Studying the Whole Through the Sum of its Parts

Comparing CD8⁺ Tumour Infiltrating Lymphocytes in Three Types of Human Carcinoma

Att studera helheten genom summan av delarna

Jämförelse av CD8⁺ tumörinfiltrerande lymfocyter i tre sorters mänskligt karcinom

Author: Gustav Christensson

Affiliations:

Department of Medical and Surgical Sciences, University of Modena and Reggio Emilia, Via Giuseppe Campi 287, 41125 Modena, Italy

RNA and Stem Cell Biology, Division of Molecular Haematology (DMH), Department of Laboratory Medicine, Faculty of Medicine, Lund University, Sölvegatan 17, Lund, Sweden

Telephone number: +46 (0) 72 448 7878

Email address: gu5456ch-s@student.lu.se

Supervisors:

Domenico Lo Tartaro, University of Modena and Reggio Emilia
Andrea Cossarizza, University of Modena and Reggio Emilia
Cristian Bellodi, Lund University

Keywords:

Cancer, immunotherapy, scRNA-seq, tumour-infiltrating lymphocyte, CD8⁺ T cells

Last updated 5 February 2020

Declaration of Original Work

This thesis is a quantitative comparison of raw data from three published studies which I chose and whose raw data I downloaded from the NCBI Gene Expression Omnibus database. These studies are listed in the bibliography as:

Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell. 2018;174(5):1293-308 e36.

Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med. 2018;24(7):978-85.

Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell. 2017;169(7):1342-56.e16.

After having imported these datasets I used the commands in Seurat and Monocle, two packages written for R Studio, to elaborate, analyse and compare them. I received help from the postgraduate student Domenico Lo Tartaro at the University of Modena and Reggio Emilia to import the data into R Studio; he also taught me how to use the commands. With this guidance and with the tutorials I found on the developers' websites I was able to write a code to process the data. I independently formulated the research questions and set the parameters in the code. Finally, my supervisor in Lund offered me comments and suggestions on a first and revised draft of this thesis.

Wednesday, 05 February 2020 *Gustav Christensson*

Table of Contents

Abbreviations

BC	Breast carcinoma
CTLA-4	Cytotoxic T-lymphocyte-associated protein 4
ENTPD1	Ectonucleoside Triphosphate Diphosphohydrolase 1
FACS	Fluorescence-activated cell sorting
FGFBP2	Fibroblast Growth Factor Binding Protein 2
GEO	Gene Expression Omnibus
GZMB	Granzyme B
HAVCR2	Hepatitis A virus cellular receptor 2
HCC	Hepatocellular carcinoma
HOBIT	Homolog of Blimp-1 In T Cells
ITGAE	Integrin Subunit Alpha E
LAG3	Lymphocyte-activation gene 3
MAIT	Mucosal-associated invariant T (cell)
NCBI	National Center for Biotechnology Information
NKG7	Natural Killer Cell Granule Protein 7
NSCLC	Non-small cell lung cancer
PCA	Principal component analysis
PD-1	Programmed cell death protein 1
PRF1	Perforin-1
scRNA-seq	Single-cell RNA sequencing
TIGIT	T cell immunoreceptor with Ig and ITIM domains
TIL	Tumour-infiltrating lymphocyte
TIM-3	T-cell immunoglobulin and mucin-domain containing-3
T_{reg}	Regulatory T cell
T_{RM}	Tissue-Resident Memory T cell
UMAP	Uniform manifold approximation and projection
UMI	Unique molecular identifier
ZNF683	Zinc Finger Protein 683

Abstract

Background: Tumour-infiltrating lymphocytes (TILs) that persist in chronic infections gradually lose their effector functions and become "exhausted". Immune checkpoint blockade is an emerging cancer treatment which aims to induce exhausted TILs into regaining their effector function to fight neoplastic cells. However, response to treatment varies significantly between different cancer types and may be the result of cancer specific TIL populations within the tumour microenvironment.

Purpose: to investigate and compare $CD8^+$ TIL populations in three carcinomas (hepatocellular, non-small-cell lung and breast carcinoma) and their relative prevalence.

Materials & Methods: Single-cell RNA sequencing (scRNA-seq) datasets from three studies concerning TIL gene expression were imported into Seurat, a scRNA-seq analysis package written for R Studio. Datasets were integrated and clustered to identify differentially expressed genes for each cluster. A pseudotemporal analysis was performed to suggest the evolutionary pathway for the TILs.

Results: $CD8^+$ TIL populations with 'effector', 'effector-memory' and proliferating signatures were found in all three carcinomas. 'Pre-exhausted' cells previously defined in non-small-cell lung carcinoma were shown to be more common in breast carcinoma; conversely 'exhausted', 'naïve-like' and $CD20^+CD8^+$ TILs were more common in lung carcinoma. 'MAIT' lymphocytes were particularly enriched in hepatocellular carcinoma samples.

Conclusions: Most of the identified lymphocyte populations possess anti-neoplastic functions that may potentially be exploited for the development of future precise cancer-specific immunotherapies. However, explaining why the populations were present in different proportions in the cancer types is challenging because the original studies used different scRNA-seq technologies. Instead, the analysis provides a useful framework for future comparisons between scRNA-seq datasets once more studies will be available.

258 words

Svensk populärvetenskaplig sammanfattning

Cancer är en sjukdom där celler i kroppen för förmågan att dela sig okontrollerat, vilket bildar svulster som påverkar kroppens funktion eller som sprider sig till andra kroppsdelar. Eftersom det är ett stort hälsoproblem så har forskare försökt ta fram nya angreppspunkter mot tumörer förutom metoder som cellgifter, cellskadande strålning eller kirurgi.

Immunförsvaret är ett antal olika celler och lösta ämnen som bekämpar främmande smittämnen i kroppen, där de vita blodkropparna hör till. Cancerceller, som är egentligen främmande celler, borde dödas av de vita blodkropparna men har en unik egenskap att undvika destruktion. Detta sker bland annat genom att vita blodkroppar som kämpat länge mot cancercellerna får signaler från omgivningen att trappa ner försvaret och blir ”utmattade”. Detta behövs för att undvika förödelse under vanliga infektioner men låter dessvärre cancerceller leva vidare.

Nya läkemedel har utvecklats som kan stoppa signalerna som gör vita blodkroppar ”utmattade” så att de kan vakna igen och förgöra cancern. Däremot är det oklart varför de inte fungerar mot alla sorters cancer. Man undrade om de vita blodkropparna har olika egenskaper beroende på vilken sorts cancer man har. Förutom ”utmattade” vita blodkroppar har kroppen även ”toxiska” blodkroppar som angriper cancercellerna direkt, ”minnes-blodkroppar” som lever i flera år och reagerar ifall cancern dyker upp igen eller ”naiva” blodkroppar som aldrig stött på cancerceller tidigare.

För att iaktta om sammansättningen av de vita blodkropparna skilde sig mellan cancrar så jämförde jag vita blodkroppar från lever-, lung- och bröstcancer. Jämförelsen byggde på vilka gener blodkropparna uttryckte. En gen är ett arvanslag och olika blodkroppar uttrycker olika gener för att erhålla sin funktion. Genom att studera generna kunde jag dela in blodkropparna i olika kategorier.

Resultaten visade att ”toxiska” blodkroppar och ”minnes-blodkroppar” fanns i ungefär lika stor utsträckning i alla cancertyperna. Lungcancer innehöll mer ”utmattade” celler och celler som kunde vara ”naiva”, medan bröstcancern hade flest celler som verkade vara ”på väg att utmattas”. Eftersom studierna hade använt olika metoder för att samla in informationen om generna var det svårt att förklara skillnaderna, men resultaten visade att en jämförelse mellan helt skilda studier var möjlig. Slutsatsen var att det finns både möjlighet för universella och cancer-specifika läkemedel för vita blodkroppar; framtida studier kommer förmodligen upptäcka fler skillnader när det finns mer data att jämföra.

1 Introduction

1.1 The Role of Tumour-Infiltrating Lymphocytes in Cancer

T cells are a core part of the adaptive immune response and include naïve lymphocytes (circulating antigen-specific T cells which have never met their antigen), activated lymphocytes (T cells which have encountered their antigen and acquired effector functions) and memory lymphocytes (T cells that have responded to the antigen, that persist long-term and that can proliferate vigorously following antigen re-encounter)[1-3]. These lymphocytes defend against pathogens but can also react against cancer cells. Tumour-infiltrating lymphocytes (TILs) are cancer-reactive lymphoid T cells that migrate to and surround solid tumours. They are part of the host response against cancer and include $CD4^+$ regulatory T cells (T_{regs}), $CD4^+$ T_H1-like T cells and cytotoxic $CD8^+$ T cells[4-6].

Conversely, the ability of tumour cells to evade immune destruction is a recent addition to the classic 'hallmarks of cancer' [7]. One way this is achieved is by inducing $CD8^+$ T cells in the tumour microenvironment into gradually losing their effector functions so that they become exhausted[8]. Exhausted $CD8^+$ T cells are a class of dysfunctional T cells, together with anergic and senescent T cells. They are generated when cytotoxic lymphocytes increase the expression of inhibitory receptors, also known as immune checkpoints, after chronic exposure to tumour antigens[8,9]. This growing family of receptors include Programmed cell death protein 1 (PD-1), cytotoxic T-lymphocyte-associated protein 4 (CTLA-4), Hepatitis A virus cellular receptor 2 (HAVCR2, also known as TIM-3), T cell immunoreceptor with Ig and ITIM domains (TIGIT) and Lymphocyte-activation gene 3 (LAG3)[5,10]. This impairs effector and memory T cell functions, enabling unrestricted cancer growth.

1.2 The Foundations of Immunotherapy

Over the past decades, cancer treatment has been transformed by the introduction of immune checkpoint inhibitors[11]. These drugs seek to relieve dysfunctional T cells from the inhibitory molecules' effects, enabling them to destroy the neoplastic cells[12,13]. However, not all types of cancer respond effectively to this form of immunotherapy[14,15]. One reason for the heterogenous response between individuals and cancer types may be the potential differences in the type of TILs residing in the tumour[5].

Previous studies have illustrated the correlation between certain TIL subpopulations within a patient and their response to therapy. For instance, a high proportion of exhausted $CD8^+$ TILs in a tumour is a poor prognostic marker for immune checkpoint blockade therapy[14]. Likewise, exhausted $CD8^+$ T cells have been divided into 'progenitor' and 'terminally' exhausted T cells, but only the former have shown to proliferate after anti-PD-1 blockade[12]. Because the presence or

absence of certain TIL population affects treatment efficacy, understanding which populations can be found in cancer can predict a patient's response to immune checkpoint inhibitors.

Knowing the importance of profiling TILs, previous studies have sought to define TIL subsets in melanoma[14,15], colorectal cancer[16], non-small-cell lung cancer[17,18] and ovarian cancer[19]. However, a direct qualitative and quantitative comparison of TIL subsets across cancers is lacking. This would not only shed light on why certain cancer types respond better to immunotherapy, but also suggest which treatments may be effective for particular cancer types.

1.3 Single-cell RNA Sequencing and its Applications

Classifying heterogenous cell populations into defined subsets requires observing their components at a cellular level. Single-cell RNA sequencing (scRNA-seq) is a powerful tool to study the gene expression in multiple cells at once and has already revolutionised the field of transcriptomics[20]. Whereas RNA microarrays and bulk RNA sequencing have long been used to sequence the transcriptome, they combine genetic material from a large cell population and cannot capture cell-specific gene expression differences[21]. By using a variety of methods to separate cells, scRNA-seq allows mRNA to be isolated from one cell at a time. The mRNA can then be reverse transcribed to cDNA, sequenced and matched against a database of human genes. scRNA-seq quantifies genes expressed by a single cell to provide useful information on its origin, behaviour and pathways of differentiation. Hence, scRNA-seq is ideal for characterising TILs across cancer types.

The establishment of scRNA-seq has spurred the development of effective toolkits for analysing gene expression data such as Seurat, an analysis package written for R Studio[22,23]. This tool can process, compare and visualise scRNA-seq data, but can also group cells into clusters using algorithms such as Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP)[24,25]. Briefly, these processes identify 'principal components': sets of genes commonly expressed together. Each component separates cells based on how much they express the genes in question. The components are then superimposed to distribute cells on a two-dimensional scatter plot. Cells with similar features crowd together and can be grouped into clusters based on the distance between them. Finally, the most differentially expressed genes in each cluster can be calculated. Therefore, the multitude of commands available to the user makes Seurat a compelling choice to analyse the scRNA-seq data.

1.4 Scope of the Present Comparison

In order to construct a valid yet achievable comparison between TIL subsets present in different cancer types, I initially decided to compare three cancer types using raw data from previously published studies. Three datasets were selected from the NCBI Gene Expression

Omnibus (GEO) database: a study by Chunhong Zheng et al. profiling TILs from mostly HBV-positive hepatocellular carcinoma (HCC); a study by Xinyi Guo et al. on TILs from non-small-cell lung carcinomas (NSCLC) and a study by Elham Azizi et al. on TILs from breast carcinoma (BC)[26-28]. The authors had sequenced TILs from tumour, adjacent normal tissue, blood and, in Azizi's case, even lymph nodes from a select group of patients.

Due to time constraints, I could only compare a fraction of the data and therefore chose to focus on the $CD8^+$ TILs as they are the main target of immune checkpoint blockade and the most important effector cells against the tumour[12,14,29,30]. Besides, many subsets with useful properties have already been identified for this cell type[1,2,6,15,30-38]. Another restriction that was imposed was to only analyse TILs sampled from the tumour. Although TILs in a patient's normal tissue and peripheral blood express different genes from TILs in the same patient's tumour[5], the tumour-resident TILs were deemed of greatest interest since they interact directly with the neoplastic cells.

1.5 Purpose

Thus, the aim of my thesis was to define and compare the genes expressed in $CD8^+$ TIL populations in HCC, NSCLC and BC tumours using scRNA-seq gene expression analysis.

1.6 Research Questions

The following questions were used to guide my study:

1. Which TIL populations can be found in the three cancer types, and what gene expression signatures define them?
2. Which TIL populations are common across the cancer types, and which are unique?

2 Method

2.1 Selection of the Datasets

The datasets from the original articles had been created by analysing lymphocytes in biopsies taken from tumour, adjacent normal tissue and peripheral blood from treatment-naïve cancer patients only[26-28]. Zheng et al. had sequenced 5,063 T cells from six HCC patients and Guo et al. had analysed 12,346 T cells from fourteen NSCLC patients[27]. Although Azizi et al. had studied the greatest number of cells (47,016 cells from eight BC patients), their scope had not been limited to T cells but had extended to all resident immune cells[26]. Finally, whereas Zheng and Guo had used fluorescence activated cell sorting (FACS) data to classify their cell datasets into $CD8^+$, $CD4^+$ and T_{reg} populations, Azizi had published his raw data without labelling which cells were $CD8^+$. Therefore, to ensure that only $CD8^+$ T cells were compared across the cancer types, the relevant cells would first have to be found in Azizi's data.

