
	101927292	-3.271032	-0.620218537	17.88045323	2.35E-05	0.002080292
PDPN	10630	-3.273019	0.024078553	13.01830424	0.000308461	0.007923777
TRIM60	166655	-3.274737	-0.456670928	13.49547508	0.000239139	0.006902073
PSPN	5623	-3.276598	-0.753202735	13.42428547	0.000248388	0.007009956
	101926911	-3.296471	-0.763499677	12.9401445	0.00032161	0.008052022
AFF3	3899	-3.302705	-0.459192127	15.84388765	6.88E-05	0.003693578
GALNTL5	168391	-3.306163	1.7485087	19.62061507	9.44E-06	0.001393037
TERB1	283847	-3.307429	-0.058331022	14.17609635	0.000166472	0.005886328
SLC35F1	222553	-3.313004	0.820331127	18.05098584	2.15E-05	0.002009481
IBA57	200205	-3.319228	3.297858135	21.18071145	4.18E-06	0.001049694
ATP1A4	480	-3.326669	0.287328082	21.56379168	3.42E-06	0.000959351
CD84	8832	-3.335667	0.707048811	19.0415179	1.28E-05	0.001576209
PDDC1	347862	-3.355086	2.327544385	21.68567511	3.21E-06	0.000915808
ACOT11	26027	-3.357738	2.137505174	14.68297386	0.00012719	0.005093146
CYB561A3	220002	-3.367704	3.872018214	16.88483385	3.97E-05	0.002759877
CASP16P	197350	-3.405662	-0.910888754	19.20033461	1.18E-05	0.001544612

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
HSPG2	3339	-3.414044	0.889678749	15.92567308	6.59E-05	0.003583906
CCBE1	147372	-3.418671	0.511695579	17.71223696	2.57E-05	0.002197043
FAM179A	165186	-3.454595	1.963355833	19.78482234	8.67E-06	0.001387666
MIR3908	100500909	-3.476967	-0.917536071	12.25689239	0.000463543	0.009891705
PGF	5228	-3.47901	0.92963648	15.18481785	9.75E-05	0.004357273
ADGRG2	10149	-3.48093	-0.356929669	19.39134876	1.06E-05	0.00147992
AP4B1-AS1	100287722	-3.484704	0.151155313	19.85298018	8.36E-06	0.001387666
SLAMF7	57823	-3.487941	1.031599273	15.18720181	9.74E-05	0.004357273
KLRK1	22914	-3.498203	-0.339312499	13.60219744	0.000225921	0.006768624
CTSS	1520	-3.498221	0.508557837	14.17494317	0.000166574	0.005886328
TBXA2R	6915	-3.51956	0.837053761	15.55563533	8.01E-05	0.00400272
XKR9	389668	-3.524824	-0.366818516	16.42078438	5.07E-05	0.003107059
	101929227	-3.529207	0.024564263	14.09222082	0.000174062	0.005968147
	101926964	-3.532751	0.482430656	23.02360392	1.60E-06	0.000715267
BEST1	7439	-3.533443	0.356511637	12.98437975	0.0003141	0.007954965
SCYL3	57147	-3.535668	3.314071492	16.79735275	4.16E-05	0.002807497
PPP1R12B	4660	-3.543752	2.207029917	15.35863622	8.89E-05	0.004184572
MIR202HG	574448	-3.547929	-0.052829318	18.52379405	1.68E-05	0.001738283
	283731	-3.548282	0.064616917	27.56659709	1.52E-07	0.000179292
RNF157	114804	-3.549665	-0.02524578	20.65099294	5.51E-06	0.001163955
RPS6KL1	83694	-3.554745	-0.746749418	12.82216324	0.000342537	0.008309532
SLC25A18	83733	-3.571203	1.670113641	16.71392638	4.35E-05	0.002877127
NMNAT1	64802	-3.588662	3.999396035	15.99825675	6.34E-05	0.003506766
