Deconvolution of bulk gene expression profiles to characterize the tumour immune landscape of early onset breast cancer

by

Yong Won Jin

A thesis submitted to the Faculty of Graduate Studies at
the University of Manitoba
in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Biochemistry and Medical Genetics
University of Manitoba
Winnipeg, Manitoba, Canada

Statement of Contributions

Yong Won Jin is the sole author for Chapters 1, 2, 4, and 5 presented in this thesis, which was written under supervision by Dr. Pingzhao Hu. Results and written work in the chapters above were not submitted to or published elsewhere at the time of writing.

Published works presented in this thesis

Research work disclosed in Chapter 3 of this thesis was published previously under open access Creative Commons CC BY 4.0 license in the journal *Cancers*. **Yong Won Jin** was the primary author for the writing and research work in this publication. Dr. Pingzhao Hu provided resources, supervision, and contributed to study design as well as review of written work for publication.

Citation:

Y. W. Jin, and P. Hu, "Tumour-infiltrating CD8 T cells predict clinical breast cancer outcomes in young women," *Cancers*, vol. 12, no. 5, pp. 1076, Apr. 2020, doi: 10.3390/cancers12051076.

Published works not presented in this thesis

Research works that were outside the scope of this thesis but nonetheless conducted by Yong Won Jin during the course of the graduate program and published elsewhere are listed below.

Citations:

S. Frenkel, C. N. Bernstein, **Y. W. Jin**, M. Sargent, Q. Kuang, W. Jiang, J. Wei, B. Thiruvahindrapuram, S. W. Scherer, and P. Hu, "Genome-wide copy number variant data for inflammatory bowel disease in a Caucasian population," vol. 25, pp. 104203, Aug. 2019, doi: 10.1016/j.dib.2019.104203.

Y. W. Jin, S. Jia, A. B. Ashraf, and P. Hu, "Integrative data augmentation with U-net segmentation masks improves detection of lymph node metastases in breast cancer patients," *Cancers*, vol. 12, no. 10, pp. 2934, Oct. 2020, doi: 10.3390/cancers12102934.

Abstract

Introduction: Young age at diagnosis (age < 40) is considered to be an independent factor for poor clinical outcomes in breast cancer patients. Patterns of tumour infiltrating lymphocytes may provide insight into the underlying biology behind this disparity, which is yet to be discovered. Deconvolution algorithms can be used to characterize tumour infiltrating lymphocytes given the gene expression profile of the bulk tumour tissue sample. In this work, models of deconvolution, both novel and existing, were used to extract distinct patterns of tumour-infiltrating immune cells in early onset breast cancer that are significantly associated with clinical outcomes and other molecular signatures.

Methods: The tumour immune landscape was characterized by computational deconvolution of bulk tissue transcriptomes using an existing tool (TIMER) as well as developing a novel tool based on deep learning – neural network immune contexture estimator (NNICE). Pseudo-bulk gene expression profiles were simulated by leveraging single cell RNA-sequencing data from immune and breast cancer cells with known cell type compositions. Pseudo-bulk profiles were used to optimize deep learning model with deep quartile regression to provide estimates for cell fractions given bulk transcriptomes. Then, the characterized tumour immune landscape was associated with clinical outcomes in early onset breast cancer using large-scale breast cancer datasets.

Results: Immune cell abundance estimates from TIMER revealed that clinical outcomes of early onset breast cancer patients were more significantly affected by the abundance of CD8+ cytotoxic T cells, but to a lesser extent in the old patients. NNICE model of deconvolution produced more accurate predictions of cell type composition from bulk transcriptomes on training dataset. However, performance of NNICE model was not robust across different datasets, and the estimates on breast cancer datasets showed inconsistent results across cohorts.

Conclusion: The canonical survival disparity of early onset breast cancer patients was observed but the TIL landscape identified by current deconvolution algorithms do not give consistent results. Novel models of deconvolution built on state-of-the-art deep learning frameworks leveraging scRNA-seq data have potential to produce accurate estimates of cell type proportions; however, further research is needed to optimize these algorithms to be robust to differences between transcriptomic datasets.

Acknowledgements

First and foremost, I would like to acknowledge my supervisor Dr. Pingzhao Hu for supporting me throughout the 3 years of my Master of Science program. Under his guidance, I was able to learn and grow as a fledging scientist as well as achieving excellence in both research and academics. One thing I have always been appreciative about Dr. Hu is that he always feels proud of his students' achievements as if they were his own. For this reason, I always felt encouraged to embrace new challenges and opportunities.

I would also like to acknowledge the advisory committee members for their incredible support over the years. I would like to thank Dr. Sam Kung for providing me with new perspectives through his insightful questions and constructive feedbacks. I have always felt that his questions and comments were sincere, which I believe stem from his genuine concern and desire to understand. To Dr. Leigh Murphy, I would like to express my sincere gratitude for her continued support in my numerous applications for various opportunities. It is thanks to her guidance and encouragement that I was able to be so successful in many of my endeavors.

To my colleagues, Shuo, Nikho, Qian, Shujun, Wasif, Mohaiminul, Nikta, Svetlana, Rayhan, and Chengyou, among many others, I am thankful for time we spent together for meetings, discussions, and occasional outings. I would like to extend a special thanks to Shuo Jia for being my friend and a mentor in coding, and everyone's friend, Nikho Hizon, for bringing his positive energy and cultural interests to our regular lunch-hour discussions.

Lastly, I would like to thank the following funding agencies for providing financial support for me either directly or indirectly: Research Manitoba, CancerCare Manitoba, Canadian Breast Cancer Foundation, and the Natural Sciences and Engineering Research Council of Canada.

This thesis is especially dedicated to my beloved wife, Eun Ji Kim, who has provided strong mental and emotional support for me throughout my graduate studies.

I would also like to dedicate this thesis to my parents for their unconditional love and support throughout all the years of my life.

Finally, this thesis is dedicated to all those whose lives were affected by cancer, with hope that this research, albeit rudimentary, may provide insight for researchers and clinicians around the world who are working to defeat the disease.

Table of Contents

List of Tables

List of Equations

List of Figures

List of Abbreviations

ABIS	Absolute immune signature
AE	Autoencoder
AJCC	American Joint Committee on Cancer
ANN	Artificial neural network
APC	Antigen-presenting cell
BN	Batch normalization
BRCA	Breast cancer
CD	Cluster of differentiation
cGMP	Cyclic guanosine monophosphate
CI	Confidence interval
CIBERSORT	Cell-type identification by estimating relative subsets of RNA transcripts
CM	Central memory
COSMIC	Catalogue of Somatic Mutations in Cancer
CTLA4	Cytotoxic T-lymphocyte-associated protein 4
CV	Cross-validation
DB	Database
DC	Dendritic cell
DCIS	Ductal carcinoma *in situ*
DNN	Dense neural network
DO	Dropout
DQR	Deep quantile regression
DSF	Disease-free survival
ECM	Extracellular matrix
EM	Effector memory
EPIC	Estimating the proportion of immune and cancer cells
ER	Estrogen receptor
Ex	Exhausted
FACS	Fluorescence-activated cell sorting
FDR	False discovery rate
FoxP3	Forkhead box P3
FPKM	Fragments per kilobase million
GDC	Genomic Data Commons
GEO	Gene Expression Omnibus
GEP	Gene expression profile
GO	Gene Ontology
GSEA	Gene set enrichment analysis
GZMB	Granzyme B

HER2	Human epidermal growth factor receptor 2
HR	Hormone receptor
HR	Hazard ratio
IDC	Invasive ductal carcinoma
IHC	Immunohistochemistry
IL	Interleukin
IMS	Intrinsic molecular subtype
IntClust	Integrative clusters
KM	Kaplan-Meier
LD	Low-density
LumA	Luminal A
LumB	Luminal B
M0	Inactive macrophage
M1	Classically-activated macrophage
M2	Alternatively-activated macrophage
mAb	Monoclonal antibody
MAIT	Mucosal-associated invariant T
MCP counter	Microenvironment cell populations-counter
mDC	Myeloid dendritic cell
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
mRNA	Messenger ribonucleic acid
MSigDB	Molecular Signatures Database
MWU test	Mann-Whitney U-test
NK cell	Natural killer cell
NKT cell	Natural killer T cell
NNICE	Neural network immune contexture estimator
NSM	Non-switched memory
PAM50	Prediction analysis of microarray 50
PBMC	Peripheral blood mononuclear cell
PD-1	Programmed cell death protein 1
pDC	Plasmocytoid dendritic cell
PD-L1	Programmed death-ligand 1
PR	Progesterone receptor
PRF	Perforin
quanTIseq	Quantification of the tumour immune contexture from human RNA-seq data
R	Pearson correlation coefficient
ReLU	Rectified linear unit
RMSE	Root mean square error
RNA-seq	Ribonucleic acid sequencing

RSEM	RNA-seq by Expectation Maximization
SBS	Single base substitution
scRNA-seq	Single cell RNA sequencing
SERM	Selective estrogen receptor modulator
SM	Switched memory
TAM	Tumour-associated macrophage
TAN	Tumour-associated neutrophil
TCGA	The Cancer Genome Atlas
TE	Terminal effector
Tfh	T follicular helper
TGF-β	Transforming growth factor β
Th	T helper
TIL	Tumour-infiltrating lymphocytes
TIMER	Tumour immune estimation resource
TME	Tumour microenvironment
TNBC	Triple-negative breast cancer
TPM	Transcripts per million
Treg	Regulatory T cell
UMAP	Uniform manifold approximation and projection
VEGF	Vascular endothelial growth factor
WES	Whole-exome sequencing

Chapter 1. Background

1.1 Breast cancer

1.1.1 Epidemiology

Breast cancer is the most prevalent type of cancer in women worldwide [1]. Specifically, for the year 2020 in Canada, it is estimated that 27,400 women will be diagnosed with breast cancer, which accounts for around 25% of all new cancer cases in females [1]. In terms of mortality, breast cancer remains the leading cause of cancer-related deaths in women after lung cancer. 5,100 Canadian women were projected to die due to breast cancer in 2020, which accounts for 13% of all cancer deaths [1]. Conversely in men, breast cancer is a rare disease and was projected to account for only around 0.2% of all new cancer cases in males for the year 2020 in Canada [1]. This is because men have less breast tissue and different hormonal environment compared to women [2]. Although it is important to study various aspects of breast cancer in male patients, they have been excluded in this work because it is difficult to make meaningful associations with scarcely available data on male breast cancer patients. Therefore, further mentions of breast cancer in this work will be referring to breast cancer in women unless otherwise noted.

Numerous large-scale studies were dedicated to the advancement of early detection, accurate prognosis, and effective therapeutics for breast cancer [3,4]. These efforts are most likely responsible for the continuous reductions in breast cancer mortality rate over the past 45 years. The incidence rate for breast cancer increased during the 1990s, most likely due to the initiation of breast screening programs but has since remained relatively stable over the recent years [1,5]. However, despite extensive research and efforts over the past decades to detect, prevent, and treat breast cancer, it still remains a leading cause of cancer-related deaths in women. It shows that

although we have made significant advancements thus far, there is much work to be done to understand the disease and improve the prevention, detection, and treatment of breast cancer.

1.1.2 Subtypes

The major challenge with breast cancer is in its heterogeneity. There are a number of clinically relevant methods to classify breast cancer, and each subtype of breast cancer is associated with differential prognoses and treatment options [5]. Core needle biopsies are often performed to obtain samples of breast tumour tissue to confirm the presence of malignant neoplasms, characterize their origin (ductal vs. lobular) and grade based on microscopic morphology [6,7]. Invasive ductal carcinoma (IDC) is most common at 55% of new cases [8], followed by ductal carcinoma *in situ* (DCIS) at around 20% [9]. Invasive forms of breast cancer have worse prognosis relative to those *in situ*, which are often classified as stage 0 tumours. In addition, the Ki-67 antibody is used in immunohistochemistry (IHC) to evaluate the extent of proliferation of malignant breast tumour cells [10]. A positive test (Ki-67+) is associated with aggressive growth rates, and relatively poor prognosis. Furthermore, tumours with higher histological grade tend to be less differentiated, and therefore, more aggressive as well as being resistant to therapeutics. Samples of sentinel lymph node from regions nearby tumours can also be assessed by pathologist to characterize extent of lymph node metastasis (nodal involvement). Presence of lymph node metastasis is traditionally interpreted as indications for poor prognosis [6]. Taken together, these tumour characteristics are used to differentiate tumour stages, often with reference to the American Joint Committee on Cancer (AJCC) guidelines [11].

For breast cancer, the receptor status of tumours has different clinical implications for response to treatments and prognosis. It has been known that the breasts are sensitive to changes in levels of hormones such as estrogen and progesterone. Breast tumour cells may exhibit high levels of the

estrogen receptor (ER) and/or the progesterone receptor (PR) that facilitates the dysregulated proliferation of the neoplasm in response to their respective ligands. The extent to which these receptors are present can be evaluated with histopathology of tumour sample sections after IHC staining. Patients with hormone receptor positive (HR+) tumours, with ER+ and/or PR+ status, can be treated with hormonal therapies using selective estrogen receptor modulators (SERMs) such as tamoxifen that block the effects of estrogen in the breasts [7]. In general, HR+ breast tumours are less aggressive than HR- tumours; therefore, patients with HR+ breast cancer have relatively better prognosis. In addition, the overexpression of human epidermal growth factor receptor 2 gene (*HER2*) in breast cancer cells is known to facilitate proliferation of malignant cells. HER2+ tumours tend to be more aggressive than HER2– tumours; however, targeted treatment regimens are available that include monoclonal antibodies against the HER2 protein such as trastuzumab and pertuzumab [6]. These treatments developed in the recent years have contributed to the significant improvements in survival for a subset of patients with HER2+ breast cancer [6]. A relatively smaller subset (10-15%) of breast tumours that exhibit no signs of overexpression for ER, PR, and HER2 are classified as triple-negative breast cancer (TNBC) because of their receptor status (ER–/PR–/HER2–) [12]. Patients with TNBC have the worst prognosis, mainly because of its aggressive nature as well as the lack of available targeted treatments [12].

Advancements in human genetics and genome sequencing technologies have enabled researchers to characterize the molecular heterogeneity of breast cancer. This enabled further subtyping of breast tumours into intrinsic molecular subtypes (IMS) based on their various genomic signatures [13]. Most widely used is the Prosigna breast cancer prognostic gene signature array, formerly known as the Prediction Analysis of Microarray 50 (PAM50) test [14]. This array uses expression of 50 genes to distinguish between four IMS: luminal A (LumA), luminal B

(LumB), HER2-enriched, and basal-like [14]. Interestingly, these molecular subtypes partially overlap with IHC classifications based on receptor status [14]. Large-scale study of breast cancer by Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) further classified breast tumours into ten integrative clusters (IntClust 1–10) based on copy number aberration profiles [3]. Notably, subgroups of patients stratified by IntClust subtypes also showed distinct clinical outcomes relative to one another [3].

Year after year, numerous studies define new subtypes or refine existing classification systems for breast cancer. Altogether, these studies re-emphasize the heterogeneity of breast cancer as a complex disease and continue to inspire advancements towards personalized medicine to find cure for all breast cancer patients.

1.1.3 Risk factors

There is no single factor that is responsible for causing breast cancer, making it a disease with complex etiology. The sensitivity of breast tissues to hormonal changes in women adds to this complexity; hormonal and reproductive factors were found to be significantly associated with breast cancer risk. These risk factors include histories of parity, menarche, breastfeeding, use of oral contraceptives, and hormone replacement therapy [5]. Furthermore, lifestyle choices such as diet, alcohol consumption, and smoking have been associated with breast cancer risk [5]. Similar to most other cancer types, genetics is an important risk factor to consider, whether with or without family history of breast cancer [6]. Germline mutations in the *BRCA* genes are well known genetic risk factors for hereditary breast cancers, which account for approximately 5-10% of all breast cancer cases [5]. Most importantly, risk of developing breast cancer has been observed to increase with age [15]. Only 5-7% of all breast cancer cases occur in young women under the age 40, while

80% occur in women over the age 50 [15]; however, younger patients are known to have particularly worse prognosis compared to older breast cancer patients [5].

1.2 Early onset breast cancer

Along with increasing risk, the age-specific incidence rate for all subtypes of breast cancer increases steadily until age 50, where an inflection point occurs [15]. After this point, incidence rate for HR+ breast cancers continue to rise, whereas the incidence rate for HR– breast cancers decrease slightly [15]. Similar observations were made with regards to the PAM50 subtypes, where the incidence of luminal subtypes (LumA and LumB) increased with age while incidence of basal-like breast cancer decreased with age [16]. This marked inflection point is known to be the superposition of two different incidence rate curves for: 1) early-onset breast cancer, with modal age at diagnosis of ~50 years; and 2) late-onset breast cancer, with modal age at diagnosis of ~70 years [15]. This suggests that the biology and etiology may differ for the same disease depending on the age at onset of breast cancer.

Young age at diagnosis of breast cancer is considered to be an indicator for poor prognosis. Many studies have used various age thresholds (<35, <40, and <45 years) to define "young age" at diagnosis for breast cancer patients [15,17]. However, regardless of age threshold used, researchers have consistently shown that a disproportionate number of young breast cancer patients present with aggressive forms of the same disease and poor prognosis. In this work, "young age" or "early onset" of breast cancer is defined as patients diagnosed under age 40. Young women under the age 40 were most frequently diagnosed with HER2-enriched (16%) and basal-like (44%) subtypes that are associated with poor clinical outcomes, and with increasing age, the proportions of older patients diagnosed with these aggressive subtypes consistently decreased to 11% for HER2-enriched and 9% for basal-like subtypes [16]. Furthermore, histopathological

analysis revealed that significantly larger proportion of young breast cancer patients under age 35 presented with higher grade tumours and lymphatic-vascular invasion [18]. More importantly, young women with breast cancer under age 40 showed significantly lower disease-free survival (DFS), with a plateau at less than 50%, compared to their older counterparts with a plateau of around 65%, ten years after initial diagnosis [17]. Thus, it is important to consider age as a prognostic indicator, especially for women diagnosed with breast cancer under the age 45.

Findings above suggest that early-onset breast cancer may possess unique biological features or show distinct etiology that result in the differences observed at both population and individual levels with regards to incidence and clinical outcomes, respectively. At the subcellular level, gene expression analysis demonstrated that many age-correlated transcripts are expressed in a bi-phasic pattern, with an inflection point between 50-60 years of age in large cohorts of breast cancer patients [19]. This pattern is attributed to the physiological changes during menopause, which occurs in women around 50-52 years of age [19]. In particular, genes encoding the ER (*ESR1* and *ESR2*), were found to be differentially expressed [20] and its expression negatively correlated with age [17, 19], only in the younger cohorts of breast cancer patients. While these findings are in accord with the observations that HR– breast cancer is more frequent in younger patients, they offer little insight for the potential unique biology of early-onset breast cancers. Furthermore, while thousands of genes were found to be significantly enriched specifically in younger breast cancer patients, none remained significant after controlling for clinicopathologic features such as tumour grade and IMS [21]. Therefore, identification of differences between early- and late-onset breast cancers at the subcellular level remains challenging, despite expectations of pathophysiological differences due to aging. At the cell-level, histopathological features indicative of poor prognosis such as high grade and lymphovascular invasion of tumour cells were more frequently observed

in early-onset breast cancers [18]. Otherwise, no mentions of apparent histological differences between early- and late-onset breast cancer could be found. There is yet to be any study that investigates whether non-malignant cells in the tumour region differ at both subcellular and cell-level biology. In this regard, further research is warranted to explore the potentially unique biology of early-onset breast cancer in context of young, pre-menopausal women under the age 40, who most frequently suffer from poor clinical outcomes and remain consistently underrepresented in large cohort studies because of their relative low incidence rate.

1.3 Tumour microenvironment

One way to view the tumour is to see it as a perturbed version of normal physiological tissue. More than half of primary tumours and their metastases are non-malignant cells by mass, that are collectively referred to as the tumour-microenvironment (TME) [22]. The breast TME can consist of all the cells that belong in a normal breast tissue in addition to the tumour mass: stromal cells such as fibroblasts, adipocytes, and endothelial cells that form the vasculature and lymphatics; normal cells that form the mammary glands such as luminal epithelial cells and myoepithelial cells; and cells of the immune system [23]. While these cells seem to have normal morphology under the microscope, non-malignant cells adjacent to the cancerous cells can have both tumourigenic and/or anti-tumour functions at all stages of the disease [22].

One of the most significant clinical manifestation of the association between the breast stroma and tumourigenesis is that mammographically dense breasts have increased risk of developing breast cancer [24]. In addition, differential stromal gene expression signatures have been identified that are associated with particular clinical subtypes of breast cancer [25, 26] and clinical outcomes [27]. Allinen and colleagues discovered that the highest proportion of differentially expressed genes were associated with the myoepithelial cells between samples from normal and DCIS [25].

Furthermore, between DCIS and IDC, epithelial and stromal cell types were found to have the highest differentially expressed genes, including increased expression of matrix metalloproteinases that remodels the basement membrane [26]. However, both studies suggest these changes in gene expression and subsequent functional alterations of cells in the TME do not originate from changes in the genome but instead from the complex interactions between the tumour and its microenvironment [23]. This makes sense from the tumour perspective because it would be more efficient for them to produce minimal amount of proteins that interacts and manipulates their environment to work in favour of their growth and survival.

Interactions between the tumour and its microenvironment can include chemical interactions in the form of secreted molecules (ex. growth factors), mechanical interactions, or physical interactions between binding of cell surface markers [22, 23]. Through these various modes of interaction, the tumour growth and development can be stimulated or suppressed. As an example of chemical interaction, growing tumours can induce hypoxia in the TME and secrete vascular endothelial growth factors (VEGFs) that stimulate the growth of new vasculature in the TME [22]. At the same time, hypoxic state is detrimental to tumour growth; therefore, angiogenesis inhibitors, such as bevacizumab and ramucirumab that bind to VEGF, have shown promising results on patients with metastatic HER2-negative breast cancers [28]. However, studies on animal models have shown increased metastatic programming and invasion of cancer cells, potentially in response to these pharmaceutical agents [23]. As an important substrate for mechanical communication between cells, the extracellular matrix (ECM) also acts as a physical barrier to metastasis. Numerous proteins and substrates can be released by both malignant and non-malignant cells to physically remodel, stiffen, or degrade the ECM [29]. Modulation of the ECM is a critical step for disease progression from *in-situ* to invasive carcinoma, which includes the loss of the

myoepithelial layer and basement membrane [23]. Low coverage of blood vessels by pericytes, which play an important role in remodelling of the basement membrane during angiogenesis, is correlated with metastasis and poor outcomes [30]. In addition, mechanical signalling between cells, either directly via cell-cell junctions or indirectly via the ECM, is crucial for cell survival and proliferation. It is known that cells have intrinsic preference for rigid and stiffer ECM. Stiffer ECM can increase the metastatic potential for tumour cells as well as inducing angiogenesis [29]. Tumour cells are often capable of stiffening the TME by secreting crosslinking enzymes, increased deposition of ECM substrates, and mechanical contraction of tumour cells themselves [29].

As such, it is important to recognize the interactions between stromal cells and the adjacent tumour cells; however, this thesis will focus on a specific subset of immune cells infiltrating the TME, from henceforth referred to as the tumour immune landscape.

1.3.1 Tumour immune landscape

A critical role of the immune system is the surveillance, detection, and destruction of malignant cells within the body. Under normal physiological circumstances, cytotoxic cells of both the innate and adaptive immune systems can be expected to exert anti-tumour effects. This is one of the reasons why cancers are frequently associated with inflammation [31]. The most abundant immune cell types in human tumours are the macrophage, followed by T lymphocytes [31]. In most cancers, including breast cancer, higher levels of T lymphocytes within the TME have been associated with favourable clinical outcomes [22,23], whereas macrophages in the TME known as tumour-associated macrophages (TAMs) have been known to be pro-tumourigenic [31]. However, it is important to realize that these immune cell types can be further distinguished into more specific subsets that respond differently to malignant cells.

T lymphocytes are the second most abundant immune cell type in most human cancers, representing up to 10% of all cells in and near the tumour area [22]. Although they are generally associated with favourable prognosis, not all T lymphocyte subsets contribute to the anti-tumour response in the same way. CD8+ T cells are known to actively kill tumour cells by secreting cytotoxic molecules such as perforin (*PRF1*) and granzyme B (*GZMB*) after becoming activated with the help of CD4+ T helper 1 (Th1) cells. Within the classical Th1/Th2 framework, Th1 response is generally favourable over Th2 response in the TME; therefore, indications for T helper 2 (Th2) responses are often connected to poor prognosis [22]. A subset of T lymphocytes known as regulatory T cells (Tregs), characterized by high levels of the transcription factor FoxP3, have consistently been associated with poor clinical outcomes in various cancers including breast cancer [22,31]. This is because of the immunosuppressive nature of these cells against the tumouricidal cytotoxic immune response executed by the aforementioned CD8+ cytotoxic T and Th1 lymphocytes. Other T lymphocyte subsets such as γδ T cells and natural killer T (NKT) cells are usually linked with favourable prognosis but it is uncertain to what extent these rare subsets of T lymphocytes have on the tumour and its microenvironment [22].

Macrophages are an important component of the immune system that initiates the adaptive immune response as one of the antigen-presenting cells (APCs). They are capable of phagocytosis and presentation of neoantigens that are produced by malignant cells to initiate a tumour-specific immune response. Macrophages that follow this classical activation pathway are referred to as M1 macrophages and this subset of cells have pro-inflammatory downstream effects involving Th1 response and type 1 cytokines [32,33]. Higher abundance of these M1-polarized macrophages is associated with better prognosis in various cancers including breast cancer [30]. However, there is another subset of macrophages that follow an alternative activation pathway referred to as M2

macrophages that are anti-inflammatory and involved in wound-healing processes [34]. This M2 polarization of macrophages is the default state of tissue-resident macrophages [33], which also accumulate in response to hypoxia and necrosis – both of which are common in the TME. Within particular tumour types, macrophages can account for up to 50% of the tumour mass [34]. M2 macrophages are considered to be pro-tumourigenic because they play an immunosuppressive role as well as signalling for angiogenesis, and hence, are associated with poor clinical outcomes in most cancers [32].

Although the macrophage and T lymphocytes are found most frequently within the TME, the immune system consists of other cell types that interact with the tumour and its surrounding TME. The B lymphocyte, responsible for humoral immune response, is found in the TME but more commonly in the adjacent lymphoid tissues [22]. Although B cells do not exert cytotoxic effects directly upon the malignant cells, they are capable of producing antibodies against neoantigens. B cells have previously been associated with good prognosis in breast and ovarian cancers [22]. However, the role of B lymphocytes in the TME remain inconclusive as many other sources suggest that they promote tumour growth and cause chronic inflammation [31]. Natural killer (NK) cells of the innate immune system are capable of direct killing of malignant cells. They are associated with good clinical outcomes in cancers but compared to other immune cell types these cells are rarely observed within the TME [22, 35]. Moreover, NK cells are reported as being in a suppressed state by levels of TGF-β within the TME [22, 36]. Dendritic cells (DCs) have important roles in initiating type 1 responses as one of the APCs in the immune system. However, the lack of immune responses against tumour-associated antigens is evidence that DCs are rendered defective within the hypoxic and inflammatory TME [22]. Although traditional understanding of DCs suggest they play a critical role in anti-tumour responses, their behaviour within the TME

remains unclear [31]. Even more controversial is the role of tumour-associated neutrophils (TANs) within the TME. Pro-tumourigenic roles of TANs include supporting angiogenesis, metastasis, and immune suppression, while their anti-tumour functions include enhancing activity of $CD8^+$ T cells and effectiveness of radiotherapy [22, 31]. It is without a doubt that TANs play a significant role in cancers, but the heterogeneity of their functions within the TME make it difficult to characterize the direction of its effects [31].

In summary, the tumour immune contexture is composed of diverse immune cell types that interact with the tumour cells. Immune cells involved in type 1 response such as $CD8^+$ T cells, Th1 cells, and M1 macrophages are often associated with favourable prognosis in cancer, whereas cells involved in type 2 response or immunosuppression such as Tregs, Th2 cells, and M2 macrophages are associated with poor survival [36]. The role of cells such as B lymphocytes, NK cells, DCs, neutrophils, and other cell types within the TME are ambiguous and remains to be clearly understood in the context of different cancers.

1.3.2 Cancer immunoediting, immune evasion, and immunotherapy

As illustrated above, the immune system is capable of both inhibiting and promoting tumour growth. In highly infiltrated tissues, the immune system eliminates immunogenic malignant cells presenting with diverse neoantigens; at the same time however, it exerts a selective pressure that allows for the clonal propagation of non-immunogenic malignant cells [37]. This process called cancer immunoediting consists of three phases that occur throughout tumour development: elimination, equilibrium, and escape [37,38]. The elimination phase represents the initial tumouricidal immune responses by both the innate and adaptive immune system against the immunogenic malignant cells. When subsets of tumour cells survive the elimination phase, the tissue is considered to be in equilibrium phase, where the rates of tumour growth and elimination

by the immune system are in equilibrium. In this phase, heterogeneity of the tumour is decreased and only the populations that are immune-resistant survive and proliferate. Tumours that grow to a size that is clinically detectable are considered to be in the escape phase, during which the proliferation of cancer cells can no longer be controlled by the immune system.

For malignant tumour cells, survival depends on their ability to evade detection by immune surveillance while taking advantage of immunosuppressive mechanisms to avoid tumouricidal attacks from the immune system. To evade immune surveillance, tumour cells stimulate expression of checkpoint molecules such as programmed death-ligand 1 (PD-L1), and cytotoxic T-lymphocyte-associated protein 4 (CTLA4) within the TME [31]. Tumour cells are also able to inhibit immune response by secreting immunosuppressive cytokines such as IL-10, TGF-β, prostaglandin E2, and VEGF to modulate the immune system within the TME [31]. Under normal physiological circumstances, these mechanisms are indispensable for self-tolerance and preventing autoimmunity; however, in various cancers, these mechanisms are taken advantage of by malignant cells to support their uncontrolled survival and growth within the tissues.