2.2 Comparison of the Technology used

Prior to sequencing, the mRNA in the original studies had been prepared using different scRNA-seq protocols. Guo et al. and Zheng et al. had employed the Smart-Seq2 method devised by Picelli et al.[39] whereas Azizi had utilized the inDrop method developed by Klein et al.[40] to construct the BC dataset. Smart-Seq2 is an older protocol which places the cells into wells for sequencing; inDrop fills droplets with one cell each and a "bead" to bind the mRNA. Briefly, the two methods differ in the "depth" of the sequencing (how many mRNA molecules are sequenced per cell) and the number of cells examined. Whereas inDrop is a *high-throughput* method sequencing thousands of mRNA molecules per cell from many cells, Smart-Seq2 is a *low-throughput* method sequencing potentially millions of mRNA molecules per cell but from fewer cells[41-43].

2.3 Import of Data into R Studio

Raw gene expression matrices for the three articles were downloaded from the NCBI Gene Expression Omnibus[44-46] into R Studio version 1.2.5019, running on R version 3.6.1[47,48], and transformed into Seurat files. Only TILs from tumour (not from peripheral blood and normal tissue) were retained. For Zheng and Guo's data, only $CD8^+$ cells as labelled by FACS were selected and analysed. For Azizi's data, this selection was delayed until the cells had been pre-processed.

Moreover, since the BC dataset had been created with three types of breast cancer, only data from the most common type of cancer in the dataset ($ER^+/PR^-/HER2^-$) were included, reducing the dataset to 8187 white blood cells.

2.4 Pre-processing of Data according to Gene Expression

The number of genes expressed and the number of mRNA molecules sequenced per cell was determined. In order to exclude dying cells or empty droplets (reads with low gene counts) and so-called "doublets" (reads from droplets containing two cells)[49], the cells were ranked by the number of genes expressed per cell. The bottom and top 2.5% of the data were removed so that only cells expressing 2300 to 6080 genes for Zheng's data, 1384 to 4987 genes for Guo's data and 21 to 2500 genes for Azizi's data remained. However, given that Azizi's data was positively skewed and that 21 genes were unlikely to be enough to define the immune phenotype, the lower limit was increased to 100 genes (the 30.4^{th} percentile).

The percentage of expressed genes that were mitochondrial was examined because high percentages of mitochondrial gene expression are often a sign of cell membrane rupture in dying cells[50]. This was not possible for Zheng and Guo's studies, who had not published the mitochondrial gene counts[27,28]. In Azizi's dataset, cells expressing greater than or equal to 17.5% mitochondrial genes were discarded. This criterion removed 754 cells, compared to 2710 cells removed in the previous step.

2.5 Selection of CD8+ cells from the Breast Cancer Dataset

Seurat was employed to identify and extract the $CD8^+$ cells from Azizi et al.'s BC dataset as the authors had not used FACS to identify their TIL populations. After normalising and logarithmically scaling the data, a principal component analysis (PCA) was carried out. The principal components to be used in the clustering were selected by graphing their standard deviation onto the *y*-axis of an elbow plot, a type of scatter plot with each component on the *x*-axis sorted by the percentage of the gene expression variance that it explained. The first fifteen were used to run a Uniform Manifold Approximation and Projection (UMAP) analysis, which clustered the TILs into eight populations at a resolution of 0.1. Cluster 2, assumed to contain the $CD8^+$ TILs, was extracted from the subset and used for further analysis.

The final BC and HCC datasets contained 398 and 735 $CD8^+$ TILs respectively, compared to 2070 in the processed NSCLC dataset. Hence only a sample of 735 cells from the NSCLC dataset was used for further analysis. This was done so that the dataset would not skew the TIL clusters in the integration.

2.6 Comparison of BC Dataset with Integrated HCC and NSCLC Dataset

The HCC and the NSCLC datasets were integrated at first to assess the extent to which the BC dataset could be mapped onto the remaining data. After the integration, the combined HCC-NSCLC dataset was scaled, mapped onto a UMAP projection and clustered into eight clusters using the 15 strongest PCs and a resolution of 0.3. To test the correspondence between the combined dataset and the BC TILs, the same PCs were used to classify the BC TILs. The result of the comparison is described in section 4.2 of the results and suggested that an integration of all three datasets was possible.

2.7 Integration of HCC, NSCLC and BC Datasets

Next, the datasets were combined, scaled and subjected to PCA. Using eleven PCs and a resolution set to 0.4, eight unsupervised clusters were created. The most differentially expressed genes for each cluster were calculated using the natural logarithm of the average fold-change (ratio) between the number of gene reads for the cluster in question and the number of gene reads in all other clusters. This measure is hereafter referred to as the ln(FC) for brevity. A *p*-value calculated by the Wilcoxon Rank Sum Test with Bonferroni correction applied was also computed. The Bonferroni correction is a formula for adjusting the *p*-value, advised in cases where several null hypotheses are being tested at once; it is an appropriate test to use when comparing one cluster to numerous other clusters simultaneously[51].

2.8 Pseudotemporal analysis of the integrated HCC, NSCLC and BC Dataset

With the integrated dataset split into eight distinct clusters, the dataset was subjected to temporal analysis using Monocle version 2, a package for R Studio written by Trapnell et al[52]. The Seurat object was transformed into a Monocle CellDataSet, and the default ordering algorithm calculated the inferred order in which the cells had emerged.

3 Ethical considerations

Zheng, Guo's and Azizi's studies had been respectively approved by the Ethics Committee of Beijing Shijitan Hospital, the Ethics Committee of Peking University and the Institutional Review Board at the Memorial Sloan Kettering Cancer Center[26-28]. According to the authors, all patients had given their informed consent to participate in their studies. The published datasets contain no personal information which can be traced back to the original patients. Moreover, since the data is freely accessible on the internet, we conclude that the potential benefits of guiding immunotherapy outweigh any harm that may be done in conducting this comparative study.

4 Results

4.1 Summary of the Processed Data

Prior to filtering the datasets, several violin plots were used to visualise the number of genes expressed in each cell (Figure 1, Figure 2 and Figure 3) and the number of mRNA molecules sequenced (Figure 4, Figure 5 and Figure 6). Figure 6 shows a positive skew in the distribution of mRNA molecules per cell for the BC dataset. Since the inDrop technology employs "beads" to capture mRNA molecules, accidentally adding two beads per droplet leads to "cells" with a misleadingly high mRNA count. Hence, after observing the graph, 136 cells containing more than 2500 mRNA molecules were eliminated from the BC dataset. Figure 7 shows the original distribution of mitochondrial genes in the same dataset used to justify the 17.5% cut-off. This cut-off was also chosen because higher cut-offs made Seurat use these contaminating genes to cluster cells. After excluding the unwanted cells, the datasets contained 737, 2071 and 4587 cells for HCC, NSCLC and BC TILs respectively.

The elbow plot for the PCs used to identify the $CD8^+$ TILs in the BC dataset is shown in Figure 8: I decided to use the first fifteen as the standard deviation for the next components was deemed too low to segregate the data. The clusters from this dataset are shown in Figure 9. Figure 10 gives the expression level of CD4 and CD8 for each cluster. Cluster 2 was assumed to contain the $CD8^+$ TILs based on its high expression of CD8 and low expression of CD4. Cluster 2 also expressed high levels of $CD8^+$ T cell markers such as NKG7 (Natural Killer Cell Granule Protein 7), FGFBP2 (Fibroblast Growth Factor Binding Protein 2), PRF1 (Perforin-1) and GZMB

(Granzyme B) compared to other cells ($p < 10^{-240}$ by Wilcoxon Rank Sum test), consistent with their assumed identity[6,15]. The 4189 remaining cells were removed so that only 398 cells remained. Some characteristics of the final data is summarised in Table 1 and Table 2 on page 35.

4.2 Clusters in the Integrated HCC-NSCLC Dataset are Conserved in the BC Dataset

Prior to integrating the three datasets together I wanted to verify that the BC dataset, which had been created using a different technology, would be compatible for integration with the HCC and the NSCLC datasets. After merging and clustering the HCC and the NSCLC dataset, the 398 BC TILs were sorted into the eight clusters shown in Figure 11. 246 TILs (61.8%) were assigned to cluster 3; 137 (34.4%) were assigned to cluster 6; 12 (3.0%) were assigned to cluster 0 while 3 (0.8%) were assigned to cluster 5. Some known markers for cluster 3 and cluster 6 are presented in Table 3 and Table 4. For each gene in the table, the average ln(FC) is given. A larger number indicates that this gene was more expressed in this cluster compared to other clusters. The results show that cluster 3, where most of the BC TILs was grouped, showed a cytotoxic T cell signature whereas cluster 6 expressed several naïve T cell signature genes.

4.3 Eleven PCs Define Eight Clusters in the Integrated HCC, NSCLC and BC Dataset

Prior to clustering the integrated dataset, an elbow plot was used to visualise the standard deviations for the suggested PCs (Figure 12). As the standard deviation was constant after the 11th PC onwards, only the first 11 were used for dimensional reduction. PC #2 was particularly interesting as it selected positively for exhaustion-related genes such as HAVCR2, TIGIT and ENTPD1 (Ectonucleoside Triphosphate Diphosphohydrolase 1). The first two genes code for known immune checkpoints whereas the third is an enzyme capable of producing the immunosuppressant adenosine[6]. On the contrary, nine of the ten most negatively selected genes were ribosomal genes of unknown significance. However, the PC was nevertheless chosen. PC #4 was also noteworthy as it segregated naïve CD8$^+$ TIL markers such as CCR7 and IL7R from memory TIL markers FGFBP2 and CX3CR1[1,34]. A more comprehensive list of genes generated by the PCA can be found in Table 9 in the appendices.

Figure 13 shows the integrated dataset on a UMAP plot with cells coloured by their original dataset; Figure 14 shows the same plot after clustering. The proportion of cells that were assigned to each cluster is shown in Table 5 on page 36. To quantify the most differentially expressed genes in each cluster, the ln(FC) was calculated for a selection of markers. The most differentially expressed genes for each are presented as a heatmap on Figure 15 and a second heatmap with only known TIL markers is given in Figure 16. Cells of a cluster where a gene was highly expressed are shown in yellow.

4.4 Cluster 0, 3 and 6 TILs are Common to HCC, NSCLC and BC

As shown in Table 5, cluster 0 was the most prominent type of TIL in the HCC and the NSCLC datasets (accounting for 51.3% and 25.9% respectively), and the second most common type in the BC dataset. This cluster had a significant higher expression of effector/cytotoxic genes such as GZMK, CCL4, CCL3, CCL20, FOS and JUNB, even if other cytotoxic genes such as S100A4, GZMB or GNLY were downregulated ($p < 0.01$, see Table 6 on page 37). The cluster also had a significantly higher expression of IFNG ($p = 0.008$) and a slightly increased expression of TNF ($p = 0.11$), suggesting cells in this cluster were capable of cytokine production. The cluster's effector capability is confirmed by its upregulation of SLAMF7 (average ln(FC) 0.24, $p = 1.19 \times 10^{-11}$). Although upregulating the naïve marker SELL ($p = 5.40 \times 10^{-71}$), the average ln(FC) for this gene was only 0.06. The results of the pseudotemporal analysis indicate that this cell type persisted throughout pseudotime and was neither an early nor a late TIL population (Figure 20).

Another cluster conserved across cancer types was cluster 3, which made up between one-tenth and one-fourteenth of all TILs in the datasets. Cluster 3 expressed effector/memory genes such as FGFBP2, CX3CR1, ITGAM, KLRD1, KLRG1 ($p < 0.0003$), as well as cytotoxic-related genes PRF1, NKG7 and CST7. The cells in the cluster downregulated naïve genes ($p < 0.004$), resident memory genes such as CD44, CD69 and CXCR3 and exhaustion-related genes ($p < 0.0001$, except for TIGIT which had $p = 0.13$), see Table 7.

Cluster 6 was also well-conserved across cancer types. This cluster upregulated genes involved in proliferation such as MKI67 (average ln(FC) 1.53, $p = 7.35 \times 10^{-15}$) and tubulin genes TBA1B/TUBB ($p < 3.82 \times 10^{-17}$). Yet this cluster also had a significantly higher expression for 'exhausted' genes such as ENTPD1 and TIGIT ($p < 0.02$).

4.5 Cluster 2, 4 and 7 TILs are More Common in NSCLC Tissue

Cluster 2 varied broadly in proportions between the cancer types. It accounted for 3.8% of BC TILs but 18.6% of NSCLC TILs where it was the second largest cluster. As shown in Figure 16 and Table 8, cluster 2 principally expressed exhaustion-related genes such as ENTPD1, HAVCR2, TIGIT, TOX and PDCD1 ($p < 0.0003$), but also expressed several interferon-inducible genes such as IFI44L, IFI44 and IFI35 ($p < 0.002$). Genes associated with naïve $CD8^+$ cells such as TCF7, SELL and IL7R were noticeably downregulated ($p < 0.003$), see Table 8. Cluster 2 also significantly downregulated 'effector' genes such as MYB, MYC, KLF2 and FOS ($p < 0.007$), consistent with exhausted $CD8^+$ T cells. KLF3, JUNB, FOSB, KLF13 and JAK3 were also downregulated, albeit not significantly. Moreover, the pseudotemporal analysis suggested that cluster 2 was late to develop (Figure 20) and that exhausted markers such as ENTPD1 became more expressed in the integrated dataset over pseudotime (Figure 21).

Another cluster that was frequent in the NSCLC dataset was cluster 4, containing every seventh NSCLC TIL. As shown in Figure 16 and Figure 17, both cluster 4 and cluster 5 had a significant upregulation of the naïve T cell markers IL7R and CCR7. This made both clusters candidates for naïve T cells. Cluster 4 had a strong expression of IL7R (average ln(FC) 1.22, $p = 1.44 \times 10^{-16}$), yet 43 of the 50 most differentially expressed genes were ribosomal genes from the RPL and the RPS families as hinted at on the *y*-axis in Figure 15. Compared to cluster 5, cluster 4 had a higher expression of naïve TIL markers such as SELL, LEF1, IL7R, CCR7 and TCF7 (ln(FC) > 0.28) however in none of these cases was the difference significant ($p = 1.00$). Compared to all other clusters, cluster 4 significantly downregulated exhausted, cytotoxic and effector markers (such as HAVCR2, PDCD1, TIGIT, ENTPD1, GZMB, GZMA, PRF1, KLRG1) and was exclusively found in the beginning of the pseudotemporal analysis (Figure 20), when IL7R expression was high (Figure 21).

Cluster 7 was ten times as common in the NSCLC dataset compared to either the BC or the HCC datasets. It strongly expressed the canonical B-cell marker MS4A1 (CD20) with average ln(FC) 2.00, $p = 1.50 \times 10^{-3}$. However, it did not upregulate other typical B-cell markers such as CD19 or CD40 (average ln(FC) –0.001 and –0.02 respectively, $p = 1.00$ in both cases). Moreover, there was no significant difference in CD3 and CD8 expression compared to the other clusters ($p = 1.00$). Hence, it was theorised that this was a unique $CD20^{+}CD8^{+}$ T cell population.