TPGS1	91978	-3.594644	3.251265777	18.39911847	1.79E-05	0.001791449
ZNF835	90485	-3.635983	-0.057939037	13.31968487	0.000262634	0.00729991
SLC52A1	55065	-3.637379	0.944586773	13.88907215	0.000193923	0.006287256
PADI2	11240	-3.643616	-0.380378937	24.08240973	9.23E-07	0.000532617
NUGGC	389643	-3.683894	0.032014889	15.03634558	0.00010546	0.004540957
FRMPD1	22844	-3.687233	0.990300819	19.87059349	8.29E-06	0.001387666
ATCAY	85300	-3.691587	-0.44000244	25.01477705	5.69E-07	0.000427417
PCDH11Y	27328	-3.692035	2.396837942	14.69814435	0.000126171	0.005089291
GUCY1B2	2974	-3.706346	-0.011501054	15.596699	7.84E-05	0.003989033
ALS2CR12	130540	-3.713906	0.062405531	12.5301115	0.000400446	0.009012015
CLK1	1195	-3.719009	8.189697877	16.11001549	5.98E-05	0.003392124
GAS6-AS2	100506394	-3.720549	0.182225297	27.63417229	1.47E-07	0.000179292
OTUD6A	139562	-3.743592	0.062623129	13.1361549	0.000289652	0.007707464
WDR90	197335	-3.763945	-0.165041575	19.40321358	1.06E-05	0.00147992
MLLT4-AS1	653483	-3.764885	1.03877246	14.2154609	0.000163026	0.005854625
HRK	8739	-3.770029	0.167127152	15.65082821	7.62E-05	0.00390047
HOGA1	112817	-3.776053	0.405922142	16.60309914	4.61E-05	0.002930767
BHLHE40	8553	-3.80378	8.400929952	13.78443859	0.000205027	0.006520662
LINC01285	101928287	-3.812175	0.676911046	14.54409404	0.000136917	0.005262231
ST8SIA6-AS1	100128098	-3.829805	-0.642227289	13.1657786	0.000285108	0.007666863
SHANK2-AS1	100874198	-3.84971	-0.597671393	17.62719561	2.69E-05	0.002244474
MAP1LC3C	440738	-3.857731	2.666384489	17.94186805	2.28E-05	0.002080292

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
MORC4	79710	-3.877851	5.492677287	18.9241218	1.36E-05	0.001582074
EPB41	2035	-3.888086	1.421722421	19.66453585	9.23E-06	0.001387666
TRIB1	10221	-3.896834	9.471287668	17.90517697	2.32E-05	0.002080292
WNT9B	7484	-3.897793	-0.511431877	13.66392455	0.000218614	0.006638293
ACBD7	414149	-3.913509	0.001370552	17.52246799	2.84E-05	0.002281333
FBXO43	286151	-3.930642	-0.240709855	15.26188839	9.36E-05	0.004263699
SLC2A1-AS1	440584	-3.976824	0.68469937	15.4191151	8.61E-05	0.004136991
ATP1A3	478	-3.987638	0.646491127	18.91036781	1.37E-05	0.001582074
FMN1	342184	-3.995153	2.697509665	22.38170024	2.23E-06	0.00088036
DSEL	92126	-3.996477	4.17133835	13.42774818	0.00024793	0.007008989
LINC00954	400946	-4.026687	2.483930119	16.10696047	5.99E-05	0.003392124
FANCF	2188	-4.036209	1.876991911	12.32382759	0.000447213	0.009642769
TRPM3	80036	-4.04747	0.41570249	18.51288523	1.69E-05	0.001738283
LINC01502	100130954	-4.061271	0.882831964	15.87407752	6.77E-05	0.003670876
OPA3	80207	-4.084005	1.029181959	15.51940837	8.17E-05	0.004050421
WDR31	114987	-4.112884	0.551975894	17.38075305	3.06E-05	0.002399233