Most recent advances in cancer treatments include the immunotherapy regimens that were designed to combat the tumour cells armed with these defenses against the immune system within the TME. In reference to the cancer immunoediting model, cancer immunotherapies are aimed at forcing the tumour and TME in the escape phase back to the equilibrium phase (partial response) or the elimination phase (complete response) [37]. Different modalities of cancer immunotherapy include the anti-tumour vaccines, adoptive cell transfers, oncolytic viruses, and tumour-targeting immunotherapies [38]. The first successful immunotherapy approved for treatment of HER2 over-expressing metastatic breast cancer was trastuzumab – a monoclonal antibody (mAb) against the HER2 protein [38]. By binding to HER2 overexpressed on surface of tumour cells, these mAbs

can block signals for survival and proliferation. More recently, the atezolizumab was approved for treatment of metastatic TNBC, which makes it the first immune checkpoint inhibitor treatment available for breast cancer [38]. Atezolizumab is an mAb that binds to PD-L1 and allows for anti-tumour responses by cytotoxic cells by disabling PD-1/PD-L1 interaction. Cancer immunotherapy ultimately relies on the immune system to eliminate tumour cells, thus, only a subset of patients with an inflammatory/immune-infiltrating ("hot") TME responds to immunotherapies [37]. Since breast cancer is generally characterized as "cold" tumours with minimal infiltration by effector immune cells [38], it is important to distinguish those small subsets of patients with significant immune infiltration in the TME when administering immunotherapy. Refer to **Figure 1.1** for an illustrative example of tumour responses to cancer treatments dependent on extent of immune infiltration.

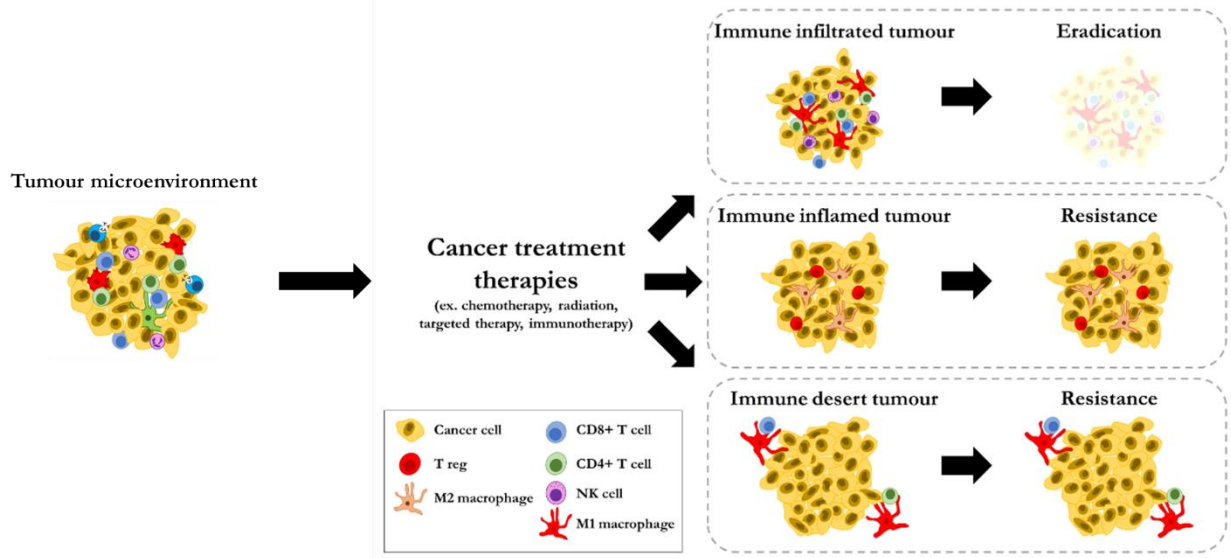


Figure 1.1 Illustration of the potentially differing responses to cancer treatments by tumours with various levels of immune infiltration.

1.3.3 Immunosenescence

Immunosenescence refers to the changes in the immune system with aging. Usually these changes entail a decrease in lymphocyte potential and increase in myeloid potential. Analysis of

peripheral blood mononuclear cells (PBMCs) from healthy adults aged between 22 and 93 years revealed that frequencies of both $CD4^+$ and $CD8^+$ T cells were affected by age in both men and women, especially significant decreases in naïve $CD8^+$ T cells with age [39]. Meanwhile, no significant age-related changes were observed for $CD14^+$ monocytes [39], suggesting an increase in myeloid-to-lymphocyte ratio. Immunosenescence is considered to be a contributing factor to a process referred to as "inflammaging" resulting in a systemic increase in low-grade chronic inflammations with age [40]. This is thought to be due to the mild increases in innate immune activity that serves to compensate for the declines in the adaptive immune system [41]. This chronic inflammation due to inflammaging leads to tissue damage and interferes with acute inflammatory processes [40,41]. Interestingly, in a mouse model of TNBC treated with anti-PD-L1 or anti-CTLA4 treatments, tumours in younger mice showed significant regression whereas older mice did not respond to given treatments [40]. Given these differences exist in breast cancer between young and old patients, characterization of their distinct tumour immune landscapes is necessary to investigate whether the age-related differences in the immune contexture of the TME has impact on clinical outcomes in breast cancer.

1.4 Characterizing the tumour microenvironment

Several methods can be used to characterize the immune contexture within the TME. Traditionally, cell phenotyping was done through histology – by observation of cell and tissue morphology using light microscopy. With identification of cell-type-specific surface markers, antibodies could be utilized to identify target cell populations within the tissues using immunohistochemistry and immunofluorescence [42]. Furthermore, advancements in flow cytometry enabled high-throughput enumeration of cells as well as developments in cell sorting technique such as fluorescence-activated cell sorting (FACS). Experimentally, FACS is considered

to be the gold standard method of quantifying cell type abundance within bulk tissues [43]. However, FACS remains as an expensive, and labour-intensive technique as well as being limited to prior knowledge for cell type identification by known surface markers [44]. Meanwhile, numerous computational cell sorting algorithms have emerged over the recent decades that provide alternative methods to quantify cell type abundance using genomic data. Bulk genomic data such as RNA-seq (RNA sequencing) represent a mixture of signals from bulk tissue with between hundreds to thousands of cells of diverse cell types. The computational cell sorting methods is commonly referred to as "deconvolution" as coined by pioneering researchers [45], "cell type decomposition", "computational microdissection", or "digital cell sorting".

1.4.1 Gene expression deconvolution

Gene expression deconvolution algorithms view the bulk mixture expression profile as a linear combination between cell fractions and a signature matrix (or basis matrix) [42]. This signature matrix is usually designed by researchers based on curated gene expression profiles (GEPs) from bulk tissue samples purified for individual cell types of interest. One way to interpret signature matrix is that it is the average expression value of each gene in the cell type of interest. Therefore, the total signal – represented by read counts in RNA-seq and fluorescence intensity in expression microarray – is a sum of the mixture of signals from various cell types in accordance with their respective proportions within the tissue. Refer to **Figure 1.2** for a graphical abstract of how gene expression deconvolution models are developed.

The first work that formulated the expression deconvolution model was by Venet and colleagues in 2001 where they proposed that the computational model as a solution to "deconfound" bulk transcriptomics analyses from the effects of tissue cell type heterogeneity [46]. Since then, the same model was proposed as a solution for diverse problems in estimating tumour purity [47-

49], cell type proportions [44], stromal content [48], and cell-type-specific marker genes [50] from bulk transcriptomics data. The unifying purpose of these models is to solve for the system of linear equations that represent the relationship between the mixture GEP, signature matrix, and cell type fractions within bulk tissue. To solve for this ill-conditioned problem, researchers have deployed various regression-type algorithms such as non-negative least squares [48], robust linear regression [51] and quadratic programming [52], as well as non-negative matrix factorization [53]. Previous studies have demonstrated that computational expression deconvolution models are capable of producing robust estimates that are highly correlated with cell type proportions within bulk tissue as quantified by FACS [51], histology [48,49], or known proportions used to simulate pseudo-bulk GEPs [54]. Here, "robustness" of estimates refer to the desirable capabilities of the computational models to produce precise and accurate estimates across different datasets.

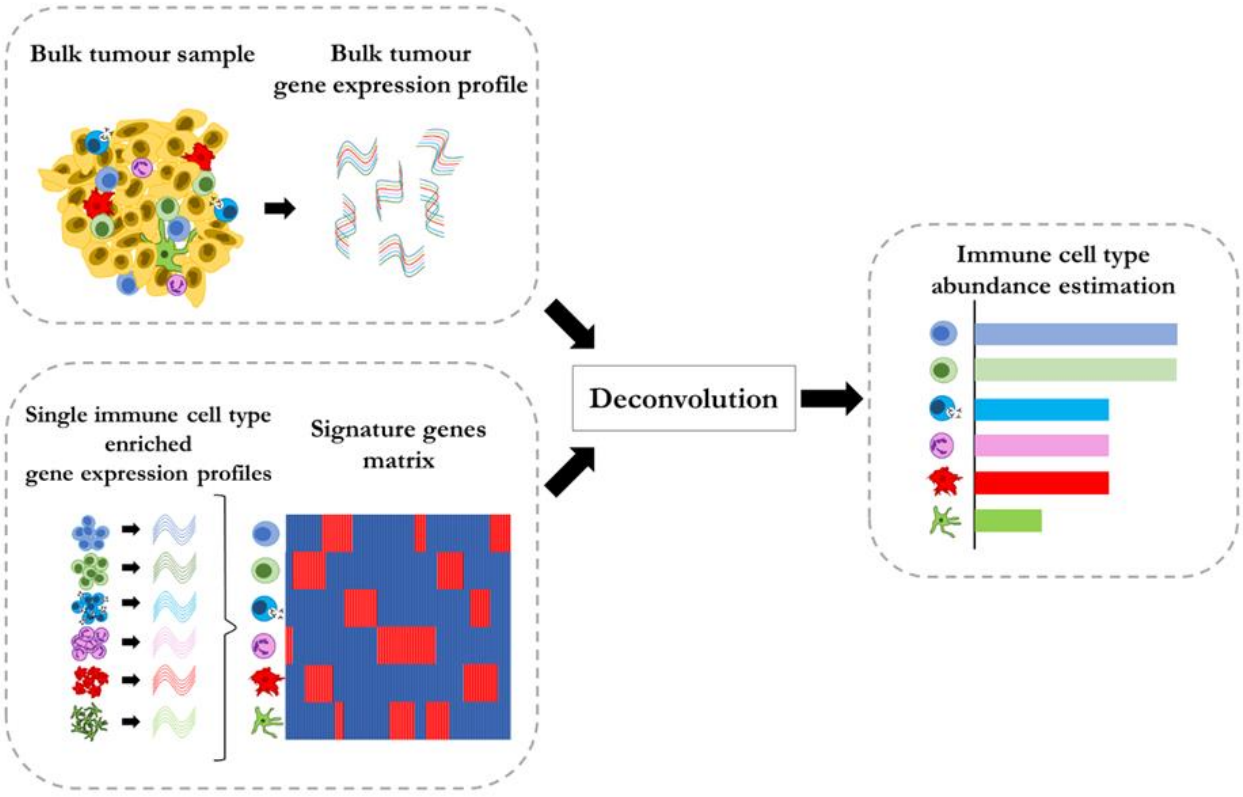


Figure 1.2 Illustration of how conventional gene expression deconvolution algorithms are developed

Chapter 2. Rationale, Hypothesis, and Objectives

2.1 Rationale

Young age (under 40) at diagnosis is known to be an independent factor for higher risk of recurrence and poor survival in breast cancer patients. Many factors may contribute to this disparity in clinical outcomes, all of which have yet to be explored in depth. Many physiological processes undergo changes as we age; therefore, it is reasonable to speculate that these age-related changes contribute to the differences in clinical outcomes between early and late onset breast cancers. Furthermore, a thorough study of the immune system and the extent to which it contributes to clinical outcomes in early onset breast cancer is warranted, recognizing that the immune system deteriorates with age – a process known as immunosenescence – as well as playing critical roles in tumour development, progression, and regression within the tumour microenvironment.

Diversity and composition of cell types within a bulk tissue sample can be identified in a number of different ways. Traditional histological techniques and modern flow cytometers are examples of experimental methods that can be used to characterize the tumour immune contexture. However, recent developments in computational methods, known as deconvolution, are capable of providing such information in a fast and cost-effective way given existing genomic data.

2.2 Hypothesis

Distinct patterns of tumour infiltrating immune cells, identified by computational expression deconvolution of bulk gene expression profiles, are associated with clinical outcomes in early onset breast cancer.

2.3 Research objectives

Aim 1. Use existing tools for expression deconvolution to find distinct tumour immune landscapes in early onset breast cancer patients.

Aim 2. Develop a novel model for expression deconvolution based on the deep learning framework.

Aim 3. Extract and evaluate markers of tumour immune infiltration from estimates of immune cell type composition made by computational deconvolution.

Aim 4. Investigate the associations between levels of tumour infiltrating immune cells and clinical outcomes.

Chapter 3. Characterizing tumour immune landscape of early onset breast cancer using existing tools for expression deconvolution

3.1 Motivations

There are clinical disparities in pathological features and disease outcomes between younger (age at diagnosis <40) versus older (age at diagnosis ≥40) breast cancer patients [17,21,55]. Young women with breast cancer are more commonly diagnosed with aggressive, invasive types of breast cancer that are difficult to treat. Studies found survival to be inversely associated with age at diagnosis [17,56,57]. The underlying biological cause for this age-dependent disparity in survival outcome is still unknown [55].

Recent studies have provided accumulating evidence that the presence of tumour-infiltrating lymphocytes (TILs) and their composition show significant associations with prognosis and response to cancer treatments [58,59]. Specific TILs such as CD8+ cytotoxic T lymphocytes, T helper 1 cells, M1 macrophages, natural killer (NK) cells, and T-follicular helper (Tfh) cells have been reported as exhibiting anti-tumour activities, whereas T-regulatory (Treg) cells and M2 macrophages are known for their immune-inhibitory and thus pro-tumour activities [58]. Furthermore, measures of TILs have been shown to be markers for pathological complete remission, chemosensitivity, and improved recurrence-free survival, especially in non-luminal, receptor-negative breast cancers [60-62].

Today, there is an abundance of bulk transcriptomic data available publicly online [3,4,63]. These bulk transcriptomes often represent the average gene expression across a heterogeneous mixture of cells. If cellular components and their proportions can be identified from bulk

transcriptomic data by computational methods, such in-silico methods can be used to characterize and quantify immune infiltrates in a cost-, time-, and labor-effective manner.

Deconvolution is a computational problem of simplifying a complex mixture into its individual constituents. In brief, most deconvolution algorithms see bulk transcriptome as a mixture where one gene of the mixture is a linear combination of that gene expressed across different cell types, weighted by the proportions of those cell types [43]. Deconvolution algorithms were previously applied to a large breast cancer transcriptomic dataset by others to characterize immune infiltrates across multiple samples [64,65]. After stratification by estrogen receptor (ER) status, Ali et al. found CD8+ T cells and activated memory T cells to be associated with favorable clinical outcomes in ER-negative tumours, which supported similar findings in the literature [36,58,64].

The tumour immune landscape, as predicted by computational deconvolution, has the potential to provide prognostic information as well as providing insight into immune functions within solid tumours. If age group differences in tumour immune landscape exist, these differences may be able to explain the age-dependent disparities in clinical outcomes in breast cancer.

3.2 Materials and methods

3.2.1 Data

From The Cancer Genome Atlas (TCGA) database, gene-level RNA-seq expression data of patients with primary breast cancer (BRCA) were downloaded from Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov/) through TCGA-Assembler 2 [4,80]. The gene expression data had previously been processed by the RNAseqV2 pipeline, providing estimated counts and scaled estimate values from RNA-seq by Expectation Maximization (RSEM) [81]. Scaled estimates were converted to transcripts per million (TPM) values by multiplying by one million, resulting in a bulk RNA-seq dataset of 20,501 genes from 1095 breast tumour samples.

We used RSEM-processed TPM measure because it was used previously to develop the TIMER method [49]. Mutations data from whole-exome sequencing (WES) and clinical data associated with the TCGA-BRCA cohort were downloaded from the cBioPortal for Cancer Genomics (https://www.cbioportal.org/) [82]. Samples from male patients (n = 12) were excluded from analyses, as well as any samples without associated age or survival data, resulting in final sample size of 989. There are inconsistencies in age thresholds used to define early onset breast cancer [17,55-57]. Here, we used a relatively conservative, but commonly used threshold (age at diagnosis = 40) to define it. Hence, of the 989 samples, 70 were identified as young/early-onset (age <40), and 919 were identified as old (age ≥40).

From the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort, gene-level microarray expression data and associated clinical data were downloaded from cBioPortal [3,82]. The gene expression data had previously been processed to log2 intensity values as described in the original publication [3]. Samples without associated age or survival data were removed, resulting in bulk microarray dataset of 24,360 genes from 1903 breast tumour samples, of which 116 were young and 1787 were old.

3.2.2 Immune cell type deconvolution

Tumour immune estimation resource (TIMER) was used to deconvolute bulk gene expression data to estimate immune cell abundance [49]. In brief, most deconvolution algorithms see bulk transcriptome as a mixture where one gene of the mixture is a linear combination of that gene expressed across different cell types, weighted by the proportions of those cell types [43]. This can be represented by a simple equation of a linear model:

$$y_i = \hat{\beta}_{i1}x_{i1} + \hat{\beta}_{i2}x_{i2} + \hat{\beta}_{i3}x_{i3} + \cdots + \hat{\beta}_{ik}x_{ik} \quad 3.1$$

where y_i is the gene expression level of a single gene i in the mixture data of n number of genes, $x_1, x_2, \ldots, x_k$ are gene expressions of the one gene across k cell types, and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are relative fractions of those cell types [43,49]. Given a signature expression matrix X, consisting of n rows for number of genes and k columns for number of cell types, the deconvolution algorithm produces estimates of cell fractions ($\hat{\beta}$). Specifically, TIMER utilizes constrained linear regression with non-negativity constraint on cell fractions to estimate the abundance of six immune cell types: B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells [49]. For settings, the sample tumour type was set to "BRCA" for breast cancer, and the default reference data and gene set were used for both cohorts.

The TIMER method was available for quick implementation through the immunedeconv package [43,49]. Expression data with rows as genes and columns as samples was provided to the algorithm as input, with which the algorithm estimated abundance of six immune cell types. Results were visualized as heatmaps and violin plots using the pheatmap and ggpubr packages, respectively.

3.2.3 Survival analyses

Kaplan–Meier (KM) survival curves were visualized using survival and survminer packages. Maximally selected rank statistics (maxstat) as implemented in the survminer package was used to determine the optimal cut-off to binarize the immune cell abundance estimates from TIMER into "high" and "low" groups which produces the maximum log-rank statistic on disease-free survival (DFS) [83].

Cox regression was used to compute hazard ratios for each of the six immune cell types for which abundance was estimated by TIMER. Analyses were conducted separately for each age group to consider the potential violation of the Cox proportional-hazards assumption [64]. Similar

to previous approach by Ali *et al.*, estimates of immune cell abundance were converted to quartiles from 1-4 [64]. Results were visualized using the meta package.

3.2.4 Single base substitution mutational signatures

From the TCGA-BRCA cohort, 981 samples had mutation data available. Number of mutations for each of 96 possible single base substitution (SBS) types were summed using deconstructSigs package [84]. With the resulting matrix of SBS counts by samples, contributions of each of the 30 SBS signatures from the Catalogue of Somatic Mutations in Cancer (COSMIC) database [85] were estimated for each sample using the R package MutationalPatterns [86]. Instead of excluding SBS signatures previously not extracted from breast cancer cohorts, all 30 signatures were incorporated in our analysis because the young subset of breast cancer patients is largely underrepresented in many cohorts that have been used to extract those signatures [75]. Samples with pairwise cosine similarity between the reconstructed mutation counts matrix and the original input less than 0.5 were excluded ($n = 23$). Samples without gene expression or relevant clinical data were also excluded, resulting in 858 samples, of which 57 were young and 801 were old. Estimated contribution from each signature was compared by Mann–Whitney U (MWU) test between young and old age groups.

3.2.5 Gene set enrichment analysis

To find genes positively associated with CD8+ T cell abundance estimated by TIMER in each cohort, simple linear regression models were fit for expression of each gene separately, with expression values treated as the continuous predictor variable and TIMER estimates as the continuous outcome variable. Models were fit separately for each age group, and the t-statistic value for each of the resulting regression coefficients were calculated as a measure of association to the CD8+ T cell abundance estimates. The list of genes ordered by decreasing t-statistic values for each age group was used as input for pre-ranked gene set enrichment analysis (GSEA) on the

C5:BP (Gene Ontology biological processes) gene sets downloaded from the Molecular Signatures Database (MSigDB; http://software.broadinstitute.org/gsea/msigdb/index.jsp) [87,88]. Gene sets of size < 2 and > 500 were excluded, resulting in total of 7094 gene sets evaluated in the analysis. Results from GSEA were visualized by Cytoscape version 3.7.2 [89] using the Enrichment Map plugin [90]. Enriched gene sets with false discovery rate (FDR) q value < 0.25 were represented as nodes and any overlaps > 0.3 between nodes were represented as edges in the resulting network diagram. Nodes were grouped together and labelled by the clusterMaker2 [91], AutoAnnotate [92], and WordCloud [93] plugins, and resulting annotations were manually corrected.

3.2.6 Data analysis software

All analyses and visualizations were performed in R Project for Statistical Computing version 3.6.1 and RStudio version 1.2.1335 (Boston, MA, United States). Unless otherwise noted, all statistical analyses were conducted using the stats package, and two-sided $p < 0.05$ was considered significant.

3.3 Results

3.3.1 Estimates of immune cell abundance by TIMER

Computational deconvolution of bulk gene expression data by the tumour immune estimation resource (TIMER) method, which used the constrained linear least-squares regression approach, allowed for abundance quantification of six immune cell types in each sample [49]. Differences in estimated immune cell abundance between age groups could not be distinguished by visual inspection of heatmaps for either cohort (**Figure 3.1**). The absolute differences in median between age groups in the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort ranged only from 0.0065 for CD8+ T cells to 0.0715 for dendritic cells.

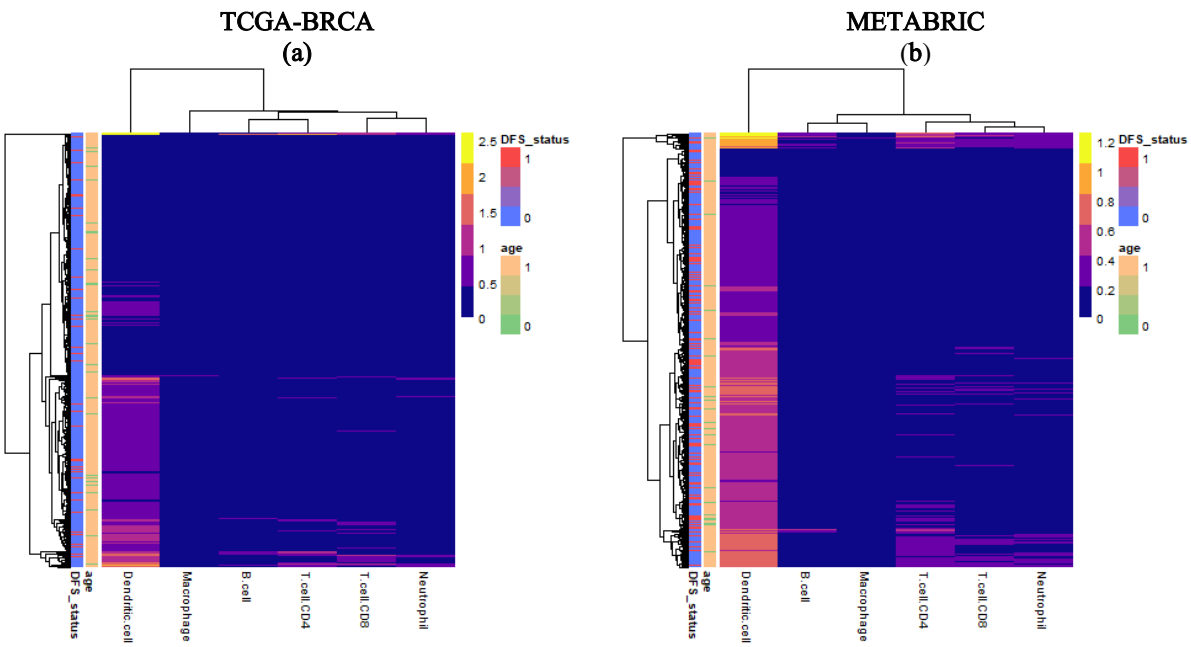


Figure 3.1 Heatmaps of abundance estimates for immune subsets predicted by computation deconvolution using the tumour immune estimation resource (TIMER) algorithm for the two cohorts: (a) The Cancer Genome Atlas (TCGA)-Breast Cancer (BRCA); and (b) Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). Row and column dendrograms show clustering of cases and cell types, respectively, according to Euclidean distance.

3.3.2 Immune cell type abundance associated with disease-free survival

The Kaplan–Meier (KM) survival curve was used to associate abundance of each of the six immune cell types estimated by TIMER with disease-free survival (DFS) time and status separately for each age groups. **Figure 3.2** shows that samples with high estimated $CD8^+$ T cell abundance in The Cancer Genome Atlas (TCGA)-Breast Cancer (BRCA) cohort had significantly better prognosis (log-rank $p = 0.019$), which was also replicated in the METABRIC cohort (log-rank $p = 0.04$). Similar trends were also visible in the old age group; however, the differences in survival were less substantial between samples with high and low levels of $CD8^+$ T cells. For the other immune cell types, results were discordant between the two cohorts; however, macrophages seemed to consistently demonstrate little to no significant associations with DFS in the young age group in both the TCGA-BRCA (log-rank $p = 0.94$) and METABRIC (log-rank $p = 0.28$) cohorts.

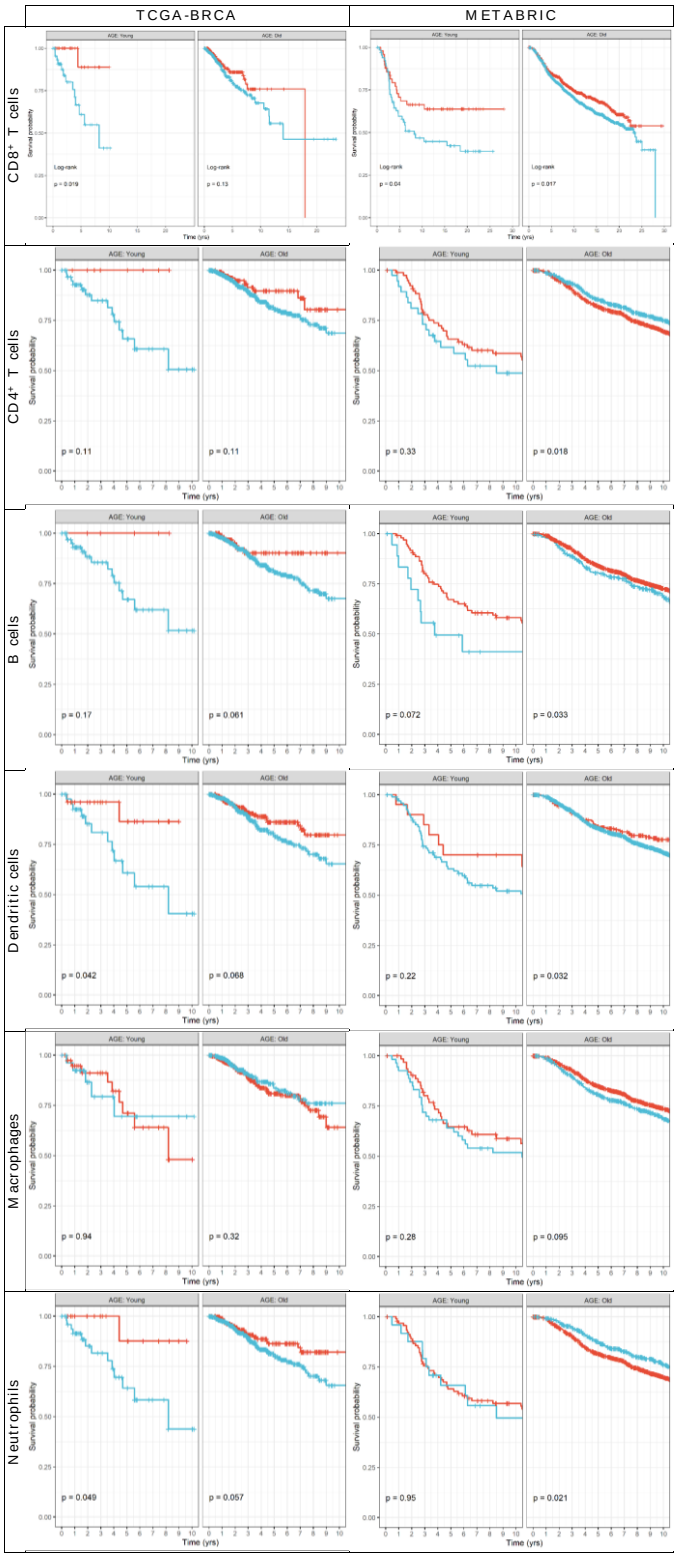


Figure 3.2 Disease-free survival Kaplan–Meier (KM) curve for: (a) TCGA-BRCA; and (b) METABRIC cohorts, grouped by age groups and stratified by high (red) and low (blue) $CD8^+$ T cell levels estimated by TIMER and binarized by the maximally selected ranked statistics algorithm. Depicted *p*-values are from log-rank tests.

To quantify associations with survival, a Cox proportional hazards regression model was fit for each immune cell type to estimate a hazard ratio (HR), visualized as a forest plot for each age group and cohort in **Figure 3.3**. The trend across both cohorts for the young age group was that higher estimates of immune cell types resulted in lower HR, albeit with little significance, with the exception of macrophages. This trend, however, was not observed in samples from the old age groups. In particular, $CD8^+$ T cells were consistently associated with better clinical outcome in the young age group in both TCGA-BRCA (HR 0.69; $p = 0.150$) and METABRIC (HR 0.81; $p = 0.110$) cohorts, as compared to the old age group. High levels of $CD4^+$ T cells were also associated with lower HR for the young age group in both TCGA-BRCA (HR 0.58; $p = 0.150$) and METABRIC (HR 0.83; $p = 0.130$) cohorts but not in the old age group. Abundance of macrophages was consistently shown as having little relationship with survival in all samples.