4.6 Cluster 5 TILs are More Common in HCC Tissue

Like cluster 4, cluster 5 expressed some naïve-related genes such as IL7R (average ln(FC) 0.49, $p = 2.77 \times 10^{-19}$) and CCR7 (0.35, $p = 2.15 \times 10^{-11}$). Other naïve genes were downregulated and/or not significant, such as SELL, (–0.43, $p = 1.00$). Instead, cluster 5 had a higher expression of genes associated with mucosal-associated invariant T (MAIT) cells, such as KLRB1, RORC, CCR6, ZBTB16, IL18R1 and SLC4A10[28,53]. Compared to cluster 4, all genes had a ln(FC) greater than 0.60 except for IL18R1 which had 0.15, and three of these six genes (KLRB1, RORC and SLC4A10) were significantly upregulated with $p < 0.00002$. When compared to all clusters, even CCR6 and ZBTB16 were found to be significantly upregulated ($p < 0.008$). This cluster, which vanished halfway through pseudotime (Figure 20), was the second largest cluster in the HCC dataset.

4.7 Cluster 1 TILs are More Common in BC Tissue

Cluster 1 was the largest cluster in the BC dataset, accounting for 40.2% of the BC TILs. Likewise, it was also the second largest cluster in the NSCLC dataset, where it contained 20.7% of TILs. This cluster significantly upregulated the tissue-resident memory T cell (T_{RM}) marker ITGAE (average ln(FC) 0.34, $p = 6.98 \times 10^{-11}$ and visualised in Figure 16)[37,38], but also expressed the naïve

marker IL7R (ln (FC) 0.30, p < 0.0003). The most differentially expressed genes included ZNF683, coding for the transcription factor Homolog Of Blimp-1 In T Cells, or HOBIT for short (average ln(FC) 0.74, $p = 4.39 \times 10^{-19}$, see Figure 15).

4.8 Pseudotemporal Analysis Distinguishes Cluster 3, Cluster 4, Cluster 5 and Cluster 7 TILs from Cluster 2 and Cluster 6 TILs

Using the Monocle toolkit for R Studio, the integrated dataset was plotted across two principal components and visualised in "pseudotime", summarised by Kumar et al. as "an artificial ordering of cells based upon a statistically inferred trajectory often interpreted as time." [54] The toolkit suggested a single pathway of development ending in two branches (Figure 18), which could be seen in all datasets regardless of technology (Figure 19). This suggests that the cells could be mapped onto the same trajectory regardless of how their transcriptome had been sequenced. By filtering the trajectory by cluster, it appears that cluster 3, 4, 5 and 7 were early populations whereas cluster 2 and 6 emerged later (Figure 20). Cluster 0 and 1 were present throughout development. The results also suggest that cluster 6 may consist of two diverging populations.

A few differentially expressed genes have also been plotted across pseudotime in Figure 21. The results indicate that CX3CR1, GZMK and IL7R expression declined with pseudotime, being replaced with ENTPD1 and later, MKI67. MS4A1 rose early on but declined soon after.

5 Discussion

5.1 Integration of Datasets Identifies Conserved $CD8^+$ TIL Populations

Immunotherapy has had varying degrees of success in treating cancer and since $CD8^+$ TILs in tumour are currently the target cells of this treatment, understanding which $CD8^+$ TIL populations are consistently found and which populations are specific to certain tumours is of great use. This study has integrated $CD8^+$ TIL gene expression data obtained through single-cell RNA sequencing from hepatocellular carcinoma, non-small cell lung carcinoma and breast carcinoma with the aim of defining common TIL populations. Although the BC dataset had been created using a different technology from the HCC/NSCLC datasets, a comparison between the BC dataset and an integrated HCC-NSCLC dataset illustrated that the BC $CD8^+$ TILs shared properties with the HCC-NSCLC cells such as cytotoxicity or naïve T cell gene expression, validating an integration of all three datasets.

5.2 Clusters 0 and 3 are Conserved Across Cancer Types and Correlate with Established TIL Phenotypes

The eight TIL populations may can be grouped into three categories: well-defined common clusters, well-defined cancer-specific clusters and doubtful clusters. Clusters 0 and 3 were

conserved across the cancer types examined and demonstrated a clear 'effector' and an 'effector-memory' phenotype.

Cluster 0 was amongst the largest TIL subsets in the integrated dataset and expressed several effector or cytotoxic genes. This makes the cluster alike the $CCR7^{-}PTPRC^{+}CD27^{low}CD28^{low}KLRG1^{+}$ "long-lived effector-type T cells" (T_{EMRA}) cells reviewed by Braun et al[55] or alike the $SELL^{-}Slamf7^{hi}CX3CR1^{-}PD\text{-}1^{-}CD8^{+}$ cluster described as "memory-precursor-like TILs" by Kurtulus et al[30]. Kurtulus' cells produced various cytokines, underwent a TCF7-dependent proliferation after combined HAVCR2/PD-1 blockade in mice and were associated with a good prognosis. Thus, given cluster 0's omnipresence across all cancer types and their ability to respond to immunotherapy, further studies comparing cluster 0 with Kurtulus' population would be valuable. Studies to determine why this 'effector' cluster was enriched in the HCC dataset would also be of interest.

A less common but nevertheless conserved cluster was cluster 3, which expressed effector and memory genes. This cluster phenotypically resembles the "memory T cells with effector function" described in mice by Böttcher et al[34]. The pseudotemporal analysis found that this 'effector-memory' cluster was more prevalent in early cells (Figure 20) and that CX3CR1 expression decreased with pseudotime (Figure 21): an unusual pattern for memory T cells. On the contrary, Böttcher claims that "[v]irus-specific $CX3CR1^{+}$ memory $CD8^{+}$ T cells are scarce during chronic infection in humans and mice but increase when infection is controlled spontaneously or by therapeutic intervention." Indeed, a study by Yan et al. found that CX3CR1 identified PD-1 therapy-responsive $CD8^{+}$ T cells that could fight neoplastic cells following chemotherapy combined with PD-1 blockade[56]. Both studies described the expression of GZMB in this cell line, which was indeed expressed in cluster 3 (average ln(FC) 0.35, $p = 2.58 \times 10^{-12}$). Having outlined the potential of memory T cells in successful immunotherapy, finding this 'effector-memory' cluster across all cancer types is thus important.

5.3 Clusters 1, 2 and 5 are Cancer-Specific TIL Subsets

Cluster 1, 2 and 5 were also well-defined clusters but were more specific to certain cancer types. Cluster 2 was specific to NSCLC tissue, where it was two times more common compared to HCC tissue and five times more common compared to BC tissue. This cluster closely resembled exhausted $CD8^{+}$ T cells, previously described as antigen-specific T cells that have lost effector functions such as IL-2 production, cytotoxicity and proliferation[3]. Moreover, its late emergence in the pseudotemporal analysis is consistent with how exhausted T cells develop by gradually losing their effector functions after chronic antigenic exposure[8]. The expression of interferon-inducible genes is unexpected and hints at an ongoing interferon-stimulation along with the inhibition. Why

this cluster was enriched in NSCLC cannot be answered, as it contrasts with a review concluding that HCC had more exhausted $CD8^+$ T cells than NSCLC[5]. This highlights the possibility that a shortage of cells made the clustering more sensitive to parameters such as the numbers of PCs and the resolution.

Cluster 1 was twice as common in BC tumours as in NSCLC tumours and six times more common when compared to HCC tumours. This cluster was marked by its expression of ITGAE and ZNF683 and closely resembles a tissue-resident memory T cell subset found by Guo et al., who labelled it "pre-exhausted" [27]. Cluster 1 was placed adjacent to the 'exhausted' cluster 2 on the UMAP plot in Figure 14, but did not flag positive for the typical exhaustion-genes in the heatmap in Figure 15. This evidence suggests that cluster 1 can be interpreted as a 'pre-exhausted' population. Like the 'effector' cluster 0, cluster 1 was found throughout pseudotime (Figure 20). The decline in the expression of the 'effector' marker GZMK across the whole dataset paralleled the increased expression of the 'exhausted' marker ENTPD1 (Figure 21), implying a continuous turnover from 'effector' to 'pre-exhausted' TILs. In the original study, Guo et al. claimed that a high ratio of "pre-exhausted" to 'exhausted' was associated with better prognosis for lung adenocarcinoma[27]. Given that this ratio varied twenty-fold from 0.6 in HCC to 10.7 in the BC dataset, further studies could elucidate the effect of the 'pre-exhausted' cluster on survival.

The 'MAIT' cell cluster 5 was the third largest cluster in the HCC dataset, where it was four times more common than the NSCLC dataset (mirroring previous studies[5]) and six times were common than the BC dataset. This accurately reflects how MAIT cells are enriched in HCC compared to other carcinomas, but also reflects their how they are scarce when compared to normal liver tissue where they make up between 20% and 40% of T lymphocytes[57,58]. MAIT cells may be useful against cancer as they express cytotoxic effector molecules[59]. On the other hand, they have also been shown to promote tumour growth, for example by secreting IL17A which promotes angiogenesis[58,59]. Hence, it is unclear how this 'MAIT' cluster could be exploited in immunotherapy.

Although being rather cancer-specific, discovering the 'pre-exhausted' cluster 1 and the 'MAIT' cluster 5 in smaller proportions in other datasets was noteworthy because it challenges the notion that TILs are completely unique to cancer types. Concerning cluster 1, Guo et al. had defined their subset of "pre-exhausted" $ZNF683^+CD8^+$ tissue-resident memory T cells from the NSCLC data, but this cluster was in fact twice as prominent in the BC dataset. Secondly, the 'MAIT' cluster had been found by Zheng in HCC tissue and to a smaller extent by Guo in NSCLC tissue but had not been mentioned in the BC study[26-28]. This highlights the clinical possibility of

using a cluster-specific immunotherapy drug against multiple cancers or using multiple cluster-specific drugs against the same cancer for an effective treatment.

5.4 Clusters 4, 6 and 7 Define Unfamiliar Immune Phenotypes

On the other hand, not all clusters were straightforward to define. One cluster that proved challenging to identify was cluster 6, a relatively conserved cluster across the cancer types. The co-expression of exhausted and proliferating genes suggested that cluster 6 may contain the 'progenitor exhausted' $CD8^+$ TILs described by Miller et al., who used scRNA-seq to distinguish these from 'terminally exhausted' $CD8^+$ TILs in an experimental chronic viral infection model. The 'progenitor exhausted' TILs demonstrated "improved proliferative capacity, survival and ability to differentiate into cytotoxic terminally exhausted $CD8^+$ T cells" [12]. This proliferating group was also found by Li et al. who used scRNA-seq to sequence TILs from melanoma patients and who labelled these cells 'transitional' on a spectrum between "cytotoxic" and "dysfunctional" [60]. As described elsewhere, 'progenitor' exhausted cells are associated with response to immunotherapy[12,14], thus discovering a similar cluster conserved across the cancer datasets is of great significance.

However, challenges to treating cluster 6 as a 'progenitor exhausted' population include downregulation of the TCF7 gene (–0.31, $p < 0.0002$), which was previously linked to the 'progenitor' blockade-responding $CD8^+$ TILs[12,14]. Instead, the cluster upregulated the 'exhausted' transcription factor BATF (0.43, $p < 0.00007$), suggesting that cluster 6 may be closer to the 'terminally exhausted' state than to a 'progenitor exhausted' state. The pseudotemporal analysis is difficult to interpret: it is hard to accept that MKI67 was only expressed at the end of the development (Figure 21). Moreover, if this cluster were indeed to represent the 'progenitor exhausted' T cells, then one may ask why the proportion of 'progenitor exhausted' TILs was almost identical across cancer types whereas the proportion of 'exhausted' cells varied. Indeed, the ratio of 'exhausted' to 'progenitor exhausted' varied from 1.50 in the BC dataset, to 1.90 for the HCC dataset and 5.71 in the NSCLC dataset. The reasons for this are not clear.

The '$CD20^+$' Cluster 7 was an unusual find but resembled a $CD20^+CD8^+$ T cell population reported by Schuh et al. in thymus, bone marrow and secondary lymphatic organs, and making up 3–5% of circulating human T cells[61]. This cluster was especially frequent in the NSCLC population, where it made up 3.7% of the TILs compared to 0.3% in both the BC and HCC datasets. Since the authors claimed that the $CD20^+CD8^+$ cells produced more IL-4, IL-17, IFN-γ and TNF-α than $CD20^-$ T cells, this suggests these "double-positive" $CD8^+$ TILs as targets for immunotherapy in the context of NSCLC.

Cluster 4, an early population assumed to contain 'naïve-like' cells, was enriched in the NSCLC dataset, expressed many ribosomal genes but was difficult to distinguish from the 'MAIT' cluster 5. Although the earlier HCC-NSCLC integration found a naïve cluster (Table 4

Table 4), it is unclear whether cluster 4 is indeed a naïve cluster or an artificial cluster created by Seurat grouping together cells expressing ribosomal genes. The identity of this cluster remains thus uncertain.

5.5 Pseudotemporal Analysis is Consistent with Cluster Identity

As briefly mentioned above, six of the eight clusters were assigned to extremes of the inferred evolutionary pathway by the Monocle toolkit. This distinguished the early TIL populations ('effector-memory', 'naïve-like', 'MAIT' and 'CD20^{+}') from the two late populations ('exhausted' and the ambiguous 'proliferating' cluster 6). For all clusters except cluster 6, the temporal findings are consistent with the accepted pathway of differentiation of the clusters' proposed TIL identity, and the gene expression signature's evolution from naïve to effector to exhausted (Figure 21) recapitulates the normal journey for TILs.

Recently, much work has been done to reverse T cell exhaustion using immune checkpoint blockade[11,30,62]. However, the results suggest that future immunotherapies could instead attempt to prevent transformation to the 'exhausted' state, since this appeared inevitable for all populations present at an early stage except for the 'effector' and the 'pre-exhausted' cell states.

Nevertheless, the pseudotemporal analysis should be interpreted with caution as it is heavily variable. For example, Zheng's HCC study also considered the CX3CR1^{+} 'effector-memory' cluster 3 an early population[28], whereas Guo's NSCLC analysis placed them on a late branch opposite to the exhausted TILs[27]. This again highlights the consequences of having too few cells to compare against.

5.6 Limitations of the Study and Weaknesses of the Analysis

Before concluding, a few limitations in this study merit further discussion. Firstly, the number of CD8^{+} TILs used in this study was low compared to previous publications. Although expensive to achieve, future studies focusing on CD8^{+} TILs could prepare more cells before removing CD8^{-} cells using FACS. This would make it possible to determine whether difficult-to-define clusters, such as the 'naïve-like' cluster 4, were in fact mixed pools of different immune phenotypes.

Secondly, the authors had processed their datasets with different criteria before publishing them: even after filtering the cells, Azizi's TILs expressed an *average* of 11.2% mitochondrial genes, which was still higher than Guo's 10% threshold[27]. Hence, when more datasets become available future studies should consider the quality of the datasets to be compared.

Most importantly, this comparative study was inherently limited by using datasets created with different technologies and different sequencing depths. By sequencing at a lower depth, inDrop data offers less "technical noise" from housekeeping genes compared to Smart-Seq2[41,42]. This is also due to inDrop using unique molecular identifiers (UMIs) which are DNA fragments that "barcode" original mRNA molecules to differentiate them from unwanted copies made during the conversion of mRNA to cDNA[41]. On the contrary, SmartSeq2 offers greater sensitivity in finding weakly expressed genes[41]. Although the integration algorithm performed well despite the BC dataset expressing on average five times fewer genes per cell, this precluded comparing gene expression in the same cluster across cancer types after integration. Hence it was not possible to investigate if different transcription factors could explain the differences in cluster sizes between cancer types. Although this study was helpful in testing the feasibility of a cross-technology comparison, only comparisons between the same technology can uncover potential mechanisms behind these patterns.

