Content available at: https://www.ipinnovative.com/open-access-journals

Southeast Asian Journal of Health Professional

Journal homepage https://sajhp.com/

Original Research Article

Utilising the IMMUcan database for meta- assessment of human cancer single-Cell RNA-seq databases

Tshetiz Dahal[1]*, Suyash Saurabh[2]

[1]General Physician, Clinical Researcher, Lugansk State Medical University, Lypnia St. Rivne, Ukraine.
[2]Dept of Physiology, Rajendra Institute of Medical Sciences, Bariatu, Ranchi. Jharkhand, India.

Abstract

Understanding the tumour microenvironment (TME) has been made possible in large part by the advancement of single-cell RNA sequencing (scRNA-seq) technology. Numerous independent scRNA-seq studies have been published, which is a great resource that offers chances for meta-analysis research. However, there are significant barriers to fully utilising scRNA-seq data due to the vast amount of biological information, the notable diversity and heterogeneity within studies, and the technical difficulties in processing diverse datasets. We created IMMUcan scDB, a fully integrated scRNA-seq database that is open to nonspecialists and solely focused on human cancer. The 144 datasets on 56 distinct cancer types in the IMMUcan scDB are annotated in 50 domains with detailed biological, clinical, and technological data. Four steps comprised the development and organization of a data processing pipeline: (i) data collection; (ii) data processing (including sample integration and quality control); (iii) supervised cell annotation using a TME cell ontology classifier; and (iv) an interface to analyse TME globally or in relation to a particular cancer type. This framework was utilized to do meta-analysis research, such as rating immune cell types and genes linked to malignant transformation, and to investigate datasets across tumour locations in a gene-centric (CXCL13) and cell-centric (B cells) manner. An unparalleled degree of thorough annotation is provided by this integrated, publicly available, and user-friendly resource, which opens up a plethora of opportunities for the downstream exploitation of human cancer scRNA-seq data for discovery and validation investigations.

Significance: The IMMUcan scDB database is a user-friendly resource for interpreting and analysing tumour -associated single-cell RNA sequencing data, enabling researchers to make the most of this information to offer fresh perspectives on the biology of cancer.

Keywords: Tumour microenvironment (TME), CHETAH, CXCL13

Received: 10-04-2025; **Accepted:** 21-06-2025; **Available Online**: 04-09-2025

1. Introduction

Aside from the tumour cells themselves, the tumour microenvironment (TME) has been shown to strongly influence the clinical outcome of immunotherapies. Therefore, better characterising the cellular composition and molecular characteristics of the TME remains an important and challenging task that could help not only develop novel anticancer strategies but also identify biomarkers, better predict outcome to current immunotherapies, and lead to optimised personalised treatment strategies. Tumor immunology has taken center stage in cancer research due to the relative success of immunotherapy in a large number of malignancies, but despite recent advancements, the majority of cancer patients still either do not respond to therapy or eventually relapse and succumb to the disease. Technologies for single-cell RNA sequencing, or scRNA-seq, are specially suited to investigate the variety of cellular phenotypes and molecular pathways found in the TME. They can assist in answering a variety of scientific problems, such as determining the cell states linked to cancer, forecasting intercellular communication, figuring out the mechanisms underlying disease resistance, and identifying new drug targets. Since the authors of the majority of published articles have only addressed a small number of hypotheses and have not integrated their data with other complementary studies, the ever-increasing number of cancer-related scRNA-seq

*Corresponding author: Tshetiz Dahal
Email: dahaltshetiz21gmail.com

https://doi.org/10.18231/j.sajhp.v.8.i.3.4

datasets published in recent years represent a highly valuable but under-utilised treasure trove for biomedical research.

Although it would be easier to retrieve, reanalyse, and compare published scRNA-seq datasets if tumor-derived single-cell data were integrated into a searchable database, this is difficult for a number of reasons: (i) the wide variety of cancer types and clinical contexts, including tumor location and treatment type, make cancer-related datasets extremely diverse; (ii) the use of single-cell technologies, experimental protocols, and data analysis techniques; and (iii) the biological and clinical interpretation of the findings. Recently, scRNA-seq data portals such as scRNASeqDB,[1] SCPortalen,[2] PanglaoDB,[3] and JingleBells[4]have been developed to address this difficulty. Nevertheless, only two databases—CancerSEA and TISCH—are devoted to storing information about tumours. The goal of CancerSEA, which has integrated 41,900 single cancer cells from 25 different cancer types,[5] is to find functional states linked to particular gene signatures. It incorporates data from patient-derived xenografts, cancer cell lines, and human malignancies. Only the tumor type annotation is included in the scant clinical data. To describe the different cell types that make up the TME and examine the expression of target genes and signatures, TISCH allows users to search through cancer scRNA-seq datasets from both humans (74 datasets) and mice (5 datasets).[6] Only tumor type, primary versus metastatic illness, and treatment are included in the clinical annotation. Comparing the target gene expression and cellular composition across different datasets is made possible by the database functionalities. Our goal was to surpass existing initiatives and create a comprehensive, integrated, and fully annotated scRNA-seq database devoted just to human cancer. In order to create and integrate the clinical, cellular, and molecular profiles of various tumor types and their microenvironment, this study was conducted as part of the European Innovative Medicine Initiative 2 program's "Integrated iMMU no profiling of large adaptive CANcer patient cohorts" (IMMUcan) consortium. Cell types and gene expression patterns can be linked to certain clinical patterns using the comprehensive clinical annotation provided by the IMMUcan database. Additionally, it provides a wide range of features for analysing various datasets. In order to promote cancer biomedical research in the early discovery, hypothesis-generating, and validation stages, we believe the database will establish itself as the gold standard reference tool.

2. Materials and Methods

2.1. Literature search and dataset selection

We used the keywords ((cancer [Title/Abstract]) AND (patient)) AND (single cell RNA sequencing) to search PubMed (ncbi.nlm.nih.gov/pubmed/) for peer-reviewed published datasets, and "human cancer single-cell rna-sequencing" as a free-text keyword to search the bioRxiv database (www.biorxiv.org) for non-peer-reviewed studies. After applying a filter to choose journals published between 2016 and 2021, we manually examined each article's title and abstract to see if scRNA-seq data was available. This produced 103 publications total, covering 144 datasets. We retrieved the data from Gene Expression Omnibus (GEO), ArrayExpress, EGA, and BioProject for all datasets from human cancer patients with more than a thousand cells and at least 10 samples. We obtained the information from BioProject, ArrayExpress, EGA, and Gene Expression Omnibus (GEO). Datasets that concentrated on extra biology, such as various biopsy sites, treatment details, or longitudinal samples, were exempt from the rule regarding the quantity of patient samples. We ultimately found 73 datasets encompassing 56 distinct cancer indications using these filters, and they were added to the IMMUcan scDB.