FAM73A	374986	-4.13499	4.230822895	22.02574437	2.69E-06	0.000883454
CX3CL1	6376	-4.135389	0.019168452	13.1120793	0.000293398	0.007753959
LRRC37A	9884	-4.1424	1.117701445	24.6775564	6.78E-07	0.000466983
ETV3L	440695	-4.146793	-0.653607293	12.48786714	0.000409604	0.009117125
ASB8	140461	-4.152424	0.143628616	14.90190478	0.000113249	0.004765662
MYC	4609	-4.18582	8.107184051	19.12553171	1.22E-05	0.001557072
ZYG11A	440590	-4.204527	2.758763894	14.18476486	0.000165707	0.005886328
14-Sep	346288	-4.216995	-0.208867543	12.78110044	0.000350139	0.008426237
PCDHA9	9752	-4.222042	-0.571347386	16.47977398	4.92E-05	0.003075646
CYP46A1	10858	-4.237933	0.704335112	26.01744537	3.38E-07	0.000349723
MAP3K15	389840	-4.249577	0.00021484	13.95313874	0.000187424	0.006162278
TCAM1P	146771	-4.25248	0.113456924	13.47046262	0.000242349	0.006920307
LINC00471	151477	-4.256091	1.792962755	12.50723564	0.000405379	0.009093264
GRK3	157	-4.276015	4.000413353	13.88573822	0.000194267	0.006287256
	101927973	-4.28765	0.449339677	19.26937687	1.14E-05	0.001526235
CHEK2	11200	-4.302493	0.792202461	12.55899448	0.000394304	0.008992712
	102723766	-4.34004	3.216179941	15.23700512	9.48E-05	0.004296571
	101928567	-4.370201	-0.213560278	15.95185419	6.50E-05	0.003554403
ATP13A2	23400	-4.437404	3.045937688	14.16846009	0.000167149	0.005888917
	100996291	-4.437515	0.345253898	19.75988553	8.78E-06	0.001387666
LAIR1	3903	-4.442186	3.494939254	23.54961469	1.22E-06	0.00061079
	440446	-4.483544	0.228012071	20.86987496	4.92E-06	0.001098581
DMC1	11144	-4.487979	0.62910621	22.09479447	2.60E-06	0.000883454
	283299	-4.520214	0.7638767	22.33219803	2.29E-06	0.00088036
TOR3A	64222	-4.521834	4.817395126	13.3505622	0.000258345	0.007217082
PDE4C	5143	-4.523079	0.828136207	14.43209067	0.000145305	0.005424503
IL6R	3570	-4.534984	6.420929844	13.56785669	0.000230092	0.006783202
EIF2B5-AS1	100874079	-4.539599	-0.081524795	12.85167587	0.000337177	0.00826109
	101928118	-4.586833	1.462139897	14.49993926	0.000140164	0.005328812

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
CELF5	60680	-4.647665	1.037643604	12.67366046	0.000370843	0.008674679
NKX3-1	4824	-4.658622	6.197520027	12.22389171	0.000471814	0.009978084
RPS6KB2	6199	-4.662378	1.09523961	17.58068968	2.75E-05	0.002274862
MOG	4340	-4.68508	0.157079752	16.17643533	5.77E-05	0.003356438
SYDE2	84144	-4.688391	1.923916618	16.36706618	5.22E-05	0.003149713
CNBD2	140894	-4.699015	0.406183347	15.16645151	9.84E-05	0.004364474
MCM3AP-AS1	114044	-4.716589	-0.156569457	17.56099534	2.78E-05	0.002277964
ZBP1	81030	-4.722309	0.294529668	13.70560697	0.000213815	0.006621864
MCF2L2	23101	-4.725039	3.783858183	17.67378616	2.62E-05	0.00221251
ATF3	467	-4.766379	7.925290432	14.06749405	0.000176366	0.005968147