Finally, this study also had weaknesses that could have been corrected by further analysis given more time. For example, the inclusion of all proliferating cells into one cluster may have been caused by not regressing out cell cycle markers when scaling the data. Seurat could have "normalised" cells by their progression in the cell cycle to minimise interference with the clustering. Another step that could have been implemented was to use databases of TIL markers to numerically score the signature of a cell population, as opposed to manually checking for known markers. These databases, available from sites such as Gene Ontology, could have helped to systematically identify clusters.

6 Conclusions

The prospect of treating cancer with immunotherapy is an emerging alternative to chemotherapy or radiotherapy, but more work is needed to understand why the efficacy varies between patients and cancer type. This study has compared $CD8^+$ tumour-infiltrating lymphocytes from three carcinomas and identified eight conserved populations. 'Effector', 'effector-memory' and proliferating TIL populations were found in all three carcinomas. 'Pre-exhausted' cells defined in NSCLC were shown to be more common in BC; conversely 'exhausted', 'naïve-like' and $CD20^+CD8^+$ TILs were more common in NSCLC. 'MAIT' cells were particularly enriched in HCC samples. Most of these populations possess anti-neoplastic functions that could be used in cancer therapy. Although the results' value is compromised by comparing cells sequenced with different technologies, the comparison nevertheless provides a useful framework for future studies when scRNA-seq will be more accessible.

7 Acknowledgements

I would like to thank Domenico Lo Tartaro for teaching me how to use R Studio and Seurat, as well as for inspiring the innovative aim for this study. I would also like to thank Professor Andrea Cossarizza and his team for welcoming me to the University of Modena and Reggio Emilia. Lastly, I would like to thank Professor Cristian Bellodi for his correspondence and assistance in refining the work.

8 Bibliography

1. Samji T, Khanna KM. Understanding memory CD8+ T cells. Immunol Lett. 2017;185:32-9.
2. Martin MD, Badovinac VP. Defining Memory CD8 T Cell. Front Immunol. 2018;9(2692).
3. Wherry EJ, Ha S-J, Kaech SM, Haining WN, Sarkar S, Kalia V, et al. Molecular Signature of CD8+ T Cell Exhaustion during Chronic Viral Infection. Immunity. 2007;27(4):670-84.
4. Lee N, Zakka LR, Mihm MC, Schatton T. Tumour-infiltrating lymphocytes in melanoma prognosis and cancer immunotherapy. Pathology. 2016;48(2):177-87.
5. Zhang L, Zhang Z. Recharacterizing Tumor-Infiltrating Lymphocytes by Single-Cell RNA Sequencing. Cancer Immunol Res. 2019;7(7):1040-6.
6. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. Nat Med. 2019;25(8):1251-9.
7. Hanahan D, Weinberg Robert A. Hallmarks of Cancer: The Next Generation. Cell. 2011;144(5):646-74.
8. Wherry EJ, Kurachi M. Molecular and cellular insights into T cell exhaustion. Nat Rev Immunol. 2015;15(8):486-99.
9. Alfei F, Kanev K, Hofmann M, Wu M, Ghoneim HE, Roelli P, et al. TOX reinforces the phenotype and longevity of exhausted T cells in chronic viral infection. Nature. 2019;571(7764):265-9.
10. Sukari A, Nagasaka M, Al-Hadidi A, Lum LG. Cancer Immunology and Immunotherapy. Anticancer Res. 2016;36(11):5593-606.
11. Pardoll DM. The blockade of immune checkpoints in cancer immunotherapy. Nat Rev Cancer. 2012;12(4):252-64.
12. Miller BC, Sen DR, Al Abosy R, Bi K, Virkud YV, LaFleur MW, et al. Subsets of exhausted CD8(+) T cells differentially mediate tumor control and respond to checkpoint blockade. Nat Immunol. 2019;20(3):326-36.
13. Haanen JB, Robert C. Immune Checkpoint Inhibitors. Prog Tumor Res. 2015;42:55-66.
14. Sade-Feldman M, Yizhak K, Bjorgaard SL, Ray JP, de Boer CG, Jenkins RW, et al. Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. Cell. 2018;175(4):998-1013 e20.
15. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352(6282):189.
16. Zhang L, Yu X, Zheng L, Zhang Y, Li Y, Fang Q, et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature. 2018;564(7735):268-72.
17. Thommen DS, Koelzer VH, Herzig P, Roller A, Trefny M, Dimeloe S, et al. A transcriptionally and functionally distinct PD-1(+) CD8(+) T cell pool with predictive

potential in non-small-cell lung cancer treated with PD-1 blockade. Nat Med. 2018;24(7):994-1004.
18. Clarke J, Panwar B, Madrigal A, Singh D, Gujar R, Wood O, et al. Single-cell transcriptomic analysis of tissue-resident memory T cells in human lung cancer. The Journal of experimental medicine. 2019;216(9):2128-49.
19. Workel HH, Lubbers JM, Arnold R, Prins TM, van der Vlies P, de Lange K, et al. A Transcriptionally Distinct CXCL13(+)CD103(+)CD8(+) T-cell Population Is Associated with B-cell Recruitment and Neoantigen Load in Human Cancer. Cancer Immunol Res. 2019;7(5):784-96.
20. Picelli S. Single-cell RNA-sequencing: The future of genome biology is now. RNA Biol. 2017;14(5):637-50.
21. Kolodziejczyk Aleksandra A, Kim JK, Svensson V, Marioni John C, Teichmann Sarah A. The Technology and Biology of Single-Cell RNA Sequencing. Mol Cell. 2015;58(4):610-20.
22. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411-20.
23. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902 e21.
24. Ringnér M. What is principal component analysis? Nat Biotechnol. 2008;26(3):303-4.
25. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38-44.
26. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. Cell. 2018;174(5):1293-308 e36.
27. Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med. 2018;24(7):978-85.
28. Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, et al. Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing. Cell. 2017;169(7):1342-56.e16.
29. Oja AE, Piet B, van der Zwan D, Blaauwgeers H, Mensink M, de Kivit S, et al. Functional Heterogeneity of CD4+ Tumor-Infiltrating Lymphocytes With a Resident Memory Phenotype in NSCLC. 2018;9(2654).
30. Kurtulus S, Madi A, Escobar G, Klapholz M, Nyman J, Christian E, et al. Checkpoint Blockade Immunotherapy Induces Dynamic Changes in PD-1−CD8+ Tumor-Infiltrating T Cells. Immunity. 2019;50(1):181-94.e6.
31. Mach N, Gao Y, Lemonnier G, Lecardonnel J, Oswald IP, Estellé J, et al. The peripheral blood transcriptome reflects variations in immunity traits in swine: towards the identification of biomarkers. BMC Genomics. 2013;14(1):894.
32. Hervas-Stubbs S, Riezu-Boj J-I, Gonzalez I, Mancheño U, Dubrot J, Azpilicueta A, et al. Effects of IFN-α as a signal-3 cytokine on human naïve and antigen-experienced CD8+ T cells. Eur J Immunol. 2010;40(12):3389-402.
33. Hu G, Chen J. A genome-wide regulatory network identifies key transcription factors for memory CD8+ T-cell development. Nat Commun. 2013;4(1):2830.
34. Böttcher JP, Beyer M, Meissner F, Abdullah Z, Sander J, Höchst B, et al. Functional classification of memory CD8+ T cells by CX3CR1 expression. Nat Commun. 2015;6(1):8306.
35. Kaech SM, Cui W. Transcriptional control of effector and memory CD8+ T cell differentiation. Nat Rev Immunol. 2012;12(11):749-61.

36. Dotiwala F, Mulik S, Polidoro RB, Ansara JA, Burleigh BA, Walch M, et al. Killer lymphocytes use granulysin, perforin and granzymes to kill intracellular parasites. Nat Med. 2016;22(2):210-6.
37. Hu X, Li Y-Q, Li Q-G, Ma Y-L, Peng J-J, Cai S-J. ITGAE Defines CD8+ Tumor-Infiltrating Lymphocytes Predicting a better Prognostic Survival in Colorectal Cancer. EBioMedicine. 2018;35:178-88.
38. Duhen T, Duhen R, Montler R, Moses J, Moudgil T, de Miranda NF, et al. Co-expression of CD39 and CD103 identifies tumor-reactive CD8 T cells in human solid tumors. Nat Commun. 2018;9(1):2724.
39. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nat Protoc. 2014;9(1):171.
40. Klein Allon M, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015;161(5):1187-201.
41. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017;65(4):631-43.e4.
42. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, et al. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. Mol Cell. 2019;73(1):130-42.e5.
43. Ding J, Adiconis X, Simmons SK, Kowalczyk MS, Hession CC, Marjanovic ND, et al. Systematic comparative analysis of single cell RNA-sequencing methods. bioRxiv2019.
44. Azizi E, Carr AJ, Plitas G, Cornish AE, Konopacki C, Prabhakaran S, et al. Single-cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment 3' RNA Sequencing. 417 E 68th St, New York, NY, 10065 USA: Memorial Sloan Kettering Cancer Center; 2018.
45. Guo X, Zhang Y, Zheng L, Zheng C, Song S, Zhang Q, et al. T cell landscape of non-small cell lung cancer revealed by deep single-cell RNA sequencing. Yiheyuan Road, 100871 Beijing, China,: Peking University; 2018.
46. Zheng C, Zheng L, Yoo J, Guo H, Zhang Y, Guo X, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. Yiheyuan Road, 100871 Beijing, China: Peking University; 2017.
47. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computinga; 2019.
48. RStudio Team. RStudio: Integrated Development for R. RStudio, Inc. Boston, MA, 2019.
49. McGinnis CS, Murrow LM, Gartner ZJ. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. 2019;8(4):329-37.e4.
50. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17(1):29.
51. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34(5):502-8.
52. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381-6.
53. Park D, Kim HG, Kim M, Park T, Ha H-H, Lee DH, et al. Differences in the molecular signatures of mucosal-associated invariant T cells and conventional T cells. Scientific Reports. 2019;9(1):7094.
54. Kumar P, Tan Y, Cahan P. Understanding development and stem cells using single cell-based analyses of gene expression. Development. 2017;144(1):17.

55. Braun J, Frentsch M, Thiel A. Hobit and human effector T-cell differentiation: The beginning of a long journey. European Journal of Immunology. 2015;45(10):2762-5.
56. Yan Y, Cao S, Liu X, Harrington SM, Bindeman WE, Adjei AA, et al. CX3CR1 identifies PD-1 therapy–responsive CD8+ T cells that withstand chemotherapy during cancer chemoimmunotherapy. JCI Insight. 2018;3(8):e97828.
57. Hinks TSC. Mucosal-associated invariant T cells in autoimmunity, immune-mediated diseases and airways disease. Immunology. 2016;148(1):1-12.
58. Duan M, Goswami S, Shi J-Y, Wu L-J, Wang X-Y, Ma J-Q, et al. Activated and Exhausted MAIT Cells Foster Disease Progression and Indicate Poor Outcome in Hepatocellular Carcinoma. Clin Cancer Res. 2019;25(11):3304.
59. Haeryfar SMM, Shaler CR, Rudak PT. Mucosa-associated invariant T cells in malignancies: a faithful friend or formidable foe? Cancer Immunol Immunother. 2018;67(12):1885-96.
60. Li H, van der Leun AM, Yofe I, Lubling Y, Gelbard-Solodkin D, van Akkooi ACJ, et al. Dysfunctional CD8 T Cells Form a Proliferative, Dynamically Regulated Compartment within Human Melanoma. Cell. 2019;176(4):775-89.e18.
61. Schuh E, Berer K, Mulazzani M, Feil K, Meinl I, Lahm H, et al. Features of Human CD3+CD20+ T Cells. J Immunol. 2016;197(4):1111.
62. Goodman A, Patel SP, Kurzrock R. PD-L1 immune-checkpoint blockade in B-cell lymphomas. Nat Rev Clin Oncol. 2017;14(4):203-20.
63. Sedelies KA, Sayers TJ, Edwards KM, Chen W, Pellicci DG, Godfrey DI, et al. Discordant Regulation of Granzyme H and Granzyme B Expression in Human Lymphocytes. J Biol Chem. 2004;279(25):26581-7.
64. Kakaradov B, Arsenio J, Widjaja CE, He Z, Aigner S, Metz PJ, et al. Early transcriptional and epigenetic regulation of CD8+ T cell differentiation revealed by single-cell RNA sequencing. Nat Immunol. 2017;18(4):422-32.
65. Carr TM, Wheaton JD, Houtz GM, Ciofani M. JunB promotes Th17 cell identity and restrains alternative CD4(+) T-cell programs during inflammation. Nat Commun. 2017;8(1):301.
66. Mueller SN, Gebhardt T, Carbone FR, Heath WR. Memory T cell subsets, migration patterns, and tissue residence. Annual review of immunology. 2013;31:137-61.
67. Reich NC. A death-promoting role for ISG54/IFIT2. Journal of interferon & cytokine research : the official journal of the International Society for Interferon and Cytokine Research. 2013;33(4):199-205.
68. Dukhanina EA, Lukyanova TI, Dukhanin AS, Georgieva SG. The role of S100A4 protein in anticancer cytotoxicity: its presence is required on the surface of CD4+CD25+PGRPs+S100A4+ lymphocyte and undesirable on the surface of target cells. Cell Cycle. 2018;17(4):479-85.