2.2. Metadata extraction

Influenced by the criteria for publishing scRNA-seq experiments,[7] we extracted the following metadata categories from the 144 chosen studies in order to organise the data and facilitate effective database searches. The first category records study-wide data, such as the number of patients and samples, abstract, DOI, title of the publication, and data access details. Sample-specific characteristics such cancer kind, cancer localisation, response, and treatment are the focus of the second category. A third category includes all information pertaining to the workflow of applied single-cell technology, such as tissue dissociation, cell type enrichment, single-cell isolation, library construction, end bias, library layout, reference genome, read alignment, read counting, and expression value format. Lastly, we also annotated and standardised information about the single-cell data in the study, such as the tissue, patient ID, time-point and location of the biopsy, author annotations, cancer (sub) type, cancer stage, enrichment strategy, and treatment information like time-point, drug, and response, whenever it was supplied. Every metadata, including cell type, cancer type, and treatment, was standardised and mapped to ontologies whenever feasible. Depending on the type of information the metadata can be either free text, a list from a controlled vocabulary, Boolean values, or quantitative information.

2.3. Analysing data

We downloaded raw read counts whenever possible to improve comparison across studies and because they are more appropriate for the majority of single-cell analysis procedures.[8] Each dataset that included several tests or indicators of cancer was then divided into distinct files. In order to process all downloaded single-cell data efficiently, we created a pipeline using the R language called scProcessoR. It primarily uses functions from the Seurat package (version 3; ref. [9]) to log normalise the data, select the most variable genes, transform and scale the data using principal component analysis, build knn neighbourhoods for each cell, cluster the data using graphs, and generate UMAP dimensionality reduction plots (Supplementary **Figure 1 A**).

Every stage is carried out semiautomatically in accordance with industry best practices.[8,10] An expression matrix with cells as columns and genes as rows, together with a metadata file with the cleaned and standardised data for the samples, such as patient information and cell annotations, are inputs to the procedure. We used the dataset SC_UNB_10X_GSE134520, which contains the single-cell transcriptome profiles of nine individuals with early stomach carcinoma, to demonstrate the workflow. We eliminated all cells with fewer than 250 genes with mapped reads and/or, depending on the type of tumor[10] include more than 5% to 20% of mitochondrial specific reads for each dataset in order to preserve only high quality data. In order to determine

$$H_{\mathrm{TC}} = -\sum_{b=1}^{B} q_{Cb} log\ q_{Cb},$$

whether a dataset has significant technical batch effects, we calculated the Shannon entropy, HTC, for each cell in the dataset, C, and each possible batch effect type, T (such as patient ID), as follows:

B is the total number of type T batches in a dataset (all

$$H_{TCnorm} = \frac{H_{TC}}{H_{Ttotal}}$$

$$H_{Ttotal} = -\sum_{b=1}^{B} q_b log\ q_b,$$

Where qb is the proportion of cells from a specific batch b, and B is the total number of batches of type T in a dataset (for example, all patients).

A cell with an HTCnorm value of 0 is surrounded entirely by cells from the same batch, while a cell with a value of 1 has 30 nearest neighbors of C that appear equally frequently from all batches of type T in the dataset. We used the Harmony package (version 0.1; ref. [11]) with the default parameters provided to adjust for the appropriate batch effects if the median entropy across all cells in a dataset had a value of ≤ 0.5 for a particular type of batch effect.

2.4. Cell type annotation and cell clustering

With the resolution parameter set to 1, Louvain graph-based clustering, which is implemented in the Seurat package[12] was used to cluster cells from a given dataset unsupervisedly. Using the supervised CHETAH method [13] which compares each cell in a dataset to a predetermined reference compendium using the 500 most variable genes, we first carried out automatic cell annotation to assign cell types to each cluster. With the exception of the categorisation confidence level, which we set to a more permissive value of 0.05, we employed standard settings. We then manually annotated each cluster using these automatic annotations, which included the most common CHETAH annotation per cluster and aneuploidy over diploidy levels from copyKAT, along with a list of cell type-specific markers obtained from bibliographic searches. As a result, we followed three cell type resolution levels. Ten major cell types, including T cells and fibroblasts, were categorised in the lowest resolution, known as "annotation major." We increased the resolution of immune cell types in "annotation immune," such as CD4 and CD8 T cells.

Lastly, we applied even higher resolution to myeloid and lymphoid cell subtypes in "annotation minor" resulting in a total of 17 cell subtypes. To be loaded into and shown by the web site outlined below, all normalised and annotated datasets were saved as Seurat objects and transformed into h5ad files by sceasy[15]

2.5. Ranking by gene, cell cluster, and dataset

We calculated three parameters in order to rank genes according to their specificity for a particular cell cluster. The first metric compares the expression of each gene in the cells from a cluster of interest to the expression of each gene in all other cells in the sample using Holm-corrected nonparametric Wilcoxon rank sum test P values. In order to expedite the computation of related P values, bigger datasets were downsampled to 20,000 cells at random. We calculated log fold changes for each gene as a second metric, comparing the average expression of the gene across cells from a cluster of interest to the average expression of the gene across all other cells in the dataset. The R-based genesorteR tool (bioRxiv 10.1101.676379) was used with default options to enable users to find datasets where a gene of interest is specifically expressed in a cluster or cell type. First, GenesorteR calculates an entropy-based score for each gene and each cluster across all datasets in the database. The closer the score is to 0, the more exclusively a gene is expressed in all cells from a single cluster. The program then assigns an entropy score to each cluster, ranking every gene. The best rank that the gene of interest attained across all of the dataset's clusters is then returned for each dataset. A dataset that contains a cluster for which the gene of interest received an entropy score close to 0 will be ranked near the top, while a dataset where the gene of interest is widely expressed across all of its clusters will be ranked near the bottom. These best ranks are then used to sort the datasets. Cell count, entropy gene index, expression and differential expression findings for CHETAH, main, immune, minor, and authors annotation, metadata of the entire dataset and subsampled h5ad, and metadata object have all been precomputed for quick data browsing.

2.6. Analysis of statistics

In every box-plot, the median value is shown by a horizontal line, and the boxes reflect the interquartile range. Outliers are shown by colored dots, while whiskers reach the furthest data point within a maximum of 1.5 times the interquartile range.