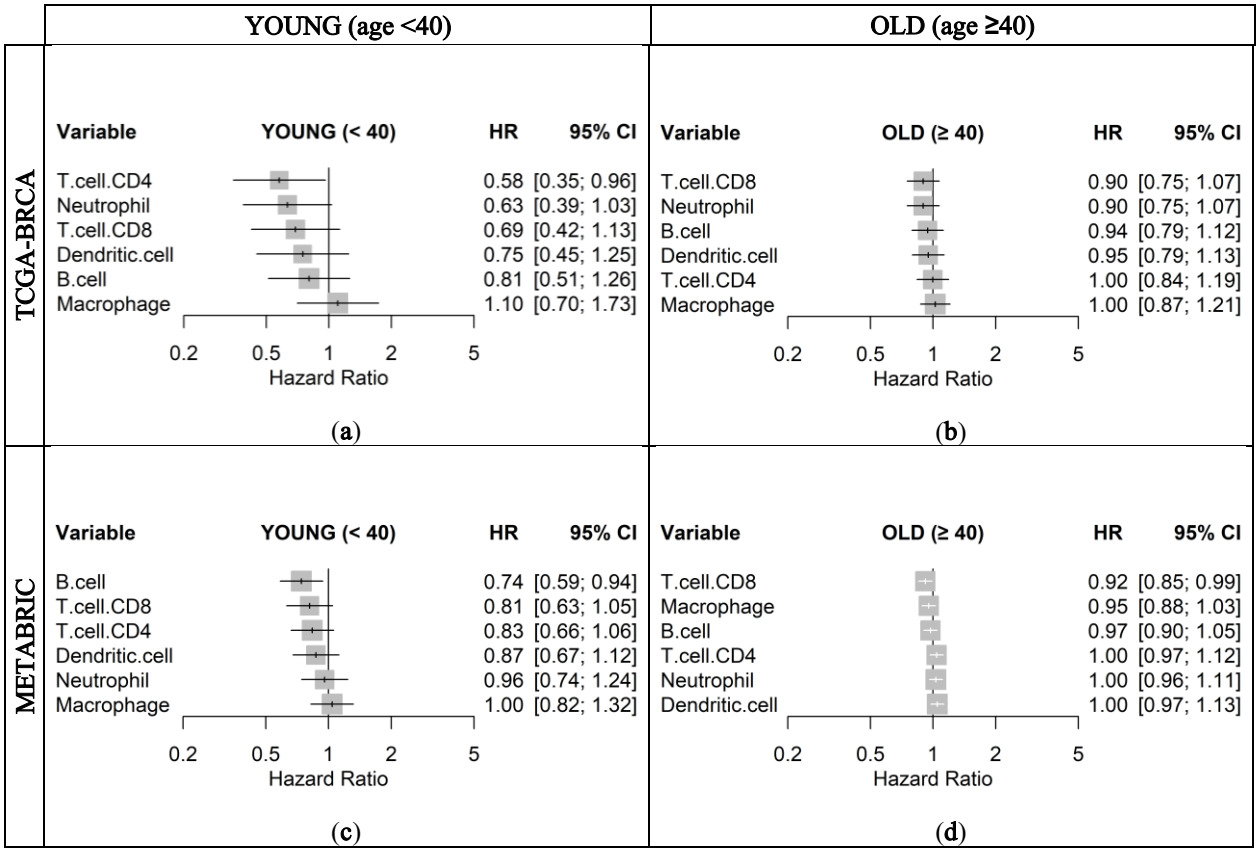


Figure 3.3 Unadjusted hazard ratios of each immune cell type quantified by TIMER, as individually estimated by univariable Cox regression models on disease-free survival for: (a) young, and (b) old patients in the

TCGA-BRCA cohort; and (c) young, and (d) old patients in the METABRIC cohort with 95% confidence intervals (CIs).

3.3.3 Mutational signatures of high TILs differ across age groups

We investigated whether mutational burdens from particular mutational signatures were significantly different between the age groups in the TCGA-BRCA cohort by analyzing the associated whole-exome sequencing (WES) data. **Table 3.1** Differences in mutational burden from 30 SBS signatures from COSMIC between young and old age groups in the TCGA-BRCA cohort. **Table 3.1** shows that contributions from signature 12 were significantly higher in the young age group (Mann–Whitney U, MWU $p = 0.00145$), with a median fold change of 1.48, as well as signature 14 (MWU $p = 0.0425$), with a median fold change of 2.18. None of the signatures were found to be significantly higher in the old age group compared to the young. Mutation data associated with the METABRIC cohort from targeted sequencing of 172 genes were also analyzed; however, results were largely discordant with TCGA-BRCA data. Reconstruction accuracy between estimated mutational burden and actual represented by mean cosine similarity was lower in the METABRIC (0.538) compared to TCGA-BRCA (0.775) cohorts, likely due to differences in the total number of mutations captured from sequencing.

Table 3.1 Differences in mutational burden from 30 SBS signatures from COSMIC between young and old age groups in the TCGA-BRCA cohort.

Signature	*n*	*p*	YOUNG mutational burden	OLD mutational burden	Fold change
1	767	0.217108	10.34976	10.59553	0.976804901
2	549	0.325358	3.888289	3.996138	0.973011829
3	343	0.690164	17.75103	8.407852	2.111244358
4	211	0.901561	3.976115	4.530104	0.87770942
5	24	0.185634	NA	2.53213	NA
6	374	0.335386	4.918544	5.84488	0.841513243
7	441	0.367437	3.826739	3.4218	1.118341061
8	120	0.128638	5.379963	3.999362	1.345205173
9	42	0.417925	1.868344	0.93043	2.00804291
10	426	0.068364	1.563409	2.486163	0.628843997
11	174	0.379684	2.325416	2.157148	1.078005214
12	149	0.001447	3.469254	2.341764	1.481470608
13	567	0.515672	3.193183	3.748332	0.851894393
14	88	0.042543	3.95495	1.810553	2.184388059
15	283	0.256524	2.316104	4.389657	0.527627423
16	164	0.832406	2.515793	3.27996	0.767019314
17	276	0.944968	3.159945	1.774596	1.780655868
18	214	0.192837	3.529686	2.599131	1.358025173
19	130	0.479199	2.194205	3.119958	0.703280388
20	229	0.964294	2.374497	2.731509	0.86929863
21	236	0.646418	1.912902	1.848762	1.034693208
22	304	0.76474	1.815304	1.421824	1.276743206
23	151	0.84864	0.940823	1.727247	0.544694984
24	306	0.285911	4.629516	3.332312	1.389280494
25	25	0.767103	8.925052	2.703811	3.300916033
26	144	0.757082	1.104366	2.569928	0.429726538
27	126	0.705574	0.16551	0.80792	0.204859272
28	157	0.283617	0.443062	1.248088	0.354992647
29	235	0.676288	2.515517	3.203889	0.785144874
30	139	0.102988	1.305804	2.927002	0.446123188

[1] *n* shows number of samples with mutational burden greater than zero from that particular signature.
[2] *p* denotes the p value from MWU test between young and old groups.

The presence and abundance of TILs, especially CD8+ T cells, have been frequently and consistently indicated as an important factor to consider for prognosis and treatment of breast cancer [36,59,62]. Therefore, we referred to the abundance of CD8+ T cells estimated by TIMER as a measure of TILs for subsequent analyses. For each age group independently, we investigated

whether contributions from particular mutational signatures were significantly different between high and low CD8+ T cell groups as estimated by TIMER in the TCGA-BRCA cohort. None of the mutational signatures were associated with TIL levels in the young age group. However, in the old age group, mutational contributions from signature 1 (MWU $p = 0.00504$; fold change 1.23), signature 2 (MWU $p = 0.0395$; fold change 0.77), signature 17 (MWU $p = 0.00319$; fold change 0.99), signature 26 (MWU $p = 0.0326$; fold change 2.02), and signature 30 (MWU $p = 0.0455$; fold change 1.36) were significantly associated with TIL levels estimated by TIMER.

3.3.4 Gene set enrichment for high TILs

Following our results from survival analyses which demonstrated this pattern in the young age group but not in the old, we sought to validate TIMER estimates by examining which gene sets from the Gene Ontology project were enriched in samples with high TIL independently for each age group. **Table 3.2** shows the number of gene sets enriched for genes positively associated with TIL at various significance levels. At false discovery rate (FDR) $q < 0.05$, there was one overlapping positively enriched gene set between TCGA-BRCA and METABRIC cohorts for each age group: "cotranslational protein targeting to membrane" in the young age group, and "cilium movement" in the old age group. Gene set enrichment analysis (GSEA) results from the young TCGA-BRCA cohort showed enrichment of gene sets related to the mitochondria and cellular respiration, among several others (**Table 3.3**). In contrast, in the young age group from the METABRIC cohort, many of the enriched gene sets were related to the adaptive immune response, most notably T cell proliferation, selection, and regulation of cytotoxicity (**Table 3.4**). In the old age group from both cohorts, many of the top positively enriched gene sets were related to cilium assembly, movement, and ciliary transport (**Table 3.5** and **Table 3.6**). For enrichment maps of enriched gene sets for the young patient group in each cohort, refer to **Figure 3.4** and **Figure 3.5**.

Table 3.2 Number of positively enriched gene sets at different significance levels for each age group in each cohort from pre-ranked gene set enrichment analysis (GSEA).

Age Group	Dataset	Number of Features†	Number of positive gene sets	Nominal $p < 0.01$	FDR $q < 0.25$	FDR $q < 0.05$	# of overlaps at FDR $q < 0.05$
Young	TCGA-BRCA	19879	1370	126	135	33	1
Young	METABRIC	24360	3765	246	204	30	
Old	TCGA-BRCA	20201	2801	42	3	0	1
Old	METABRIC	24360	3793	95	8	1	

† Denotes the number of genes in the ranked gene list used as input for the analyses.

Table 3.3 GSEA pre-ranked results for gene list ranked by correlation with TIL levels in the young age group from TCGA-BRCA cohort.

NAME	SIZE[1]	ES[2]	NES[3]	NOM *p*[4]	FDR *q*[5]	FWER *p*[6]
GO_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	97	0.544279	2.847659	0	0	0
GO_ESTABLISHMENT_OF_PROTEIN_LOCALIZATION_TO_ER	110	0.492764	2.655034	0	4.77E-04	0.001
GO_STRESS_RESPONSE_TO_METAL_ION	15	0.780708	2.597907	0	6.39E-04	0.002
GO_ATP_SYNTHESIS_COUPLED_ELECTRON_TRANSPORT	79	0.518184	2.561809	0	9.47E-04	0.004
GO_COFACTOR_CATABOLIC_PROCESS	64	0.522318	2.551812	0	7.57E-04	0.004
GO_RESPIRATORY_ELECTRON_TRANSPORT_CHAIN	97	0.47257	2.499912	0	0.001425	0.009
GO_DETOXIFICATION	112	0.454115	2.460509	0	0.001641	0.012
GO_KILLING_OF_CELLS_OF_OTHER_ORGANISM	54	0.506229	2.428837	0	0.00192	0.016
GO_HYDROGEN_PEROXIDE_CATABOLIC_PROCESS	32	0.60728	2.420802	0	0.001814	0.017
GO_CORNIFICATION	111	0.445246	2.417347	0	0.001829	0.019
GO_PROTEIN_LOCALIZATION_TO_ER	134	0.425255	2.323629	0	0.005919	0.065
GO_OXIDATIVE_PHOSPHORYLATION	101	0.43471	2.318714	0	0.005665	0.068
GO_ANTIBIOTIC_CATABOLIC_PROCESS	50	0.484244	2.300764	0	0.006411	0.08
GO_NUCLEAR_TRANSCRIBED_MRNA_CATABOLIC_PROCESS_NONSENSE_MEDIATED_DECAY	114	0.420275	2.282805	0	0.007047	0.095
GO_GAS_TRANSPORT	19	0.628352	2.273192	0	0.00773	0.112
GO_ELECTRON_TRANSPORT_CHAIN	163	0.405188	2.223335	0	0.013208	0.197
GO_REGULATION_OF_SERINE_TYPE_PEPTIDASE_ACTIVITY	9	0.771998	2.216882	0	0.01328	0.21
GO_CELLULAR_RESPONSE_TO_ZINC_ION	21	0.606071	2.19978	0	0.016612	0.269
GO_KERATINIZATION	178	0.377058	2.185627	0	0.018429	0.303
GO_ACYLGLYCEROL_ACYL_CHAIN_REMODELING	7	0.83337	2.18484	0	0.017607	0.305
GO_MITOCHONDRIAL_ELECTRON_TRANSPORT_UBIQUINOL_TO_CYTOCHROME_C	12	0.660834	2.147358	0	0.024977	0.419
GO_C21_STEROID_HORMONE_METABOLIC_PROCESS	38	0.499533	2.134394	0	0.027856	0.466
GO_AEROBIC_ELECTRON_TRANSPORT_CHAIN	16	0.623273	2.120486	0	0.031017	0.517
GO_PEPTIDE_CROSS_LINKING	53	0.45384	2.110602	0	0.033145	0.563
GO_NEGATIVE_REGULATION_OF_SERINE_TYPE_PEPTIDASE_ACTIVITY	7	0.814052	2.101609	0	0.034936	0.601
GO_MITOCHONDRIAL_ELECTRON_TRANSPORT_NADH_TO_UBIQUINONE	45	0.480546	2.094323	0	0.03611	0.631
GO_RIBOSOMAL_LARGE_SUBUNIT_ASSEMBLY	28	0.537186	2.089743	0	0.036627	0.645
GO_PROGESTERONE_METABOLIC_PROCESS	10	0.710203	2.081615	0	0.038603	0.676
GO_OXYGEN_TRANSPORT	15	0.632096	2.078758	0	0.038627	0.691
GO_XENOBIOTIC_METABOLIC_PROCESS	107	0.380285	2.060239	0	0.045989	0.756
GO_BENZENE_CONTAINING_COMPOUND_METABOLIC_PROCESS	11	0.673088	2.057808	0	0.045593	0.765
GO_CELLULAR_DETOXIFICATION	100	0.384201	2.056978	0	0.0445	0.769
GO_FORMATION_OF_CYTOPLASMIC_TRANSLATION_INITIATION_COMPLEX	14	0.628908	2.050842	0.004255	0.04569	0.792

[1] Size indicates number of genes in the gene set.

[2] ES is the enrichment score.

[3] NES is the ES value normalized by size.

[4] NOM *p* denotes nominal *p* value

[5] FDR *q* denotes false discovery rate *q*-value. Only gene sets with FDR *q*-value < 0.05 are shown.

[6] FWER *p* denotes family-wise error rate *p*-value.

Table 3.4 GSEA pre-ranked results for gene list ranked by correlation with TIL levels in the young age group from METABRIC cohort.

NAME	SIZE[1]	ES[2]	NES[3]	NOM *p*[4]	FDR *q*[5]	FWER *p*[6]
GO_ALPHA_BETA_T_CELL_PROLIFERATION	28	0.66127	2.348763	0	0.001097	0.001
GO_B_CELL_RECEPTOR_SIGNALING_PATHWAY	53	0.57866	2.300078	0	0.002778	0.005
GO_NEGATIVE_T_CELL_SELECTION	12	0.838532	2.254999	0	0.005192	0.014
GO_DENDRITIC_CELL_MIGRATION	24	0.662275	2.224096	0	0.007778	0.028
GO_INACTIVATION_OF_MAPK_ACTIVITY	24	0.659955	2.221114	0	0.006223	0.028
GO_ADAPTIVE_IMMUNE_RESPONSE	350	0.396757	2.159699	0	0.01592	0.084
GO_POSITIVE_T_CELL_SELECTION	32	0.588408	2.137951	0	0.020617	0.122
GO_PROTEIN_ACTIVATION_CASCADE	25	0.618532	2.136991	0	0.018177	0.123
GO_REGULATION_OF_B_CELL_RECEPTOR_SIGNALING_PATHWAY	24	0.640197	2.130815	0	0.017994	0.137
GO_REGULATION_OF_ANTIGEN_RECEPTOR_MEDIATED_SIGNALING_PATHWAY	53	0.522912	2.129687	0	0.01653	0.14
GO_POSITIVE_REGULATION_OF_ALPHA_BETA_T_CELL_PROLIFERATION	16	0.711918	2.125998	0	0.01563	0.145
GO_LYMPHOCYTE_COSTIMULATION	54	0.521161	2.089008	0	0.025422	0.24
GO_REGULATION_OF_CYTOPLASMIC_MRNA_PROCESSING_BODY_ASSEMBLY	8	0.870165	2.087923	0	0.024146	0.246
GO_T_CELL_SELECTION	44	0.542041	2.079368	0	0.024476	0.263
GO_THYMIC_T_CELL_SELECTION	20	0.635614	2.070601	0	0.025954	0.293
GO_LYMPHOCYTE_DIFFERENTIATION	312	0.386417	2.06839	0	0.024889	0.299
GO_POSITIVE_THYMIC_T_CELL_SELECTION	12	0.743443	2.056719	0.002024	0.027521	0.339
GO_REGULATION_OF_T_CELL_MEDIATED_CYTOTOXICITY	28	0.585715	2.043598	0	0.032096	0.406
GO_REGULATION_OF_LEUKOCYTE_APOPTOTIC_PROCESS	76	0.465984	2.031467	0	0.035965	0.463
GO_ANTIGEN_RECEPTOR_MEDIATED_SIGNALING_PATHWAY	216	0.388313	2.01726	0	0.041657	0.53
GO_PROTEIN_DEGLYCOSYLATION	25	0.592343	2.01484	0	0.041258	0.543
GO_POSITIVE_REGULATION_OF_LEUKOCYTE_CELL_CELL_ADHESION	197	0.398304	2.011936	0	0.040939	0.555
GO_POSITIVE_REGULATION_OF_T_CELL_MEDIATED_CYTOTOXICITY	21	0.616534	2.011214	0	0.039835	0.558
GO_IMMUNE_RESPONSE_REGULATING_CELL_SURFACE_RECEPTOR_SIGNALING_PATHWAY	355	0.364463	2.008047	0	0.039702	0.576
GO_LEUKOCYTE_PROLIFERATION	262	0.381872	2.003348	0	0.040511	0.599
GO_POSITIVE_REGULATION_OF_ADAPTIVE_IMMUNE_RESPONSE	88	0.447851	1.994216	0	0.044336	0.644
GO_HEMOGLOBIN_METABOLIC_PROCESS	5	0.939277	1.99229	0	0.043678	0.651
GO_POSITIVE_REGULATION_OF_CELL_ACTIVATION	289	0.374459	1.990124	0	0.04311	0.658
GO_URONIC_ACID_METABOLIC_PROCESS	9	0.788019	1.987935	0	0.042694	0.665
GO_ESTABLISHMENT_OF_LYMPHOCYTE_POLARITY	10	0.747246	1.974582	0	0.048517	0.725

[1] Size indicates number of genes in the gene set.

[2] ES is the enrichment score.

[3] NES is the ES value normalized by size.

[4] NOM *p* denotes nominal *p* value

[5] FDR *q* denotes false discovery rate *q*-value. Only gene sets with FDR *q*-value < 0.05 are shown.

[6] FWER *p* denotes family-wise error rate *p*-value.

Table 3.5 GSEA pre-ranked results for gene list ranked by correlation with TIL levels in the old age group from TCGA-BRCA cohort.

NAME	SIZE[1]	ES[2]	NES[3]	NOM p[4]	FDR q[5]	FWER p[6]
GO_INTRACILIARY_TRANSPORT	49	0.531478	2.095876	0	0.198538	0.211
GO_CILIUM_MOVEMENT	45	0.549711	2.084047	0	0.118814	0.247
GO_OUTFLOW_TRACT_SEPTUM_MORPHOGENESIS	27	0.599063	2.020211	0	0.21298	0.538

[1] Size indicates number of genes in the gene set.

[2] ES is the enrichment score.

[3] NES is the ES value normalized by size.

[4] NOM p denotes nominal p value

[5] FDR q denotes false discovery rate q-value. Only gene sets with FDR q-value < 0.05 are shown.

[6] FWER p denotes family-wise error rate p-value.

Table 3.6 GSEA pre-ranked results for gene list ranked by correlation with TIL levels in the old age group from METABRIC cohort.

NAME	SIZE[1]	ES[2]	NES[3]	NOM p[4]	FDR q[5]	FWER p[6]
GO_REGULATION_OF_CILIUM_ASSEMBLY	44	0.605123	2.31246	0	0.006294	0.006
GO_HEPATOCYTE_GROWTH_FACTOR_RECEPTOR_SIGNALING_PATHWAY	16	0.70647	2.066047	0	0.135248	0.228
GO_EPITHELIAL_CILIUM_MOVEMENT	16	0.675739	2.020608	0.001938	0.179515	0.399
GO_CILIARY_BASAL_BODY_PLASMA_MEMBRANE_DOCKING	80	0.465679	2.0138	0	0.151112	0.436
GO_HYALURONAN_BIOSYNTHETIC_PROCESS	12	0.743555	1.994188	0.001862	0.16514	0.549
GO_NEGATIVE_REGULATION_OF_MRNA_PROCESSING	28	0.57346	1.963609	0	0.22187	0.717
GO_POSITIVE_REGULATION_OF_CILIUM_ASSEMBLY	20	0.63506	1.96157	0.001855	0.195556	0.724
GO_CILIUM_MOVEMENT	40	0.520157	1.944506	0	0.207414	0.791

[1] Size indicates number of genes in the gene set.

[2] ES is the enrichment score.

[3] NES is the ES value normalized by size.

[4] NOM p denotes nominal p value

[5] FDR q denotes false discovery rate q-value. Only gene sets with FDR q-value < 0.05 are shown.

[6] FWER p denotes family-wise error rate p-value.

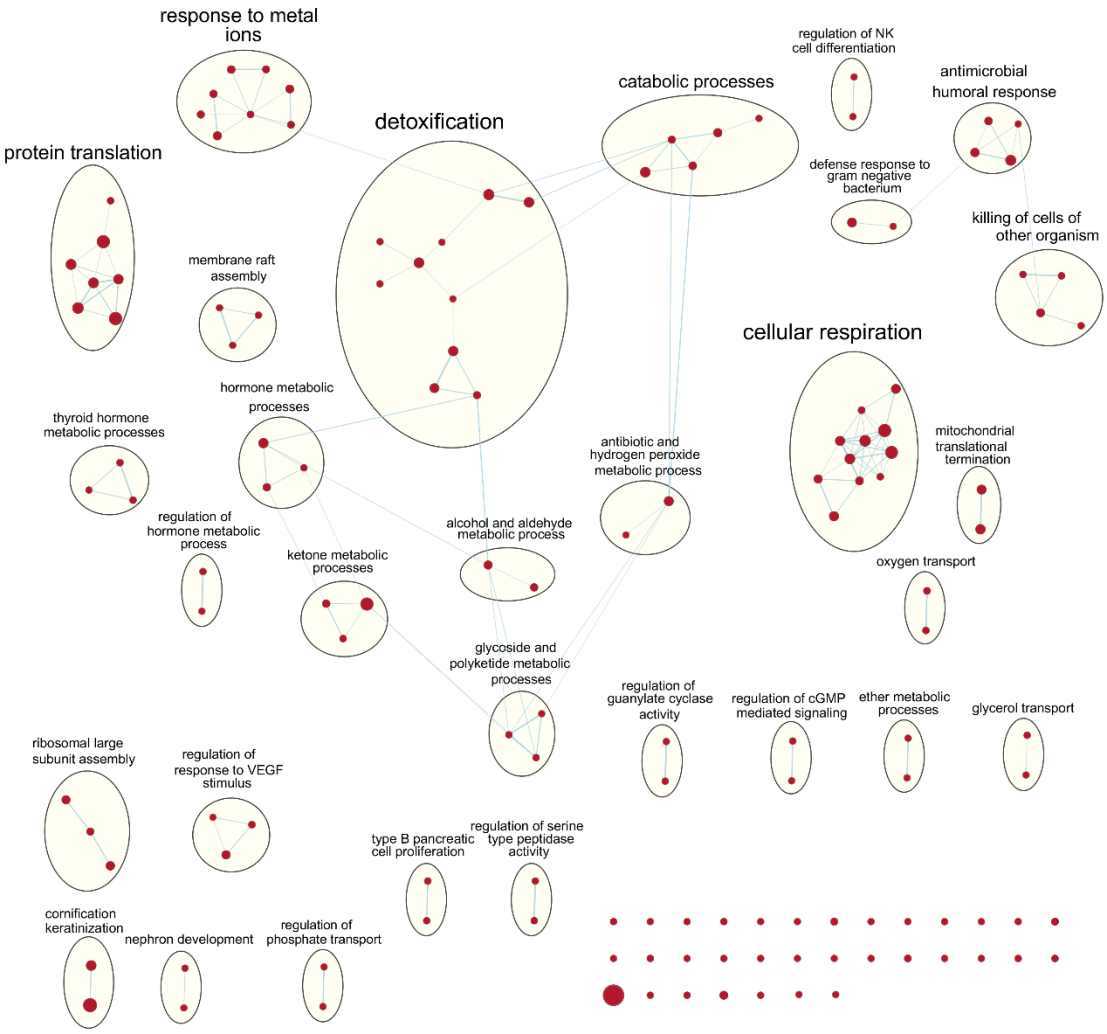


Figure 3.4 Enrichment map of results from pre-ranked GSEA on the ranked gene list from the young TCGA-BRCA cohort. Nodes represent gene sets significant at FDR q-value < 0.25 and edges are drawn between nodes with similarity coefficient > 0.5. NK: natural killer; cGMP: cyclic guanosine monophosphate; VEGF: vascular endothelial growth factor.

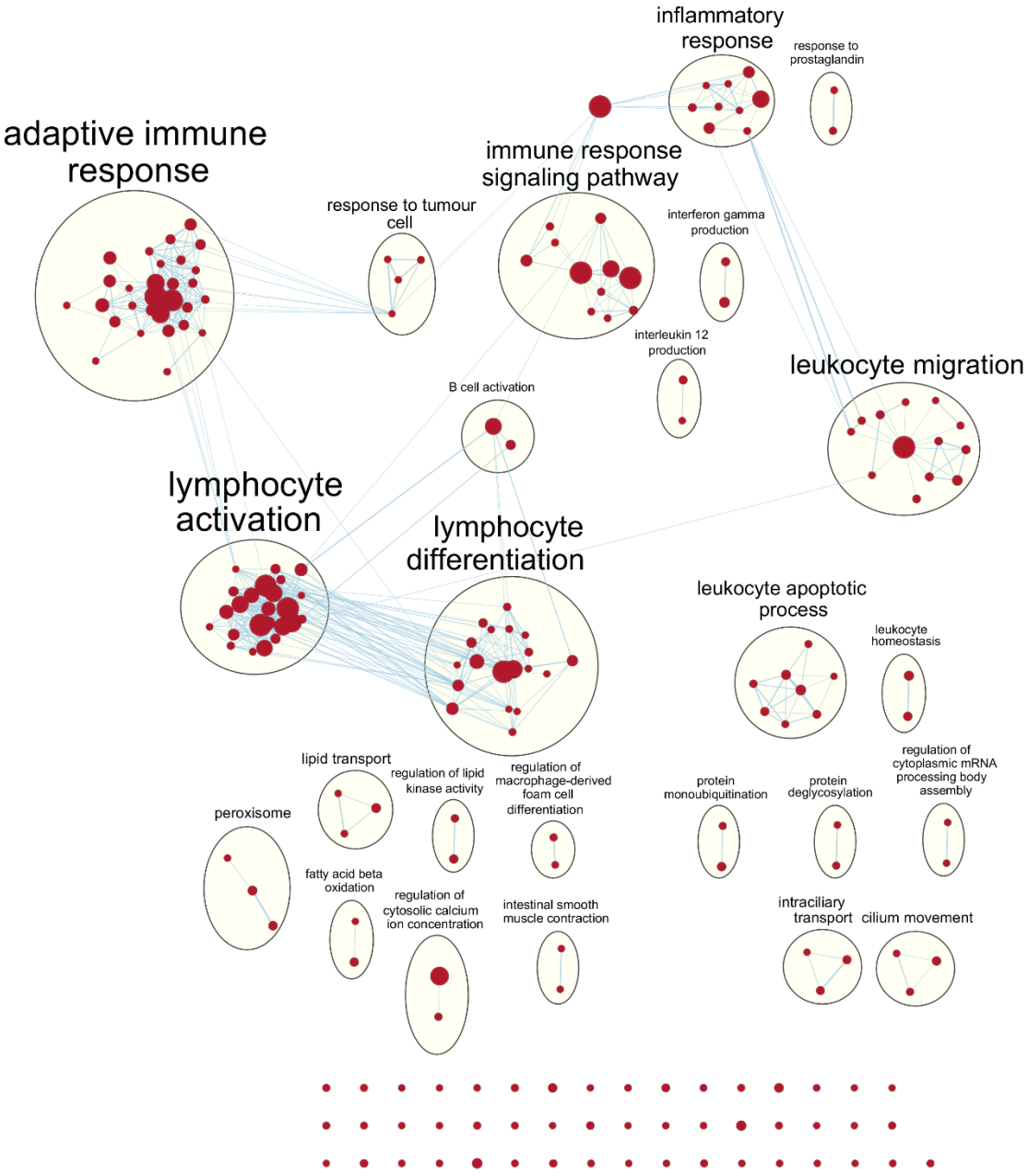


Figure 3.5 Enrichment map of results from pre-ranked GSEA on the ranked gene list from the young METABRIC cohort. Nodes represent gene sets significant at FDR *q*-value < 0.25 and edges are drawn between nodes with similarity coefficient > 0.5. mRNA: messenger ribonucleic acid.

3.4 Discussion

Despite several reports in various parts of the world echoing the conclusion that young age at diagnosis is an indication for poor prognosis of breast cancer [17,57,66,67], researchers have been

unsuccessful in identifying significant biological differences between young and old breast cancer patients. For example, Anders et al. found a number of genes that were differentially expressed between younger and older breast cancer patients, but none remained significant after correcting for subtype and other clinicopathological features [21,56]. Our results are consistent with previous findings in that we were also unable to uncover significant differences in immune cell compositions estimated by gene expression deconvolution between young and old patients.

Various immune subsets should decline with age in a phenomenon known as immunosenescence [68]. The older patient population was also expected to have greater myeloid potential and lesser lymphocyte potential due to involution of the thymus [68]. These observations were not apparent in our analyses of differential immune profiles between age groups, which may be due to tumour-associated immune responses masking the global age-related changes in the immune system. Immunosenescence in women can also be attributed to deprivation of estrogen, which has immune-enhancing activities, during menopause [69]. In the datasets we used, 95.8% and 100% of the young patients (age at diagnosis <40) were pre-menopausal, whereas in the older patients (age at diagnosis ≥40) 20.3% and 16.5% were pre-menopausal in the TCGA-BRCA and METABRIC cohorts, respectively. Hence, menopausal state was a potential confounding variable in our study. Therefore, we fit a multiple linear regression model to further explore broader associations between estimated TIL, age at diagnosis (continuous variable), and menopausal state (categorical variable) (refer to **Table 3.7** Coefficients and statistical results of multiple linear regression between the response variable: estimated TIL level; and predictor variables: age at diagnosis, and menopausal state.). We found that TIL levels estimated by TIMER were significantly lower with older age at diagnosis (p-value: 0.0258 and 0.000282 for TCGA-BRCA and METABRIC cohorts, respectively). This finding is consistent with a recent meta-analysis

exploring tumour-infiltrating lymphocytes and prognosis of early-stage triple-negative breast cancers [70]. However, menopausal state was not associated with TIL levels in both cohorts.

Table 3.7 Coefficients and statistical results of multiple linear regression between the response variable: estimated TIL level; and predictor variables: age at diagnosis, and menopausal state.