TRPM6	140803	-4.77147	0.669733141	29.12028863	6.80E-08	0.000170277
CFAP69	79846	-4.795469	0.0485773	23.03461981	1.59E-06	0.000715267
GPRIN3	285513	-4.806506	0.812685079	19.17046102	1.20E-05	0.001544612
LINC00649	100506334	-4.812559	3.114961709	15.33068446	9.02E-05	0.004203858
CXCL2	2920	-4.812591	7.165527576	19.66482624	9.23E-06	0.001387666
ZNF37BP	100129482	-4.817106	2.642656888	14.00814188	0.000182021	0.006069071
C4orf36	132989	-4.817508	2.558511198	15.51457229	8.19E-05	0.004050421
RGPD5	84220	-4.841817	6.063685229	16.98419553	3.77E-05	0.002662756
ANKMY1	51281	-4.856705	0.499856777	18.12500841	2.07E-05	0.001954949
	100129917	-4.873734	1.532630917	14.33189693	0.000153246	0.00559467
ICOSLG	23308	-4.932024	1.850388696	15.10988503	0.000101431	0.004461335
MLXIP	22877	-4.959525	3.361625721	22.64306244	1.95E-06	0.000827168
C12orf76	400073	-4.99084	3.23062363	15.0999837	0.000101964	0.0044729
FAM71D	161142	-5.01523	1.563150266	15.05272382	0.000104549	0.004526272
ARHGAP44	9912	-5.030905	3.658713965	21.45340516	3.63E-06	0.000975649
DTNA	1837	-5.098352	3.744462508	14.39779531	0.000147975	0.00546446
FGFR2	2263	-5.105154	0.495653034	19.05002072	1.27E-05	0.001576209
SNHG20	654434	-5.151696	4.20723041	14.81884463	0.000118347	0.004909276
FLVCR1	28982	-5.239668	6.658924625	21.30032243	3.93E-06	0.001022818
BBS1	582	-5.276197	0.647165564	14.39713369	0.000148027	0.00546446
FBRS	64319	-5.283566	1.642085348	15.06533131	0.000103853	0.00450794
METTL12	751071	-5.314914	4.153285146	20.01672536	7.68E-06	0.001387666
FAM86B3P	286042	-5.321754	-0.014125255	13.0478311	0.000303636	0.007866287
FAM153B	202134	-5.351187	0.552948244	17.06323619	3.62E-05	0.002611089
LOXL1-AS1	100287616	-5.359962	3.015177134	17.1878782	3.39E-05	0.002568624
CLNK	116449	-5.383978	0.364041749	19.90877685	8.12E-06	0.001387666
HYPK	25764	-5.418138	4.926087575	12.82190343	0.000342585	0.008309532
ZBTB16	7704	-5.422035	0.78354867	14.13515925	0.000170134	0.005911082
	441072	-5.481064	2.623199692	13.47680588	0.000241531	0.006920307
	100506083	-5.485485	1.45561287	14.10704999	0.000172695	0.005950073
ESAM	90952	-5.502859	0.55386337	13.72736882	0.000211352	0.006621864
SNHG23	79104	-5.519062	1.379880507	20.32654911	6.53E-06	0.001265698
DOC2A	8448	-5.530024	0.708938141	15.48330379	8.32E-05	0.004060701
TSIX	9383	-5.615211	3.31572938	21.47408772	3.59E-06	0.000975649
RGS16	6004	-5.662557	4.321704979	15.85198079	6.85E-05	0.003689795

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
VGLL3	389136	-5.690017	0.747126178	12.22699055	0.000471031	0.00997428
SPDYE1	285955	-5.728661	3.354428928	14.42465999	0.000145879	0.005433682
WNT2B	7482	-5.780266	4.985659849	18.26098653	1.93E-05	0.001862815
TUBGCP4	27229	-5.786456	5.71160392	17.07817167	3.59E-05	0.002611089