2.7. Data availability

Access to data GEO, ArrayExpress, EGA, and BioProject provide all of the public dataset's that we collected for IMMUcan scDB. The public datasets that were processed and incorporated into the database, together with all accession codes.

3. Results

3.1. Development of the IMMUcan scRNA-seq database based on literature

The creation of the IMMUcan scRNA-seq database (scDB) involved four primary steps: In order to find human cancer scRNA-seq studies, (i) a thorough literature search was conducted; (ii) each pertinent article was manually reviewed and curated; (iii) the corresponding datasets were gathered via web repositories or by contacting the authors; (iv) the datasets and all related metadata were processed and integrated into the IMMUcan scDB (**Figure 2 A**).

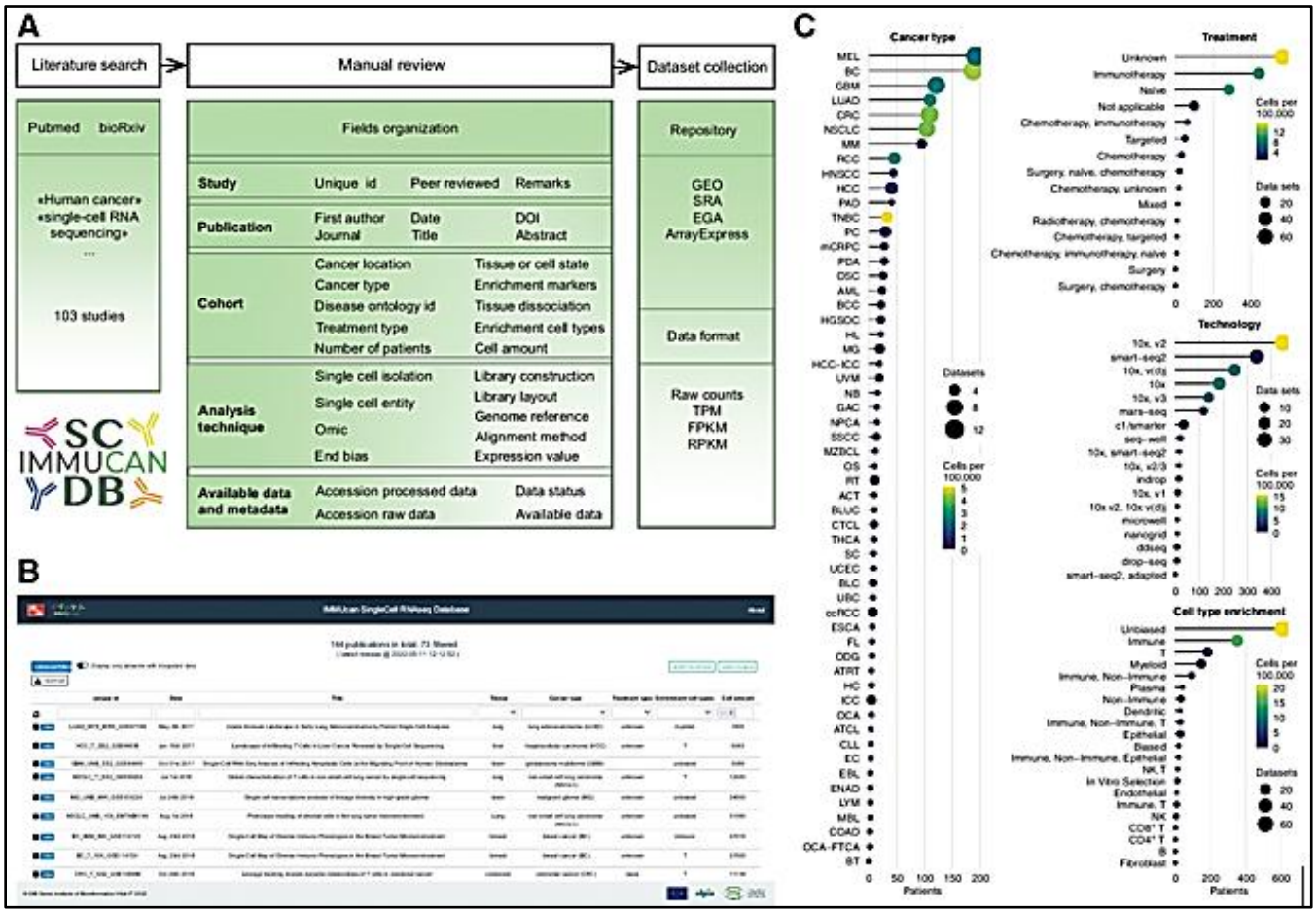

Figure 1: Development of the IMMUcan scRNA-seq database based on literature The creation of the IMMUcan scRNA-seq database (scDB) involved four primary steps: In order to find human cancer scRNA-seq studies, (i) a thorough literature search was conducted; (ii) each pertinent article was manually reviewed and curated; (iii) the corresponding datasets were gathered via web repositories or by contacting the authors; (iv) the datasets and all related metadata were processed and integrated into the IMMUcan scDB (**Figure 1 A**).

A user-friendly interface allows users to search the IMMUcan scDB by integrating all available annotations, data, and metadata. This allows users to query datasets based on annotated metadata, such as cancer type, treatment type, or the presence of a specific cell type (**Figure 1 B**). Datasets can also be ranked by the user according to the precise expression of a particular gene of interest in a subset of cell types or clusters. After selecting a dataset, the user can display the expression of several genes across the clusters or against one another, as well as mine the data and visualise the cell types and clusters it contains. Additionally, the standardised and normalised (batch corrected) files are available for download. The successful integration of 103 publications and their associated information into our database resulted in 144 datasets (**Figure 1 B**). There were fifty-six forms of cancer. The most common cancers were

glioblastoma (10 datasets, 121 patients; **Figure 1 C**), breast cancer (12 datasets, 187 patients), and melanoma (13 datasets, 192 patients with melanoma). Acute T-cell leukaemia, renal cell carcinoma, and other juvenile tumours such as medulloblastoma were less common tumor forms. T-cell enrichment (15 datasets; **Figure 1 C**), immune cell selection via CD45 enrichment (23 datasets), and single-cell suspensions without any prior enrichment (unbiased; 61 datasets) comprised the majority of the datasets.

In total, 21 distinct kinds of enrichment methods were used in the various investigations. 56% of the datasets, or 61% of the patients, had known and detailed patient treatment. This data makes it possible to perform particular analysis, such determining the type of cell and transcriptome alterations linked to particular cancer therapies. Last but not least, the database includes data produced by eleven distinct single-cell sequencing methods, with the majority of research using 10X Genomics or SMART-seq2 single-cell sequencing (**Figure 1 C**).

