Meta-Analysis of Human Cancer Single-Cell RNA-Seq Datasets Using the IMMUcan Database

Jordi Camps[1], Floriane Noël[2], Robin Liechti[3], Lucile Massenet-Regad[2,4], Sidwell Rigade[2], Lou Götz[3], Caroline Hoffmann[5], Elise Amblard[2,6], Melissa Saichi[2], Mahmoud M. Ibrahim[7], Jack Pollard[8], Jasna Medvedovic[2], Helge G. Roider[9], and Vassili Soumelis[2,10,11]

ABSTRACT

The development of single-cell RNA sequencing (scRNA-seq) technologies has greatly contributed to deciphering the tumor microenvironment (TME). An enormous amount of independent scRNA-seq studies have been published representing a valuable resource that provides opportunities for meta-analysis studies. However, the massive amount of biological information, the marked heterogeneity and variability between studies, and the technical challenges in processing heterogeneous datasets create major bottlenecks for the full exploitation of scRNA-seq data. We have developed IMMUcan scDB (https://immucanscdb.vital-it.ch), a fully integrated scRNA-seq database exclusively dedicated to human cancer and accessible to nonspecialists. IMMUcan scDB encompasses 144 datasets on 56 different cancer types, annotated in 50 fields containing precise clinical, technological, and biological information. A data processing pipeline was developed and organized in four steps: (i) data collection; (ii) data processing (quality control and sample integration); (iii) supervised cell annotation with a cell ontology classifier of the TME; and (iv) interface to analyze TME in a cancer type–specific or global manner. This framework was used to explore datasets across tumor locations in a gene-centric (*CXCL13*) and cell-centric (B cells) manner as well as to conduct meta-analysis studies such as ranking immune cell types and genes correlated to malignant transformation. This integrated, freely accessible, and user-friendly resource represents an unprecedented level of detailed annotation, offering vast possibilities for downstream exploitation of human cancer scRNA-seq data for discovery and validation studies.

Significance: The IMMUcan scDB database is an accessible supportive tool to analyze and decipher tumor-associated single-cell RNA sequencing data, allowing researchers to maximally use this data to provide new insights into cancer biology.

Introduction

Tumor immunology has taken central stage in cancer research due to the relative success of immunotherapy in a large number of malignancies. However, despite recent progress, the majority of cancer patients still either does not respond to therapy or eventually relapses and succumbs to disease. Aside from the tumor cells themselves, the tumor microenvironment (TME) has been shown to strongly influence clinical outcome of immunotherapies. Better characterizing the cellular composition and molecular characteristics of the TME thus remains an important and challenging task that could help not only develop novel anticancer strategies but also to identify biomarkers, better predicting outcome to current immunotherapies, leading to optimized personalized treatment strategies.

Single-cell RNA sequencing (scRNA-seq) technologies are uniquely suited to explore the diversity of cellular phenotypes and molecular pathways present in the TME. They can help addressing an array of biomedical questions, ranging for identifying cancer-associated cell states, to predicting intercellular communication, disease resistance mechanisms, and discovering novel drug targets. The ever-growing number of cancer related scRNA-seq datasets published in recent years represent a highly valuable but only partially explored treasure trove for biomedical research, given that in most published articles the authors have addressed only a limited number of hypotheses and have not integrated their data with other complementary studies.

While integrating tumor derived single-cell data into a searchable database would facilitate access, reanalysis, and comparing published scRNA-seq datasets doing so is challenging for several reasons: (i) cancer related datasets are highly heterogenous due to the large number of different cancer types and clinical contexts such as treatment type and tumor location; (ii) applied single-cell technologies, experimental protocols, and data analysis methods; (iii) biological and clinical interpretation of the results.