RPARP-AS1	100505761	-5.822226	4.254540496	19.51453598	9.98E-06	0.001435744
SLC25A34	284723	-5.915867	3.683949205	18.8999726	1.38E-05	0.001582074
REXO1L1P	254958	-5.968681	2.004936849	13.55612721	0.000231534	0.006783202
FAM217B	63939	-5.993685	4.375371129	13.59823974	0.000226398	0.006770643
GSN-AS1	57000	-6.001878	0.472769279	16.1299083	5.91E-05	0.003392124
ZNF596	169270	-6.003161	1.812800967	16.65307576	4.49E-05	0.00288027
LACTB2-AS1	286190	-6.062908	5.324223622	18.70415217	1.53E-05	0.001660869
C1orf132	407025	-6.100325	1.890576066	13.05142859	0.000303053	0.007866287
GIT2	9815	-6.179879	3.590518489	15.03779927	0.000105379	0.004540957
IRF1	3659	-6.23193	2.928529639	19.20670722	1.17E-05	0.001544612
AKAP5	9495	-6.269642	3.666260067	17.83508592	2.41E-05	0.002096844
MMACHC	25974	-6.274479	3.306524131	17.78947644	2.47E-05	0.002125342
NFS1	9054	-6.286629	2.458782266	18.41779724	1.77E-05	0.001788861
CCDC84	338657	-6.323064	2.507937963	15.26259211	9.36E-05	0.004263699
ZNF83	55769	-6.332075	5.054968449	19.03051467	1.29E-05	0.001576209
ETFBKMT	254013	-6.332264	5.935745914	20.14136869	7.19E-06	0.001321645
FAAP24	91442	-6.347168	2.650011856	17.03990851	3.66E-05	0.002620477
COBLL1	22837	-6.393453	2.641396327	13.71269498	0.00021301	0.006621864
ITGA10	8515	-6.43062	2.129322767	13.62754195	0.000222892	0.006738903
SLC16A4	9122	-6.467666	4.162922726	19.76208583	8.77E-06	0.001387666
	102724532	-6.485947	5.181890689	14.50571039	0.000139735	0.005324753
PINK1-AS	100861548	-6.491	2.349361471	14.53418066	0.000137639	0.005262231
AOC3	8639	-6.642884	1.952085536	13.09000304	0.000296876	0.007780877
MATN2	4147	-6.658558	4.031785356	15.28171439	9.26E-05	0.004262059
SNHG12	85028	-6.681623	6.289302077	37.36620323	9.79E-10	1.62E-05
SPATA21	374955	-6.748577	2.257509442	16.82175707	4.11E-05	0.002789352
TRIM65	201292	-6.832591	2.748611874	12.50023886	0.0004069	0.009093665
ZNF584	201514	-6.853105	2.801508921	14.60562499	0.000132518	0.005233142
ZNF700	90592	-6.877556	6.782150941	13.17737869	0.000283349	0.007666863
ZNF561-AS1	284385	-6.953768	4.8915147	15.03184577	0.000105712	0.004540957
ZFP82	284406	-6.960809	5.131701406	12.92212234	0.000324721	0.008099915
RAPGEF4	11069	-7.078777	1.215355687	14.01814748	0.000181055	0.006061303
CPA4	51200	-7.136833	5.948087699	15.01365107	0.000106736	0.004573069
FAM174B	400451	-7.185559	2.416233454	13.30956073	0.000264056	0.007327117
ZNF292	23036	-7.226714	7.051376656	16.18846784	5.73E-05	0.003356438
ARHGAP28	79822	-7.268756	2.538813016	14.19902068	0.000164456	0.005874243
TCEANC	170082	-7.32122	4.236484343	14.75222732	0.000122603	0.005006425
ZNF285	26974	-7.374588	4.042459086	15.56080579	7.99E-05	0.00400272
ANKRD46	157567	-7.51377	4.14591672	14.75683738	0.000122303	0.005006425