3.2. Investigation of the IMMUcan scDB

Depending on cell type we first concentrated on the cell type-specific use case of identifying cell types overrepresented in ICI treatment responders versus non-responders in order to illustrate the value of the IMMUcan scDB. In order to achieve this, we looked through the scDB interface for datasets pertaining to patients who had received immunotherapy.

The melanoma dataset MEL_IMM_SS2_GSE120575 was chosen since it has a thorough set of about 17,000 TME cells from patients both before and after anti-PD-1 therapy. When a dataset is selected, a panel displaying a UMAP visualisation of the data is immediately displayed. Cells in this figure can be colored based on different levels of annotation, such as the patient's treatment, the tissue from whence they originated, or the automatic CHETAH cell type assignments (**Figure 2 A**).

Cell groups can be chosen and deselected, and group names and sizes are shown in the interactive legend. A pulldown menu adjacent to the UMAP plot enables you to restrict the display to a random selection of 10,000 cells, which speeds up plotting. A stacked bar chart displays the relevant cell type composition of each study sample next to the UMAP plot for a specific annotation, such as CHETAH cell type assignments. The cell type composition of responders and non-responders can be compared by creating numerous plots using a clinical annotation, such as "treatment response," on top of the bar chart (**Figure 2 B**). B cells were elevated in melanoma patients responding to anti-PD-1 therapy, according to our bar chart presentation.

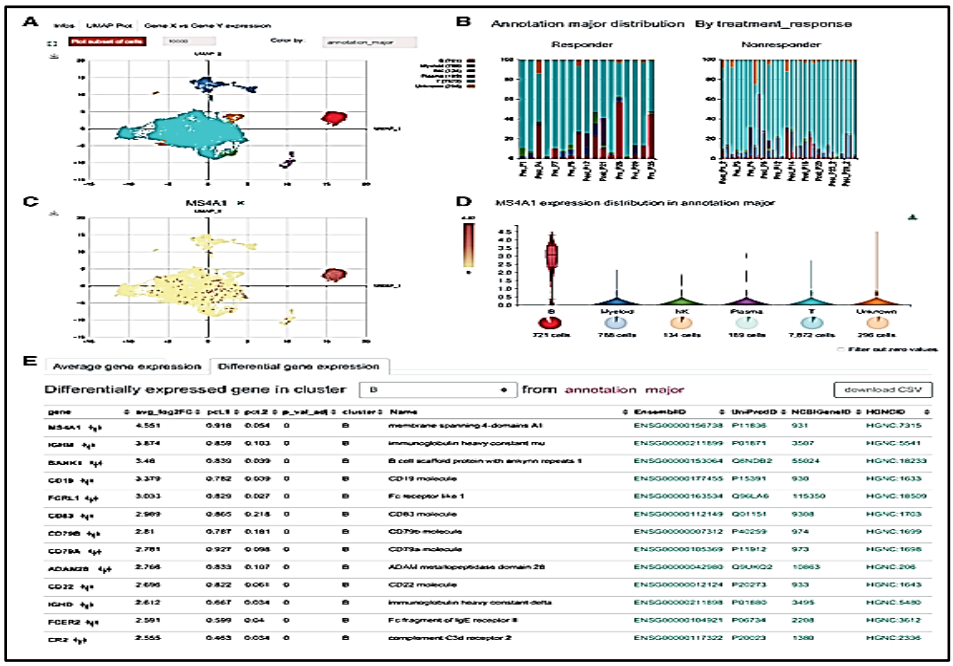

Figure 2: Cell-based exploration of IMMUcan scDB looking at B cells involvement in melanoma patients treated with anti–PD-1. A; UMAP plot of MEL_IMM_SS2_GSE120575 dataset. The cells are colored according to their major annotation. B, Bar plots of the percentage of cells per cell types in the whole dataset and per response to treatment status. The cell types are colored according to the major annotation. C and D, Expression of MS4A1, a marker genes of B cells, visualised on a UMAP plot **(**C**)** and violin plot **(**D**).** Below violin plots are pie charts representing the proportion of expressing cells (“non-zero”) and below the absolute number. The colours correspond to major annotation. **E,** Table of the average expression of genes and differential expression of genes. Here, the top differentially expressed genes of B cells are ranked by descending average log2-fold change compared with the rest of the cells in the dataset.

The average expression of each gene for each cell type and the degree of differential expression of each gene between a user-selected cell type and every other cell in the dataset are displayed in two gene tables that are automatically loaded at the bottom of the page. Each gene's expression can be shown as a violin plot (**Figure 2 D**) and on an extra UMAP plot (**Figure 2 C**).

The proportion of cells expressing the gene is shown in a mouseover (**Figure 2 C**), and the absolute cell number is shown as a pie chart beneath the violin plot to enhance the readability of these plots. Additionally, a different UMAP plot shows the expression of the chosen gene. The most selective gene expression marker for B cells is MS4A1 (CD20), as demonstrated by our visualisation of the expression of the top marker gene of B cells (**Figure 2 E**).

3.3. Searching the IMMUcan scDB using genes

CXCL13 and CXCL9 may be employed as predictive biomarkers for checkpoint immunotherapy response, according to a recent analysis of many bulk transcriptomic cancer datasets.[16] We used IMMUcan scDB as a use case for a gene-centric search to determine which cell types expressed these two genes in various cancer types. The IMMUcan scDB shows a heat-map of the gene's average expression in each cell type in each dataset in which the gene is expressed when a gene is entered in the corresponding search field located at the upper right corner of the entry page.

We looked for CXCL13 and used "annotation minor" as a cell type resolution. We found that it is most highly expressed in exhausted CD8+ T cells (CD8+ T ex) and T follicular helper cells (Tfh) in the majority of cancer indications, such as non-small cell lung cancer (NSCLC; Fig. 3A–C), melanoma (MEL), and basal cell carcinoma (BCC). In contrast, and consistent with current research, CXCL9 was discovered to be expressed in all myeloid cells, with the highest amounts observed in LAMP3-positive dendritic cells and macrophages[17]from a variety of causes, such as melanoma, NSCLC, and hepatocellular carcinoma (HCC).