TIL ~ age_at_diagnosis + menopausal_state[1]

	Coefficients:	Estimate	Std. Error	*t* value	Pr(>|t|)
TCGA-BRCA	(Intercept)	0.2579713	0.0354456	7.278	1.09e-12
	age_at_diagnosis	-0.0016823	0.0007642	-2.201	0.0281
	menopausal_state = "Peri"	-0.0599772	0.0364901	-1.644	0.1008
	menopausal_state = "Post"	0.0210642	0.0224287	0.939	0.3480
	F-statistic: 2.891 on 2 and 590 DF, *p*-value: 0.03151				
	Coefficients:	Estimate	Std. Error	*t* value	Pr(>|t|)
METABRIC	(Intercept)	0.1474797	0.0094865	15.546	< 2e-16
	age_at_diagnosis	-0.0007406	0.0002036	-3.638	0.000282
	menopausal_state = "Post"	0.0064385	0.0064213	1.003	0.316143
	F-statistic: 9.864 on 2 and 1900 DF, *p*-value: 5.474e-05				

[1] Menopausal state is a categorical variable with three levels ("Pre", "Peri", and "Post") provided for TCGA-BRCA and two levels ("Pre", and "Post") provided for METABRIC data. For both cohorts, the first level ("Pre") was used as reference.

We found that specific immune subsets, in particular the CD8+ T cells, were significantly associated with disease-free survival in the young breast cancer patients but not in their older counterparts. This is in line with past studies that showed that higher infiltration of cytotoxic immune cells is indicative of better prognosis and greater odds of response to treatment [59,62]. Notably, Ali et al. showed that CD4+ and CD8+ T cells were more closely associated with favorable outcomes in ER-negative tumours than in ER-positive tumours [64]. Considering the fact that receptor status-negative tumours are much more common in the younger subset of breast cancer patients, our results are consistent with previous findings [17,58].

GSEA results showed that gene sets related to adaptive immune response and cytotoxic T cell processes were most obviously enriched in the young age group from the METABRIC cohort. On the other hand, gene sets related to the mitochondria and oxidative phosphorylation were positively

enriched in the young age group from the TCGA-BRCA cohort. It may not be obvious, but mitochondria are important for T cell activation, proliferation, and differentiation because they are involved in processes such as immune synapse functions, production of reactive oxygen species, and various metabolic processes [71]. In the old age group from both cohorts, gene sets related to cilium assembly, movement, and ciliary transport were positively enriched for the high TIL phenotype, although with lower significance. Ciliary processes have recently been proposed to be involved in the formation of the immunological synapse between T cells and antigen-presenting cells, which is necessary for T cell activation and downstream functions [72,73]. This suggests that the TIMER algorithm may focus on different subset of genes for different datasets to provide estimates of immune cell composition.

Mutational signatures analysis showed that signatures 12 and 14 were significantly enriched in young compared to old breast cancer patients; however, the two signatures have not been previously associated with breast cancer [74,75]. Signature 12 was originally identified in liver cancer and is characterized by T > C substitutions with a transcriptional-strand-bias, which is indicative of being associated with transcription-coupled nucleotide excision repair [74]. Signature 14 was first identified in uterine cancer and is characterized by C > A and C > T substitutions, with a recent study suggesting that it is associated with microsatellite instability due to defects in mismatch repair [76]. However, our current knowledge on mutational signatures specific to breast cancer is likely reflective of the majority of breast cancer cases that are diagnosed in those over the age 40 (~93%). The landscape of mutational signatures specific to early onset breast cancer (~7%) has not been characterized previously. Further research is necessary to validate if the signatures previously not associated with breast cancer or novel signatures are responsible for mutations in the young subset of breast cancer patients. Within the young age group, none of the

30 single base substitution (SBS) signatures were significantly associated with TIL levels estimated by TIMER. However, within the old age group, mutational burdens from signatures 1, 26, and 30 were significantly higher in samples with high TIL levels, whereas signatures 2 and 17 were significantly associated with low TIL levels, although to varying degrees. Moreover, mutational signatures that were significantly associated with TIL levels in the old age group were all previously reported in breast cancer patients, which once again suggests that TIL levels estimated by TIMER may also be indicative of tumour purity within samples.

TIMER has been frequently used on RNA-seq data from the TCGA database, since it was originally developed on TCGA data to work with bulk RNA-seq data [49,77,78]. However, it must be noted that the reference gene sets used in TIMER were curated from microarray gene expression data. To our knowledge, we are the first to employ the method for analyzing microarray data from the METABRIC cohort. Microarray gene expression data suffers from artifacts, such as limited profiling of genes with very low expression levels or saturation at very high expression levels, which poses potential challenges for data analysis compared to bulk RNA-seq data [49,79]. However, authors of TIMER method have stated that both RNA-seq and microarray data may be used as input to estimate the abundance of immune cell types in the tumour microenvironment [49,79]. They showed that TIMER is capable of producing highly concordant estimates between RNA-seq and microarray gene expression data generated from the same tumour samples [49]. This demonstrated that the algorithm used for deconvolution, constrained least-square regression (**Equation 3.1**), is robust regardless of input data type [49,79].

TIMER only provides estimates for six broad immune cell types: B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells [49]. This can be advantageous because limiting number of cell types interrogated to those that are linearly separable prevents instability

of estimates due to statistical co-linearity between cell types with very similar gene expression [49,78]. However, this is at the expense of resolution of the output. Even though they may originate from the same progenitor cell, different immune cell types and states can play vastly different roles within the tumour tissue; for example, M0 and M2 macrophages have traditionally been associated with pro-tumour responses whereas M1 macrophages have been associated with anti-tumour responses [36,58]. Our results from TIMER that found macrophages consistently insignificant to clinical outcomes in both cohorts may stem from the fact that TIMER cannot distinguish between the different activation states of macrophages. Despite its limitations, we were able to use TIMER as a tool to estimate the composition of immune cells within breast tumour samples and find specific immune subsets that showed significant trends with regards to disease-free survival in the young age group of breast cancer patients but not in the old.

3.5 Conclusions

In summary, we determined that a particular immune cell type, the cytotoxic CD8+ T cell, was significantly associated with disease-free survival in young breast cancer patients under the age 40, but not in their older counterparts. Furthermore, our analyses showed that single base substitution mutational signatures 12 and 14 were significantly enriched in the young patient group compared to the old, that were not previously associated with breast cancer. We also highlighted some potential limitations of our data and methods, especially with TIMER and its lack of resolution. Nonetheless, our work suggests that the underlying biological differences may stem from more abstract relationships involving multiple levels of tumour physiology and not age alone.

Chapter 4. Neural network algorithm for expression deconvolution

4.1 Motivations

There are numerous expression deconvolution tools that have been developed over the recent decades [42–54]; however currently, there is not a single "gold-standard" tool for expression deconvolution. Traditional expression deconvolution algorithms such as TIMER use a previously curated reference dataset in conjunction with a set of marker genes to construct a signature matrix [49]. However, this approach relies on previous knowledge and is limited in its ability to leverage new information such as data generated from scRNA-seq experiments. Also, generation of signature matrix does not take into account several confounding factors such as tissue type, stage of disease, population characteristics, and/or differences between male and females. Another main challenge with research in expression deconvolution is that these studies require an interdisciplinary approach with considerations for two ever-changing fields in science: machine learning, and molecular biology. Just within the last few decades, genomics research has already gone through several technological revolutions starting with the microarrays, then to next-generation sequencing, and most recently with scRNA-seq [94]. On the other hand, the machine learning field has advanced just as fast with developments in artificial intelligence – most notably deep learning and artificial neural network-based algorithms [95].

Despite the huge success of neural network algorithms in computer vision and recently in bioinformatics [95], only a few studies have explored how such algorithms should be applied to the problem of expression deconvolution. A recently developed tool based on artificial neural network called digitalDLsorter was trained on simulated pseudo-bulk data generated from scRNA-seq datasets [96]. The digitalDLsorter showed predictions that correlated with estimates from other methods as well as providing prognostic values in breast and colorectal cancer patients [96].

However, the authors did not compare the performance against existing tools for expression deconvolution on any benchmark datasets, which makes the robustness of the tool questionable. Another recent neural network-based tool, Scaden [97], was trained rigorously on both simulated and real GEPs. While demonstrating seemingly robust performances in recovering immune cell type fractions, even against existing methods, as well as being available as a Python package, authors do not provide the weight kernels used in the study. This requires the user to re-train the Scaden model from scratch using scRNA-seq from their cell types of interest, which makes implementations challenging. Therefore, the main motivation for this chapter was to develop a novel expression deconvolution tool based on artificial neural network architecture that is more robust, accessible, and accurate than the existing tools. The focus will remain on characterizing the immune landscape given bulk tissue profiles because immune cell types were shown to greatly influence patient outcome and response to various treatments [37].

4.2 Materials and methods

4.2.1 Single cell RNA-sequencing data of immune cells

Single cell RNA sequencing (scRNA-seq) datasets from immune cells isolated by FACS was downloaded from the 10X Genomics website [98]. Specifically, datasets for the following 10 immune cell types were curated: $CD14^+$ monocytes, $CD19^+$ B cells, $CD34^+$ cells, $CD4^+$ helper T cells, $CD4^+/CD25^+$ regulatory T cells, $CD4^+/CD45RA^+/CD25^-$ naïve T cells, $CD4^+/CD45RO^+$ memory T cells, $CD56^+$ NK cells, $CD8^+$ cytotoxic T cells, $CD8^+/CD45RA^+$ naïve cytotoxic T cells; refer to **Table 4.1**.

Table 4.1 List of scRNA-seq data curated from the 10x Genomics database.

Cell type	Marker	FACS purity	# of cells detected	# of cells after filtering	Reads per cell
B cells	$CD19^+$	~100%	~10,000	9879	~25,000
Monocytes	$CD14^+$	98%	~2,600	2553	~100,000
Hematopoietic stem and progenitor cells	$CD34^+$	45%	~9,000	9080	~24,700
Helper T cells	$CD4^+$	99%	~11,000	11168	~21,000
Regulatory T cells	$CD4^+/CD25^+$	95%	~10,000	10201	~27,000
Naïve T cells	$CD4^+/CD45RA^+/CD25^-$	98%	~10,000	10454	~19,000
Memory T cells	$CD4^+/CD45RO^+$	98%	~10,000	10191	~24,000
Natural Killer cells	$CD56^+$	92%	~8,000	8238	~29,000
Cytotoxic T cells	$CD8^+$	98%	~10,000	10187	~28,600
Naïve cytotoxic T cells	$CD8^+/CD45RA^+$	99%	~12,000	11891	~20,000

From the original dataset preprocessed by Cell Ranger 1.1.0, only those cells expressing between 200 and 3,000 different features with positive counts, and proportion of total counts from mitochondrial genes under 5% were retained. Joint plot of scRNA- datasets before (top) and after (bottom) processing is shown in **Figure 4.1** visualized by scanpy [99] package in Python.

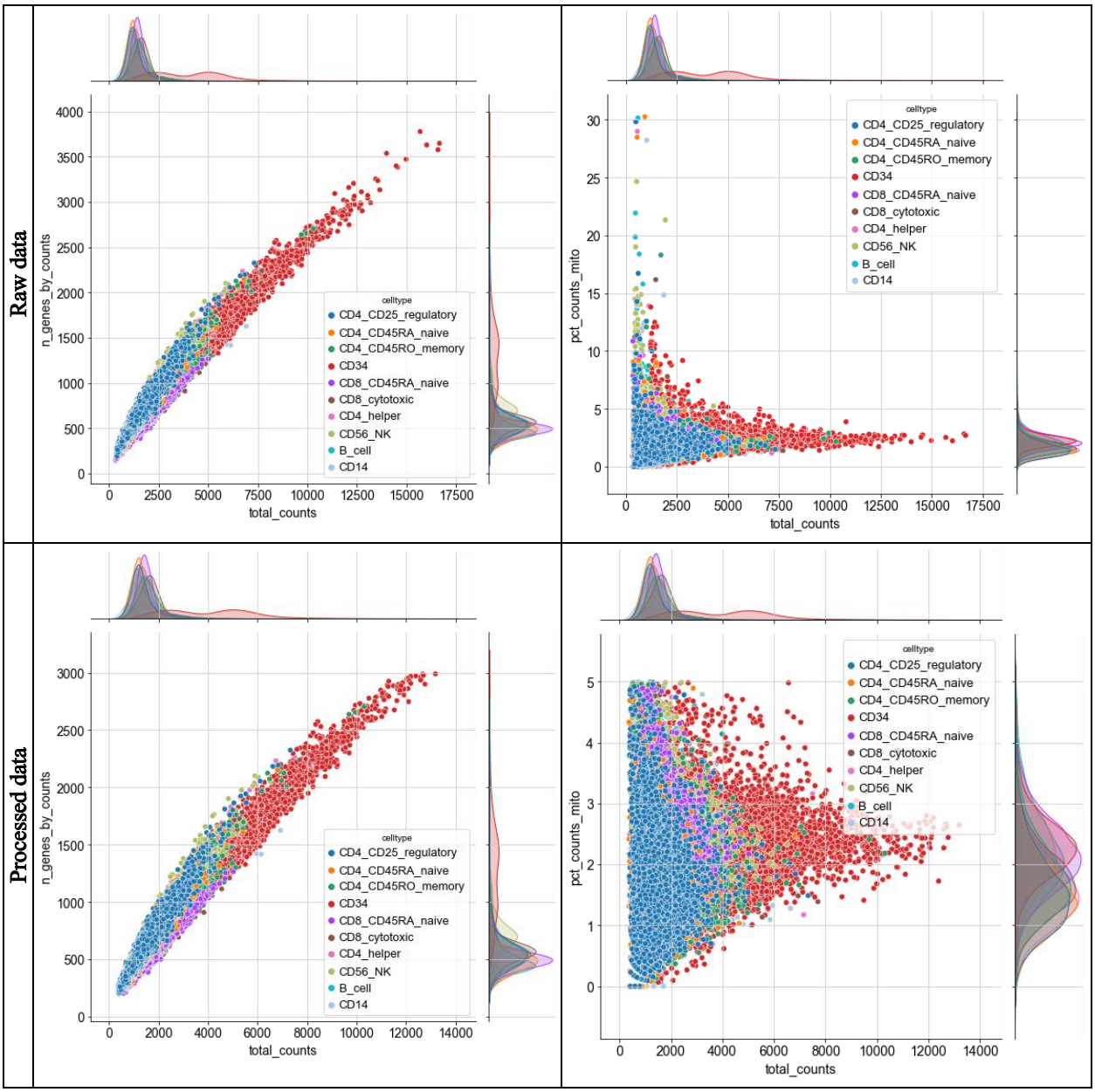


Figure 4.1 Joint plot of dataset quality control metrics before (top panels) and after (bottom panels) processing the collected scRNA-seq datasets. "n_genes_by_counts" equals the number of genes/features with positive (non-zero) count within the single cell profile. "total_counts" refers to the sum of read counts across all features for a given single cell profile. "pct_counts_mito" refers to the percentage of total read counts that are mitochondrial genes/transcripts.

To further explore the scRNA-seq dataset, and whether computational algorithms would be able to distinguish single cell GEPs from one cell type to another, uniform manifold approximation and projection (UMAP) algorithm was used to cluster the single cell profiles into 10 groups as shown below in **Figure 4.2**.

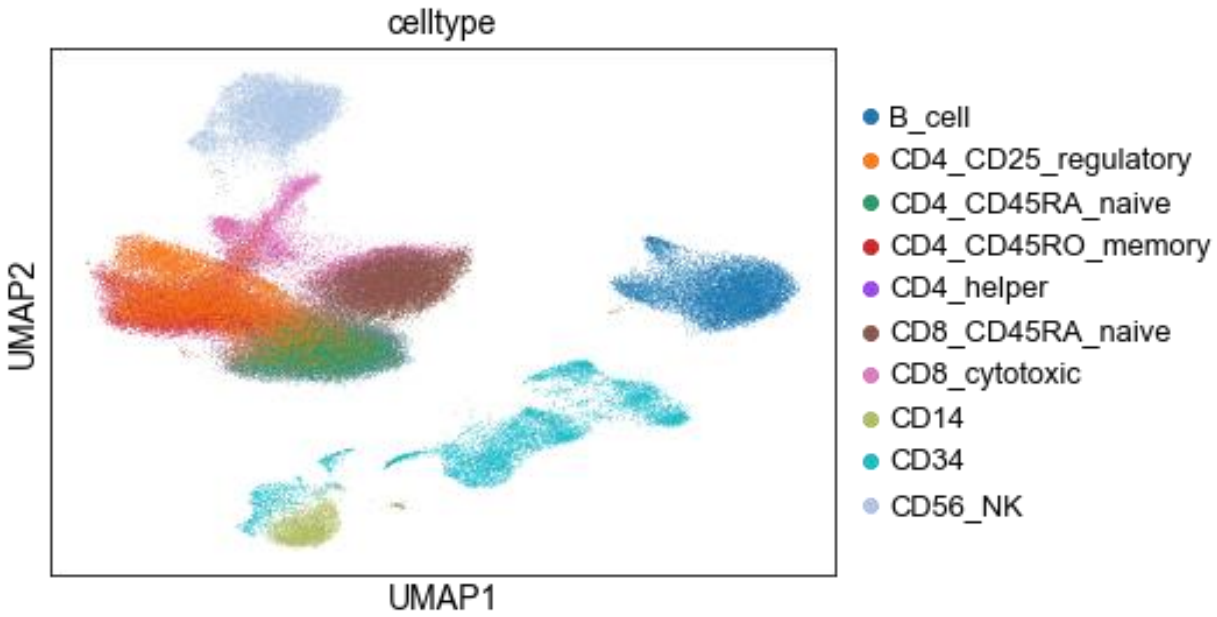


Figure 4.2 UMAP dimensionality reduction of single cell gene expression profiles into 10 clusters, coloured by the known cell type annotations.

The resulting combined scRNA-seq dataset contained GEPs of a total of 93,842 cells for 20,024 genes/transcripts. Features (genes/transcripts) were further filtered by using SCTransform method [100] provided in Seurat v3 R package [101] to normalize read counts, compensating for percentage of mitochondrial genes, and selecting the top 3000 most variable features, which are standard preprocessing steps in scRNA-seq pipelines to discard poor-quality cells and housekeeping genes [100].

4.2.2 Bulk tissue RNA-sequencing data of immune cells with FACS-quantified cell proportions

Bulk GEP data from mRNA-seq (single-end Illumina HiSeq 2000) of PBMCs isolated from 468 whole blood samples obtained from healthy participants aged between 50-74 was downloaded from the ImmPort database (accession SDY67) [102]. Briefly, the sequenced reads were aligned to hg19 using TopHat and Bowtie, followed by gene counting with HTSeq, resulting in raw read count matrix of 23,113 genes and 468 samples [102]. From here onwards, this dataset will be referred to as "SDY67" dataset. For each sample of whole blood, the cell type proportions

quantified by flow cytometry panels was downloaded from supplementary materials of Monaco *et al.* [51]. From this dataset, proportions for 14 immune cell types were available: B naïve, B exhausted (Ex), B non-switched memory (NSM), B switched memory (SM), plasmablasts, $CD4^+$ T cells, $CD8^+$ T cells, basophils, myeloid dendritic cells (mDCs), plasmocytoid dendritic cells (pDCs), natural killer (NK) cells, and the classical (C), intermediate (I) and non-classical (NC) subsets of monocytes [51]. The known proportion of the cells from cell types of interest within each GEP sample will be referred to as "true fractions" or simply, "labels". Missing flow cytometry data for specific cell types were filled with zero and given values for each sample were normalized to sum to one by dividing each value by the sample total.

Additional bulk GEP data from mRNA-seq (paired-end Illumina HiSeq 2000) of PBMCs in whole blood samples drawn from 13 healthy young adults (age 20-35) was obtained to be used as an external validation dataset from the Gene Expression Omnibus (GEO) database (accession GSE107011) [51]. Briefly, the sequenced reads were preprocessed using FastQC, aligned to GRCh38.p10 transcriptome using kallisto, and aggregated into transcripts per million (TPM) for each gene with tximport, resulting in final dataset of 55221 features/transcripts and 13 samples. This dataset will be referred to as "ABIS" dataset because it was used to develop a deconvolution algorithm known as ABIS by the authors [51]. One sample without associated FACS sorting data was removed, so only 12 samples with both bulk GEP and FACS quantified cell proportions data were used for subsequent analyses. Feature annotations were provided as Ensembl IDs, which were converted to HGNC gene symbols using the *EnsDb.Hsapiens.v79* package in R. Cell type proportions for 29 immune-related cell types were available: $CD4^+$ T lymphocyte subsets (naïve, terminal effector (TE), T follicular helper (Tfh), Tregs, T helper (Th) 1, Th1/Th17, Th17, Th2); $CD8^+$ T lymphocyte subsets (naïve, central memory (CM), effector memory (EM), TE); mucosal-

associated invariant T (MAIT); γδ T lymphocyte subsets (Vδ2 and non-Vδ2); B lymphocyte subsets (naïve, SM, NSM, Ex, and plasmablasts); low-density (LD) granulocytes (LD basophils, LD neutrophils); NK cells; DCs (pDCs and mDCs); and monocyte subsets(C, I, and NC) [51].

4.2.3 Immune cell type pooling

Since the number of cell types were not the same across datasets (10 for scRNA-seq; 14 for SDY67; and 29 for ABIS), the cell types were pooled into the following cell types whenever appropriate: B cells, $CD4^+$ T cells, $CD8^+$ T cells, myeloid cells, NK cells, and other cells as illustrated in **Figure 4.3**. For the scRNA-seq dataset, number of cells after cell type pooling for $CD4^+$ T cells and $CD8^+$ T cells was 42,014 and 22,078, respectively. Number of cells for the remaining cell types that were not pooled are shown in **Table 4.1**. Cell type lineages were drawn with reference to cell type mappings available in previous studies [43,51].

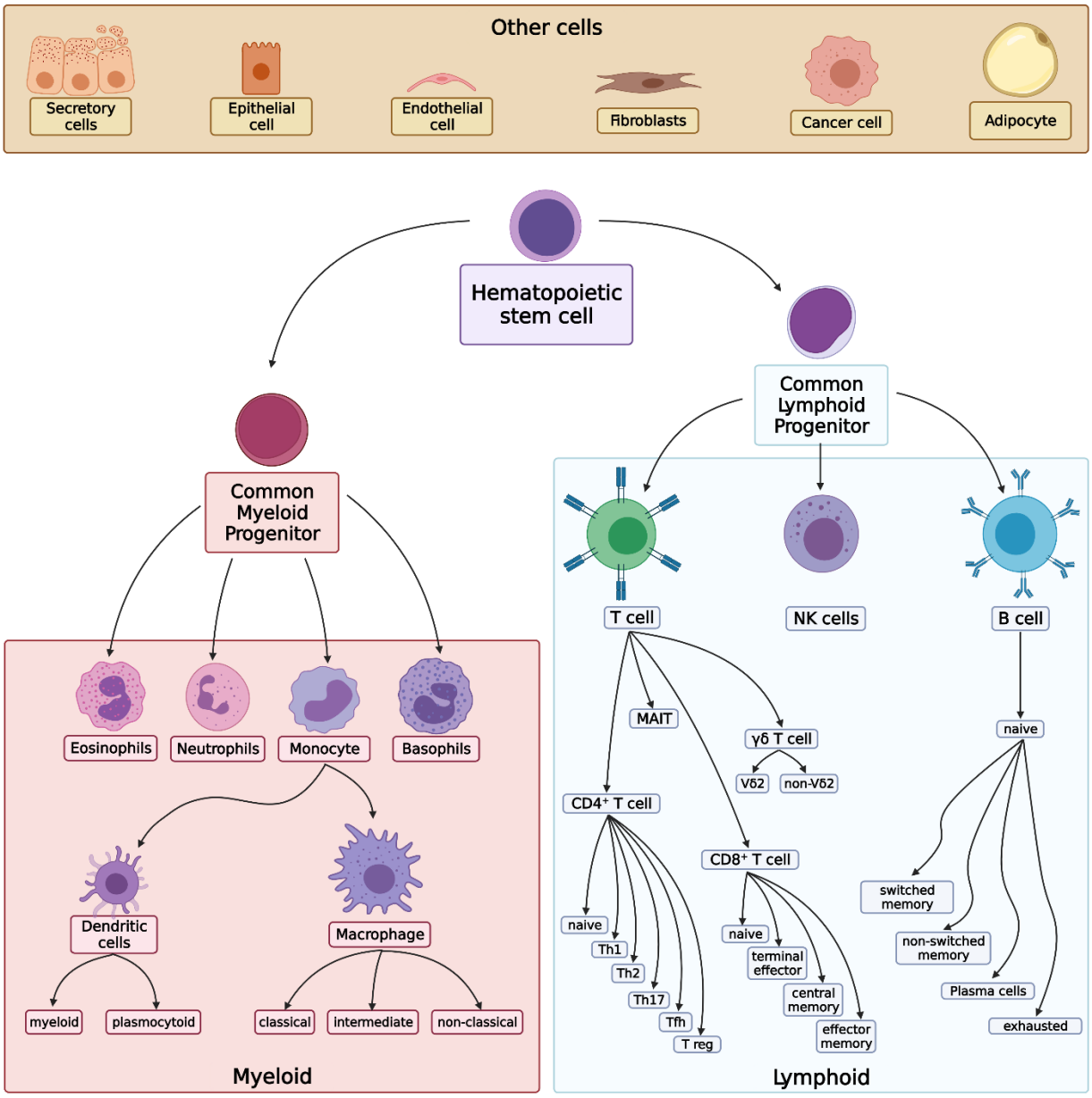


Figure 4.3 Lineage of immune cells used to pool cell types into broader cell type categories.

4.2.4 Bulk tissue RNA-sequencing data of breast tumours

As previously done in Chapter 3, bulk GEPs of breast tumours from TCGA-BRCA [4] and METABRIC [3] cohorts were used.

The TCGA-BRCA dataset from "HTSeq – FPKM" workflow was downloaded using the GDCquery function in TCGAbiolinks R package [103]. The read counts matrix provided as fragments per kilobase million (FPKM) values was converted to transcripts per million (TPM)

values that is conventionally used for expression deconvolution [42,49]. Associated clinical data including patient survival, age at diagnosis, menopause status, and other relevant variables were downloaded from the cBioPortal for Cancer Genomics [82]. As previously done in Chapter 3, only data from female patients with associated age at diagnosis and survival data were used, resulting in final sample size of 997. Of these patients, 70 were identified as young/early-onset (age <40), and 927 were identified as old/late-onset (age ≥40) patients.

The bulk breast tumour microarray GEPs and clinical data from patients in the METABRIC cohort were downloaded from cBioPortal as previously done in Chapter 3 [82]. Samples without associated age or survival data were removed, resulting in bulk microarray dataset of 24,360 genes from 1903 breast tumour samples, of which 116 were young and 1787 were old.

4.2.5 Feature selection

In order for the expression deconvolution model to be applied to a variety of datasets, it was clear from the beginning that there was a need to provide the model with consistent number of input features. Initially, to prevent potential loss of information, all features were provided to the model as input if they were represented consistently across the datasets being used. Interestingly, the model performance did not suffer when only the top 3,000 most variable genes selected by SCTransform method [100] on the scRNA-seq dataset were provided. Furthermore, performance of the trained models significantly improved when the input features were further selected for only those based on the 4,815 immunologically related genes from the ImmPort through the InnateDB [104]. Therefore, the final model accepted input features that satisfied all of the following criteria:

1. Included in the ImmPort list of immunologically related features (4,815 genes).
2. Included in all of the datasets used in the study.
3. Top 3,000 most variable features determined by SCTransform on scRNA-seq dataset.

Consequently, total of 3,000 features were used as input into the final model.

4.2.6 Pseudo-bulk simulation with known cell type proportions

For the simulation of pseudo-bulk GEPs using scRNA-seq data with G number of features (genes/transcripts), the cell type fractions for each pseudo-bulk sample were simulated first by sampling vectors of numbers from the Dirichlet distribution for K cell types that add up to 1, which represents the proportion of given cell types in each sample. Therefore, N by K matrix was simulated for N pseudo-bulk samples and K cell types of interest. Then cells from each of the K cell types were randomly subsampled according to the simulated proportion of a total number of cells (C) in each pseudo-bulk tissue. Following previous studies that simulated pseudo-bulk GEPs from scRNA-seq data [96,97], the total number of cells in each pseudo-bulk sample (C) was set to be 500 by default. Then the GEPs of C subsampled cells were summed together across each feature, resulting in the final pseudo-bulk GEP of 1 by G. Graphical summary of this simulation process is shown below in **Figure 4.4**.

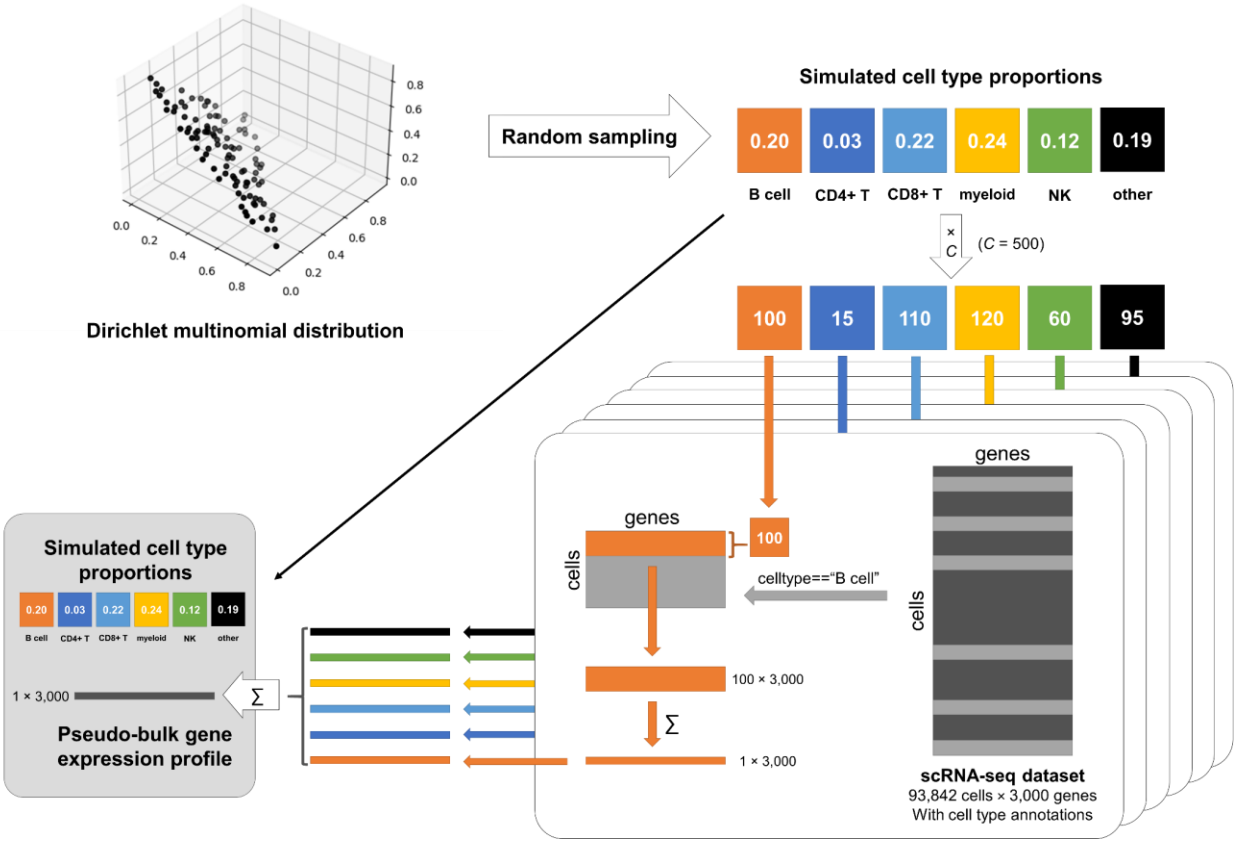


Figure 4.4 Graphical representation of the pseudo-bulk gene expression profile simulation process using scRNA-seq dataset.