To address this challenge, scRNA-seq data portals have been created recently, including scRNASeqDB (1), SCPortalen (2), PanglaoDB (3), and JingleBells (4). However, only two databases, CancerSEA and TISCH, are dedicated to hosting tumor related data. CancerSEA has integrated 41,900 single cancer cells from 25 cancer types (5) and is focused on identifying functional states associated with specific gene signatures. It combines datasets from human tumors, but also from cancer cell lines and patient-derived xenografts. Clinical information is

[1]Biomedical Data Science, Research & Early Development Oncology, Bayer AG, Berlin, Germany. [2]Université de Paris, Institut de Recherche Saint-Louis, INSERM U976, Paris, France. [3]Vital-IT group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland. [4]Université Paris-Saclay, Saint Aubin, France. [5]Institut Curie, INSERM U932 Research Unit, Department of Surgical Oncology, PSL University, Paris, France. [6]Université de Paris, Centre de Recherches Interdisciplinaires, Paris, France. [7]Biomedical Data Science, Research & Early Development Premedical, Bayer AG, Wuppertal, Germany. [8]Sanofi Research and Development, Cambridge, Massachusetts. [9]Oncology Precision Medicine, Research & Early Development Oncology, Bayer AG, Berlin, Germany. [10]Assistance Publique-Hôpitaux de Paris (AP-HP), Hôpital Saint-Louis, Laboratoire d'Immunologie, Paris, France. [11]Owkin, Paris, France.

J. Camps and F. Noël contributed equally to this article.

Corresponding Authors: Vassili Soumelis, Institut de Recherche St Louis (IRSL), Inserm U976, 26 rue d'Ulm, Paris 75005, France. Phone: 677-721-530; E-mail: vassili.soumelis@aphp.fr; and Helge G. Roider, Bayer AG, Müllerstraße 178, Berlin 13353, Germany. Phone: 152-068-42034; E-mail: helge.roider@bayer.com

Cancer Res 2023;83:363–73

doi: 10.1158/0008-5472.CAN-22-0074

minimal and restricted to tumor type annotation. TISCH enables to browse through cancer scRNA-seq datasets from both human (74 datasets) and mouse (5 datasets) to characterize the various cell types composing the TME and analyzing the expression of target genes and signatures (6). Clinical annotation is limited to tumor type, primary versus metastatic disease, and treatment. The database functionalities allow comparison of the cellular composition and target gene expression across various datasets.

Our objective was to go beyond these efforts, and to build an in-depth, fully annotated, and integrated scRNA-seq database exclusively dedicated to human cancer. This study was performed in the context of the "Integrated iMMUnoprofiling of large adaptive CANcer patient cohorts" (IMMUcan) consortium, as part of the European Innovative Medicine Initiative 2 program and aims at creating and integrating the clinical, cellular and molecular profile of different tumor types and their microenvironment. The IMMUcan database offers detailed clinical annotation allowing to connect cell types and gene expression patterns to specific clinical patterns. It further offers a large number of functionalities for the analysis of multiple datasets. We hope the database will become the gold standard reference tool to support cancer biomedical research, in the early discovery, hypothesis-generating, as well as validation settings.

Materials and Methods

Literature search and dataset selection

We searched for peer-reviewed published datasets in PubMed (ncbi.nlm.nih.gov/pubmed/) using ((cancer[Title/Abstract]) AND (patient)) AND (single cell RNA sequencing) as key words as well as for non–peer-reviewed studies in the bioRxiv database (www.biorxiv.org) using "human cancer single-cell rna-sequencing" as free-text keywords. We applied a filter to select articles published from 2016 to 2021 and then manually reviewed all the resulting article titles and abstracts to check for the availability of scRNA-seq data, which resulted in a total of 103 publications spanning 144 datasets. For all datasets from human cancer patients with more than a thousand cells and at least ten samples, we downloaded the data from Gene Expression Omnibus (GEO), ArrayExpress, EGA, and BioProject. An exception concerning the number of patient samples was made when datasets focused on additional biology such as multiple biopsy sites, treatment information or longitudinal samples. Using these filters, we finally arrived at 73 datasets covering 56 different cancer indications that were integrated into the IMMUcan scDB.