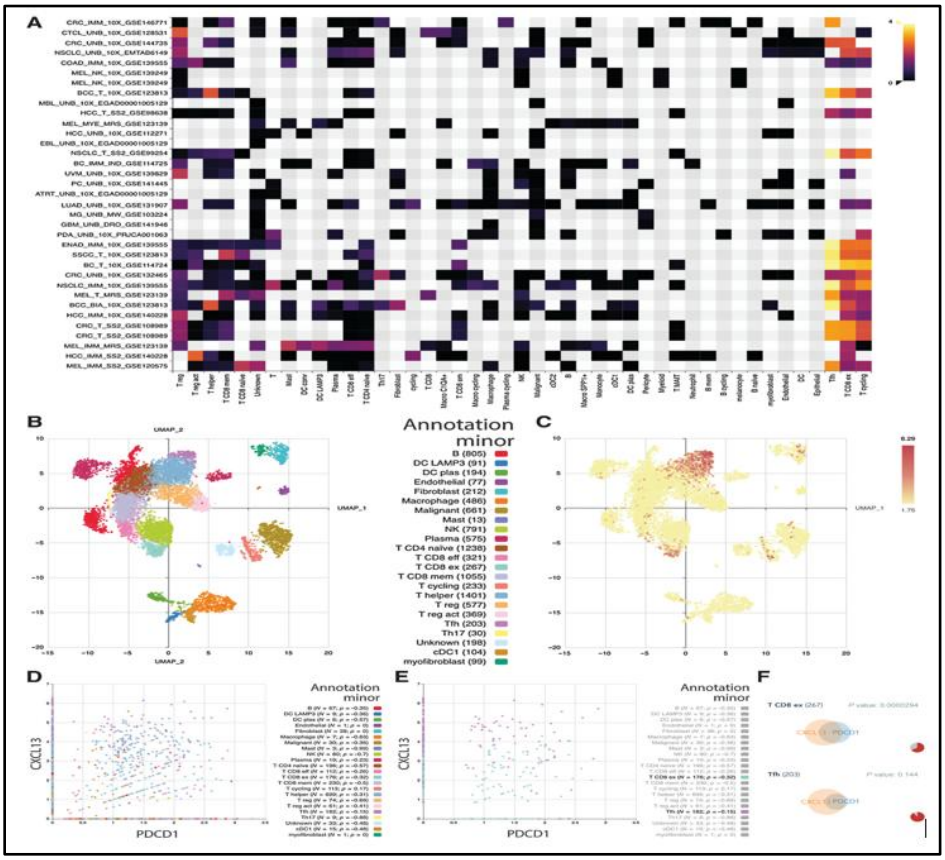

Figure 3: Gene-based exploration of the IMMUcan scDB using CXCL13, a predictive biomarker for immunotherapy response. A, Heat-map of CXCL13 expression across datasets (y-axis) and annotation minor cell types (x-axis). B and C, UMAP plots of BCC_BIA_10X_GSE123813 dataset colored by cell type (minor annotation; B) and CXCL13 expression (C). D and E, Coexpression plot of CXCL13 and PDCD1 (PD1), cells are colored according to the minor annotation displaying all cell types (D) and only exhausted CD8+ T cells (T CD8 ex) and Tfh (E). The legend indicates the cell type with the number of expressing cells and the Pearson correlation coefficient in brackets. F, Venn diagram showing the coexpression of CXCL13 and PDCD1 by T CD8 ex (top) and Tfh (bottom). The P value of a hypergeometric test is shown in the top-right corner of each plot; a pie chart representing the proportion of expressing cells for one of the two genes is in the bottom-right corner of each plot.

3.4. Assessing the IMMUcan scDB's gene coexpression

The co-expression of two genes can also be measured using the IMMUcan scDB. In order to accomplish this, the user can choose the "Gene X vs. Gene Y expression" panel and input the names of two genes after choosing a dataset on the entry page. One point is then added to each individual cell in a scatter plot. The chosen cell type resolution level determines the colour of the cells. All cell types that express both genes are listed in the legend along with the number of cells that correspond to each gene and the corresponding Pearson correlation coefficients. To illustrate the coexpression features, we chose the BCC study BCC_BIA_10X_GSE123813 because of the high expression of CXCL13 in exhausted CD8+ T cells in this dataset. We also looked into the coexpression of CXCL13 with PD1 (PDCD1), another well-known marker for T-cell exhaustion, and found highly significant coexpression of the two genes in exhausted CD8 T cells with an overlap P value of $2.9\times10-5$ (**Figure 3 D–F**). As CXCL9 expression was limited to cells of myeloid origin, CXCL9 and CXCL13 were not coexpressed in any of the cell types from the BCC_BIA_10X_GSE123813 dataset, which is consistent with the findings of the aforementioned gene-based search. Consequently, the scatter plot reveals no cells expressing both genes, and the Venn diagrams reveal no overlap between the CXCL9 and CXCL13 positive cells.

3.5. Finding typical alterations in the frequencies of cell types linked to malignant transformation

Next, we took use of the availability of a sizeable collection of harmonised single-cell datasets to find T-cell and macrophage compartment alterations that were consistently seen across the TME of various cancer types. The 25 datasets from the IMMUcan scDB that included both tumor and normal tissue were chosen for this purpose. We next compared the variations in T-cell and macrophage subtype frequencies with the corresponding variations in gene expression patterns linked to malignant transformation. With log-fold change differences larger than 1 and a multiple testing corrected P value of 0.001, we discovered 705 up-regulated and 611 down-regulated genes across 11 cell types. Numerous top DE genes were associated with tissue-specific genes like alveolar and surfactant genes, as well as dissociation-artefacts[18] like heat shock proteins. In order to eliminate all genes linked to these effects from further analysis, we developed a gene blacklist. Heat-shock proteins, immunoglobins, mitochondrial genes, tissue-specific genes, ribosomal genes, ERCC spike-ins, and other dissociation-associated genes like DNAses, FOS, and JUN were all included in the gene blacklist. Genes that were detected in fewer than 20% of the datasets were excluded, and the genes that resulted from differential expression were ranked according to the average fold change across all datasets (**Figure 4 A**).

The TME is linked to a sharp rise in regulatory T cells (Treg) in a number of cancer types, including NSCLC, CRC, and HCC, as illustrated in **Figure 4 A**. Activated Tregs, in particular, seem to be almost entirely found in the TME and essentially nonexistent in the comparable normal tissues. On the other hand, the proportion of naïve CD4 and CD8 T cells in the TME is consistently lower than that of normal tissue (**Figure 4 A**). According to the observed shifts in cell type frequencies, we find that genes linked to naïve T cells, like TCF7, have much lower expression in the TME, whereas genes linked to T-cell activation, like TNFRSF4, TNFRSF9, and TNFRSF18, and T-cell exhaustion, like HAVCR2 and CTLA4, seem to have higher expression (**Figure 4 B**).