Different values were tested for parameters in pseudo-bulk simulation. For the number of pseudo-bulk samples N, values 100, 1,000, and 10,000 were tested, while for the number of cells per pseudo-bulk sample C, values 500, 1,000, and 2,500 were tested. Simulation of pseudo-bulk GEPs using each combination of parameters above was repeated 20 times to assess the performance of deconvolution model trained on each simulated pseudo-bulk GEP dataset. As shown below in **Figure 4.5**, value of C higher than 500 has no impact on model performance whereas the number of training samples Nhas a much greater influence on performance. Finalized pseudo-bulk dataset was simulated with N= 10,000 for training set and N= 1,000 for test set; G = 3,000; C= 500; and number of cell types K= 6 (B cell, $CD4^+$ T cell, $CD8^+$ T cell, myeloid cell, NK cells, and other cells)

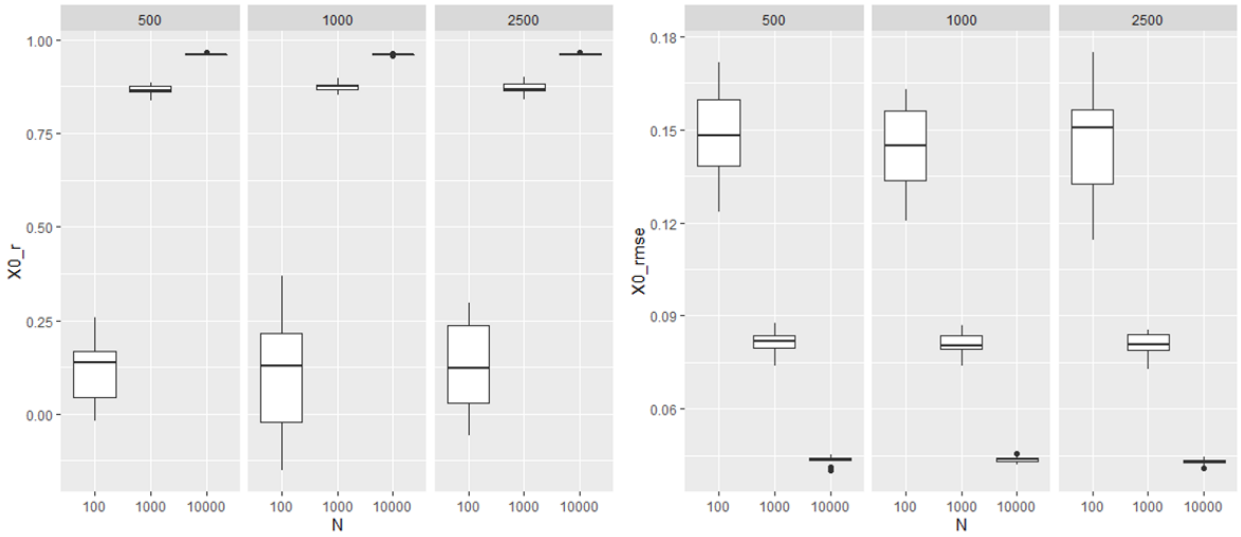


Figure 4.5 Performance metrics on pseudo-bulk training datasets simulated with different combination of parameters replicated 20 times for each combination. Left panel shows performance in Pearson correlation (X0_r) and right panel shows performance in RMSE (X0_rmse). For each panel, columns show different number of cells (*C*) per sample, x-axis shows different number of samples (*N*) in each replicate.

All feature values in input data were scaled sample-wise to a range between 0 and 1 by using the MinMaxScaler function in scikit-learn package in Python, which is described as:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad 4.1$$

where x_{scaled} is the scaled feature values for a single GEP sample (x), and x_{min} and x_{max} are the minimum and maximum feature values within the GEP sample.

4.2.7 Neural network model architecture and optimization

A variety of model architectures were tested for expression deconvolution on TensorFlow (version 2.3.0) using the Python programming language (version 3.8.8). Recent works in expression deconvolution involving deep learning frameworks used a series of densely connected layers followed by a softmax activation function shown below [96,97].

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad 4.2$$

Here, given the i^{th} element of vector $\vec{z}$ with length K, the softmax function (σ) provides the normalized probabilities of the exponentials of input vector with base $e \approx 2.718282$.

Preliminary model architectures that were tested consisted of autoencoders or densely connected layers with rectified linear unit (ReLU), batch normalization (BN), dropout, softmax, or linear activation functions with or without regularization (L1 or L2). Equations for each layer are listed below with x representing a batch of input vectors.

$$\mathrm{ReLU}(x) = max(0, x) \tag{4.3}$$

$$\mathrm{BN}(x) = \gamma\left(\frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}}\right) + \beta \tag{4.4}$$

In **Equation 4.4** above: $\mathrm{E}[x]$ represents the expected value of x (in other words, the mean); $\mathrm{Var}[x]$ represents the variance of x; γ and β are learned scaling and centering parameters; and a small value ϵ is added to the denominator for numerical stability (to avoid division by zero).

$$\mathrm{Dropout}(x) = \delta x \tag{4.5}$$

In **Equation 4.5** above: δ takes the value of 0 or 1 depending on a pre-defined drop-out frequency.

$$\lambda \sum \|\omega\| \tag{4.6}$$

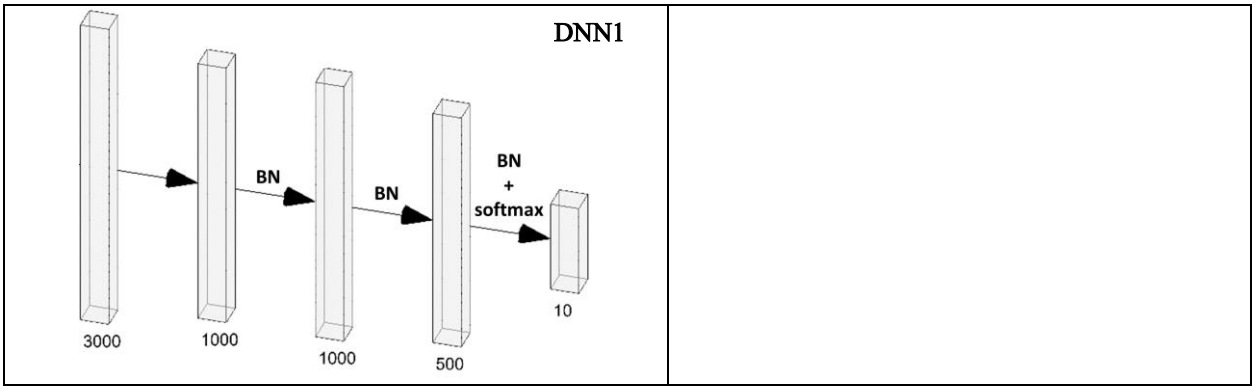
$$\lambda \sum \omega^2 \tag{4.7}$$

In **Equations 4.6** and **4.7** above that represent L1 and L2 regularization terms, respectively: λ is the pre-defined regularization factor; and ω is the weight kernel of the layer.

Different loss functions for optimizing these models were tested, including root mean square error (RMSE), Pearson correlation loss (1 – *R*), as well as custom loss function from OnionNet [105]. Custom loss function is defined in **Equation 4.8**.

$$\text{Loss} = \alpha(1 - R) + (1 - \alpha)\text{RMSE} \quad 4.8$$

In machine learning, it is desirable to be able to provide stable estimates that do not underfit or overfit to the training dataset. Models that underfit fail to accurately represent the relationship between input and output values; hence, often show poor performance during optimization. For this reason, different combinations of model architectures and hyperparameters were tested as shown in **Figure 4.6**. First architecture tested was the pre-trained autoencoder (AE) model, in which a single hidden layer was used to recover the input values after reducing the dimensions of input features from 3,000 to 150. Then the pre-trained hidden layer was used with a 10-neuron output layer with L1 regularization for the expression deconvolution. Subsequent architectures – named "DNN" models for dense neural network and numbered in chronological order – largely consisted of a series of densely connected layers, linear activation functions, and intermittent batch normalization or drop out layers. Two types of output layers were tested, including softmax activation function and a simple linear activation function with either L1 or L2 regularization. Additional details on the models' hyperparameters and performance are shown in **Table 4.3**.



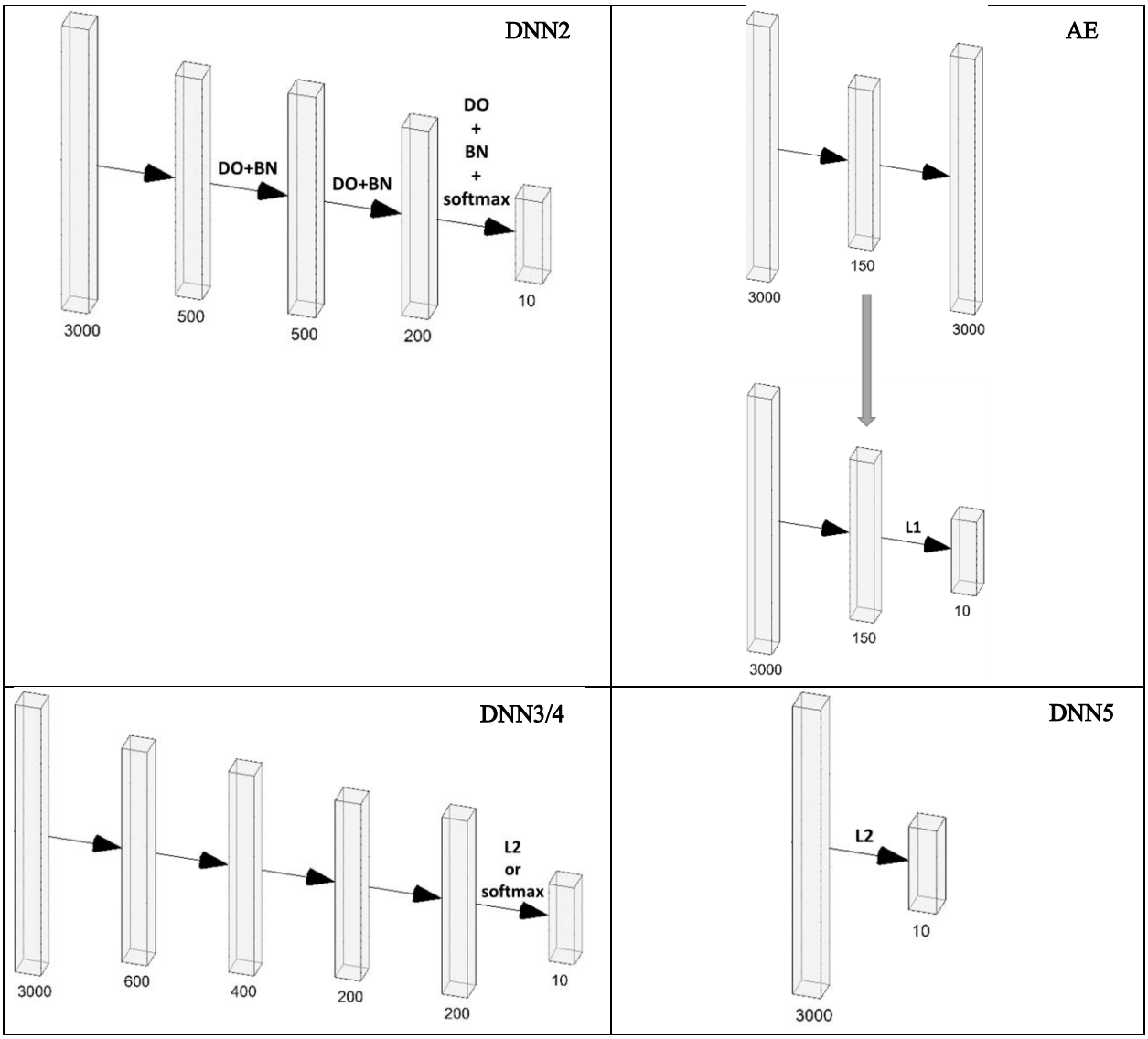


Figure 4.6 Neural network model architectures tested. "DNN" is short form for densely connected neural network; and AE stands for autoencoder model. Layers (rectangles) are scaled and do not represent true dimensions labelled under each layer as numbers. Unless otherwise stated, activation functions between each layer is a linear function. "BN" denotes bath normalization operation; "DO" denotes drop out operation; "softmax" denotes softmax activation function; "L1" and "L2" denote linear activation with L1 or L2 regularization, respectively.

4.2.8 Deep quantile regression model for expression deconvolution

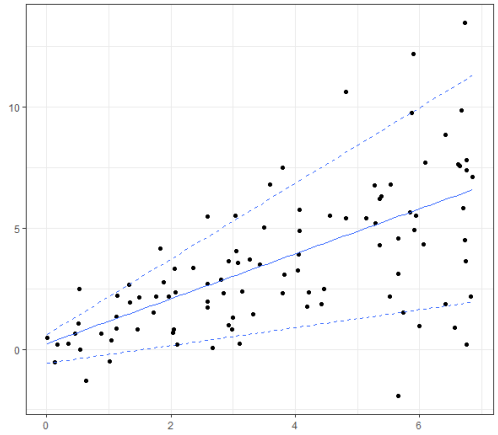
For many real-world problems, we are often interested in the uncertainty or the confidence interval of the estimates provided by the prediction model as much as the accuracy of the estimates themselves. However, one limitation of traditional regression models as well as the conventional neural network models is that they do not provide any measure of uncertainty on the predictions.

One way to obtain the range of uncertainty from these models is to fit the predicted values on a number of conditional quantiles [106]. Whereas conventional regression models that predict the mean value of the distribution at each data point are fitted on the residuals (difference against the expected value of the mean) as is, quantile regression can provide the interval between which a percentage of all data points exist within. **Figure 4.7** Graphical representations of quantile regression and tilted loss: A) An example of quantile regression – solid line shows model fit against the median (50^{th} quantile) whereas the dashed lines show models fit against the 10^{th} and 90^{th} quantiles; therefore, the interval between the dashed lines contain 80% of data points. B) Tilted loss function plotted with different quantiles (0.10, 0.25, 0.50, 0.75, 0.90).**Figure 4.7A** shows an example of quantile regression with the uncertainty in prediction represented by the dashed lines representing predictions by quantile regression models fitted against the 10^{th} and the 90^{th} quantiles. This is made possible by using the tilted loss function visualized in **Figure 4.7B**. Given an error metric or residuals (ξ) between true and predicted fractions (ex. Pearson correlation loss ($1 - R$), root mean square error (RMSE), or custom loss), and q from the set of quantiles {0.10, 0.25, 0.50, 0.75, 0.90}, the tilted loss is represented below as a mathematical expression in **Equation 4.9**.

$$\text{TiltedLoss} = max((q \times \xi), (q - 1) \times \xi)) \quad 4.9$$

This approach is referred to as deep quantile regression (DQR) when combined with neural network model. Multiple instances of the same model were trained on different tilted loss function fit on different quantiles (10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th}), which allowed the visualization of uncertainty in the cell type abundance estimates provided by the model.

A



B

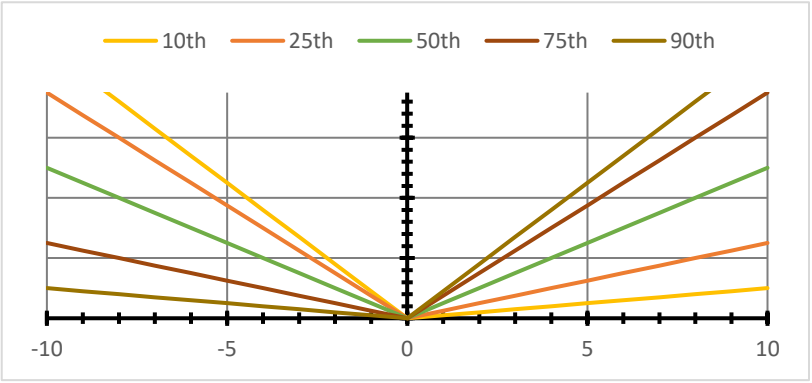


Figure 4.7 Graphical representations of quantile regression and tilted loss: A) An example of quantile regression – solid line shows model fit against the median (50th quantile) whereas the dashed lines show models fit against the 10th and 90th quantiles; therefore, the interval between the dashed lines contain 80% of data points. B) Tilted loss function plotted with different quantiles (0.10, 0.25, 0.50, 0.75, 0.90).

4.2.9 Neural network immune contexture estimator for expression deconvolution

During model optimization, it was also realized that instead of training a single model with one input and *K* outputs for *K* number of cell types of interest, models were able to provide better estimates when their number of outputs were reduced to a single cell type during optimization. Hence, the final expression deconvolution model was based on DQR and consisted of 6 sub-models, one for each cell type of interest (B cell, CD4$^+$ T cell, CD8$^+$ T cell, myeloid cells, NK cells, and other cells). Each sub-model consisted of an input layer for 3,000 input features connected to a densely connected layer by linear activation functions with 300 neurons, followed by a 1 neuron output layer with linear activation functions with non-negativity constraint on the bias kernel and L2 regularization on the weight kernel with λ=0.0001. Models were optimized by multiple

iterations, or epochs, through a training dataset, minimizing the custom loss function (**Equation 4.8**) nested inside the tilted loss function (**Equation 4.9**) across 5 quantiles (0.10, 0.25, 0.50, 0.75, 0.90). The final model will be referred to as "NNICE" – Neural Network Immune Contexture Estimator; refer to **Figure 4.8** and **Algorithm** 1 below.

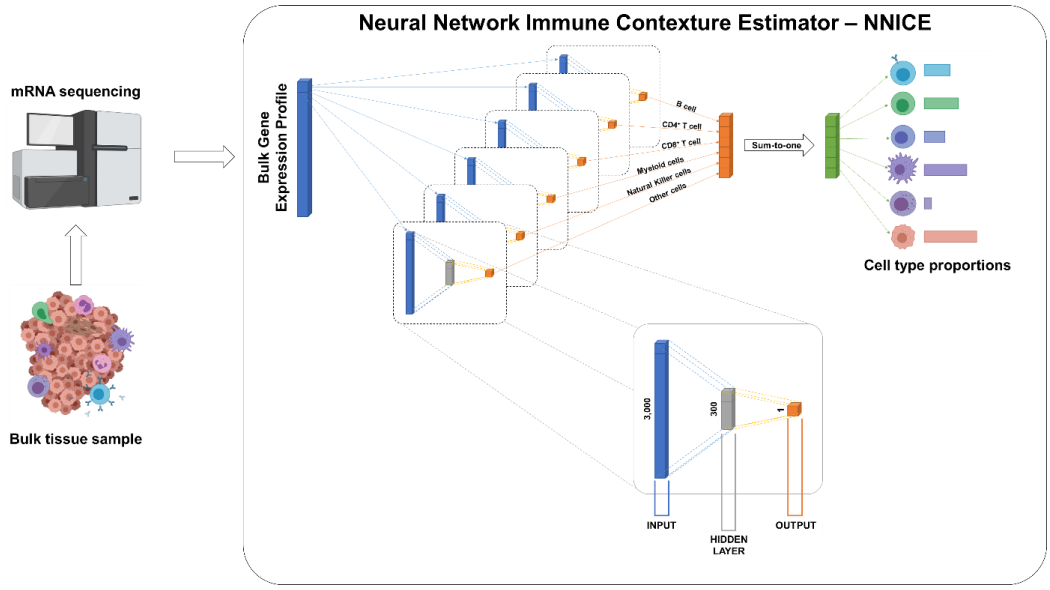


Figure 4.8 Graphical description of NNICE expression deconvolution model.

Algorithm 1 Pseudocode for NNICE expression deconvolution

w^l: weight kernel for l^{th} layer
b^l: bias vector for l^{th} layer
Z^l: output of l^{th} layer

Input:
Gene expression dataset: $\mathbf{M} = [m_{n,g}]_{N \times G}$
N: # of samples in the input dataset
G: # of features/genes/transcripts in the input dataset
Output:
Cell type fractions: $\mathbf{F} = [f_{n,k}]_{N \times K}$
K: # of cell types of interest

Algorithm:
for $n = 1$ to N
for $k = 1$ to K
for $q = \{0.10, 0.25, 0.50, 0.75, 0.90\}$
$Z^0 \leftarrow LinearActivation(w^0, m_n, b^0)$
$f_{n,k,q} \leftarrow LinearActivation(w^1, Z^0, b^1)$
end for
$f_{n,k} \leftarrow mean(f_{n,k,q=0.10}, f_{n,k,q=0.25}, f_{n,k,q=0.50}, f_{n,k,q=0.75}, f_{n,k,q=0.90})$
end for
$f_{n,k} \leftarrow sum\text{-}to\text{-}one(f_{n,k})$
end for

4.2.10 Benchmarking against existing expression deconvolution tools

When developing a new computational tool, it is important to compare its performance against existing tools. Six existing expression deconvolution tools available in the immunedeconv [43] R package were used for performance comparisons against NNICE on benchmark datasets: pseudo-bulk, SDY67 and ABIS. Refer to **Table 4.2** below for the descriptions of expression deconvolution tools:

Table 4.2 List of existing deconvolution tools used to compare performance against.

Tool name	Acronym	Algorithm	Cell types	Year	Reference
CIBERSORT	CBS	ν-SVR	22	2015	[44]
TIMER	TMR	Constrained linear least square regression	6	2016	[49]
MCP-counter	MCP	Geometric mean of expression marker genes	9	2016	[107]
EPIC	EPC	Constrained least square regression	9	2017	[108]
quanTIseq	QTS	Constrained least square regression	11	2019	[54]

4.2.11 Model evaluation

We first evaluated our model performance using the simulated pseudo-bulk dataset, which consisted of 10,000 training samples and separately simulated 1,000 test/validation samples. Next, we evaluated our model performance on two real datasets: the larger SDY67 dataset (468 samples), and the smaller ABIS dataset (12 samples). Since there is only limited number of samples from the real datasets, 10-fold cross-validation (CV) approach was used to train and validate the model on the SDY67 dataset. This means that a small portion of the dataset (10% in this study) was set aside as internal validation data to validate performance of the model that is trained on a larger subset of the dataset referred to as the training data (90% of samples). Therefore, 10 instances of the model were created – each with trained with 90% of the dataset and tested on a unique remaining subset (10%) of the dataset as illustrated in **Figure 4.9**. Next, another instance of the model was trained on the entirety (100%) of the SDY67 dataset and then tested using the ABIS dataset as an external validation set.

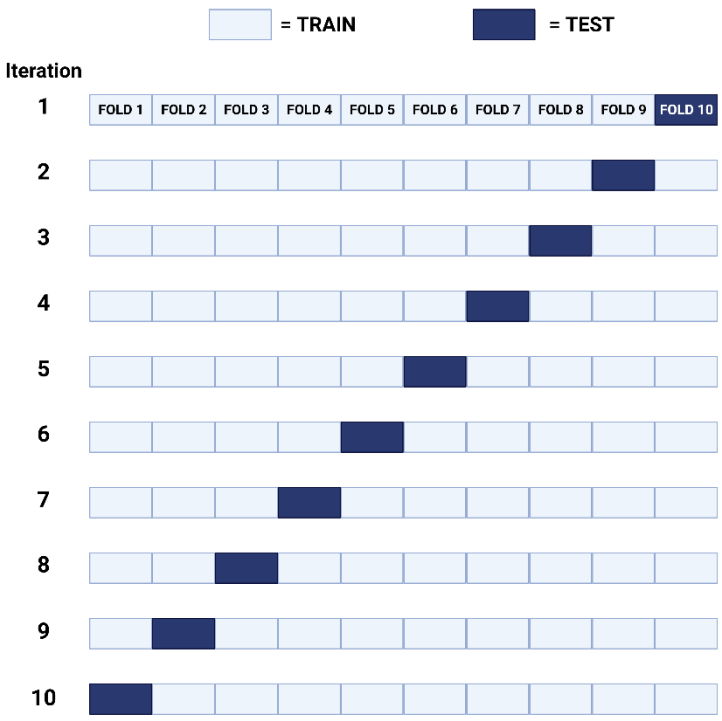


Figure 4.9 Illustration of the 10-fold cross validation technique for model evaluation.

For performance metrics, Pearson correlation (R) and root mean square error (RMSE) were used. Formula for Pearson correlation (R) where y represents true fractions (labels), $\hat{y}$ represents estimated fractions, and the bar above variables ($\bar{y}, \bar{\hat{y}}$) represents the mean of the given vector, is as follows:

$$R = \frac{\sum(\hat{y} - \bar{\hat{y}})(y - \bar{y})}{\sqrt{\sum(\hat{y} - \bar{\hat{y}})^2 \sum(y - \bar{y})^2}} \quad 4.10$$

For RMSE, the formula is as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad 4.11$$

where n represents total number of samples in the minibatch.

4.2.12 Model prediction and statistical analyses

To obtain immune contexture estimates for TCGA-BRCA and METABRIC cohorts, only those features that fulfilled the criteria outlined in **Section 4.2.5** were included. Then the input expression values were scaled to a range between 0 and 1 using the MinMaxScaler function (**Equation 4.1**). Scaled input GEPs were provided to NNICE model trained on the 468 bulk GEPs from the SDY67 dataset for prediction.

Subsequent statistical analyses followed methods as described previously in Chapter 3 for the analysis of differences in immune subsets between early- and late-onset breast cancer patients as well as the survival analyses including KM survival analyses and Cox proportional hazards regression. Unless noted otherwise, statistics were computed using packages such as stats, survival, survminer, and maxstat in the R programming language.

4.3 Results

4.3.1 Model optimization

Results from preliminary and exploratory evaluation of different model architectures is shown in **Table 4.3**. Most of the architectures initially tested, including AE and DNN2 models, showed underwhelming performances in both training and validation datasets ($R < 0.50$). Overfitting occurs when models demonstrate very high performance on the training data but do poorly on all other datasets. This is usually because the model has learned too much of the patterns underlying the training dataset, as demonstrated by results from the preliminary model DNN1 in **Table 4.3** with $R = 0.9754$ on training data but only $R = 0.3005$ on the validation set. Therefore, developing successful models in machine learning involves achieve a fine balance between underfitting and overfitting to the training data. There are several possible interventions that can be used to avoid the two extremes. Applying regularization, and dropout layers during the training phase can reduce overfitting. For example, when dropout layers were included in DNN2, the validation R increased

to 0.4674 compared to $R = 0.3005$ in DNN1 without dropout layers; however, this approach often times also involve a trade-off in performance or convergence on the training dataset. Ultimately, changes to the number of parameters in the hidden layers was made to achieve robust performance across training and validation datasets; refer to results for DNN5 in **Table 4.3**.

Table 4.3 Description of models trained during preliminary exploratory analyses, hyperparameters, and performance (R) on training and validation datasets.

MODEL	Hidden layer dimensions	Output layer	Loss function	Learning rate	R (train)	R (valid)
DNN5	None	L2 (λ=0.2)	Custom (α=0.8)	1e-3	0.7587	0.7587
DNN4	600→400→200→200	softmax	Custom (α=0.8)	1e-3	0.6282	0.6044
DNN3	600→400→200→200	L2 (λ=0.2)	Custom (α=0.8)	1e-3	0.6333	0.6134
DNN2	500→DO(0.2)→BN→500→ DO(0.2)→BN→200→DO(0.2)→BN	softmax	Pearson correlation	1e-4	0.4292	0.4674
DNN1	1000→BN→1000→BN→500→BN	softmax	Pearson correlation	1e-4	0.9754	0.3005
AE	150	L1 (λ=0.1)	Pearson correlation	1e-3	0.3201	0.3203

By using grid search algorithm to find the optimal hyperparameters for layers that provide the best performance, it was discovered that non-negativity constraint on weight kernels had negative impact on performance, while the same constraint placed on the bias kernels seemed to have slightly positive impact. In addition, 0.0001 was selected as the optimal λ term for L2 regularization. Various batch sizes were also tested; however, it became clear that when training the models with real data (ex. SDY67), they generally performed better when provided with the entire training set instead of in mini-batches in sizes of 1, 32, or 100.

4.3.2 Evaluation of model trained on pseudo-bulk GEPs

First, the performance of the NNICE model after training on 10,000 pseudo-bulk GEPs was evaluated. As shown below in **Figure 4.10**, a convergence was reached for simulated pseudo-bulk within around 10 epochs – meaning that for all cell types, little-to-no improvement in performance

could be expected after 10 epochs. For most cell types except $CD4^+$ T cell, NNICE model was able to provide estimates with well over 90% correlation on truth values. Estimates for $CD4^+$ T cell was the lowest at R = 0.776 for training set and R = 0.797 for validation set. Exact values of performance metrics for each cell type is provided in **Table 4.4**. Performance metrics were similar on training and validation data, which indicated that the model showed no performance deficits due to overfitting on the training set, at least on the simulated pseudo-bulk dataset. **Figure 4.11** shows performance on previously unseen simulated pseudo-bulk data (n = 1,000) by quantiles, by cell types, and after averaging estimates across the five quantiles.