9 Figures and Tables

9.1 Gene Expression in Raw Datasets

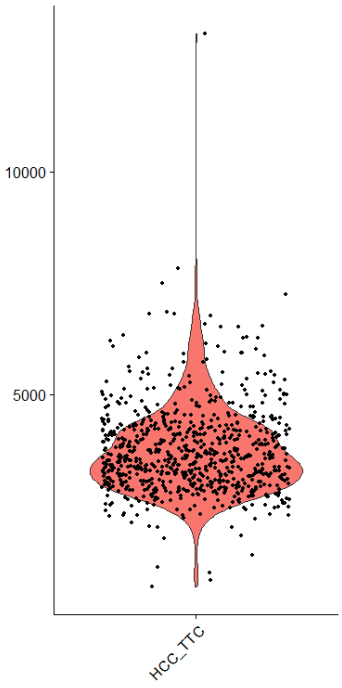

Figure 1: Distribution of number of unique genes expressed per cell prior to filtering for HCC dataset

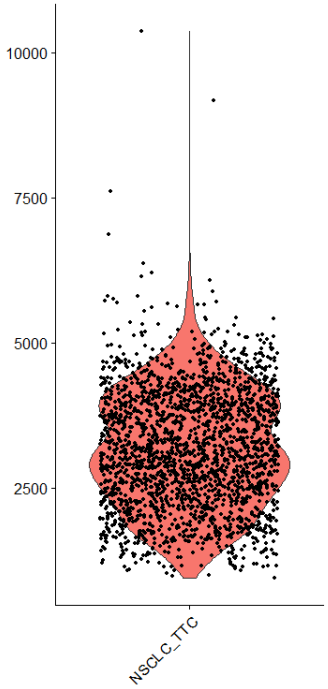

Figure 2: Distribution of number of unique genes expressed per cell prior to filtering for NSCLC dataset

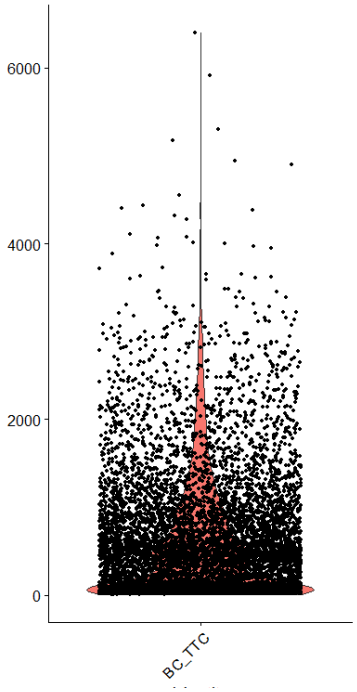

Figure 3: Distribution of number of unique genes expressed per cell prior to filtering for BC dataset

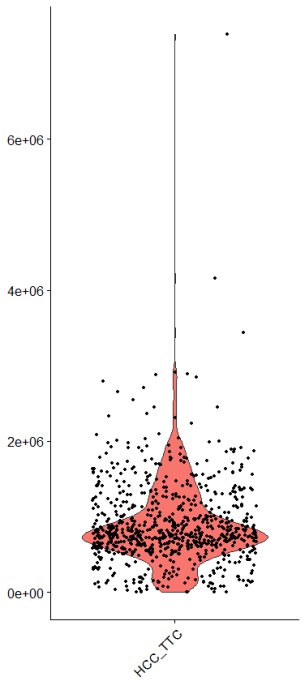

Figure 4: Distribution of mRNA molecule count per cell prior to filtering for HCC dataset

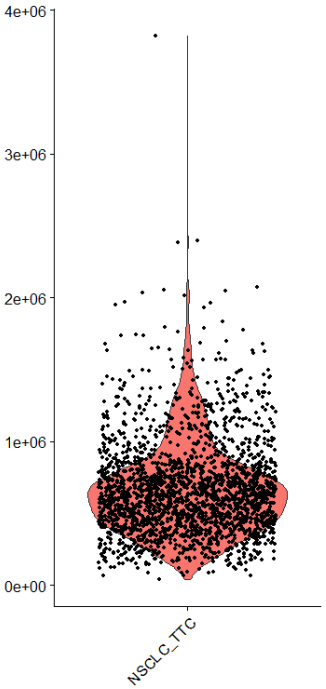

Figure 5: Distribution of mRNA molecule count per cell prior to filtering for NSCLC dataset

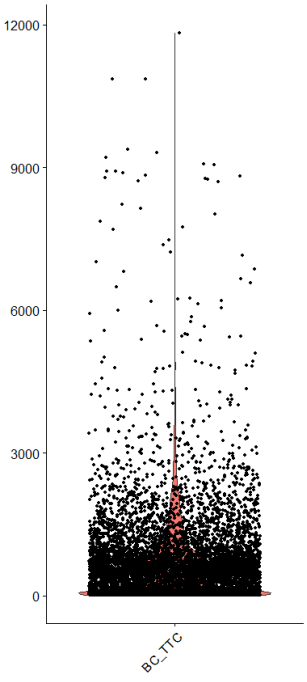

Figure 6: Distribution of mRNA molecule count per cell prior to filtering for BC dataset

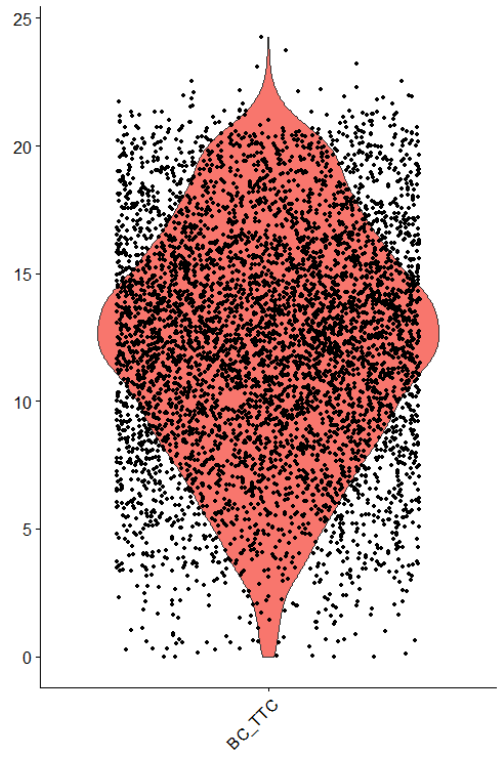

Figure 7: Distribution of the percentage of unique genes that were mapped to mitochondrial genome per cell in the BC dataset

9.2 Clustering of Breast Cancer data

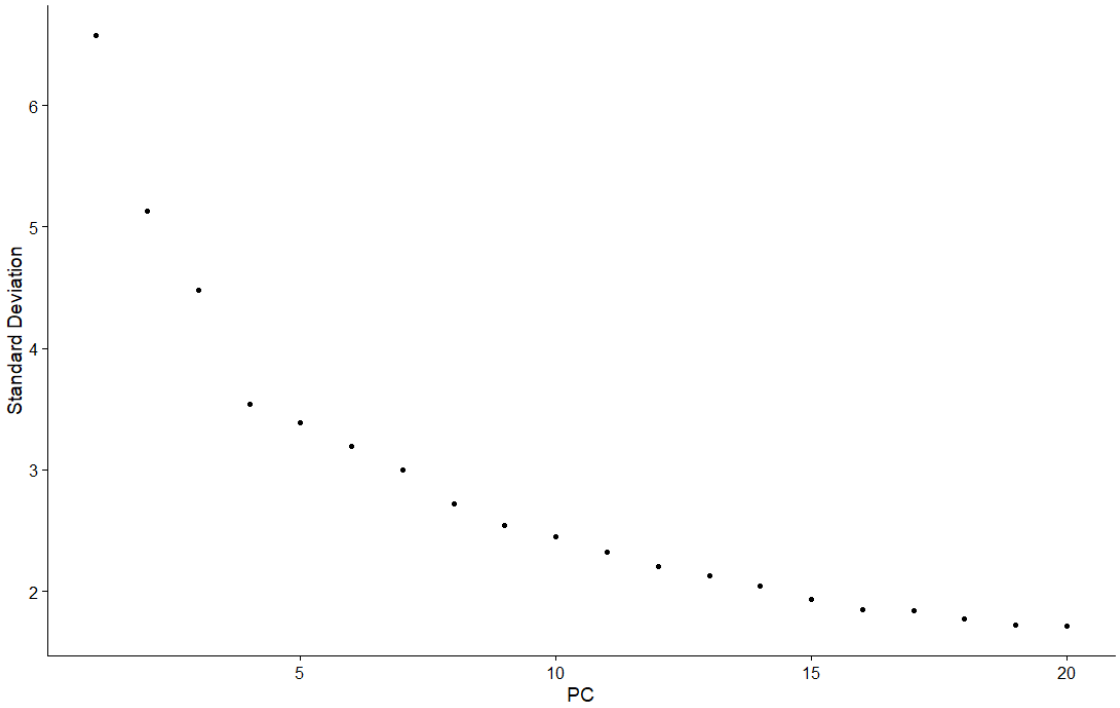

Figure 8: Elbow plot for the most significant PCs proposed by Seurat for identification of $CD8^+$ cells in BC TIL sample. The *y*-axis graphs the standard deviations of the twenty most significant principal components. The “elbow” around the fifteenth PC indicates that further PCs can be excluded from the analysis, as the standard deviation is too low to distinguish cells from different populations.

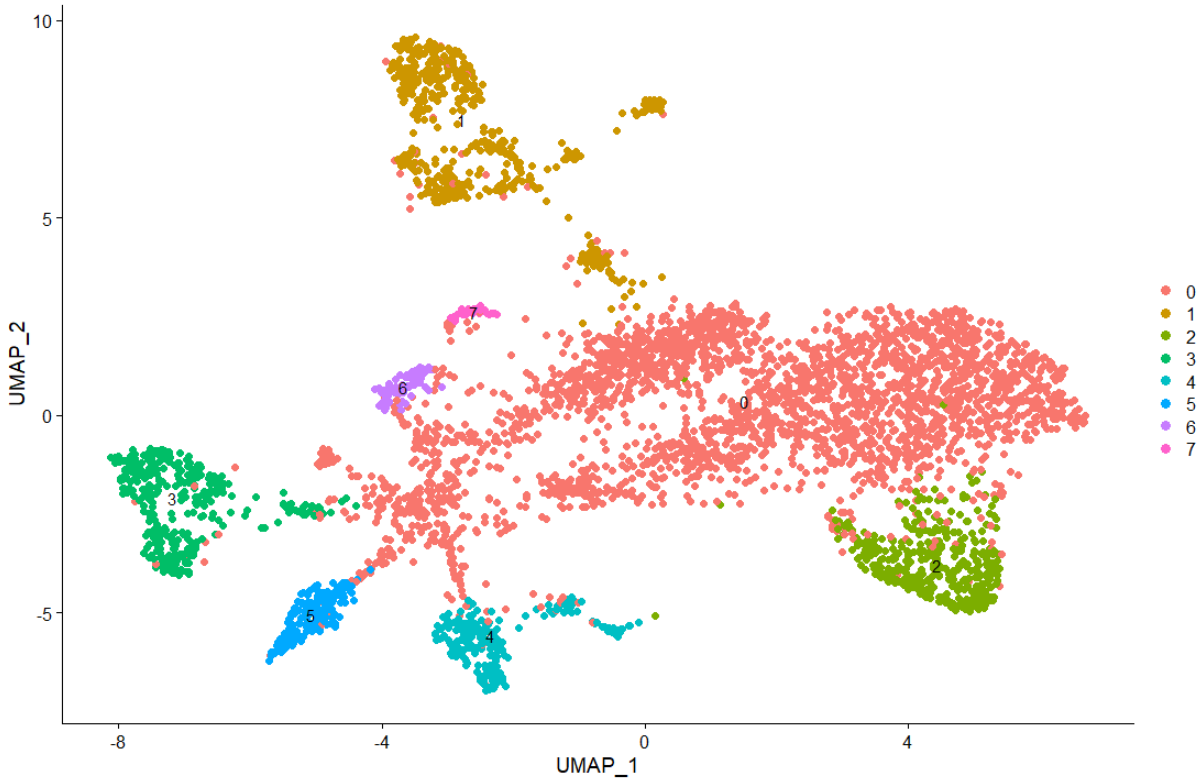

Figure 9: UMAP projection of the BC TIL dataset (15 PCs used, resolution set to 0.1)

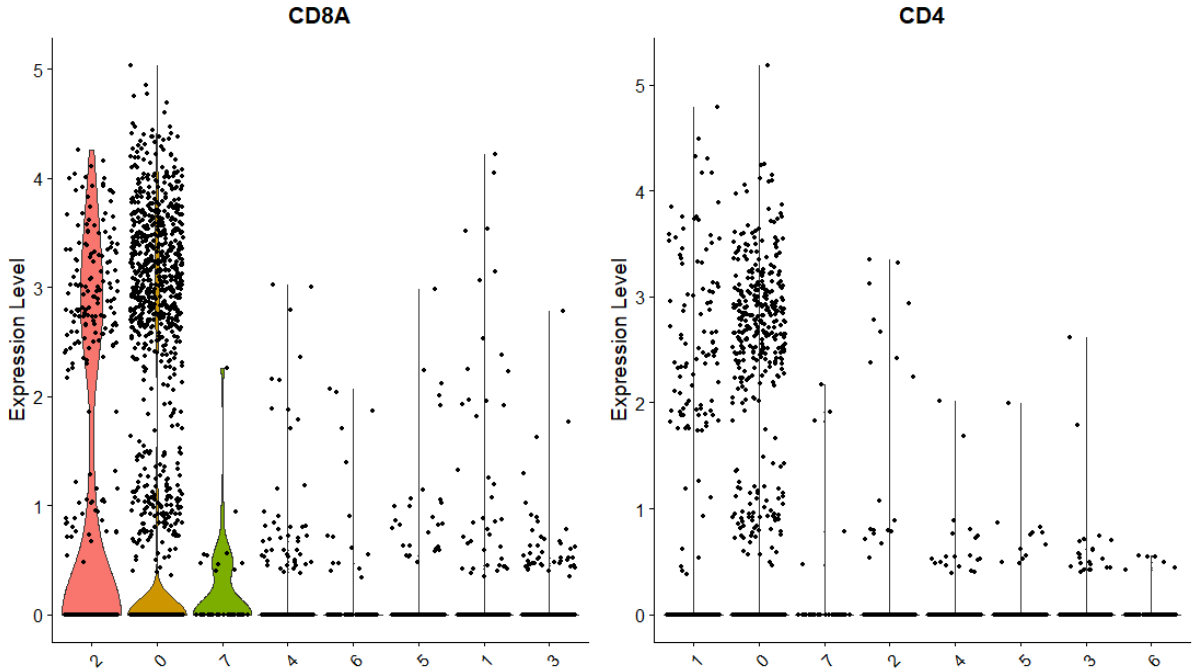

Figure 10: Expression level (the logarithm of the ratio between the number of mRNA molecules found in a cell and the total number of mRNA molecules in that cell) for CD4 and CD8A genes across clusters (*x*-axis) for BC TILs, sorted by highest to lowest expression. The graph illustrates that cluster 2 had the highest expression of CD8A, and hardly any expression of CD4. Cluster 0 featured more cells, but had a lower average CD8A expression as well as a noticeable CD4 expression. It was therefore deemed to contain double-positive T lymphocytes.