Furthermore, in addition to CXCL13, which attracts B cells, we also uncover very tumor-specific markers for activated Tregs, such as CCR8 and LAYN. Consistent with this finding, anti-CCR8 antibodies are presently being evaluated in clinical trials for Treg depletion strategies after CCR8 was recently discovered to be a tumor Treg-specific target.[19] SPP1 and APOE are strongly up-regulated in the TME of macrophages (**Figure 4 B**). It's interesting to note that while most other indications, such CRC and breast cancer, exhibit sharp up-regulations of both genes, this up-regulation is absent in tumor indications like HCC and glioblastoma (**Figure 4 C**). In NSCLC, SPP1 is thought to mediate pro-inflammatory pathways and the immunotherapy response[20] whereas in pancreatic cancer, APOE is thought to encourage immune suppression.[21] Given that we find these genes to be significantly unregulated in tumor-associated macrophages across almost all datasets, it is possible that their impact on immune suppression is more extensive than previously thought.

Finally, we present the IMMUcan scDB, a curated, easily searchable, and explorable database of scRNA-seq studies of the human TME. We demonstrated through three use cases that the IMMUcan scDB is a useful tool for generating new ideas, validating observations from the literature, and offering fresh biological insights.

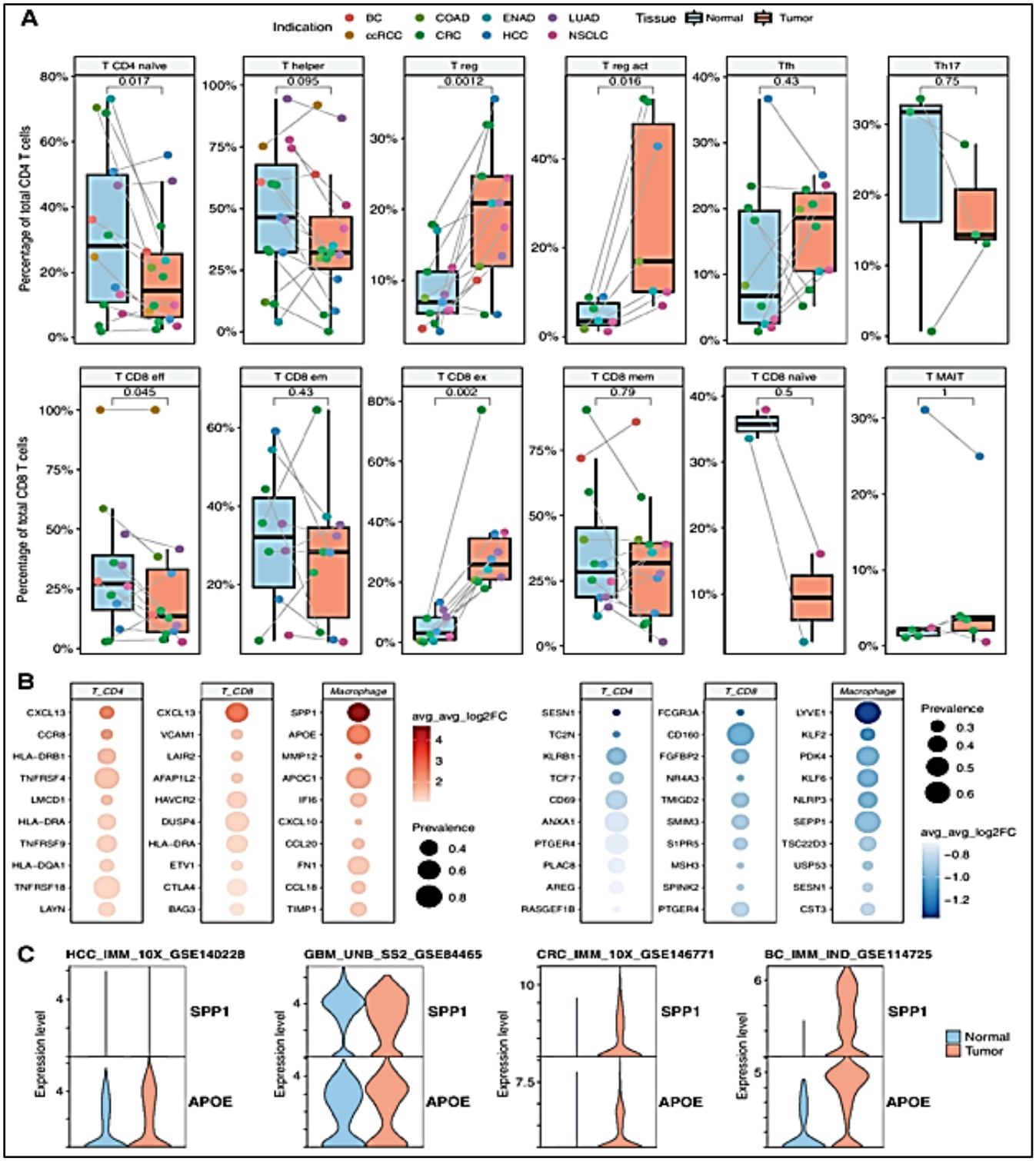

Figure 4: Transcriptional changes between normal and tumor-associated immune cells of 25 datasets from the IMMUcan scDB. **A,** Composition of CD8 (top) and CD4 (bottom) T-cell subtypes in normal tissue and TME. Every dot represents one dataset, and the gray line represents samples from the same dataset. Paired Wilcoxon-ranked sum test, Bonferroni corrected. **B,** Top 10 up-regulated and down-regulated genes between normal and tumor-associated cells in a selection of cell types. Genes ranked by the average $\log_2$-fold change over all datasets and filtered for a prevalence of detection as differentially expressed gene in at least 20% of the datasets. **C,** Log normalised expression of SPP1 and APOE between macrophages from matched normal and tumor samples in four selected datasets. CRC, colorectal cancer.