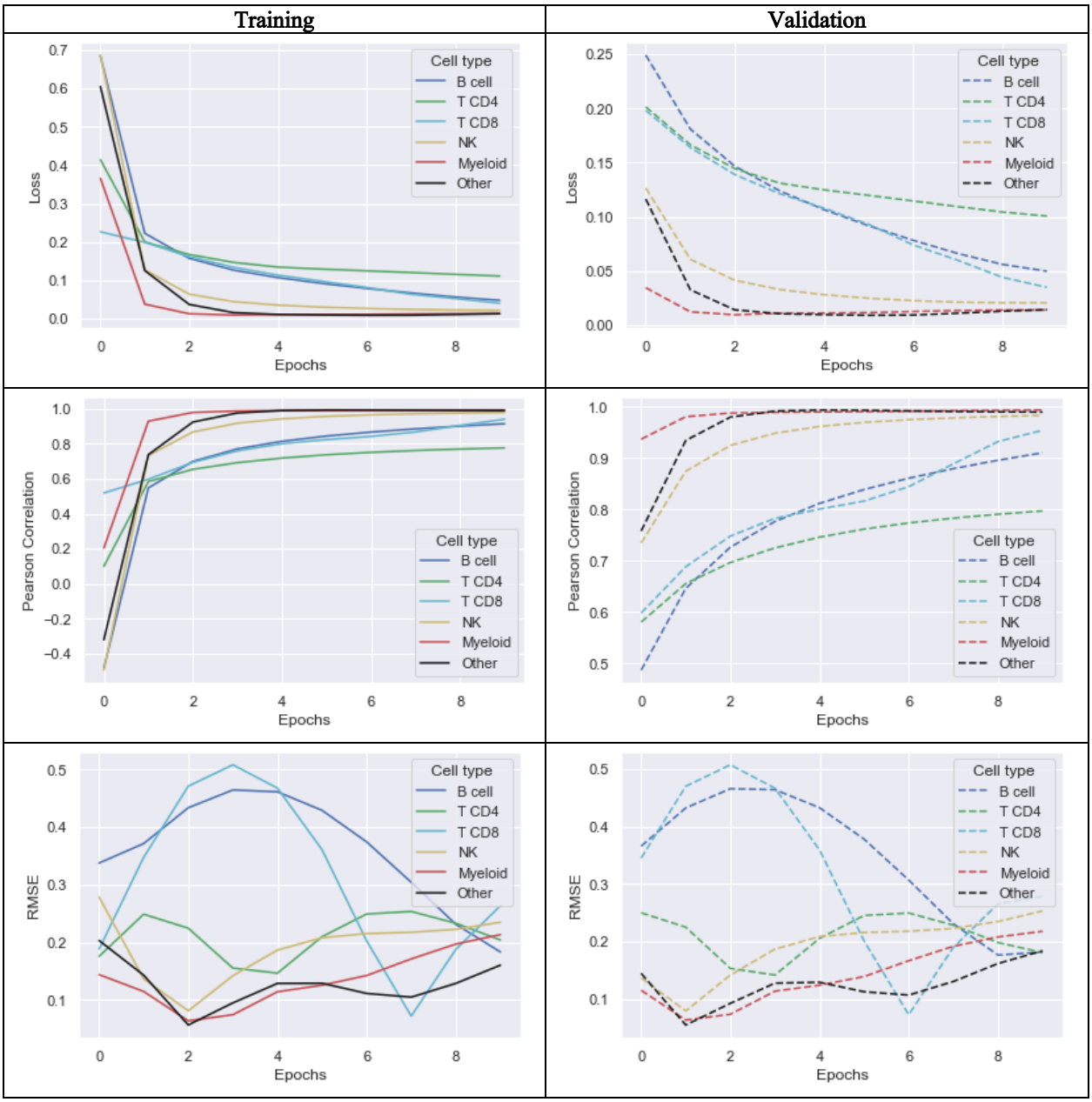


Figure 4.10 Training history of performance metrics (loss, Pearson correlation, RMSE) for the NNICE model for training (left) and validation (right) data.

Table 4.4 Performance of NNICE expression deconvolution model by cell types.

Cell type	Loss (training)	Loss (validation)	Pearson correlation (training)	Pearson correlation (validation)	RMSE (training)	RMSE (validation)
B cell	0.047868	0.049630	0.914451	0.910366	0.183390	0.181860
T CD4	0.111183	0.100479	0.776069	0.797260	0.204266	0.180917
T CD8	0.040219	0.034704	0.940366	0.954283	0.263763	0.278719
NK	0.021716	0.020406	0.978335	0.983276	0.235035	0.253303
Myeloid	0.013878	0.013869	0.993336	0.993851	0.213518	0.217974
Other	0.012828	0.014040	0.989757	0.989689	0.160532	0.184164

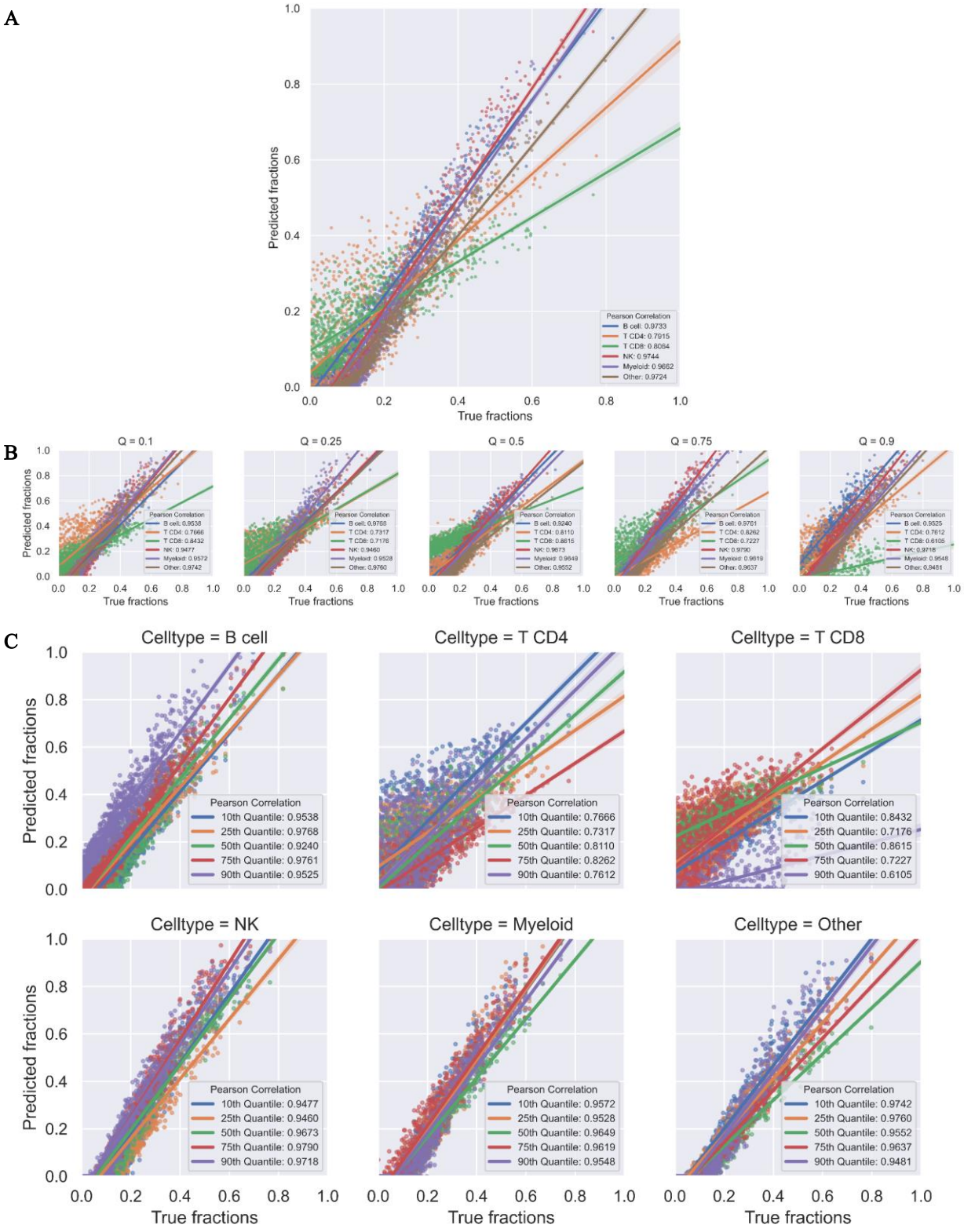
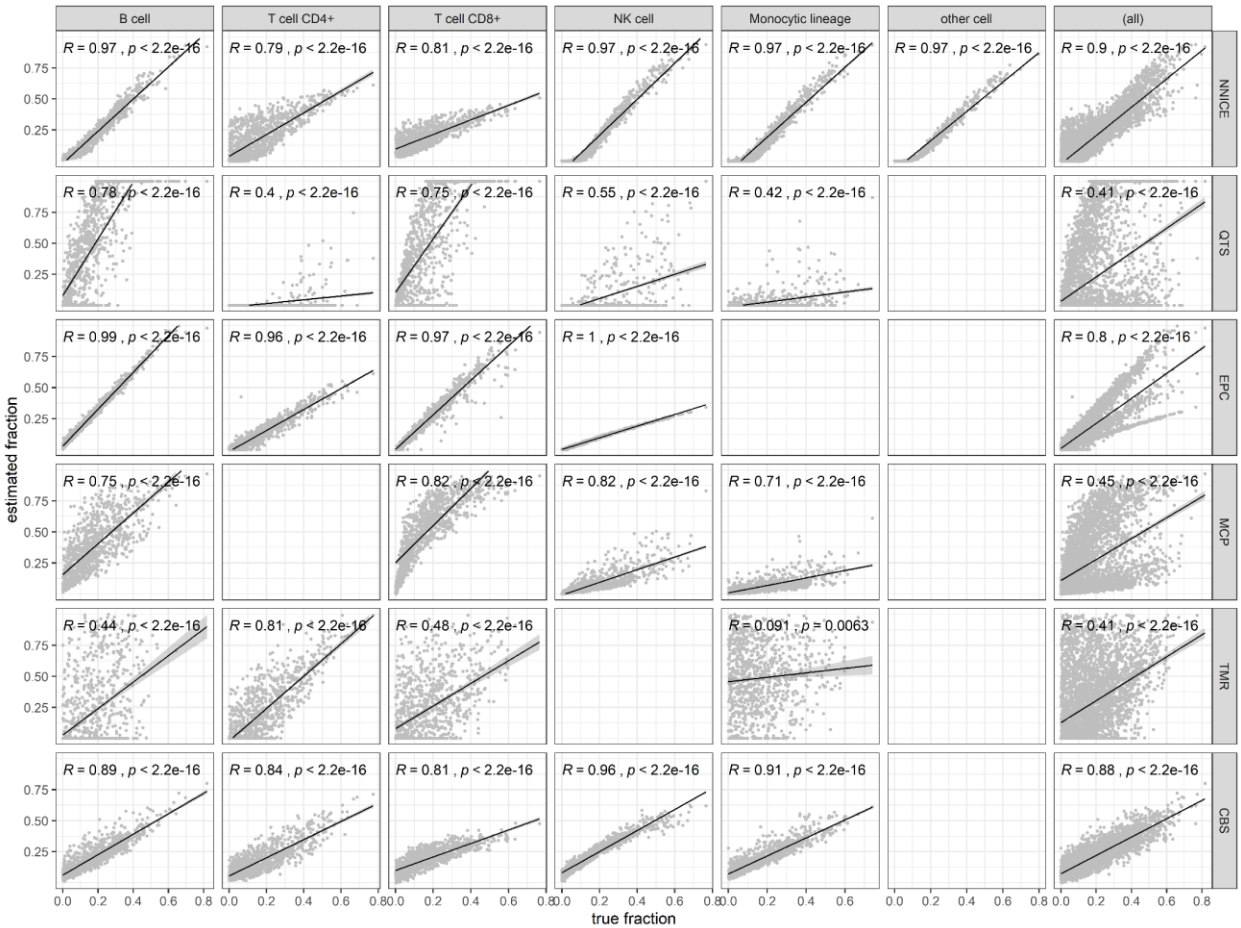


Figure 4.11 Scatter plots of true cell type fractions (x-axis) by fractions predicted by NNICE (y-axis) on the previously unseen 1,000 simulated pseudo-bulk GEPs. Each panel shows results from (**A**) average across all quantiles; (**B**) each quantile; and (**C**) each cell type. Solid lines show simple linear regression lines for each quantile or cell type.

The performance of NNICE was compared against existing methods on previously unseen 1,000 pseudo-bulk GEPs; refer to **Figure 4.12**. NNICE trained on 10,000 pseudo-bulk GEPs showed the best performance among all existing methods at overall Pearson correlation $R = 0.9$. The second-best performing tool on pseudo-bulk GEP was CIBSERSORT at $R = 0.88$.

Figure 4.12 Comparison of prediction performance of NNICE and existing expression deconvolution methods on previously unseen 1,000 pseudo-bulk GEPs. Each row shows results from a single method from the following: NNICE (trained on simulated data); QTS (quanTIseq); EPC (EPIC); MCP (MCP-counter);



TMR (TIMER); and CBS (CIBERSORT). Each column represents results for a single cell type from the following list: B cell, $CD4^+$ T cell, $CD8^+$ T cell, NK cell, cells of monocytic lineage, other cells, and all cell types combined. For each scatter plot, x-axis is the true fraction and the y-axis is the estimated fraction by the methods. Blank plots indicate that the specific method provides no estimates for the particular cell type. Solid black line shows the computed Pearson correlation (R) between the true and estimated fractions between -1 and 1 with p-value showing probability that the correlation in data is due to chance.

4.3.3 Evaluation of model trained on real bulk tissue GEPs with FACS-quantified cell type proportions

To evaluate the performance of NNICE model trained with real data, 10-fold cross-validation approach was used to train on the available 468 real bulk tissue GEP from the SDY67 dataset to evaluate its performance. **Figure 4.13** below shows prediction results from 10-fold cross-validation of NNICE model on SDY67 dataset. Correlation between the predicted and true fractions was similar across the 5 quantiles but the model did not perform consistently across all six cell types. Estimates from NNICE on SDY67 dataset was most accurate on the NK cells ($R = 0.745$) and least on the $CD4^+$ T cells ($R = 0.3833$). Across all cell types, however, predictions from the NNICE model showed greater correlation with truth ($R = 0.7$) when compared with existing methods, of which predictions from EPIC had the highest correlation ($R = 0.65$); refer to **Figure 4.14**.

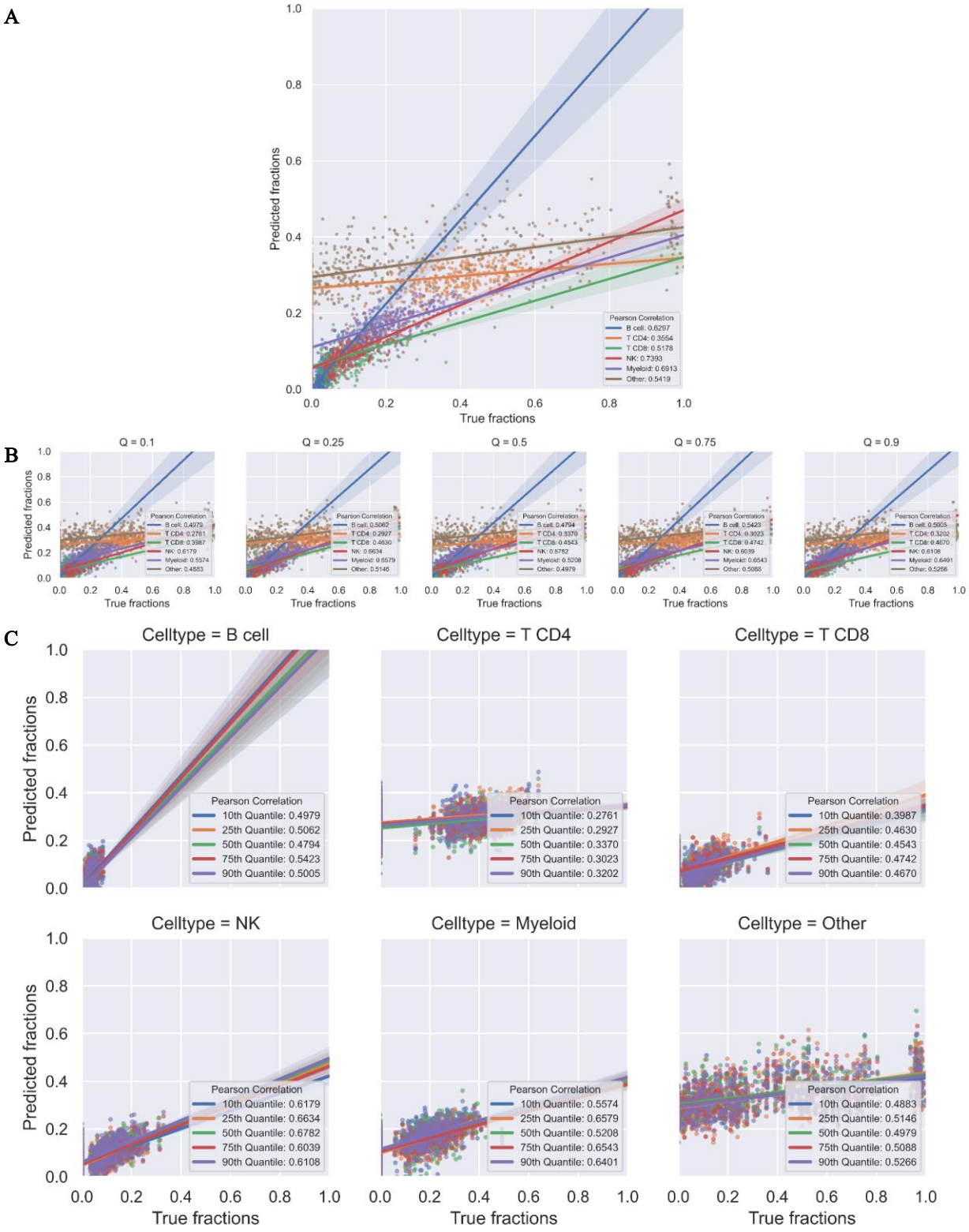


Figure 4.13 Scatter plots of true cell type fractions (x-axis) by fractions predicted by NNICE (y-axis) on the previously unseen 468 GEPs from SDY67 dataset. Each panel shows results from (**A**) average across all quantiles; (**B**) each quantile; and (**C**) each cell type. Solid lines show simple linear regression lines for each quantile or cell type.

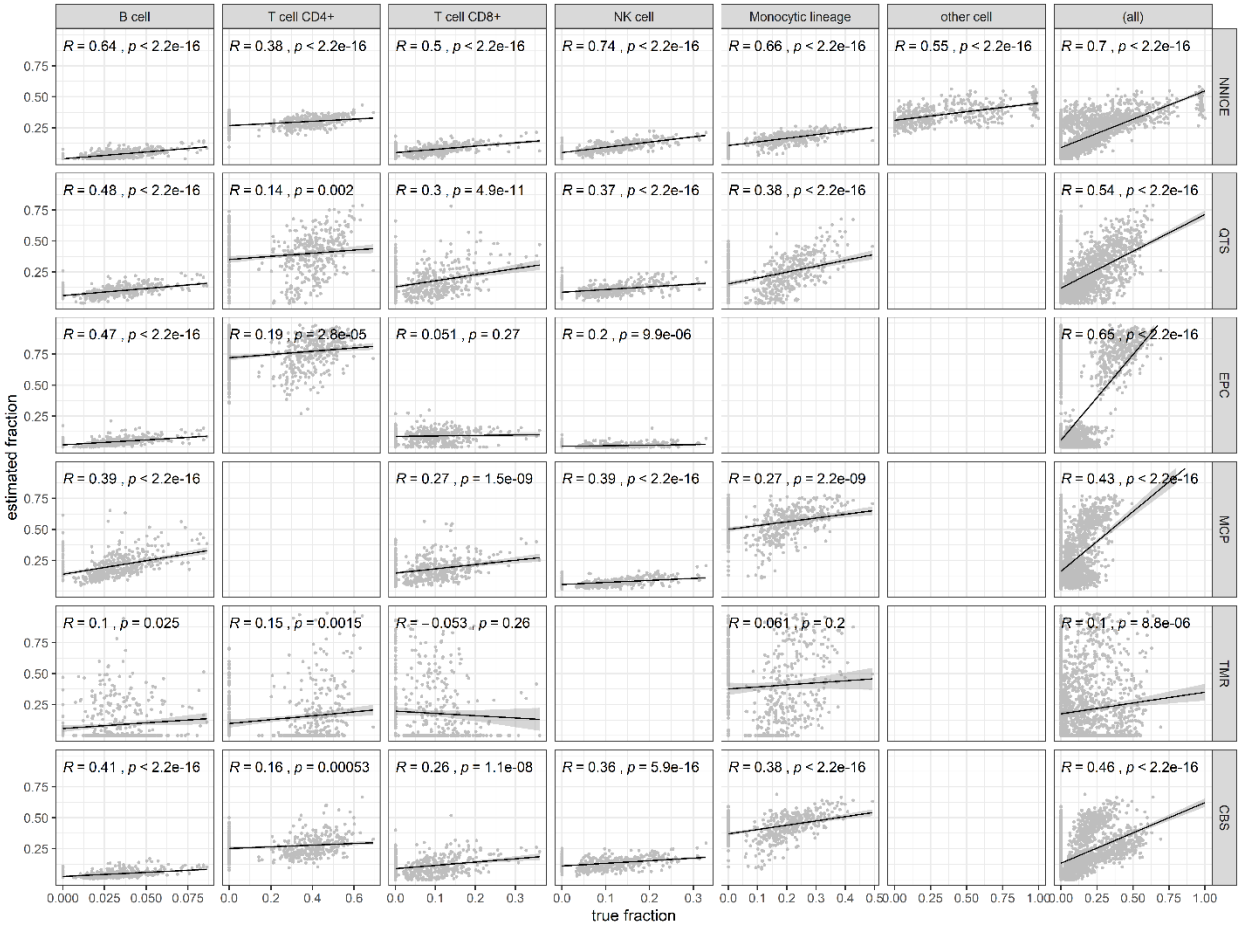


Figure 4.14 Comparison of prediction performance of NNICE and existing expression deconvolution methods on previously unseen 468 GEPs from SDY67 dataset. Each row shows results from a single method from the following: NNICE (trained on real data); QTS (quanTIseq); EPC (EPIC); MCP (MCP-counter); TMR (TIMER); and CBS (CIBERSORT). Each column represents results for a single cell type from the following list: B cell, $CD4^+$ T cell, $CD8^+$ T cell, NK cell, cells of monocytic lineage, other cells, and all cell types combined. For each scatter plot, x-axis is the true fraction and the y-axis is the estimated fraction by the methods. Blank plots indicate that the specific method provides no estimates for the particular cell type. Solid black line shows the computed Pearson correlation (*R*) between the true and estimated fractions between -1 and 1 with *p*-value showing probability that the correlation in data is due to chance.

Upon confirming that NNICE could be expected to attain robust performance when trained with only a subset of the real dataset, another instance of NNICE model was trained on the entire 468 samples of the SDY67 dataset for 100 epochs. This instance of NNICE model was used to validate its performance on an external real dataset (ABIS) as well as the simulated pseudo-bulk GEPs. Record of training (SDY67) and validation (ABIS) metrics – loss, Pearson correlation, and RMSE – throughout the 100 epochs of training is shown on **Figure 4.15**. In contrast to training

metrics that reached convergence between 80 and 100 epochs, validation metrics fluctuated throughout the training process and did not plateau. Consequently, NNICE model performed poorly on the external validation (ABIS) dataset. Overall correlation across all cell types was $R = 0.35$ with best estimates for $CD4^+$ T cell at $R = 0.2067$; refer to **Figure 4.16**. and **Figure 4.17**.

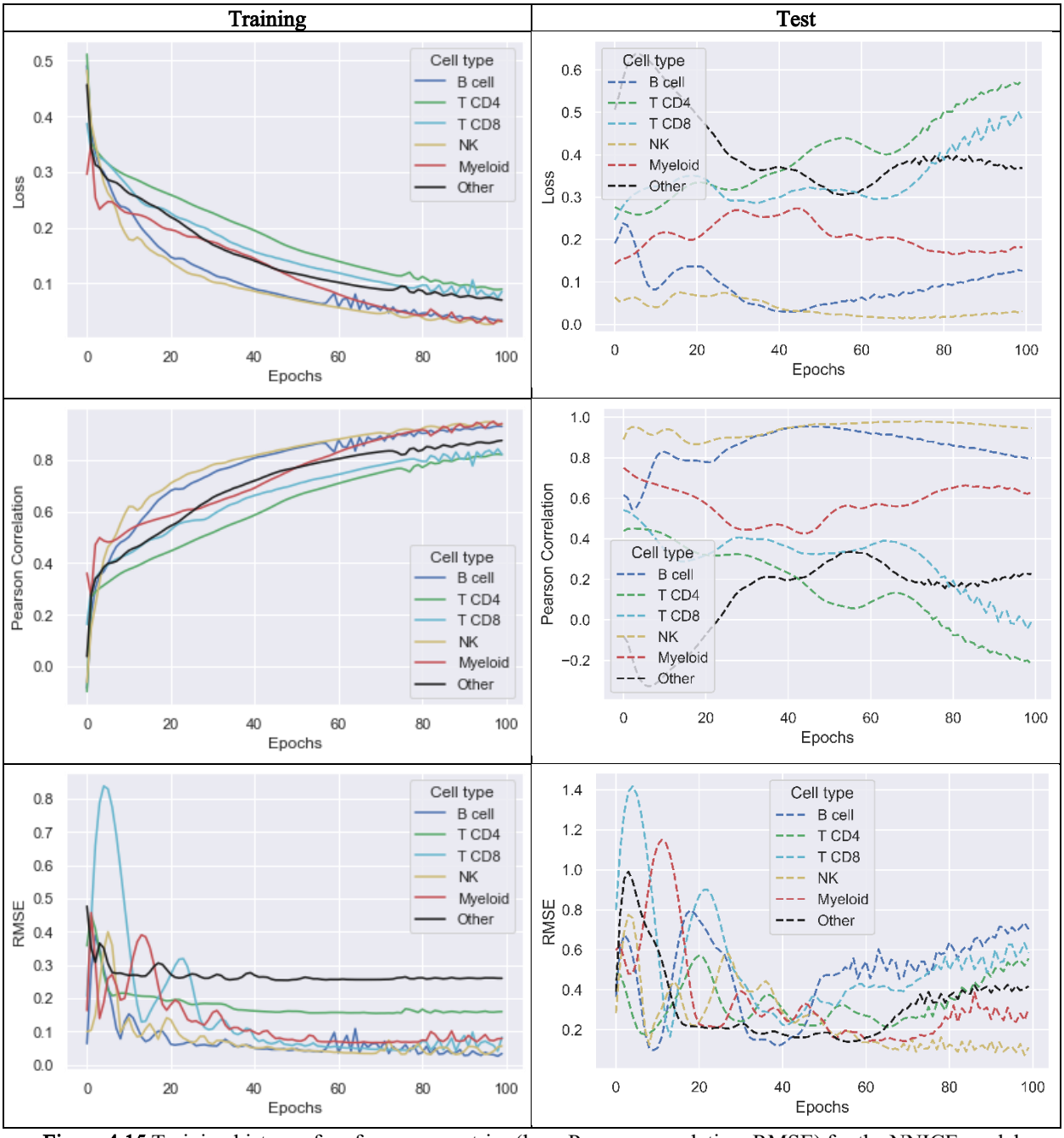


Figure 4.15 Training history of performance metrics (loss, Pearson correlation, RMSE) for the NNICE model for training (left) and validation (right) data (ABIS; $n = 12$).

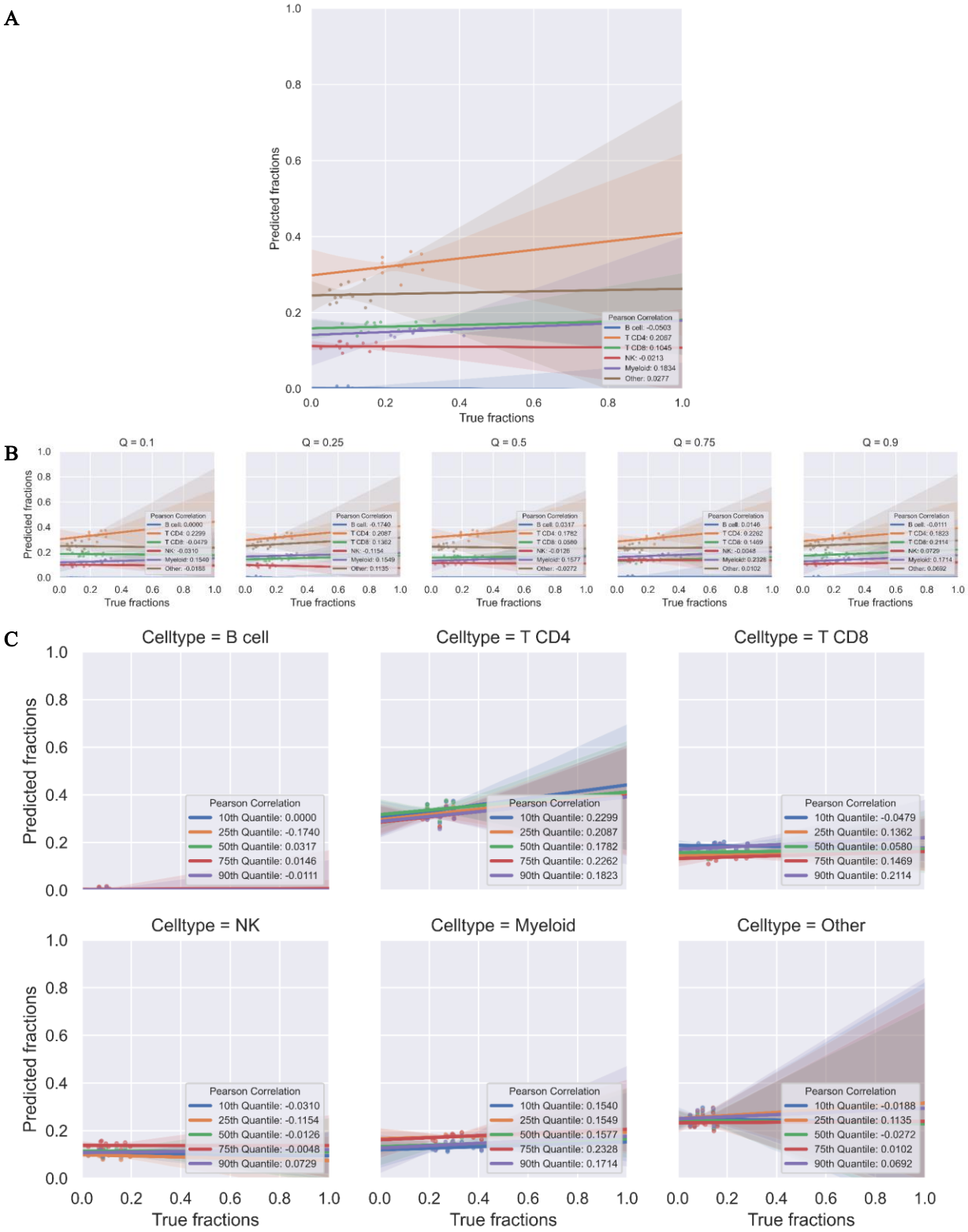


Figure 4.16 Scatter plots of true cell type fractions (x-axis) by fractions predicted by NNICE (y-axis) on the previously unseen 12 GEPs from ABIS dataset. Each panel shows results from (**A**) average across all quantiles; (**B**) each quantile; and (**C**) each cell type. Solid lines show simple linear regression lines for each quantile or cell type.