9.3 Integration of HCC and NSCLC datasets

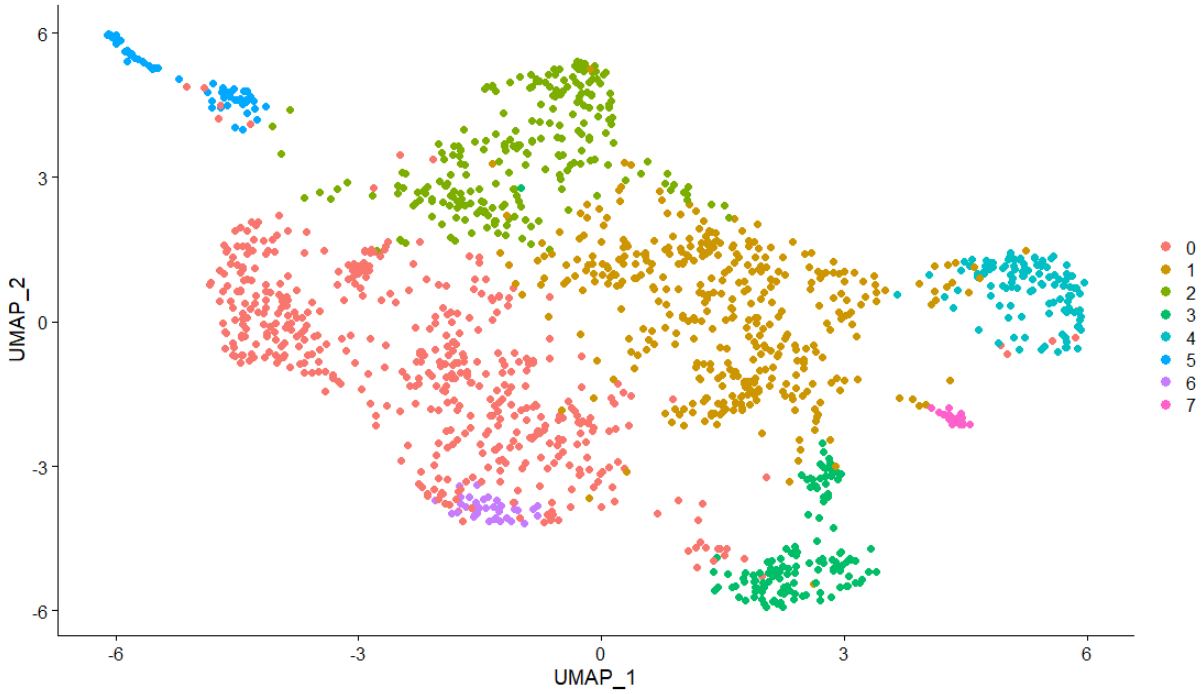

Figure 11: UMAP plot of the integrated HCC-NSCLC dataset, coloured by cluster (15 PCs, resolution 0.3)

9.4 Determining PCs to use for integrating the HCC, NSCLC and BC datasets

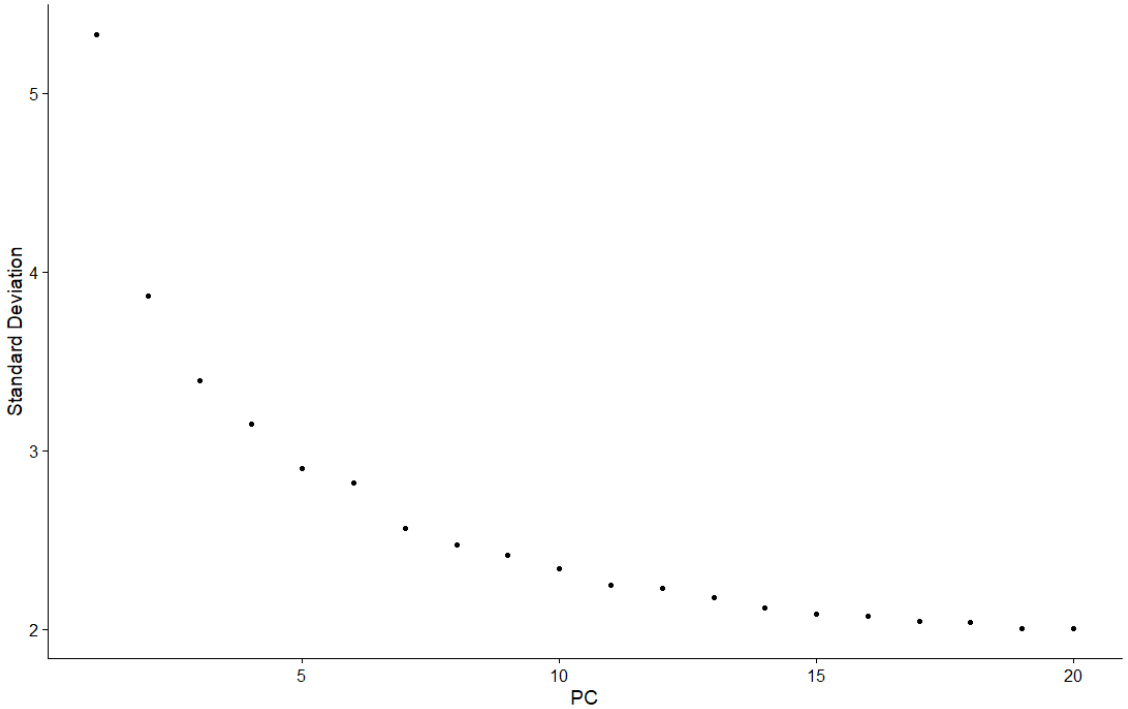

Figure 12: Elbow plot of the 20 most significant PCs for the integrated HCC, NSCLC and BC datasets. The curve flattens out at the 11th PC, indicating that the first 11 PCs were useful for discriminating between clusters.

9.5 Integration of the HCC, NSCLC and BC Datasets

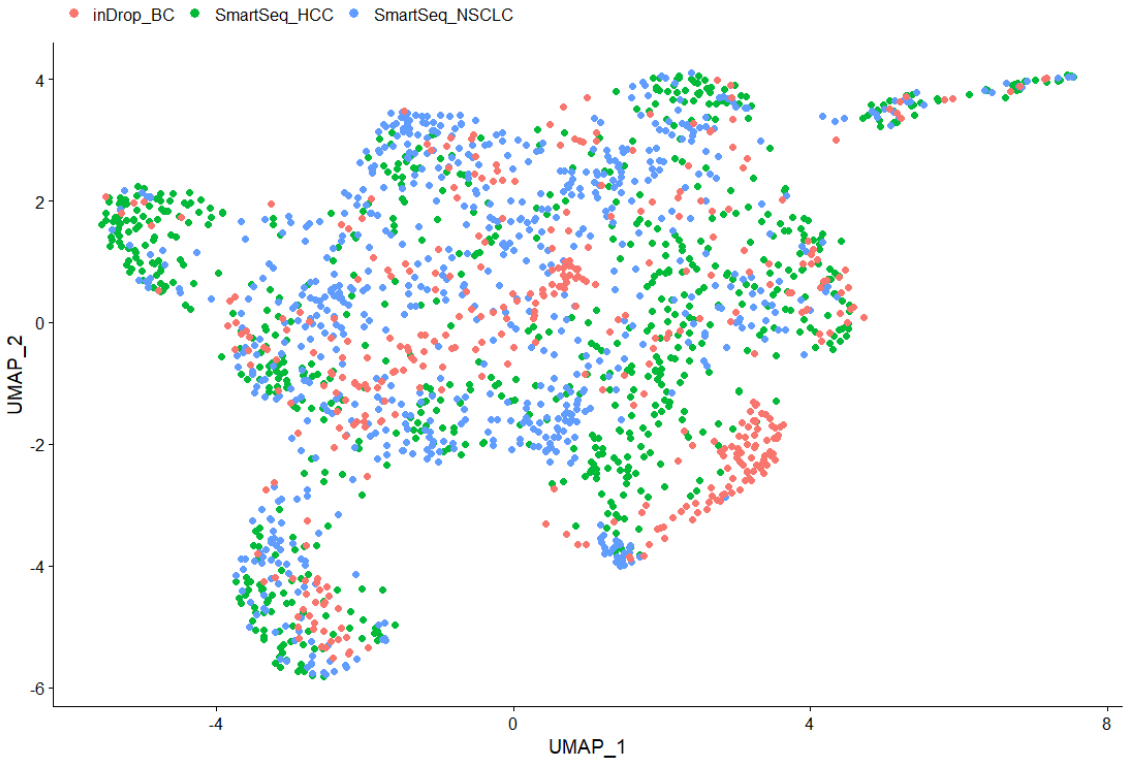

Figure 13: UMAP plot of the integrated HCC, NSCLC and BC dataset (11 PCs)

9.6 Clustering of the HCC, NSCLC and BC Datasets

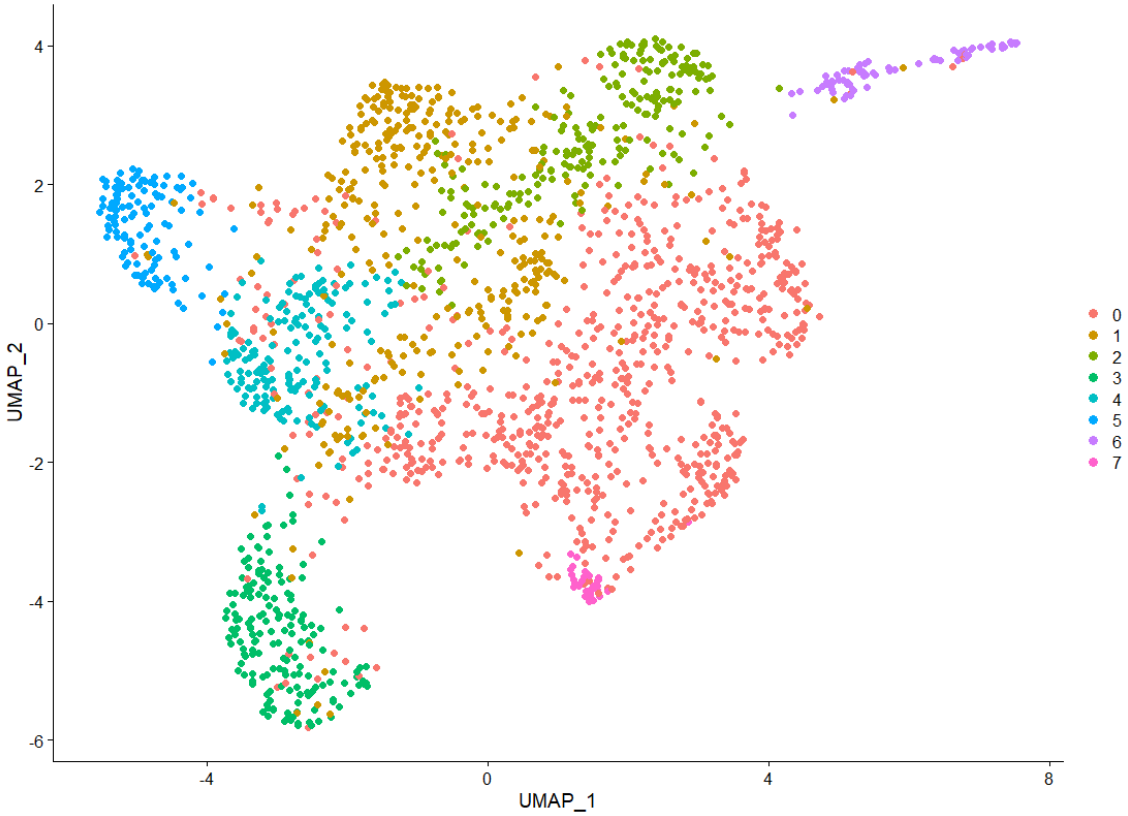

Figure 14: Clusters in the integrated HCC, NSCLC and BC dataset (11 PCs, resolution 0.4)

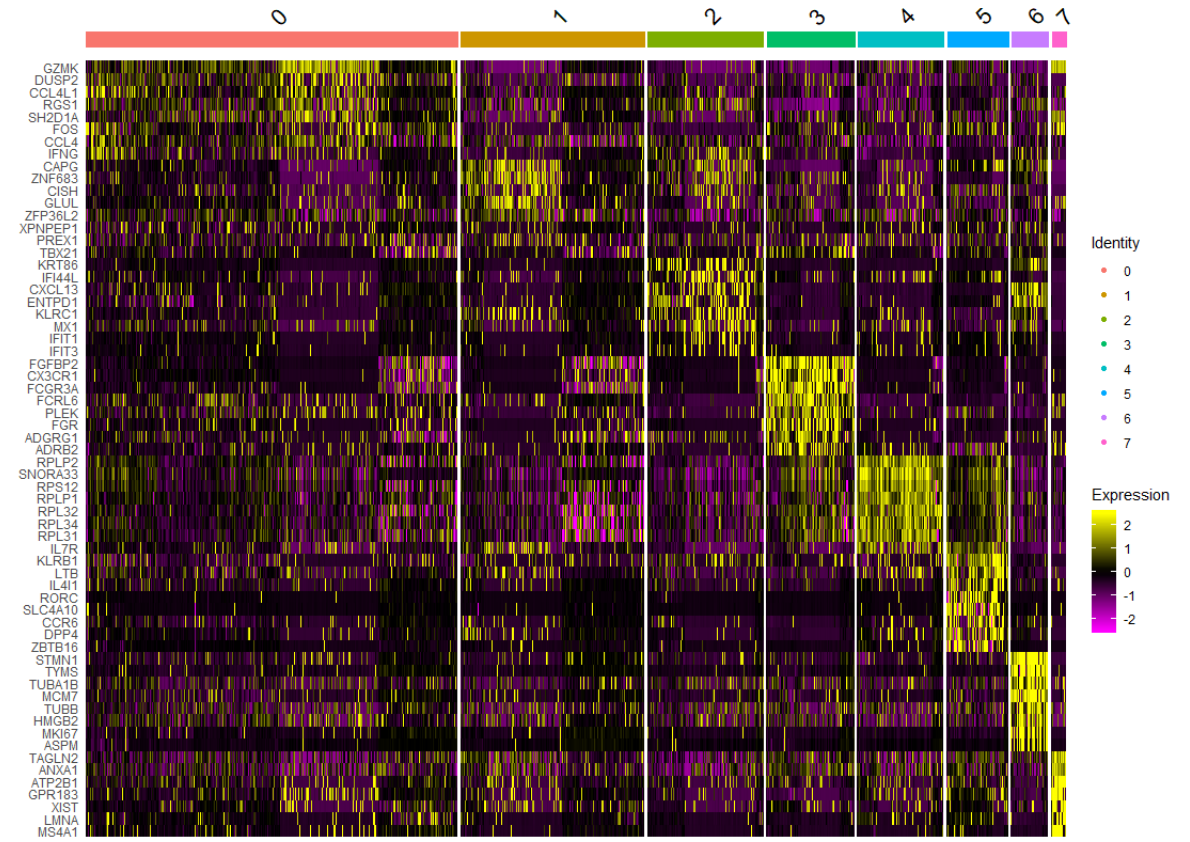

Figure 15: Heatmap illustrating the gene expression for the eight most differentially expressed genes for each cluster in the integrated HCC, NSCLC and BC dataset

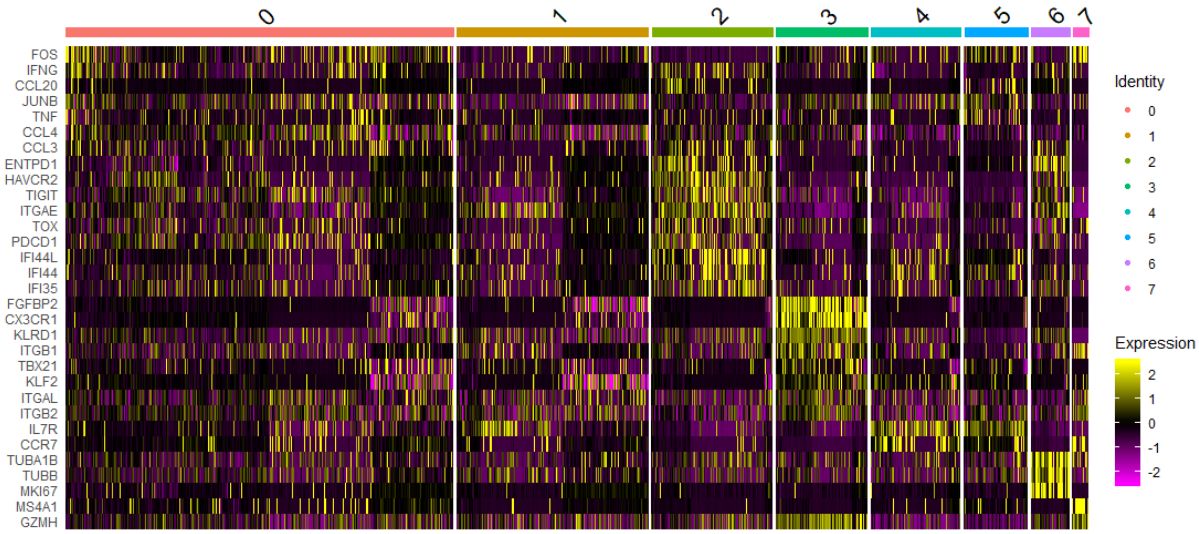

Figure 16: Heatmap illustrating the gene expression across clusters for a selection of known effector, cytotoxic, exhausted and naïve TIL markers. The figure shows a moderate cytotoxic signature in cluster 0, a strong exhausted/interferon-inducible signature in cluster 2, a strong memory signature in cluster 3, a sparse naïve signature across clusters 4 and 5 and a proliferation signature in cluster 6.