4. Discussion

In recent years, the quantity of scRNA-seq research on human cancer has skyrocketed. From common tumor types (melanoma, breast cancer, and non-small cell lung cancer) to uncommon cancers like atypical teratoid rhabdoid tumor[22] or less common molecular subtypes like triple-negative breast cancer,[23] the initial studies offered a comprehensive description of tumor cells and TME (also known as the "atlas" view). As more patients, samples, and cells are included in scRNA-seq "atlas" studies, we expect them to progressively concentrate on an even wider range of tumor types. In parallel with these descriptive investigations, scRNA-seq has recently been used to determine immune checkpoint inhibitor responsiveness,[25] or mechanisms of resistance.[24] As scRNA-seq technologies become more widely used and accessible, the quantity and scope of these hypothesis-driven investigations should likewise rise. Comparing several anatomical areas, such as the location of primary vs metastatic tumours, is another kind of study design.[26]

The quantity and variety of scRNA-seq research warrants a website that is entirely devoted to human cancer datasets, with comprehensive annotation, user-friendly and effective search features, and several applied meta-analysis techniques. In our opinion, this is the best approach to handle the hundreds of datasets that are expected to be generated in the upcoming years. The projected integration of recently released datasets in accordance with the standardised approach that we have developed will receive special attention in this regard. We will update the database once a month as part of the IMMUcan consortium. Large-scale ("omics") datasets, particularly those related to transcriptomics and genomics, are becoming more and more accessible through public data repositories. However, clinical annotation—such as the tumor type—is frequently absent or restricted to a minimal amount of information. This significantly reduces the potential for combining biological and clinical data in the analysis and interpretation process.

This kind of annotation is absent from single-cell portals like the Broad Institute Single Cell Portal, UCSC Cell Browser,[27] and the single-cell expression atlas.[28] Tumor type, primary or metastatic stage, and treatment type are the only clinical details included in cancer scRNA-seq databases like CancerSea (5) or TISCH (6). In our research, we manually extracted and mapped comprehensive clinical features (9 items) linked to each patient cohort and dataset to reference ontologies. IMMUcan scDB is the sole database devoted exclusively to human cancer single-cell transcriptomic datasets when compared to the other resources (Table 1). It incorporates data from 144 investigations, comprising 73 datasets. Of all the resources now available, the tumor clinical annotations are among the most comprehensive. It is the sole database with interactive graphs (enabling to display graph according to clinical features of relevance, such as splitting graph according to tissue type, treatment, or patients) and offers the majority of the functionalities provided by other resources. This should enable physicians and biologists to compare datasets across clinical contexts and concentrate on datasets that match to a specific clinical circumstance. Important information on cell kinds, cell states, and related signatures should also be included.

Even in single samples, scRNA-seq produces data from a vast number of cells, in contrast to bulk transcriptomics. This presents the potential for thorough characterisation of cellular clusters and related gene expression programs in individual patients, provided that cell counts are enough. To find unifying patterns linked to a tumor kind, a particular anatomical region, or a therapeutic effect, it is also crucial to aggregate the analysis of multiple datasets that meet common criteria. Using unique scRNA-seq datasets from four cancer types, a recent work created a "pan-cancer blueprint" of stromal cell heterogeneity.[29] It showed that invading immune cells had similar gene expression programs. To find common trends and improve statistical relevance to a particular clinical situation, we have integrated several samples using proven approaches in our IMMUcan scDB. Consequently, users can utilise targeted tactics on specific patient samples. There are several opportunities for biomedical applications with the IMMUcan scDB. Through exploratory analysis, hypotheses for additional validation can be generated during an early discovery process. When cell type-specific signatures from various clinical settings are compared, for instance, intriguing mechanisms of immune escape or activation or new therapeutic targets may become apparent.

Table 1: Comparison of content and functionalities of seven resources gathering human single-cell transcriptomics datasets.

		scRNASeqDB	*SCPortalen*	*Jingle Bells*	*Cancer SEA*	*TISCH*	*IMMUcanDB*
Data	Number of datasets	38	66	302	74	79	144 [73][a]
	Number of human oncology datasets	3	2	14	20	75	144 [73][a]
	Number of criteria for datasets query	3	5	3	2	9	7+
Sample type	Cell lines	X	X	X	X	–	–
	Xenograft	X	X	X	X	–	–
	Mouse samples	–	X	X	X	X	–
	Human tissues/blood	X	X	X	X	X	X
Clinical annotations	Tumor type	–	–	–	X	X	X
	Tissue site	–	–	–	X	X	X
	Primary vs metastatic	–	–	–	–	X	X
	Treatment type	–	–	–	–	X	X
	Response to treatment	–	–	–	–	X	X
	Cell enrichment strategy	–	–	–	–	X	X
Availability of processed data	BAM	–	X	X	–	–	–
	Average gene expression per cell type	–	X	–	X	X	X
	Differentially expressed genes	X	–	–	–	X	X
	Single-cell object	–	–			–	X
Gene specific functionality across datasets	Gene expression distribution	X	X	–	–	X	X
	Signature expression among datasets	–	–	–	–	X	–
	Dataset filtering	X	–	–	–	X	X
Visualisation	UMAP	X	X	–	–	X	X
	Cluster proportion	–	–	–	–	X	X
	Gene signature expression	–	–	–	–	X	X
	Interactive graphs	–	–	–	–	–	X

[a]*Datasets with integrated data.*
[b]*Only for integrated reference atlas.*

Finally, our database can be used to validate results established in an independent study. On the other hand, hypothesis-driven analyses may establish the expression pattern of specific genes or signatures according to different annotation terms. The growing number of scRNA-seq datasets provides unique opportunities for cross-validation of results from various technologies, including proteomics, genomics, and spatial transcriptomics. A significant step forward would be the improvement and generalisation of standardised terminologies, such as the human disease ontology[9] and cell ontology[30] as well as a more systematic and thorough clinical annotation within existing genomics data repositories, along with a unified data storage procedure. Because sample quality and dataset annotation depend on the quality of the information provided in the original publication, integrating so many scRNA-seq datasets into a single database carries potential risks and limitations. It is undoubtedly difficult and prone to technical biases to handle scRNA-seq datasets produced in separate research employing different tissue dissociation and enrichment techniques, as well as perhaps distinct technology platforms. We have incorporated reliable and verified procedures at every stage of our processing chain. Equilibrium is the technique we have used to lessen experimental bias while integrating various datasets. Reiterative clustering is used by Harmony to eliminate batch effects between patients and studies. Harmony performed well in recent benchmark studies on scRNA-seq data integration[31-33] and because of its strong performance, it is suggested as an integration method above techniques like CCA[34], Liger[35] and UMI downsampling (bioRxiv 2021.11.15.468733).