TMEM267	64417	-7.584835	2.999239868	15.40949892	8.66E-05	0.004136991
DENND6B	414918	-7.596089	4.179380235	25.36021453	4.76E-07	0.000413996

HGNC symbol	Entrez Gene ID	log Fold Change	log CPM	Log Ratio	P-value	FDR adjusted P-value
DDHD1	80821	-7.670101	3.592303715	15.60390646	7.81E-05	0.003986122
	101928053	-7.744694	3.594817287	15.21621234	9.59E-05	0.004332268
ZNF620	253639	-7.908539	5.787952804	21.11794916	4.32E-06	0.001049694
MIR1302-9	100422831	-8.005222	4.838546482	12.81432675	0.000343975	0.008310241
KIAA0922	23240	-8.103099	3.715850789	13.7116491	0.000213128	0.006621864
PTCH2	8643	-8.221238	2.684898852	12.37646638	0.000434779	0.00953935
LMBR1L	55716	-8.29643	3.20744298	14.44929286	0.000143984	0.005415927
SHISA2	387914	-8.48004	5.914753645	18.19829998	1.99E-05	0.001902887
ZKSCAN8	7745	-8.557042	5.800297152	20.27694251	6.70E-06	0.001273677
EDN2	1907	-8.576043	4.500724668	12.94709429	0.000320418	0.008045913
NEB	4703	-8.59392	4.789161409	13.06562265	0.000300765	0.007840129
UBAP1L	390595	-8.626988	4.055236717	14.63257493	0.000130637	0.005193443
TULP2	7288	-8.722725	4.246888283	19.32562487	1.10E-05	0.001506424
MBD5	55777	-8.769197	2.957199276	14.88566244	0.000114228	0.004794672
ARC	23237	-8.979471	5.796699942	18.95640852	1.34E-05	0.001582074
AP4B1	10717	-9.043399	5.885627895	23.54734692	1.22E-06	0.00061079
TRIM66	9866	-9.121437	4.702617936	19.30495776	1.11E-05	0.001510335
	400512	-9.220656	3.579697556	12.75565809	0.000354934	0.008458063
FER1L5	90342	-9.33722	3.870985489	15.58223537	7.90E-05	0.003995079
	100505658	-9.549445	3.528708609	13.04148443	0.000304667	0.007872773
	101060498	-9.577044	3.982768811	12.50284767	0.000406332	0.009093264
HIST2H2BE	8349	-9.602855	5.369999903	17.80317693	2.45E-05	0.00212114
CLN5	1203	-9.759469	4.642449808	17.04174875	3.66E-05	0.002620477
DEPDC5	9681	-9.974682	5.219061306	13.45042992	0.00024495	0.006960453
ZHX3	23051	-10.01858	4.832239026	17.15556364	3.44E-05	0.002577228
WDR47	22911	-10.02017	4.533553234	15.65833784	7.59E-05	0.003897078
THAP9	79725	-10.14389	4.755678658	23.08315965	1.55E-06	0.000715267
	101929595	-10.3839	4.932831337	22.34503079	2.28E-06	0.00088036
MAGI2	9863	-10.51061	4.865230129	16.17263338	5.78E-05	0.003356438
PCSK1	5122	-10.57115	5.008580425	12.56194316	0.000393682	0.008992712
SNORD99	692212	-10.67375	4.859883381	28.94967738	7.43E-08	0.000170277
RGS2	5997	-10.75496	7.386671094	18.50764006	1.69E-05	0.001738283
FDX1	2230	-10.75539	4.696498539	21.78401695	3.05E-06	0.000901119
	100422737	-10.7703	4.540239917	14.07725077	0.000175453	0.005968147
NUP210L	91181	-11.15984	6.170722595	27.87808073	1.29E-07	0.000179292

Table 7: Differentially expressed genes between orange and blue clusters as determined by edgeR (FDR