Capturing of metadata

To structure the data from the 144 selected studies and to allow for efficiently searching our database we extracted the following metadata categories from the studies inspired by the guidelines for reporting scRNA-seq studies (7). The first category captures study wide information including manuscript title, abstract, DOI, number of patients and samples, as well as data access information. The second category focuses on sample-specific attributes such as cancer type, cancer localization, treatment, and response. All information regarding the applied single-cell technology workflow (including tissue dissociation, cell type enrichment, single-cell isolation, library construction, end bias, library layout, reference genome, read alignment, read counting, and expression value format) are part of a third category. Finally, whenever it was provided, we also annotated and standardized information about the single-cell data contained in the study including enrichment strategy, patient ID, tissue, timepoint of biopsy, location of biopsy, author annotations, cancer (sub)type, cancer stage, as well as treatment information including timepoint, drug, and response. Where possible all metadata such as cancer type, treatment, and cell type were standardized and mapped to ontologies. Depending on the type of information the metadata can be either free text, a list from a controlled vocabulary, Boolean values, or quantitative information.

Data processing

To increase the comparability between studies and because raw read counts are better suited for most single-cell analysis workflows (8), we preferentially downloaded this data type whenever available. Every dataset that contained multiple experiments or cancer indications was then split into separate files.

To efficiently process all downloaded single-cell data, we developed an R language-based pipeline, called *scProcessoR*, that mainly uses functions form the *Seurat* package (version 3; ref. 9) for log normalizing the data, selecting the most variable genes, principal component analysis transforming and scaling of the data, building knn neighborhoods for each cell, graph-based clustering and generating UMAP dimensionality reduction plots (Supplementary Fig. S1A). All steps are performed in a semiautomatic manner based on best practices in the field (8, 10). The workflow takes as input an expression matrix formatted to have cells as columns and genes as rows as well as a metadata file containing the cleaned and standardized information available for the samples including patient data and cell annotations. To illustrate the workflow, we used the dataset SC_UNB_10X_GSE134520, which includes single-cell transcriptomic profiles of 9 patients with early stomach carcinoma. For each dataset, to only retain high quality data we removed all cells that have less than 250 genes with mapped reads and/or depending on tumor type (10) contain more than 5% to 20% of mitochondrial specific reads (Supplementary Fig. S1B). To evaluate whether a dataset suffers from severe technical batch effects we computed for each potential type of batch effect, T (e.g., patient ID), and each cell in the dataset, C, the Shannon entropy, H_{TC}, given by:

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

where B is the total number of batches of type T in a dataset (e.g., all patients), and q_C is the percentage of cells within the 30 nearest neighbors of C that come from a given batch b. To end up with a comparable entropy metric between datasets, H_{TCnorm}, we further normalized the entropy of every cell by the total entropy of the dataset H_{Ttotal}

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

H_{Ttotal} is given by:

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

where B is the total number of batches of type T in a dataset (e.g., all patients), and q_b is the percentage of cells from a given batch b.

Values for H_{TCnorm} range from 0 (corresponding to a cell being surrounded only by cells from the same batch), to 1 (corresponding to the 30 nearest neighbors of C coming with equal frequency from all different batches of type T in the dataset). If the median entropy across all cells in a dataset had a value of ≤ 0.5 for a given type of batch effect, we corrected for the corresponding batch effects using the Harmony package (version 0.1; ref. 11) with supplied default parameters (Supplementary Fig. S1C and S1D).

Cell clustering and cell type annotation

Unsupervised clustering of cells from a given datasets was performed using Louvain graph-based clustering implemented in the *Seurat* package (12) with the resolution parameter set to 1. To assign cell types to each cluster, we first performed automatic cell annotation using the supervised *CHETAH* algorithm (13), which uses the 500 most variable genes to compare each cell in a dataset to a predefined reference compendium (Supplementary Fig. S1E). We used standard parameters except for the confidence threshold for classification, which we set to a more lenient value of 0.05. As reference data, we reannotated the integrated human TME scRNA-seq compendium provided with the *CHETAH* package to now also include plasmacytoid dendritic cells as separate cell type (Supplementary Fig. S2A). To delineate malignant cells, we then applied the *CopyKAT* algorithm (14), which uses an assessment of copy number aberrations to identify malignant cells. To increase *CopyKAT*'s prediction accuracy, we provided the macrophage cluster in each dataset as healthy reference cells (Supplementary Fig. S2B).