Furthermore, we evaluated this on a range of IMMUcan scDB datasets and found no significant differences between harmony and other integration techniques (Supplementary Fig. S4). Users may employ their own cross-validation techniques to strengthen the validity of their findings, but they should be mindful of the restrictions and potential biases. In the upcoming years, enhancing the efficiency of our data processing will continue to be a primary focus. All things

considered, we think that the strength and potential provided by combining so many datasets greatly exceeds the drawbacks and restrictions that come with meta-analysis. We anticipate that this resource will make it easier to use publicly accessible scRNA-seq datasets to tackle both new and current issues in the study of human cancer.

5. Author's Disclosure

No disclosures were reported by the other authors.

6. Source of Funding

None.

7. Conflict of Interest

None.

References

1. Cao Y , Zhu J, Jia P, Zhao Z . scRNASeqDB: a database for RNA-seq based gene expression profiles in human single cells. *Genes* 2017;8:368.
2. Abugessaisa I , Noguchi S , Böttcher M , Hasegawa A , Kouno T , Kato S, et al. SCPortalen: human and mouse single-cell centric database. *Nucleic Acids Res*. 2018;46:D781–7.
3. Franzén O , Gan LM , Björkegren JLM . PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database. 2019;2019:baz046.
4. Ner-Gaon H, Melchior A, Golan N, Ben-Haim Y, Shay T. JingleBells: a repository of immune-related single-cell RNA sequencing datasets. *J Immunol*. 2017;198:3375–9.
5. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G. et al. CancerSEA: a cancer single-cell state atlas. *Nucleic Acids Res*. 2019;47:D900–8.
6. Sun D, Wang J, Han Y, Dong X, Ge J, Zheng R, et al. TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualisation of tumor microenvironment. *Nucleic Acids Res*. 2021;49:D1420–30.
7. Füllgrabe A , George N , Green M , Nejad P, Aronow B , Fexova SK. et al. Guidelines for reporting single-cell RNA-seq experiments. *Nat Biotechnol*. 2020;38(12):1384–6.
8. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019;15(6):e8746.
9. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T. et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* 2011;39:W541–5.
10. Osorio D, Cai JJ. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA sequencing data quality control. *Bioinformatics.* 2021;37(7):963–7.
11. Korsunsky I, Millard N, Fan J, Slowikowksi K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*. 2019;16(12):1289–96.
12. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. *Cell*. 2019;177:1888–902.
13. de Kanter JK , Lijnzaad P , Candelli T , Margaritis T , Holstege FCP. CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Res*. 2019;47(16):e95.
14. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumours from single-cell transcriptomes. *Nat Biotechnol*. 2021;39(5):599–608.
15. Cakir B, Prete M, Huang N, van Dongen S, Pir P, Kiselev VY. Comparison of visualisation tools for single-cell RNA- seq data. *Nucleic Acids Res.* 2020;2:lqaa05
16. Litchfield K, Reading JL, Puttick C, Thakkar K, Abbosh C, Bentham R. et al. Meta-analysis of tumor- and T cell–intrinsic mechanisms of sensitisation to checkpoint inhibition. *Cell* 2021;184(3):596–614.
17. Park MK, Amichay D, Love P, Wick E, Liao F, Grinberg A. et al. The CXC chemokine murine monokini induced by IFNγ (CXC chemokine ligand 9) is made by APCs, targets lymphocytes including activated B cells, and supports antibody responses to a bacterial pathogen in vivo. *J Immunol*. 2002;169(3):1433–43.
18. Machado L, Geara P, Camps J, Dos Santos M, Teixeira-Clerc F, Van Herck J, et al. Tissue damage induces a conserved stress response that initiates quiescent muscle stem cell activation. *Cell Stem Cell*. 2021;28(6):1125–35.
19. Campbell JR , McDonald BR , Mesko PB , Siemers NO , Singh PB, Selby M. et al. Fc-optimized anti-CCR8 antibody depletes regulatory T cells in human tumor models. *Cancer Res.* 2021;81(11):2983–94.
20. Leader AM, Grout JA, Maier BB, Nabet BY, Park MD, Tabachnikova A. et al. Single-cell analysis of human non–small cell lung cancer lesions refines tumor classification and patient stratification. *Cancer Cell.* 2021;39(12):1594–609.
21. Kemp SB, Carpenter ES, Steele NG, Donahue KL, Nwosu ZC, Pacheco A. et al. Apolipoprotein E promotes immune suppression in pancreatic cancer through NF-κB–mediated production of CXCL1. *Cancer Res* 2021;81(16):4305–18.
22. Jessa S, Blanchet-Cohen A, Krug B, Vladoiu M, Coutelier M, Faury D. et al. Stalled developmental programs at the root of pediatric brain tumors. *Nat Genet.* 2019;51(12):1702–13.
23. Kim C , Gao R , Sei E , Brandt R , Hartman J , Hatschek T. et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. Cell 2018;173(4):879–93.
24. Jerby-Arnon L, Shah P, Cuoco MS, Rodman C, Su MJ, Melms JC. et al. A cancer cell program promotes T-cell exclusion and resistance to checkpoint blockade. *Cell*. 2018;175(4):984–97.
25. Sade-Feldman M, Yizhak K, Bjorgaard SL. Defining T-cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*. 2018;175(4):998–1013.
26. Puram SV, Tirosh I, Parikh AS, Ray JP, de Boer CG, Jenkins RW. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell.* 2017;171(7):1611–24.
27. Speir ML Bhaduri A, Markov NS, Moreno P, Nowakowski TJ, Papatheodorou I. et al. UCSC cell browser: Visualise your single-cell data. Bioinformatics; 2021;37:4578–80.
28. Papatheodorou I, Moreno P, Manning J, Fuentes AMP, George N, Fexova S, et al. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* 2019;48(1):77–83.
29. Qian J, Olbrecht S, Boeckx B, Vos H, Laoui D, Etlioglu E. et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. Cell Res 2020;30(9):745–62.
30. Meehan TF, Masci AM , Abdulla A , Cowell LG , Blake JA , Mungall CJ. et al. Logical development of the cell ontology. *BMC Bioinf* 2011;12:6.
31. Luecken M , Büttner M , Chaichoompu K , Danese A , Interlandi M, Mueller MF. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods*. 2022;19(1):41–50.
32. Chazarra-Gil R, van Dongen S, Kiselev VY, Hemberg M. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic Acids Res.* 2021;49(70:e42.
33. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol*. 2020;21(1):12.
34. 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.

35. Welch JD, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko EZ. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell.* 2019;177(7):1873–87

Cite this article: Dahal T, Saurabh S. Utilising the IMMUcan database for meta- assessment of human cancer single-Cell RNA-seq databases. *Southeast Asian J Health Prof* 2025;8(3):71-82