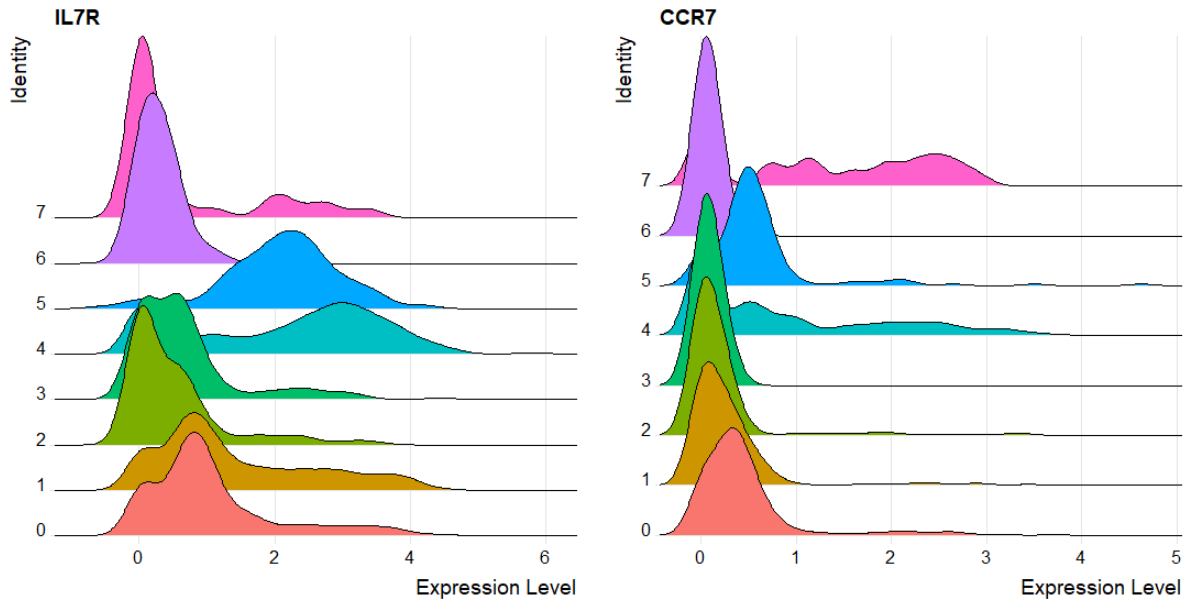

Figure 17: Ridge-plot illustrating gene expression level (logarithm of the ratio between mRNA counts for the gene and mRNA counts in the whole cell) in each cluster for two naïve TIL markers. Cluster 4 and 5 appear as likely candidates for naïve $CD8^+$ populations.

9.7 Possible Developmental Trajectory for the Integrated Dataset

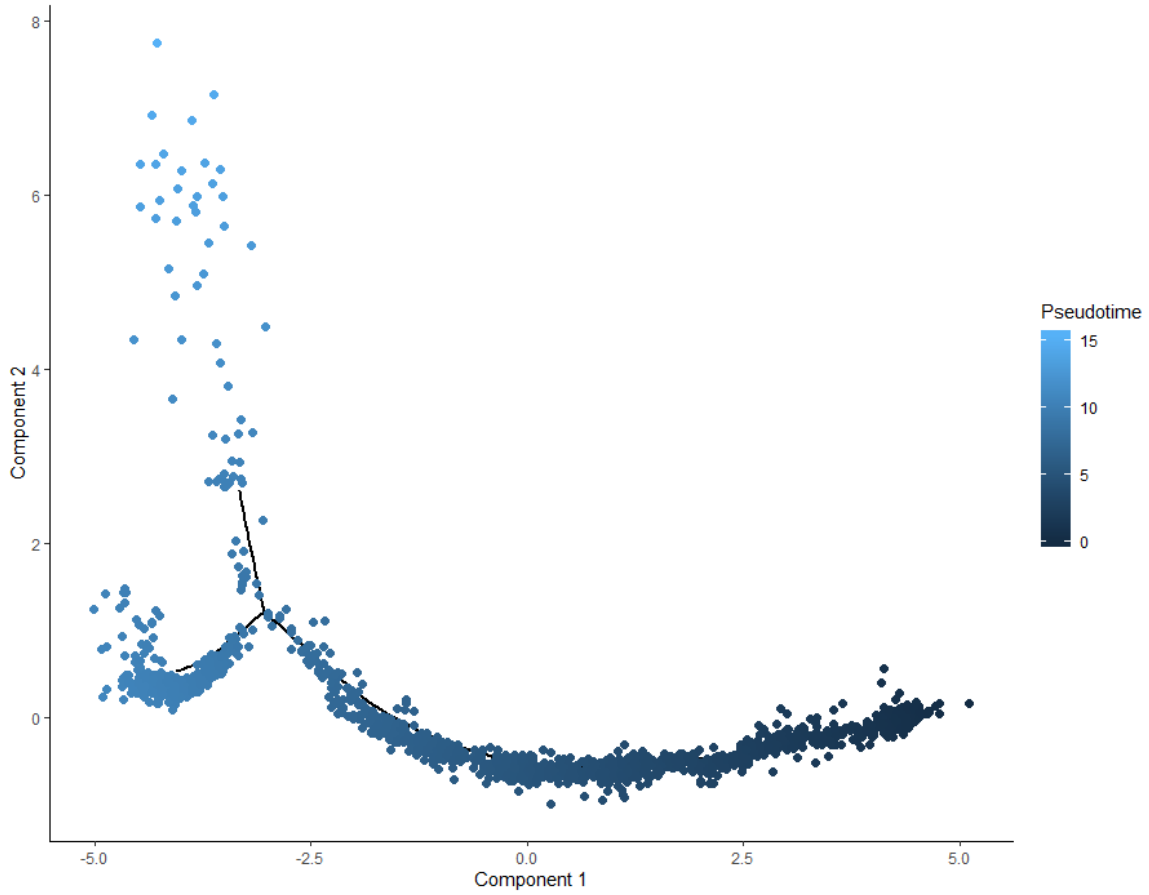

Figure 18: Proposed developmental trajectory for the integrated dataset graphed across two principal components, created using Monocle for R Studio, with individual cells coloured by “pseudotime”. This represents the most likely order in which the cells emerged. “Time” advances thus from the right to the left, finishing with the population branching into two paths.

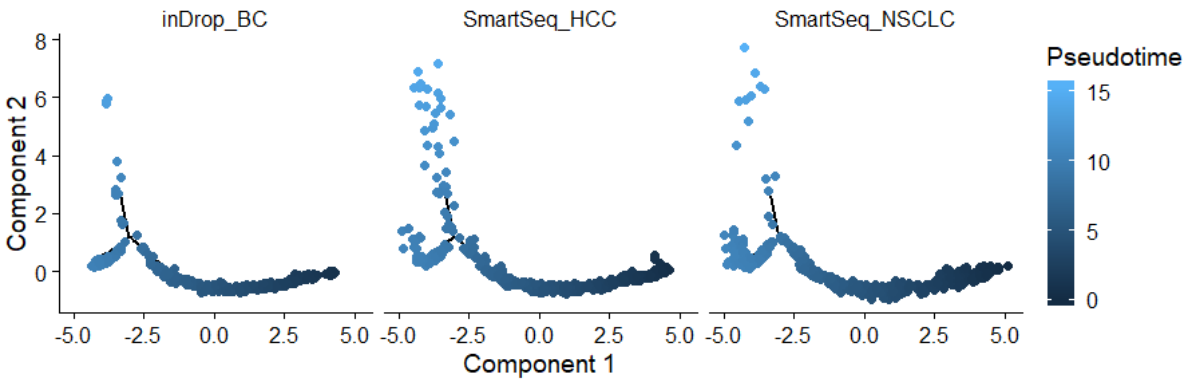

Figure 19: Proposed development for the integrated dataset, split by sequencing technology used. The consistency of the trajectory's shape suggests that sequencing technology did not affect the results in the previous figure.

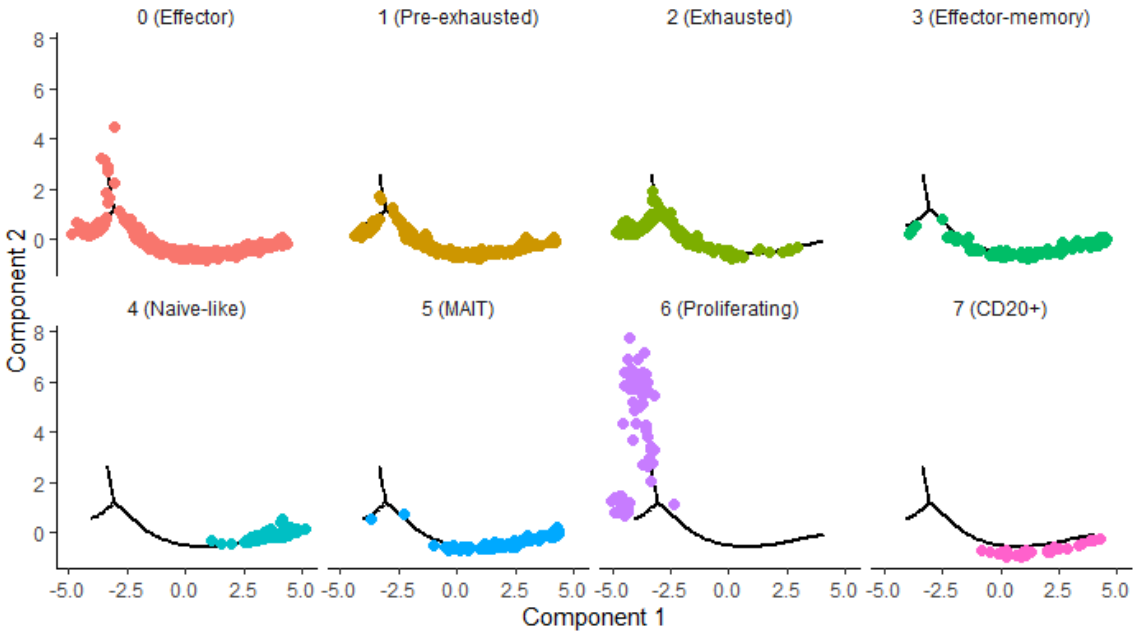

Figure 20: Proposed development for the integrated dataset in 'pseudotime' grouped by cluster. The graph suggests that cluster 3 ('Effector-memory'), 4 ('Naïve-like'), 5 ('MAIT') and 7 ('CD20^{+}') represent cells early in development, and that cluster 2 ('Exhausted') and 6 ('Proliferating') are late in their development.

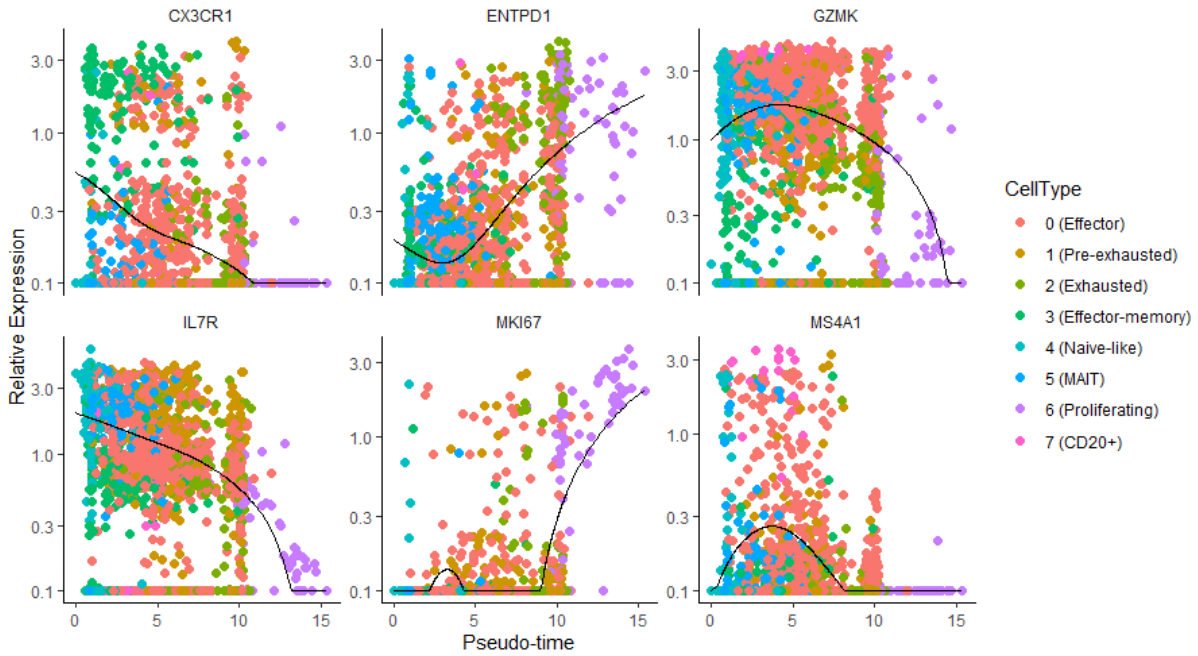

Figure 21: Relative gene expression in the integrated dataset across "pseudotime" for selected genes mentioned in the Discussion. Memory, naïve and cytotoxic genes (shown here by CX3CR1, IL7R and GZMK) decreased with pseudotime whereas ENTPD1 and MKI67 increased. MS4A1 was limited to expression on early cells.

9.8 Summary of datasets prior to and after processing

Table 1: Distribution of unique genes expressed per cell across datasets before and after data pre-processing. For the BC dataset, a summary of the extracted $CD8^+$ population is also given.

	Hepatocellular Carcinoma		Non-Small Cell Lung Cancer			Breast Cancer		
	Before	**After**	**Before**	**After**	**Sampled**	**Before**	**After**	**$CD8^+$**
Number of Cells	777	735	2182	2070	735	8187	4587	398
Minimum	665	2308	954	1385	1389	14.0	101.0	131.0
1st Quartile	3052	3080	2398	2451	2454	66.0	364.0	520.0
Median	3570	3567	3112	3111	3096	404.0	583.0	703.0
Mean	3720	3677	3161	3146	3138	600.8	727.8	778.3
3rd Quartile	4207	4170	3933	3898	3890	874.0	983.0	1017.8
Maximum	13107	6076	10358	4981	4955	6394.0	2496.0	2033.0

Table 2: Distribution of mRNA molecules per cell across datasets prior to and after data pre-processing. As above, a summary of the extracted $CD8^+$ population is also given.