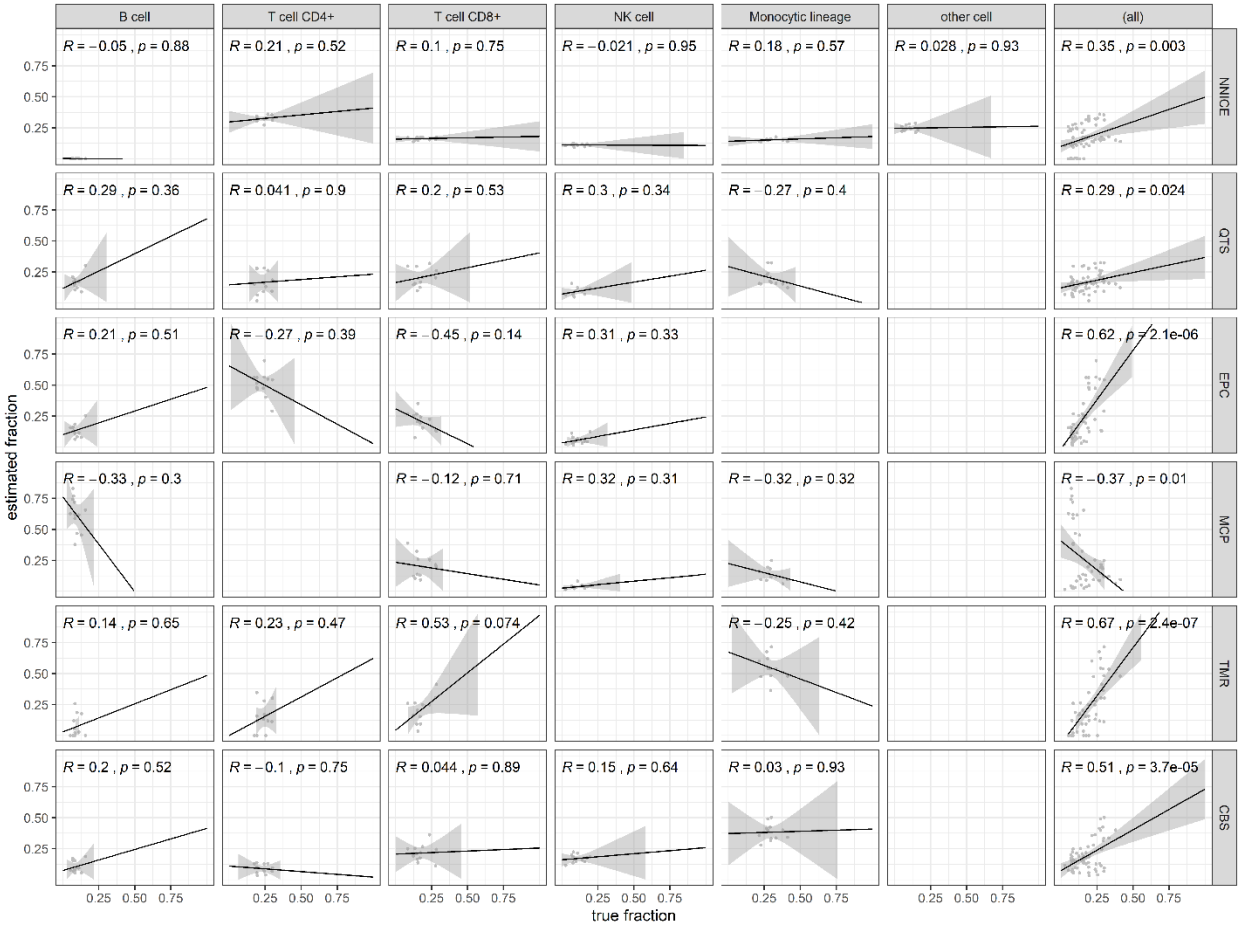


Figure 4.17 Comparison of prediction performance of NNICE and existing expression deconvolution methods on previously unseen 12 GEPs from ABIS dataset. Each row shows results from a single method from the following: NNICE (trained on real data); QTS (quanTIseq); EPC (EPIC); MCP (MCP-counter); TMR (TIMER); and CBS (CIBERSORT). Each column represents results for a single cell type from the following list: B cell, $CD4^+$ T cell, $CD8^+$ T cell, NK cell, cells of monocytic lineage, other cells, and all cell types combined. For each scatter plot, x-axis is the true fraction and the y-axis is the estimated fraction by the methods. Blank plots indicate that the specific method provides no estimates for the particular cell type. Solid black line shows the computed Pearson correlation (R) between the true and estimated fractions between -1 and 1 with p-value showing probability that the correlation in data is due to chance.

When applied to the simulated pseudo-bulk dataset, cell type fraction estimates from NNICE model trained on real data showed poor overall correlation across all cell types (R = 0.15) but showed remarkable recovery of certain cell type fractions for NK cells (R = 0.9363), $CD8^+$ T cells (R = 6406), and myeloid cells (R = 0.5199). **Figure 4.18** below shows overall prediction performance of NNICE on 1,000 pseudo-bulk GEPs, as well as by quantiles and by cell types. In addition, **Table 4.5** shows a summary of overall performance of all expression deconvolution tools compared thus far in this study on different datasets.

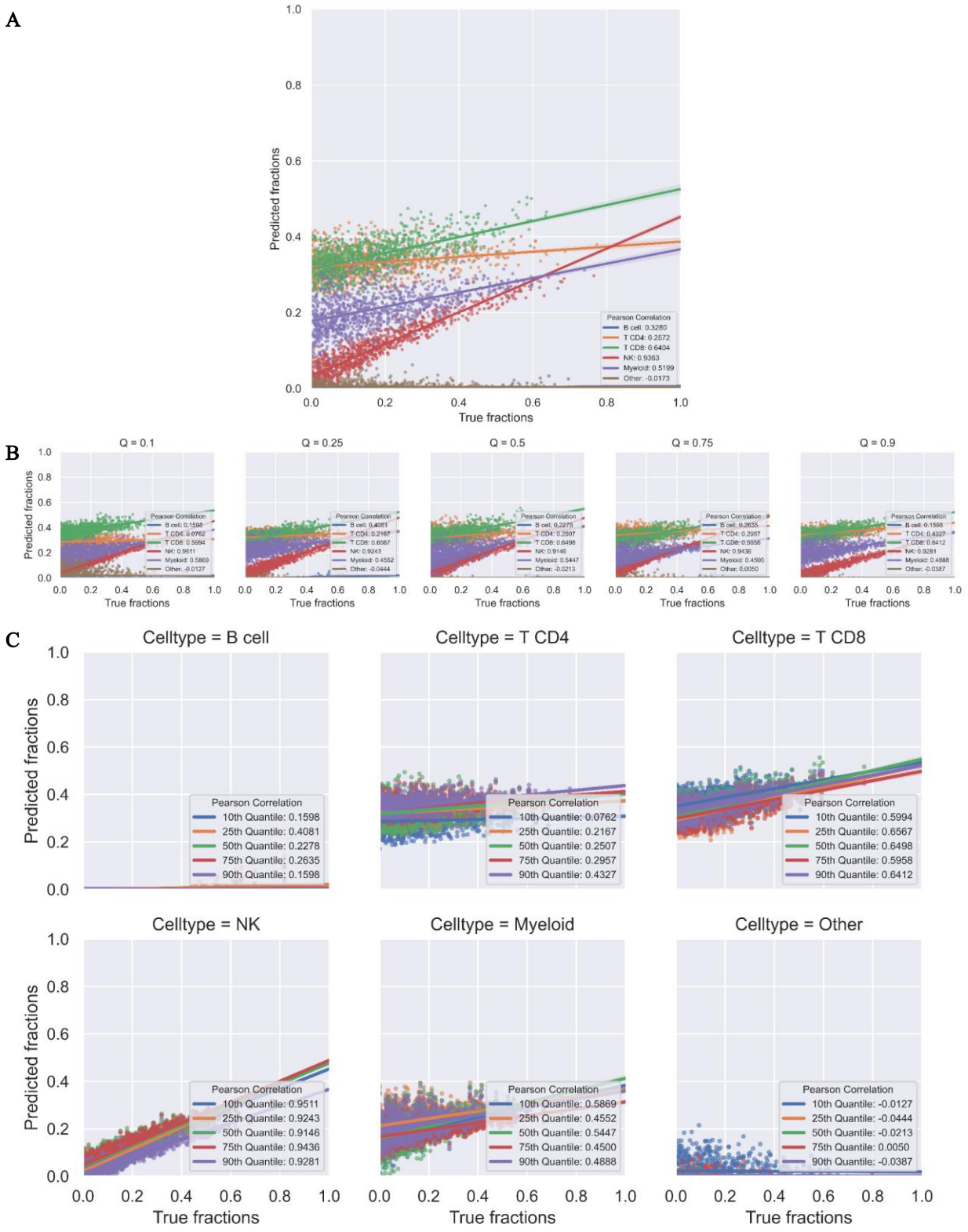


Figure 4.18 Scatter plots of true cell type fractions (x-axis) by fractions predicted by NNICE (y-axis) on the previously unseen 1,000 simulated pseudo-bulk GEPs. Each panel shows results from (**A**) average across all quantiles; (**B**) each quantile; and (**C**) each cell type. Solid lines show simple linear regression lines for each quantile or cell type.

Table 4.5 Overall performance (Pearson correlation) of expression deconvolution models on the three datasets. Performance on simulated data for NNICE is after training on 10,000 pseudo-bulk samples. Performance on SDY67 dataset for NNICE is based on the 10-fold CV approach. Performance on the ABIS dataset for NNICE is after training on 468 samples from SDY67 dataset. Best performance achieved for each dataset is bolded.

Models	Simulated (n = 1,000)	SDY67 (n = 468)	ABIS (n = 12)
NNICE	**0.9**	**0.7**	0.35
quanTIseq	0.41	0.54	0.29
EPIC	0.8	0.65	0.62
MCP-counter	0.45	0.43	-0.37
TIMER	0.41	0.1	**0.67**
CIBERSORT	0.88	0.46	0.51

4.3.4 Characterizing the breast tumour immune contexture with NNICE

NNICE model was applied to breast tumour GEP datasets (TCGA-BRCA and METABRIC) to characterize the tumour immune contexture breast cancer patients as previously done in Chapter 3.

First, immune cell type fractions estimated by NNICE was visualized by heatmap for both TCGA-BRCA and METABRIC cohorts; see **Figure 4.19**. From the heatmaps, stark visual differences could be noticed between estimates for the two cohorts – whereas most samples in the METABRIC cohort were predicted to have high proportion of "other" cell types within the tumour tissue, results for TCGA were more variable across samples.

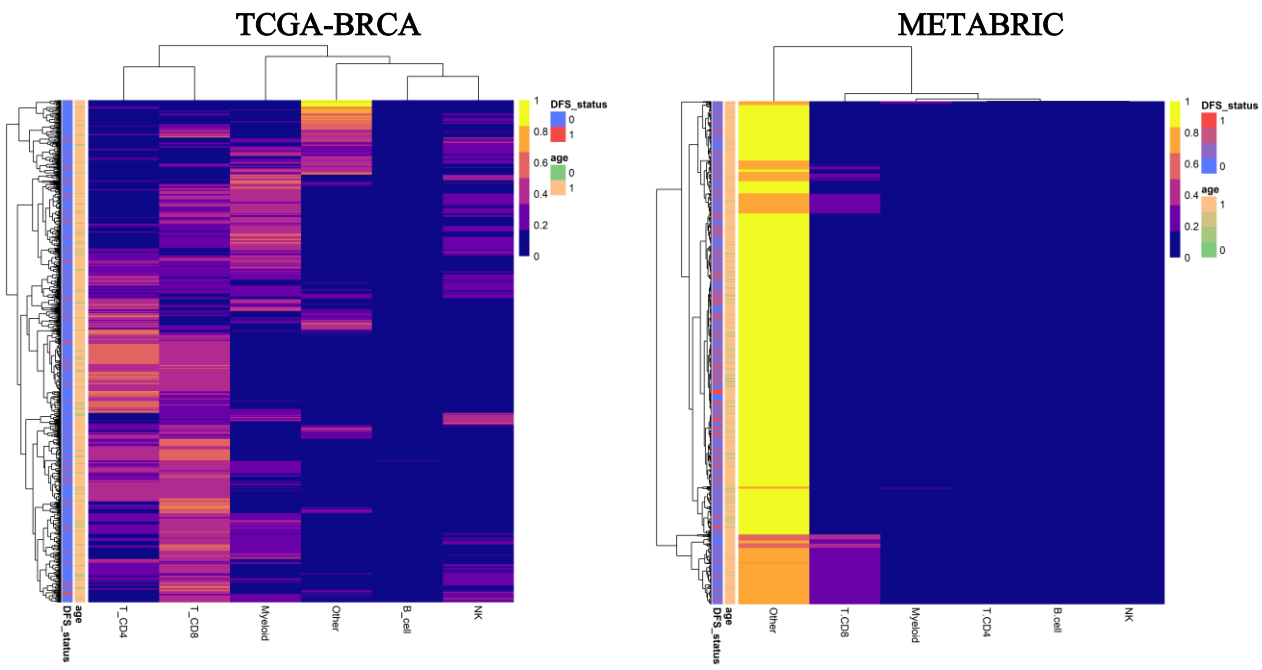


Figure 4.19 Heatmaps of abundance estimates for immune subsets predicted by computation deconvolution using NNICE model for the two cohorts: TCGA-BRCA (left); and METABRIC (right). Sum of cell type fraction values across each row is 1. Row and column dendrograms show clustering of cases and cell types, respectively, according to Euclidean distance. Status of disease-free survival within the follow-up period and age at diagnosis is indicated for each row.

Wilcoxon rank-sum statistic was computed to identify any immune subsets with significant differences in abundance between early- (age <40) and late-onset (age ≥40); refer to **Figure 4.20**. for the statistics along with the violin plot of estimated immune cell proportions for both cohorts. As expected from the heatmaps, NNICE predicted that bulk tissue samples from METABRIC cohort are primarily occupied by "other" cells with proportions >50% in most samples. Younger patients (age <40) in the METABRIC cohort had significantly higher proportion of other cells – medians of 95% in young versus 91.6% older patients (Wilcoxon $p = 1.5\times10^{-7}$). However, this was opposite in the TCGA-BRCA cohort, from which the older patients were estimated to have higher proportion of other cells with median of 4.46% versus 3.21% in young patients (Wilcoxon $p = 0.049$). For the remaining cell types and in general, there were no significant differences between the two age groups that were consistent across both cohorts.

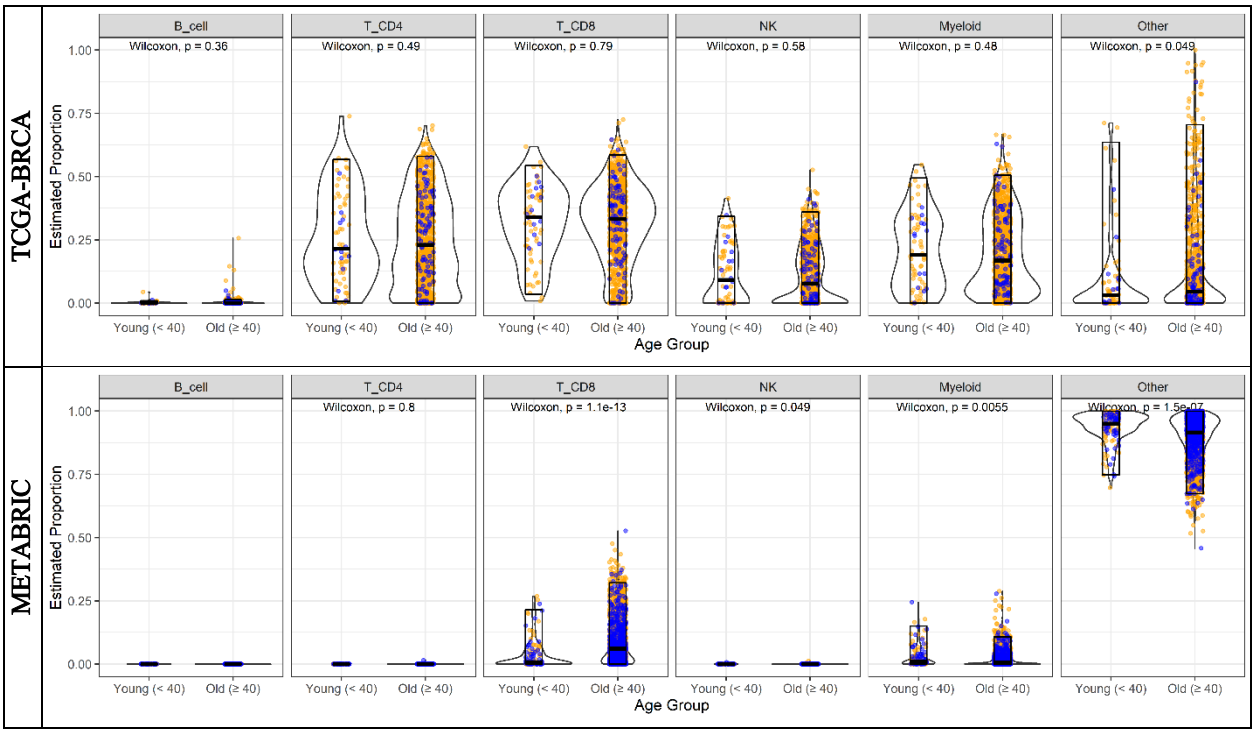


Figure 4.20 Violin plot of abundance estimates for immune subsets predicted by computation deconvolution using NNICE model after stratifying each estimate for age group and cell type for the two cohorts: TCGA-BRCA (top); and METABRIC (bottom). Each dot represents a patient from which the blue dots represents patients who experienced relapse, metastasis, or death due to disease and yellow dots represent all other patients with no record of such events during the follow-up years.

4.3.5 Immune subsets associated with clinical outcomes in breast cancer

As previously done in Chapter 3, KM survival analyses were performed to discover potential associations between proportion of immune subset estimated by NNICE and clinical outcome in early- and late-onset breast cancers. In brief, both TCGA-BRCA and METABRIC cohorts were stratified into young (<40) and old (≥40) age groups, as well as being separated into high and low abundance group for each cell type depending on the threshold value computed by the maxstat R package. **Figure 4.21** shows KM survival curves for samples split into high and low abundance of myeloid, other, and $CD8^+$ T cells as estimated by NNICE; cell types with insignificant results were excluded. Of the six cell types that NNICE provided cell type proportion estimates for, only the proportion of myeloid cell type had significant effect on disease-free survival (DFS) in the old age group. In both TCGA-BRCA (log-rank $p = 0.026$; threshold = 5.47%) and METABRIC (log-rank

$p = 0.018$; threshold = 0.17%) cohorts, late-onset patients with high proportion of myeloid cells estimated by NNICE experienced poor clinical outcomes. Similarly, patients with high proportion of "other" cells were expected to have worse survival outcomes in the METABRIC cohort (log-rank $p < 0.0001$; threshold = 91.41%) but this relationship was opposite in the TCGA-BRCA cohort (log-rank $p = 0.027$; threshold = 0.88%) from which patients with high fraction of other cells faired better in terms of disease-free survival. Furthermore, high proportion of $CD8^+$ T cells in late-onset breast cancer patients from METABRIC was significantly associated with favourable survival outcomes (log-rank $p < 0.0001$; threshold = 9.77%). This relationship was also opposite in late-onset patients from TCGA-BRCA cohort, although statistical significance was marginal at log-rank $p = 0.093$ at threshold of 31.42%.

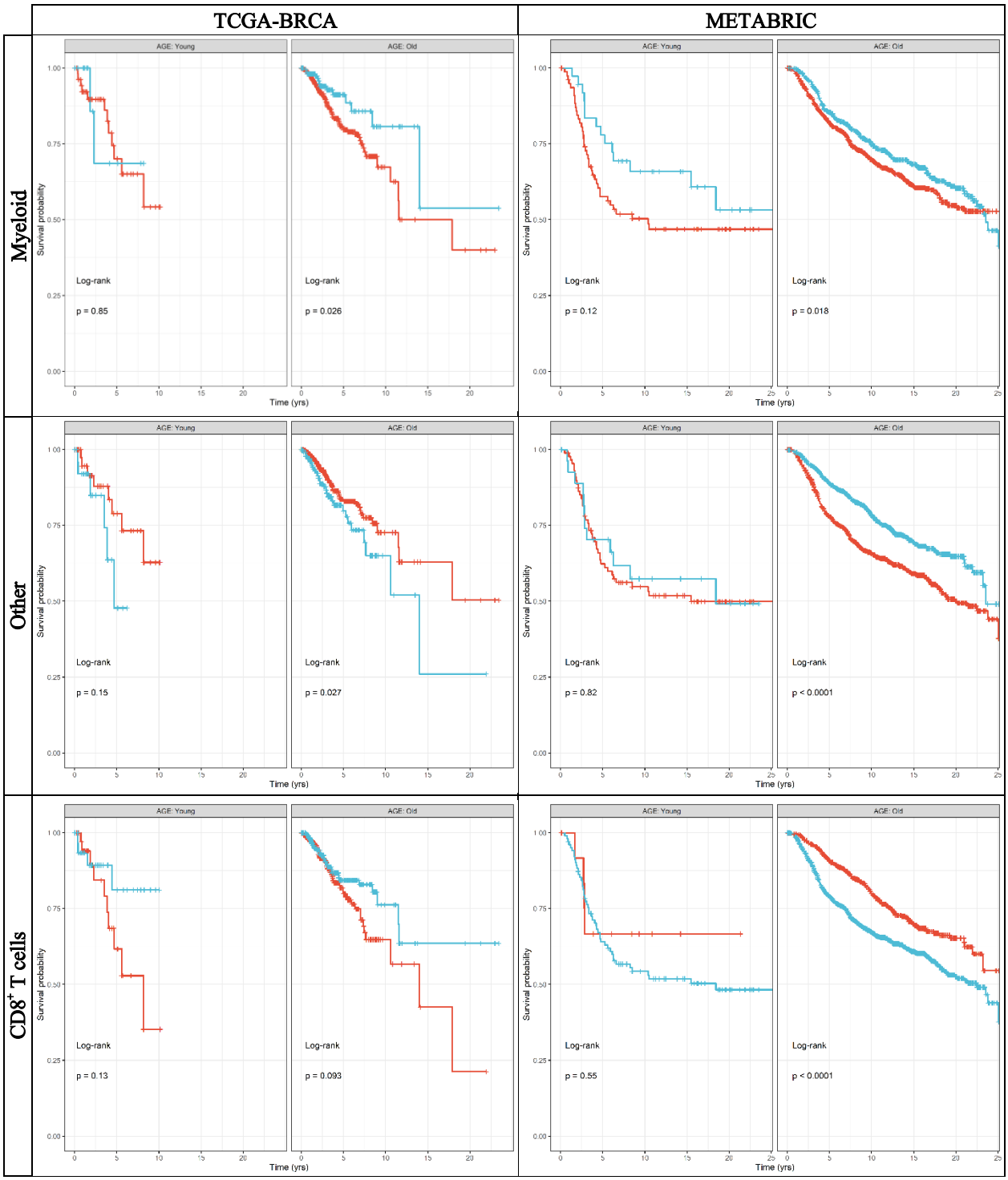


Figure 4.21 Disease-free survival Kaplan–Meier curve for: TCGA-BRCA (left panels); and METABRIC (right panels) cohorts, grouped by age groups and binarized into high (red) and low (blue) immune cell proportions for three cell types (myeloid, other, and $CD8^+$ T cells) estimated by NNICE. Threshold to binarize estimates was determined by maximally selected ranked statistics algorithm. Depicted *p*-values are from log-rank tests.

The hazard ratios (HR) for the associations between cell type fractions and DFS were quantified by Cox proportional hazards regression as described in Chapter 3. The estimated HRs and the 95% confidence intervals (CIs) are shown in **Figure 4.22**. Results from some cell types were removed if HR estimates were not meaningful with 95% CI range exceeding limits on both sides of the plot axis. No significant results were found except for the late-onset (age ≥40) group, in which proportions of $CD8^+$ T cells (HR = 0.95; Wald test $p = 3.29\times10^{-6}$) and "other" cells (HR = 1.20; Wald test $p = 1.137\times10^{-7}$) estimated by NNICE was significantly associated with DFS.

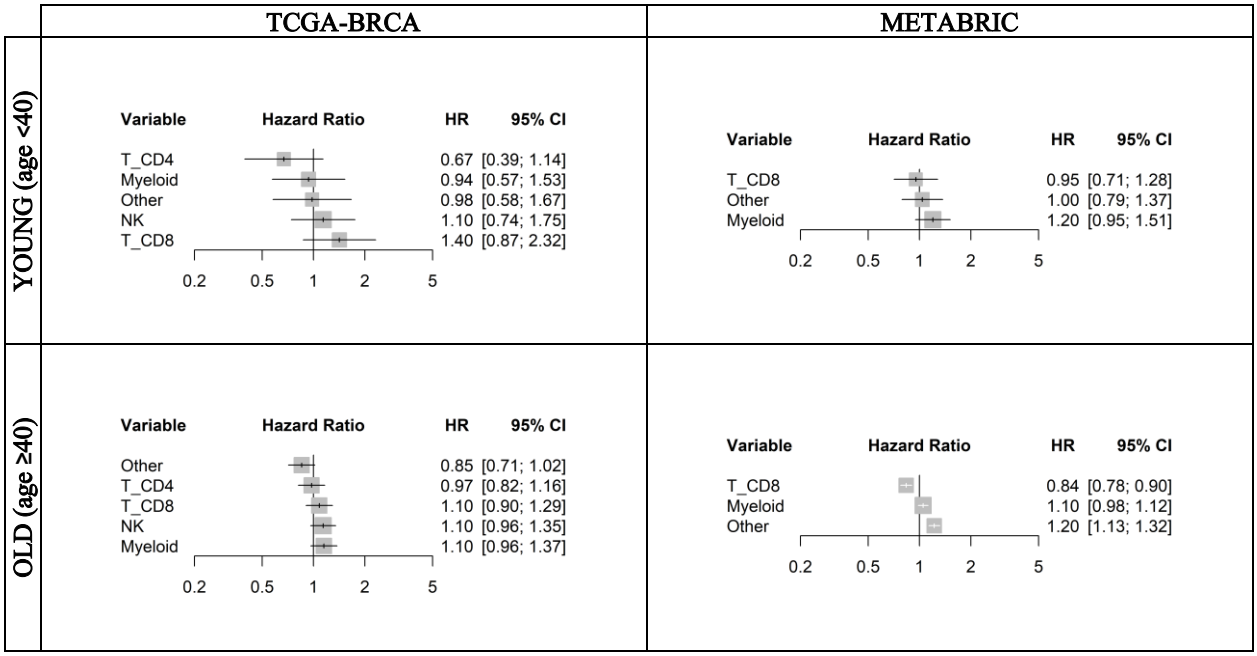


Figure 4.22 Hazard ratios of each immune cell type fraction quantified by NNICE, as individually estimated by univariable Cox regression models on disease-free survival with 95% confidence intervals (CIs). Left panels show results for TCGA-BRCA, while the right panels show results for METABRIC cohort. Top panels show results for young patients with early-onset breast cancer and bottom panels are for their older counterparts with late-onset breast cancer.

4.4 Discussion

Enormous success of artificial neural networks (ANNs) and deep learning in the fields of computer vision and natural language processing has prompted their application to various problems in computational biology over the recent decade. Central motivation of this work was to develop a novel method of expression deconvolution based on a deep learning framework to

characterize the tumour immune contexture of breast cancer patients. Because artificial neural network models have shown to outperform many of the existing algorithms, it was reasonable to expect potential improvements over existing methods of expression deconvolution. This research endeavor, however, was met with numerous challenges.

Firstly, ANNs usually require massive amounts of data to train, in orders of thousands if not millions, depending on the type of problem. It is not impossible to train on a smaller amount of data; however, ANNs often fail to achieve convergence with insufficient training data during training, leading to unstable predictions. In order to overcome this limitation, it was proposed that models would be trained on artificial gene expression data simulated from scRNA-seq datasets. This way, an arbitrarily large amount of pseudo-bulk GEPs could be generated to optimize performance of ANN models. Using simulated GEPs to benchmark and validate tools for expression deconvolution is not a novel idea – several groups have previously simulated pseudo-bulk GEP data from both RNA-seq [44,54] and scRNA-seq [96,97]. Pseudo-bulk simulation involves the following steps: 1) simulation of cell type fractions within each pseudo-bulk mixture; 2) subsampling of single cell GEPs or reads from bulk GEPs in proportions simulated in the first step; and 3) summation of reads across all features up to a maximum cell or read count. Although this simulation approach is often used, previous studies have focused on validating performance of expression deconvolution tools on these datasets; therefore, validation of the simulation approach itself was largely neglected. Artificial datasets simulated in this way may be poor representations of real data because bulk-RNA-seq profiles are not simple summations of scRNA-seq profiles that have been subject to many technical biases and errors, for example, drop-outs. This could be one of the reasons underlying poor performance of NNICE on unseen real or simulated data. There is a need for existing pseudo-bulk GEP simulation strategies to undergo

rigorous evaluations to assess to what extent simulated datasets resemble real datasets. This way, new algorithms should be developed to generate simulated datasets that are increasingly similar to real datasets. Computational approach to generate new data is resource-efficient; however, large-scale studies that involves collection of new gene expression and cell fractions data would also help overcome the current shortage of training data with known cell type proportions data.

Second, underfitting and overfitting are two common problems in machine learning. In particular, preventing overfitting was a major challenge for this project because in the end, none of the neural network models were able to provide stable and robust estimates for tissue immune contexture across the three datasets considered (simulated, SDY67, and ABIS). Models trained only on simulated pseudo-bulk GEP data failed to show robust performances when applied to GEP from real bulk tissues (SDY67). Since previous works that applied deep learning framework to expression deconvolution used simple densely connected layers, the model architectures tested in this project were mostly also variations of dense neural networks. Therefore, there it is necessary to test whether different model architectures and optimization strategies have a positive impact on the performance of expression deconvolution and generalizability of the model to unseen data.