	Hepatocellular Carcinoma		Non-Small Cell Lung Cancer			Breast Cancer		
	Before	**After**	**Before**	**After**	**Sampled**	**Before**	**After**	**$CD8^+$**
Minimum	1343	19793	41198	41198	119129	13.0	17.3	53.5
1st Quartile	599920	610930	448224	451067	449244	54.0	397.2	562.0
Median	780591	784299	621454	619357	601683	437.4	635.1	706.6
Mean	895498	888597	666934	660845	647081	655.6	743.9	735.8
3rd Quartile	1128306	1125182	806395	793149	779504	852.4	968.5	882.1
Maximum	7385913	3437635	3819276	2398056	2070682	11828.0	2497.3	2036.1

9.9 Comparison of BC Dataset with Integrated HCC and NSCLC Dataset

Table 3: Upregulated known TIL markers in cluster 3 in the integrated HCC-NSCLC dataset, into which 61.8% of the BC TILs were sorted.

	Average ln(FC)	*p*-value (Wilcoxon Rank Sum test with Bonferroni correction)	Type of Marker
GNLY	2.06	1.23×10^{-43}	Cytotoxic $CD8^+$ T cells[31,36]
CX3CR1	1.95	3.05×10^{-35}	Cytotoxic memory $CD8^+$ T cells[31,34]
FGFBP2	1.95	8.95×10^{-34}	Effector Memory $CD8^+$ T cells[6]
KLRD1	1.23	5.70×10^{-30}	Effector Memory $CD8^+$ T cells[6]
GZMH	1.14	6.88×10^{-32}	NK cells[63]
KLRG1	0.97	7.35×10^{-27}	Effector $CD8^+$ T cells[1]
GZMB	0.65	1.21×10^{-19}	Cytotoxic CD8+ Cells[15]
PRF1	0.49	9.45×10^{-16}	Cytotoxic $CD8^+$ T cells[15]
KLF2	0.34	1.00	Memory lymphocytes[64]

Table 4: Upregulated known TIL markers in cluster 6 in the integrated HCC-NSCLC dataset, into which 34.4% of the BC TILs were sorted.

	Average ln(FC)	*p*-value (Wilcoxon Rank Sum test with Bonferroni correction)	Type of Marker
SELL	1.55	3.06×10^{-16}	Naïve T cell[1,15]
FOS	1.48	8.22×10^{-14}	Activated T cells[6]
LEF1	1.26	8.06×10^{-15}	Naïve T cell[15]
CCR7	1.05	6.41×10^{-16}	Naïve T cell[6,15]
IL7R	0.56	8.18×10^{-7}	Naïve T cell[6]
KLF2	0.43	5.33×10^{-9}	Memory lymphocytes[64]
IFNG	0.25	1.00	Cytotoxic or Activated T cells[6,15]

9.10 Cluster properties in the Integrated HCC, NSCLC and BC Dataset

Table 5: Proportions of cells in the separate datasets and the integrated datasets distributed to each cluster. The proposed cluster titles from the Discussion are also given in column 1.

Cluster	Percentage of BC Cells	Percentage of HCC Cells	Percentage of NSCLC Cells	Percentage of All Cells
0 (‘Effector’)	38.7	51.3	25.9	38.6
1 (‘Pre-exhausted’)	40.2	6.3	20.7	19.2
2 (‘Exhausted’)	3.8	10.1	18.6	12.1
3 (‘Effector-memory’)	7.3	9.3	10.2	9.2
4 (‘Naïve-like’)	5.3	5.3	14.7	9.0
5 (‘MAIT’)	2.0	12.2	3.0	6.4
6 (‘Proliferating’)	2.5	5.3	3.3	3.9
7 (‘$CD20^+$’)	0.3	0.3	3.7	1.6

Table 6: Differentially expressed genes for cluster 0 in the integrated HCC, NSCLC and BC dataset, sorted by the average ln(FC). Although cytotoxic genes were both increased and decreased in expression, cytokine genes such as IFNG and TNF were moderately increased. Genes with average ln(FCs) between −0.25 and +0.25 or that had an associated *p*-value equal to 1.00 (calculated by the Wilcoxon Rank Sum test with Bonferroni correction applied) are not included here as their differential expression could not be deemed significant.

Gene	Average ln(FC)	*p*-value (Wilcoxon Rank Sum test with Bonferroni correction)	Type of Marker
GZMK	0.96	1.26×10^{-92}	Memory $CD8^+$ T cells[6]
CCL4	0.63	2.76×10^{-12}	Cytotoxic $CD8^+$ T cells[15]
FOS	0.55	2.34×10^{-16}	Activated $CD8^+$ T cells[6]
IFNG	0.43	8.43×10^{-3}	Cytotoxic $CD8^+$ T cells[15]
CCL3	0.38	7.81×10^{-5}	Cytotoxic $CD8^+$ T cells[15,32]
CCL20	0.33	3.87×10^{-9}	Cytotoxic $CD8^+$ T cells[32]
EOMES	0.31	1.64×10^{-18}	Central Memory $CD8^+$ T cells[2]
JUNB	0.30	2.05×10^{-8}	$CD4^+$ T_H17 cells[65]
CD69	0.27	1.82×10^{-9}	Resident Memory $CD8^+$ T cells[66]
TNF	0.26	1.06×10^{-1}	Activated $CD8^+$ T cells[6]
IFIT2	–0.27	7.38×10^{-10}	Interferon-inducible elements[67]
KLRB1	–0.35	1.40×10^{-7}	T_H17 $CD4^+$ cells[6]
GZMB	–0.37	4.35×10^{-5}	Cytotoxic CD8+ Cells[15]
ID2	–0.38	2.76×10^{-6}	Effector Memory $CD8^+$ T cells[2]
KLRD1	–0.52	6.13×10^{-2}	Effector Memory $CD8^+$ T cells[6]
S100A4	–0.65	1.03×10^{-26}	Cytotoxic $CD8^+$ T cells[68]
GNLY	–1.15	8.64×10^{-14}	Cytotoxic $CD8^+$ T cells[36]

Table 7: Differentially expressed genes for cluster 3 in the integrated HCC, NSCLC and BC dataset, sorted by the average ln(FC). Effector-related and memory-related genes appear to be upregulated, compared to exhaustion- or naïve-related genes. As before, statistically insignificant genes are not included.

Gene	Average ln(FC)	*p*-value (Wilcoxon Rank Sum test with Bonferroni correction)	Type of Marker
FGFBP2	1.84	5.89×10^{-58}	Effector Memory $CD8^+$ T cells[6]
CX3CR1	1.73	1.28×10^{-55}	Cytotoxic Effector $CD8^+$ T cells[34]
GNLY	0.98	6.27×10^{-35}	Cytotoxic $CD8^+$ T cells[31,36]
GZMH	0.97	3.73×10^{-32}	NK cells[63]
ITGAM	0.97	1.08×10^{-35}	Effector $CD8^+$ T cells[1]
KLRD1	0.92	1.92×10^{-32}	Effector Memory $CD8^+$ T cells[6]
KLRG1	0.81	4.01×10^{-25}	Effector Memory $CD8^+$ T cells[1]
PRF1	0.79	6.87×10^{-37}	Cytotoxic $CD8^+$ T cells[15]
NKG7	0.74	5.63×10^{-35}	Cytotoxic $CD8^+$ T cells[15]
CST7	0.71	2.92×10^{-31}	Cytotoxic $CD8^+$ T cells[15]
ITGB1	0.71	2.49×10^{-13}	Memory $CD8^+$ T cells[1]
S100A4	0.50	1.61×10^{-7}	Cytotoxic $CD8^+$ T cells[68]
ITGA4	0.48	8.84×10^{-7}	Effector $CD8^+$ T cells[1]
TBX21	0.40	3.10×10^{-7}	Effector Memory $CD8^+$ T cells[2]
KLF2	0.39	4.66×10^{-20}	Memory lymphocytes[64]
ITGAL	0.38	8.29×10^{-8}	Memory $CD8^+$ T cells[1]
GZMB	0.35	2.58×10^{-12}	Cytotoxic CD8+ Cells[15]
ITGB2	0.35	4.99×10^{-4}	Memory $CD8^+$ T cells[1]
PTPRC	0.26	3.46×10^{-4}	Effector $CD8^+$ T cells[1]
CD244	0.25	9.56×10^{-1}	Exhausted Cytotoxic $CD8^+$ T cells[3]
SELL	–0.28	3.95×10^{-3}	Naïve $CD8^+$ T cells[1,15]
CD38	–0.30	4.32×10^{-2}	Effector $CD8^+$ T cells[32]
CCL20	–0.39	4.39×10^{-3}	Cytotoxic $CD8^+$ T cells[32]
TIGIT	–0.39	1.34×10^{-1}	Exhausted $CD8^+$ T cells[15]
IFI44	–0.45	2.30×10^{-4}	Interferon-inducible elements[67]
JUN	–0.47	2.71×10^{-2}	Activated $CD8^+$ T cells[6]
CD44	–0.48	6.10×10^{-7}	Resident Memory $CD8^+$ T cells[66]
CD69	–0.52	3.34×10^{-5}	Resident Memory $CD8^+$ T cells[66]
CXCR3	–0.56	9.40×10^{-14}	Resident Memory $CD8^+$ T cells[2]
HAVCR2	–0.59	7.85×10^{-9}	Exhausted Cytotoxic $CD8^+$ T cells[6,15]
PDCD1	–0.69	1.68×10^{-10}	Exhausted Cytotoxic $CD8^+$ T cells[6,15]
CCR7	–0.78	6.56×10^{-18}	Naïve $CD8^+$ T cells[6,15]
IL7R	–0.80	5.42×10^{-7}	Naïve $CD8^+$ T cells[6]
CD27	–0.95	3.27×10^{-26}	Memory $CD8^+$ T cells[1]
GZMK	–1.08	7.91×10^{-15}	Memory $CD8^+$ T cells[6]
ITGAE	–1.16	2.30×10^{-30}	Exhausted $CD8^+$ T cells[6]

Table 8: Differentially expressed genes for cluster 2 in the integrated HCC, NSCLC and BC dataset, sorted by the average ln(FC). Several exhaustion-related or interferon-inducible genes can be found amongst the highly expressed genes, whereas naïve-related genes were lowly expressed. As before, statistically insignificant genes are not included.

Gene	Average ln(FC)	*p*-value (Wilcoxon Rank Sum test with Bonferroni correction)	Type of Marker
IFI44L	1.18	7.78×10^{-8}	Interferon-inducible elements[67]
ENTPD1	1.06	2.24×10^{-6}	Exhausted $CD8^+$ T cells[6]
HAVCR2	1.04	1.50×10^{-26}	Exhausted Cytotoxic $CD8^+$ T cells[6,15]
GZMB	0.92	7.00×10^{-25}	Cytotoxic CD8+ Cells[15]
TIGIT	0.89	2.50×10^{-20}	Exhausted $CD8^+$ T cells[15]
IFI44	0.89	2.97×10^{-8}	Interferon-inducible elements[67]
ITGAE	0.84	9.51×10^{-20}	Tissue-resident memory T cells[38]
FASLG	0.82	1.24×10^{-11}	Effector $CD8^+$ T cells[32]
IFI35	0.63	1.53×10^{-3}	Interferon-inducible elements[67]
TOX	0.55	2.88×10^{-8}	Exhausted $CD8^+$ T cells [9]
CCL3	0.55	2.47×10^{-4}	Cytotoxic or Exhausted $CD8^+$T cells[15]
GZMA	0.53	2.55×10^{-7}	Cytotoxic or Exhausted $CD8^+$T cells[15]
ID2	0.49	3.32×10^{-12}	Effector Memory $CD8^+$ T cells[2]
PDCD1	0.49	2.59×10^{-4}	Exhausted Cytotoxic $CD8^+$ T cells[6,15]
IFNG	0.44	7.02×10^{-4}	Cytotoxic $CD8^+$ T cells[15]
CD27	0.34	8.93×10^{-4}	Memory $CD8^+$ T cells[1]
STAT3	0.30	2.46×10^{-1}	Central Memory $CD8^+$ T cells[2]
CCR7	–0.28	3.73×10^{-13}	Naïve $CD8^+$ T cells[6,15]
CD44	–0.29	1.39×10^{-1}	Resident Memory $CD8^+$ T cells[66]
TCF7	–0.31	1.29×10^{-3}	Memory or Naïve $CD8^+$ T cells[15,64]
SELL	–0.31	2.58×10^{-3}	Naïve $CD8^+$ T cells[1,15]
ITGAL	–0.33	5.89×10^{-1}	Effector Memory $CD8^+$ T cells[1]
ITGB2	–0.34	7.10×10^{-3}	Effector Memory $CD8^+$ T cells[1]
KLF2	–0.36	2.51×10^{-7}	Memory lymphocytes[64]
CST7	–0.38	2.59×10^{-2}	Cytotoxic $CD8^+$ T cells[15]
SLAMF7	–0.42	2.04×10^{-7}	Effector $CD8^+$ T cells[30]
EOMES	–0.54	1.72×10^{-11}	Central Memory $CD8^+$ T cells[2]
JUNB	–0.59	7.03×10^{-1}	$CD4^+$ T_H17 cells[65]
KLRG1	–0.86	4.95×10^{-20}	Effector Memory $CD8^+$ T cells[1]
GZMK	–0.94	5.17×10^{-25}	Memory $CD8^+$ T cells[6]
IL7R	–1.10	2.38×10^{-25}	Naïve $CD8^+$ T cells[6]
FOS	–1.52	8.61×10^{-9}	Activated $CD8^+$ T cells[6]

10 Appendices

10.1 Principal Components used for clustering the HCC, NSCLC and BC dataset

Table 9: Summary of the five most positively and five most negatively selected genes for the eleven PCs used to analyse the integrated HCC, NSCLC and BC dataset

	Positively selected genes	Negatively selected genes
PC #1	TXNIP, SNORA33, IL7R, RPS6, RPLP2	BIRC5, ASPM, KIF23, MKI67, CEP55
PC #2	HAVCR2, CXCL13, TIGIT, TNFRSF9, ITGAE	SNORA33, RPLP2, RPS12, RPL27A, RPS18
PC #3	A2M, CXCR5, PDLIM1, GZMK, FCRL6	RPS19, RPLP1, RPLP2, SNORA33, RPL32
PC #4	GZMK, CXCR4, CCR7, GPR183, ITM2C	FCGR3B, FGFBP2, FCGR3A, CX3CR1, GNLY
PC #5	CCL4L1, NR4A2, DUSP2, DUSP1, NR4A1	VIM, ZNF683, TAGLN2, RSAD2, CISH
PC #6	CCNB1, IFI44L, ZNF683, CDC20, PLK1	TYMS, MCM2, FAM111B, MCM4, GINS2
PC #7	GZMK, CD74, COTL1, GZMH, EOMES	RORC, SLC4A10, KLRB1, ZBTB16, CCR6
PC #8	SPIB, HBA2, HBA1, HBB, CDCA7	TPPP, CRIP2, KRT86, CD14, PLPP2
PC #9	IFIT1, IFIT3, IFI44L, RSAD2, MX1	HTRA1, KRT86, TMEM173, TXNIP, PTGIS
PC #10	GLUL, CXCR4, COTL1, PDK4, HERPUD1	SPIB, HBA2, HBA1, HBB, GPR183
PC #11	RORC, SLC4A10, GZMK, FCRL3, IGFBP7	GLUL, ZNF683, HSPA1B, ANXA1, HSPA1A