Finally, NNICE model trained on SDY67 dataset did not perform well on the ABIS dataset. This means that the trained NNICE model could not generalize well between real datasets, which could be because of several factors. As mentioned in Section 4.2.2, while Illumina HiSeq 2000 sequencer was used to generate transcriptomic data for both datasets, different sequencing (single-end vs. paired-end) and bioinformatics workflows were used to collect and preprocess the data. These technical differences could have contributed to the poor generalization of NNICE model trained on one dataset to the other, despite transforming both datasets to the same range and scale. Moreover, it is important to consider that the SDY67 dataset was generated from PBMCs of

healthy participants between ages 50 and 74, while the ABIS dataset was generated from PBMCs of healthy participants between ages 20 and 35. As previously mentioned, aging is known to significantly effect immune composition; therefore, immunosenescence may have affected the robustness of the NNICE model trained on one dataset over the other.

With all things considered, it is difficult to conclude that estimates for immune cell type fractions by NNICE models trained on either simulated or real data are true and accurate. However, the abundance of certain immune subsets – such as myeloid cells, $CD8^+$ T cells and other cells – were significantly associated with clinical outcomes in the older subset of breast cancer patients. Therefore, although the estimates should not be interpreted as ground-truth cell type fractions within tissues, it is also important to recognize that these values have potential to serve as intermediate risk scores for prognostic use, while providing additional insight into the immune contexture within the TME in breast cancer.

Contrary to what was initially expected, no significant differences in tumour immune landscape were identified between early- and late-onset breast cancer patients. As mentioned in Chapter 3, this could reflect the actual state of the immune landscape within the context of the TME, which would be different from the global changes in immune cell abundance resulting from process of immunosenescence. Moreover, it is more likely that any significant differences in the tumour immune landscape were not detected because the estimates from NNICE cannot be relied on to be true representations of actual cell type proportions within the tissue anyway. Therefore, results from this study are inconclusive – it would be interesting to investigate experimentally whether the cell type compositions within the TME are significantly different between young and older patients in breast cancer.

4.5 Conclusions

Artificial neural network model (NNICE) was able to successfully recover ground-truth cell type fraction values given unseen bulk mixture gene expression profiles from the same dataset it was trained on. NNICE was not sufficiently robust to provide consistent performance across different datasets; however, this could be resolved either with acquisition of additional good quality bulk gene expression datasets or improvements to simulation algorithms that generate pseudo-bulk profile data. It is difficult to conclude whether estimates from NNICE should be interpreted as accurate representation of the tumour immune contexture; however, the estimated values for certain immune cell type fractions showed significant associations with disease-free survival in late-onset breast cancer patients, suggesting that these estimates may provide prognostic value.

Chapter 5. Significance, limitations, and future work

Due to the ever-growing interest to characterize the cell type contexture within bulk tissue samples for research in various fields of molecular biology, computational deconvolution of gene expression profiles has become an active area of research in the past two decades. Although experimental methods including microscopy-based methods such as immunohistochemistry, and flow cytometry-based method such as fluorescence-activated cell sorting technique, are considered to be the "gold-standard" methods of providing ground-truth values of cell type proportions, these methods remain expensive with respect to time, resources and labour. In comparison, computational expression deconvolution algorithms, once established, have potential to offer a more efficient way to quantify cell type proportions within bulk tissue samples, especially with substantial reductions in cost of next generation sequencing in the recent years. However, development of efficient deconvolution algorithm that is robust to the diversity of genomic datasets have presented itself with unique challenges of its own as discussed in this thesis.

In Chapter 3, computational estimates for specific immune subsets, such as $CD8^+$ T cells, made by TIMER deconvolution tool were found to potentially have significant associations with clinical outcomes in early-onset breast cancer patients. This immune abundance score was further validated by GSEA of list of genes significantly correlated with the $CD8^+$ T cell estimates, which showed that the list was enriched for genes related to adaptive immune response. Therefore, computational deconvolution of tumour gene expression profiles was able to characterize distinct patterns of tumour infiltrating immune cells that were significantly associated with clinical outcomes in early-onset breast cancer.

However, due to the ill-conditioned nature of the problem of expression deconvolution – meaning that there are numerous computationally-sound solutions from the same input data – it is

unseemly to expect that a single estimate would perfectly correlate with true cell type fractions within the tissue. Nonetheless, even though the estimated cell type proportions may not resemble true proportions within the bulk tissues, it was discussed that these estimates can provide prognostic information when interpreted as a type of risk scores that are associated with clinical outcomes in cancer patients.

This thesis also encompassed endeavors to apply state-of-the-art deep learning and artificial neural network models as a potential solution to the problem of expression deconvolution. Few groups have already attempted to make significant improvements over existing methods using deep learning algorithms [96,97]. However, their models as well as those tested in this thesis, were elementary in terms of the architecture because they only consisted of a series of densely connected layers. This shows that research is still nascent in this area of study, and future work should focus on testing more sophisticated model architectures from the rapidly advancing field of deep learning. Furthermore, in order to train these sophisticated models to be data agnostic or impervious to biases present in different datasets, there is a necessity for pseudo-bulk simulation algorithms to be further refined to generate higher quality of artificial datasets that are reflective of real-world data. With all things considered, even though there is not yet a gold-standard method for expression deconvolution, with future research and development, this computational approach holds enormous potential to provide novel insights into the cellular composition within various tissues for research in the diverse areas of biology to ultimately enhance the capacity for personalized medicine.

References

[1] D. R. Brenner *et al.*, "Projected estimates of cancer in Canada in 2020," *CMAJ*, vol. 192, no. 9, pp. E199–E205, Mar. 2020, doi: 10.1503/cmaj.191292.

[2] M. Yalaza, A İnan, M Bozer, "Male Breast Cancer," *J. Breast Health*, vol. 12, no. 1, pp. 1–8, Jan. 2016, doi: 10.5152/tjbh.2015.2711.

[3] C. Curtis *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," *Nature*, vol. 486, no. 7403, pp. 346–352. Apr. 2012, doi: 10.1038/nature10983.

[4] The Cancer Genome Atlas Network, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, pp. 61–70, Sep. 2012, doi: 10.1038/nature11412.

[5] K. Rojas, and A. Stuckey, "Breast cancer epidemiology and risk factors," *Clin. Obstet. Gynecol.*, vol. 59, no. 4, pp. 651–672, Dec. 2016, doi: 10.1097/GRF.0000000000000239.

[6] N. Harbeck, and M. Gnant, "Breast cancer," *Lancet*, vol. 389, no. 10074, pp. 1134–1150, Mar. 2017, doi: 10.1016/S0140–6736(16)31891-8.

[7] A. G. Waks, and E. P. Winer, "Breast cancer treatment: A review," *JAMA*, vol. 321, no. 3, pp. 288–300, Jan. 2019, doi:10.1001/jama.2018.19323.

[8] J. Makki, "Diversity of breast carcinoma: Histological subtypes and clinical relevance," *Clin. Med. Insights. Pathol.*, vol. 8, pp. 23–31, Dec. 2015, doi: 10.4137/CPath.S31563.

[9] A. V. Barrio, and K. J. Van Zee, "Controversies in the treatment of DCIS," *Annu. Rev. Med.*, vol. 68, pp. 197–211, Jan. 2017, doi: 10.1146/annurev-med-050715-104920.

[10] E. S. McDonald, A. S. Clark, J. Tchou, P. Zhang, and G. M. Freedman, "Clinical diagnosis and management of breast cancer," *J. Nucl. Med.*, vol. 57, suppl. 1, pp. 9S–16S, Feb. 2016, doi: 10.2967/jnumed.115.157834.

[11] S.B. Edge, and C.C. Compton, "The American joint committee on cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM," *Ann. Surg. Oncol.* vol. 17, pp. 1471–1474, Feb. 2010, doi: 10.1245/s10434-010-0985-4.

[12] N. Harbeck *et al.*, "Breast cancer," *Nat. Rev. Dis. Primers*, vol. 5, no. 66, Sep. 2019, doi: 10.1038/s41572-019-0111-2.

[13] C. Perou *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747–752, Aug. 2000, doi: 10.1038/35021093.

[14] J. S. Parker *et al.*, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol.*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009, doi: 10.1200/JCO.2008.18.1370.

[15] C. C. Benz., "Impact of aging on the biology of breast cancer," *Crit. Rev. Oncol. Hematol.*, vol. 66, no. 1, pp. 65–74. Apr. 2008, doi: 10.1016/j.critrevonc.2007.09.001.

[16] E. O. Jenkins *et al.*, "Age-specific changes in intrinsic breast cancer subtypes: A focus on older women," *Oncologist*, vol. 19, no. 10, pp. 1076–1083, Oct. 2014, doi: 10.1634/theoncologist.2014-0184.

[17] C. K. Anders *et al.*, "Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression," *J. Clin. Oncol.*, vol. 26, no. 20, pp. 3324–3330, Jul. 2008, doi: 10.1200/JCO.2007.14.2471.

[18] J. Kollias, C. W. Elston, I. O. Ellis, J. F. Robertson, and R. W. Blamey, "Early-onset breast cancer--histopathological and prognostic considerations," *Br. J. Cancer*, vol. 75, no. 9, pp. 1318–1323, 1997, doi: 10.1038/bjc.1997.223.

[19] T. Osako *et al.*, "Age-correlated protein and transcript expression in breast cancer and normal breast tissues is dominated by host endocrine effects," *Nature Cancer*, vol. 1, pp. 518–532, May 2020, doi: 10.1038/s43018-020-0060-4.

[20] C. Yau *et al.*, "Aging impacts transcriptomes but not genomes of hormone-dependent breast cancers," *Breast Cancer Res.*, vol 9, no. R59, Sep. 2007, doi: 10.1186/bcr1765.

[21] C. K. Anders *et al.*, "Breast carcinomas arising at a young age: Unique biology or a surrogate for aggressive intrinsic subtypes?" *J. Clin. Oncol.*, vol. 29, no. 1, pp. e18–e20, Jan. 2011, doi: 10.1200/JCO.2010.28.9199.

[22] F. R. Balkwill, M. Capasso, and T. Hagemann, "The tumour microenvironment at a glance," *J. Cell Sci.*, vol. 125, no. Pt 23, pp. 5591–5596, Dec. 2012, doi: 10.1242/jcs.116392.

[23] A. E. Place, S. J. Huh, and K. Polyak, "The microenvironment in breast cancer progression: Biology and implications for treatment," *Breast Cancer Res.*, vol. 13, no. 6, pp. 227, 2011, doi: 10.1186/bcr2912.

[24] M.W. Conklin, and, P.J. Keely, "Why the stroma matters in breast cancer: Insights into breast cancer patient outcomes through the examination of stromal biomarkers," *Cell Adh. Migr.*, vol. 6, no. 3, pp. 249-260, 2012, doi: 10.4161/cam.20567.

[25] M. Allinen *et al.*, "Molecular characterization of the tumour microenvironment in breast cancer," *Cancer Cell*, vol. 6, no. 1, pp. 17-32, Jul. 2004, doi: 10.1016/j.ccr.2004.06.010.

[26] X. J. Ma, S. Dahiya, E. Richardson, M. Erlander, and D. C. Sgroi, "Gene expression profiling of the tumour microenvironment during breast cancer progression," *Breast Cancer Res.*, vol. 11, no. 1, pp. R7, 2009, doi: 10.1186/bcr2222.

[27] G. Finak *et al.*, "Stromal gene expression predicts clinical outcome in breast cancer," *Nat. Med.*, vol. 14, pp. 518-527, Apr. 2008, doi: 10.1038/nm1764.

[28] C. O. Madu, S. Wang, C. O. Madu, and Y. Lu, "Angiogenesis in breast cancer progression diagnosis, and treatment," *J. Cancer*, vol. 11, no. 15, pp. 4474-4494, May 2020, doi: 10.7150/jca.44313

[29] S. C. Schwager, P. V. Taufalele, and C. A. Reinhart-King, "Cell-cell mechanical communication in cancer," *Cell Mol. Bioeng.*, vol. 12, no. 1, pp. 1-14, Feb. 2019, doi: 10.1007/s12195-018-00564-x.

[30] R. Baghban *et al.*, "Tumour microenvironment complexity and therapeutic implications at a glance," *Cell Commun. Signal*, vol. 18, no. 59, Apr. 2020, doi: 10.1186/s12964-020-0530-4.

[31] H. Gonzalez, C. Hagerling, and Z. Werb, "Roles of the immune system in cancer: From tumour initiation to metastatic progression," *Genes Dev.*, vol. 32, no. 19-20, pp. 1267-1284, Oct. 2018, doi: 10.1101/gad.314617.118.

[32] S. D. Soysal, A. Tzankov, and S. E. Muenst, "Role of the Tumour Microenvironment in Breast Cancer," *Pathobiology*, vol. 82, no. 3-4, pp. 142-152, Sep. 2015, doi: 10.1159/000430499.

[33] P. Italiani, and D. Boraschi, "From monocytes to M1/M2 macrophages: Phenotypical vs. functional differentiation," *Front. Immunol.*, vol. 5, no. 514, Oct. 2014, doi: 10.3389/fimmu.2014.00514.

[34] N. M. Anderson, and M. C. Simon, "The tumour microenvironment," *Curr. Biol.*, vol. 30, no. 16, pp. R921-R925, Aug. 2020, doi: 10.1016/j.cub.2020.06.081.

[35] T. L. Whiteside, "The tumour microenvironment and its role in promoting tumour growth," *Oncogene*, vol. 27, pp. 5904-5912, Oct. 2008, doi: 10.1038/onc.2008.271.

[36] D. Nagarajan, and S. E. B. McArdle, "Immune landscape of breast cancers," *Biomedicines*, vol. 6, no. 1, pp. 20, Feb. 2018, doi: 10.3390/biomedicines6010020.

[37] J. S. O'Donnell, M. W. L. Teng, and M. J. Smyth, "Cancer immunoediting and resistance to T cell-based immunotherapy," *Nat. Rev. Clin. Oncol.*, vol. 16, no. 3, pp. 151-167, Mar. 2019, doi: 10.1038/s41571-018-0142-8.

[38] M. García-Aranda, and M. Redondo, "Immunotherapy: A challenge of breast cancer treatment," *Cancers*, vol. 11, no. 12, pp. 1822, Nov. 2019, doi: 10.3390/cancers11121822.

[39] E. J. Márquez *et al.*, "Sexual-dimorphism in human immune system aging," *Nat. Commun.*, vol. 11, no. 751, Feb. 2020, doi: 10.1038/s41467-020-14396-9.

[40] M. Fane, and A.T. Weeraratna, "How the ageing microenvironment influences tumour progression," *Nat. Rev. Cancer*, vol. 20, pp. 89–106, Dec. 2020, doi: 10.1038/s41568-019-0222-9.

[41] C. Franceschi, and J. Campisi, "Chronic inflammation (inflammaging) and its potential contribution to age-associated diseases," *J. Gerontol. A. Biol. Sci. Med. Sci.*, vol. 69, suppl. 1, pp. S4-S9, Jun. 2014, doi: 10.1093/gerona/glu057.

[42] F. Finotello, and Z. Trajanoski, "Quantifying tumour-infiltrating immune cells from transcriptomics data," *Cancer Immunol. Immunother.*, vol. 67, pp. 1031–1040, Mar. 2018, doi: 10.1007/s00262-018-2150-z.

[43] G. Sturm *et al.*, "Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology," *Bioinformatics*, vol. 35, no. 14, pp. i436-i445, Jul. 2019, doi: 10.1093/bioinformatics/btz363.

[44] A. Newman *et al.*, "Robust enumeration of cell subsets from tissue expression profiles," *Nat. Methods*, vol. 12, pp. 453–457, Mar. 2015, doi: 10.1038/nmeth.3337.

[45] P. Lu, A. Nakorchevskiy, and E. Marcotte, "Expression deconvolution: A reinterpretation of DNA microarray data reveals dynamic changes in cell populations,"*Proc. Natl. Acad. Sci. U.S.A.*, vol. 100, no. 18, pp. 10370-10375, Sep. 2003, doi: 10.1073/pnas.1832361100.

[46] D. Venet, F. Pecasse, C. Maenhaut, and H. Bersini, "Separation of samples into their constituents using gene expression data," *Bioinformatics*, vol. 17, suppl. 1, pp. S279-287, Jun. 2001, doi: 10.1093/bioinformatics/17.suppl_1.s279.

[47] R. O. Stuart *et al.*, "In silico dissection of cell-type-associated patterns of gene expression in prostate cancer," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, no. 2, pp. 615-620, Jan. 2004, doi: 10.1073/pnas.2536479100.

[48] K. Yoshihara *et al.*, "Inferring tumour purity and stromal and immune cell admixture from expression data," *Nat. Commun.*, vol. 4, no. 2612, Oct. 2013, doi: 10.1038/ncomms3612.

[49] B. Li *et al.*, "Comprehensive analyses of tumour immunity: implications for cancer immunotherapy," *Genome Biol.*, vol. 17, no. 174, Aug. 2016, doi: 10.1186/s13059-016-1028-7.

[50] G. Quon, and Q. Morris, "ISOLATE: A computational strategy for identifying the primary origin of cancers using high-throughput sequencing," *Bioinformatics*, vol. 25, no. 21, pp. 2882-2889, Nov. 2009, doi: 10.1093/bioinformatics/btp378.

[51] G. Monaco *et al.*, "RNA-seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types," *Cell Rep.*, vol. 26, no. 6, pp. 1627-1640.e7, Feb. 2019, doi: 10.1016/j.celrep.2019.01.041.

[52] T. Gong *et al.*, "Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples," *PLoS One*, vol. 6, no. 11, pp. e27156, Nov. 2011, doi: 10.1371/journal.pone.0027156.

[53] D. Repsilber *et al.*, "Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach,". *BMC Bioinformatics*, vol. 11, no. 27, Jan. 2010, doi: 10.1186/1471-2105-11-27.

[54] F. Finotello *et al.*, "Molecular and pharmacological modulators of the tumour immune contexture revealed by deconvolution of RNA-seq data," *Genome Med.*, vol. 11, no. 34, May 2019, doi: 10.1186/s13073-019-0638-6.

[55] R. H. Johnson, C. K. Anders, J. K. Litton, K. J. Ruddy, and A. Bleyer, "Breast cancer in adolescents and young adults," *Pediatr. Blood Cancer*, vol. 65, no. 12, pp. e27397, Dec. 2018, doi: 10.1002/pbc.27397.

[56] R. H. Johnson, P. Hu, C. Fan, and C. K. Anders, "Gene expression in 'young adult type' breast cancer: A retrospective analysis," *Oncotarget*, vol. 6, no. 15, pp. 13688–13702, May 2015, doi:10.18632/oncotarget.4051.

[57] M. T. Martínez *et al.*, "Breast cancer in very young patients in a Spanish cohort: Age as an independent bad prognostic indicator," *Breast Cancer Basic Clin. Res.*, vol. 13, pp. 1-10, Feb. 2019, doi: 10.1177/1178223419828766.

[58] G. Pruneri, A. Vingiani, and C. Denkert, "Tumour infiltrating lymphocytes in early breast cancer," *Breast*, vol. 37, pp. 207-214, Feb. 2018, doi:10.1016/j.breast.2017.03.010.

[59] J. Ziai *et al.*, "CD8+ T cell infiltration in breast and colon cancer: A histologic and statistical analysis," *PLoS One*, vol. 13, no. 1, pp. e0190158, Jan. 2018, doi: 10.1371/journal.pone.0190158.

[60] K. Wang, T. Shen, G.P. Siegal, and S. Wei, "The CD4/CD8 ratio of tumour-infiltrating lymphocytes at the tumour-host interface has prognostic value in triple-negative breast cancer," *Hum. Pathol.*, vol. 69, pp. 110-117, Nov. 2017, doi:10.1016/j.humpath.2017.09.012.

[61] X. Yang *et al.*, "Prognostic significance of CD4/CD8 ratio in patients with breast cancer," *Int. J. Clin. Exp. Pathol.*, vol. 10, no. 4, pp. 4787–4793, 2017.

[62] H. R. Ali *et al.*, "Association between CD8+ T-cell infiltration and breast cancer survival in 12,439 patients," *Ann. Oncol.*, vol. 25, no. 8, pp. 1536–1543, Aug. 2014, doi:10.1093/annonc/mdu191.

[63] S. E. Clare, and P. L. Shaw, ""Big Data" for breast cancer: Where to look and what you will find," *npj Breast Cancer*, vol. 2, pp. 16031, Nov. 2016, doi:10.1038/npjbcancer.2016.31.

[64] H. R. Ali, L. Chlon,P. D. P. Pharoah, F. Markowetz, and C. Caldas, "Patterns of immune infiltration in breast cancer and their clinical implications: A gene-expression-based retrospective study," *PLoS Med.* vol. 13, no. 12, pp. e1002194, Dec. 2016, doi:10.1371/journal.pmed.1002194.

[65] T. O'Meara *et al.*, "Immune microenvironment of triple-negative breast cancer in African-American and Caucasian women," *Breast Cancer Res. Treat.*, vol. 175, no. 1, pp. 247-259, May 2019, doi:10.1007/s10549-019-05156-5.

[66] C. K. Anders, R. Johnson, J. Litton, M. Phillips, and A. Bleyer, "Breast cancer before age 40 years," *Semin. Oncol.*, vol. 36, no. 3, pp. 237–249, Jun. 2009, doi:10.1053/j.seminoncol.2009.03.001.

[67] S. Zeeshan, B. Ali, K. Ahmad, A. B. Chagpar, and A. K. Sattar, "Clinicopathological features of young versus older patients with breast cancer at a single Pakistani institution and a comparison with a national US database," *J. Glob. Oncol.*, vol. 5, pp. 1–6, Mar. 2019, doi:10.1200/jgo.18.00208.

[68] E. Montecino-rodriguez, B. Berent-maoz, and K. Dorshkind, “Causes, consequences, and reversal of immune system aging,” *J. Clin. Invest.*, vol. 123, no. 3, pp. 958–965, Mar. 2013, doi:10.1172/JCI64096.958.

[69] C. M. Gameiro, F. Romao, and C. Castelo-Branco, “Menopause and aging: Changes in the immune system--A review,” *Maturitas*, vol. 67, no. 4, pp. 316–320, Dec. 2010, doi:10.1016/j.maturitas.2010.08.003.

[70] S. Loi *et al.*, “Tumour-infiltrating lymphocytes and prognosis: A pooled individual patient analysis of early-stage triple-negative breast cancers,” *J. Clin. Oncol.*, vol. 37, no. 7, pp. 559–569, Mar. 2019, doi:10.1200/JCO.18.01010.

[71] G. Desdín-Micó, G. Soto-Heredero, and M. Mittelbrunn, “Mitochondrial activity in T cells,” *Mitochondrion*, vol. 41, pp. 51–57, Jul. 2018, doi:10.1016/j.mito.2017.10.006.

[72] M. Le Borgne, and A. S. Shaw, “Do T cells have a cilium?” *Science*, vol. 342, no. 6163, pp. 1177–1178, Dec. 2013, doi: 10.1126/science.1248078.

[73] C. Cassioli, and C. T. Baldari, “A ciliary view of the immunological synapse,” *Cells*, vol. 8, no. 8, pp. 789, Jul. 2019, doi:10.3390/cells8080789.

[74] S. Nik-Zainal *et al.*, “Landscape of somatic mutations in 560 breast cancer whole-genome sequences,” *Nature*, vol. 534, no. 7605, pp. 47–54, Nov. 2016, doi:10.1038/nature17676.

[75] M. Smid *et al.*, “Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration,” *Nat. Commun.*, vol. 7, pp. 12910, Sep. 2016, doi:10.1038/ncomms12910.

[76] L. B. Alexandrov *et al.*, “The repertoire of mutational signatures in human cancer,” *Nature*, vol. 578, pp. 94–101, Feb. 2020, doi:10.1038/s41586-020-1943-3.

[77] I. S. Kim *et al.*, “Immuno-subtyping of breast cancer reveals distinct myeloid cell profiles and immunotherapy resistance mechanisms,” *Nat. Cell Biol.*, vol. 21, pp. 1113–1126, Aug. 2019, doi:10.1038/s41556-019-0373-7.

[78] K. Wang *et al.*, “Identification of differentially expressed genes in non-small cell lung cancer,” *Aging*, vol. 11, no. 23, pp. 11170–11185, Dec. 2019, doi:10.18632/aging.102521.

[79] B. Li, T. Li, J. S. Liu, and X. S. Liu, “Computational deconvolution of tumour-infiltrating immune components with bulk tumour gene expression data,” *Bioinformatics for Cancer Immunotherapy*, in *Methods in Molecular Biology*, vol. 2120, S. Boegel, Ed., New York, NY, USA: Humana, 2020; pp. 249–262, doi:10.1007/978-1-0716-0327-7_18.

[80] L. Wei, Z. Jin, S. Yang, Y. Xu, Y. Zhu, and Y. Ji, “TCGA-assembler 2: Software pipeline for retrieval and processing of TCGA/CPTAC data,” *Bioinformatics*, vol. 34, no. 9, pp. 1615–1617, May 2018, doi: 10.1093/bioinformatics/btx812.

[81] F. G. Gao *et al.*, “Before and after: Comparison of legacy and harmonized TCGA Genomic Data Commons’ data,” *Cell Syst.*, vol. 9, no. 1, pp. 24–34.e10, Jul. 2019, doi: 10.1016/j.cels.2019.06.006.

[82] J. Gao *et al.*, “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal,” *Sci. Signal.*, vol. 6, no. 269, pp. pl1, Apr. 2018, doi: 10.1126/scisignal.2004088.Integrative.

[83] B. Lausen, and M. Schumacher, “Maximally Selected Rank Statistics,” *Biometrics*, vol. 48, no. 1, pp. 73–85, .Mar. 1992, doi: 10.2307/2532740.

[84] R. Rosenthal, N. McGranahan,, J. Herrero, B. S. Taylor, C. Swanton, “DeconstructSigs: Delineating mutational processes in single tumours distinguishes DNA repair deficiencies and patterns of carcinoma evolution,” *Genome Biol.*, vol. 17, pp. 31, Feb. 2016, doi: 10.1186/s13059-016-0893-4.

[85] J. G. Tate *et al.*, "COSMIC: The catalogue of somatic mutations in cancer," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D941–D947, Jan. 2019, doi: 10.1093/nar/gky1015.

[86] F. Blokzijl, R. Janssen, R. van Boxtel, E. Cuppen, "MutationalPatterns: Comprehensive genome-wide analysis of mutational processes," *Genome Med.*, vol.10, no. 1, pp. 33, Apr. 2018, doi: 10.1186/s13073-018-0539-0.

[87] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/pnas.0506580102.

[88] J. Reimand *et al.*, "Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap," *Nat. Protoc.*, vol. 14, pp. 482–517, Jan. 2019, doi: 10.1038/s41596-018-0103-9.

[89] P. Shannon *et al.*, "Cytoscape: A software environment for integrated models," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.metabolite.

[90] D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader, "Enrichment map: A network-based method for gene-set enrichment visualization and interpretation," *PLoS One*, vol. 5, no. 11, pp. e13984, Nov. 2010, doi: 10.1371/journal.pone.0013984.

[91] J. H. Morris *et al.*, "clusterMaker: A multi-algorithm clustering plugin for Cytoscape," *BMC Bioinform.*, vol. 12, pp. 436, Nov. 2011, doi: 10.1186/1471-2105-12-436.

[92] M. Kucera, R. Isserlin, A. Arkhangorodsky, and G. D. Bader, "AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations," F1000Res., vol. 5, pp. 1717, Jul. 2016, doi: 10.12688/f1000research.9090.1.

[93] L. Oesper, D. Merico, R. Isserlin, and G. D. Bader, "WordCloud: A Cytoscape plugin to create a visual semantic summary of networks," *Source Code Biol. Med.*, vol. 6, pp. 7, Apr. 2011, doi: 10.1186/1751-0473-6-7.

[94] M. Hong *et al.*, "RNA sequencing: new technologies and applications in cancer research," *J. Hematol. Oncol.*, vol.13, no. 166, Dec. 2020, doi: 10.1186/s13045-020-01005-x.

[95] B. Tang, Z. Pan, K. Yin and A. Khateeb, "Recent advances of deep learning in bioinformatics and computational biology," *Front. Genet.*, vol. 10, pp. 214, Mar. 2019, doi: 10.3389/fgene.2019.00214.

[96] C. Torroja, and F. Sanchez-Cabo, "Digitaldlsorter: Deep-learning on scRNA-seq to deconvolute gene expression data," *Front. Genet.*, vol. 10 pp. 978, Oct. 2019, doi: 10.3389/fgene.2019.00978.

[97] K. Menden *et al.*, "Deep learning-based cell composition analysis from tissue expression profiles," *Sci. Adv.*, vol. 6, no. 30, pp. eaba2619, Jul. 2020, doi: 10.1126/sciadv.aba2619.

[98] G. Zheng *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nat. Commun.*, vol. 8, no. 14049, Jan. 2017, doi: 10.1038/ncomms14049.

[99] F. Wolf, P. Angerer, and F. Theis, "SCANPY: Large-scale single-cell gene expression data analysis," *Genome Biol.*, vol. 19, no. 15, Feb. 2018, doi: 10.1186/s13059-017-1382-0.

[100] C. Hafemeister, and R. Satija, "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression," *Genome Biol.*, vol. 20, no. 296, Dec. 2019, doi: 10.1186/s13059-019-1874-1.

[101] Y. Hao *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573-3587, May 2021, doi: 10.1016/j.cell.2021.04.048.

[102] M. T. Zimmermann *et al.*, "System-wide associations between DNA-methylation, gene expression, and humoral immune response to influenza vaccination," *PLoS One*, vol. 11, no. 3, pp. e0152034. doi: 10.1371/journal.pone.0152034.

[103] A. Colaprico *et al.*, "TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data," *Nucleic Acids Res.*, vol. 44, no. 8, pp. e71, May 2016, doi: 10.1093/nar/gkv1507.

[104] K. Breuer *et al.*, "InnateDB: Systems biology of innate immunity and beyond--recent updates and continuing curation," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D1228-1233, Jan. 2013, doi: 10.1093/nar/gks1147.

[105] L. Zheng, J. Fan, and Y. Mu, "OnionNet: A multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction," *ACS Omega*, vol. 4, no. 14, pp. 15956-15965, Sep. 2019, doi: 10.1021/acsomega.9b01997.

[106] F. Rodrigues, and F. C. Pereira, "Beyond expectation: Deep joint mean and quantile regression for spatio-temporal problems," arXiv:1808.08798[stat.ML], Aug. 2018.

[107] E. Becht *et al.*, "Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression," *Genome Biol.*, vol. 17, no. 218, Oct. 2016, doi: 10.1186/s13059-016-1070-5.

[108] J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, and D. Gfeller, "Simultaneous enumeration of cancer and immune cell types from bulk tumour gene expression data," *Elife*, vol. 6, pp. e26476, doi: 10.7554/eLife.26476.