With the help of these automatic annotations, i.e., the most frequent *CHETAH* annotation per cluster and aneuploidy over diploidy levels from *copyKAT*, as well as a compendium of cell type specific markers derived from bibliographic searches, we then performed manual annotation of each cluster (Supplementary Table S1). We thereby adhered to three levels of cell type resolution. In the lowest resolution referred to as "annotation major," we classified ten main cell types such as fibroblasts and T cells. In "annotation immune," we added higher resolution to immune cell types distinguishing for instance CD4 and CD8 T cells (Supplementary Fig. S2C and S2D). Finally, in "annotation minor," we applied even greater resolution on myeloid and lymphoid cell subtypes (Supplementary Fig. S2E; Supplementary Table S2) giving rise to a total of 17 cell subtypes. All normalized and annotated datasets were stored as Seurat objects and converted to h5ad files by *sceasy* (15) to be loaded into and visualized by the web portal described below.

Gene, cell cluster, and dataset ranking

To be able to rank genes based on their specificity for a given cell cluster we computed three measures. The first metric is based on Holm-corrected nonparametric Wilcoxon rank sum test *P* values comparing for each gene its expression in the cells from a cluster of interest to its expression across all other cells in the dataset. To speed up corresponding *P* value calculations, larger datasets were randomly downsampled to 20,000 cells. As a second metric, we computed for each gene log fold changes between its average expression across the cells from a cluster of interest to its average expression across all other cells in the dataset. Finally, we also determined for each gene the percentage of cells in the cluster of interest that express the gene. In the interface, users can sort genes for each cluster or annotated cell type based on each of these three measures.

To allow users to identify datasets in which a gene of interest is specifically expressed in a cluster or cell type, we applied the R based *genesorteR* package (bioRxiv 10.1101.676379) using default parameters. *GenesorteR* first computes for every gene and every cluster from all the datasets in the database an entropy-based score, the closer this score is to 0 the more exclusively a gene is expressed in all cells from only one cluster. For each cluster, the algorithm then ranks all the genes based on their entropy scores. Finally, it returns for every dataset the best rank that the gene of interest achieved across all the clusters in the dataset. These best ranks are then used to sort the datasets, i.e., a dataset containing a cluster for which the gene of interest obtained an entropy score of close to 0 will be ranked near the top while a dataset where the gene of interest is broadly expressed across all its clusters will be ranked near the bottom.

For fast browsing of the data, the following metrics and rankings have been precomputed: cell count, entropy gene index, expression, and differential expression results for CHETAH, major, immune, minor and authors annotation, metadata of full dataset and subsampled h5ad and metadata object.

Web portal

The front-end of the web portal has been developed using the *VueJS* framework (https://vuejs.org/, version 2.6), the *Bootstrap CSS* library (https://getbootstrap.com/, version 4.6), the *echarts* visualization library (http://echarts.apache.org/, version 4.9) and the *d3js* library (https://d3js.org/, version 5.16). The back-end has been developed with *PHP* (version 7.1) and the *SLIM* framework (https://www.slimframework.com/, version 3.12). Once a dataset is selected by the user the corresponding h5ad file is parsed to the portal via a custom *Python* script using the *scanpy* library (scanpy.readthedocs.io/, version 1.7.2).

Statistical analysis

In all boxplots, boxes represent the interquartile range with a horizontal line indicating the median value. Whiskers extend to the farthest data point within a maximum of 1.5 x the interquartile range, and colored dots present outliers.

Data availability

All public datasets we gathered in IMMUcan scDB are available from GEO, ArrayExpress, EGA, or BioProject (Supplementary Table S3). All accession codes as well as the public datasets that were processed and integrated into the database are available at https://immucanscdb.vital-it.ch/.

Code availability

The source code for processing all the collected datasets is available as a repository on GitHub (https://github.com/ImmucanWP7/immucan-scdb).

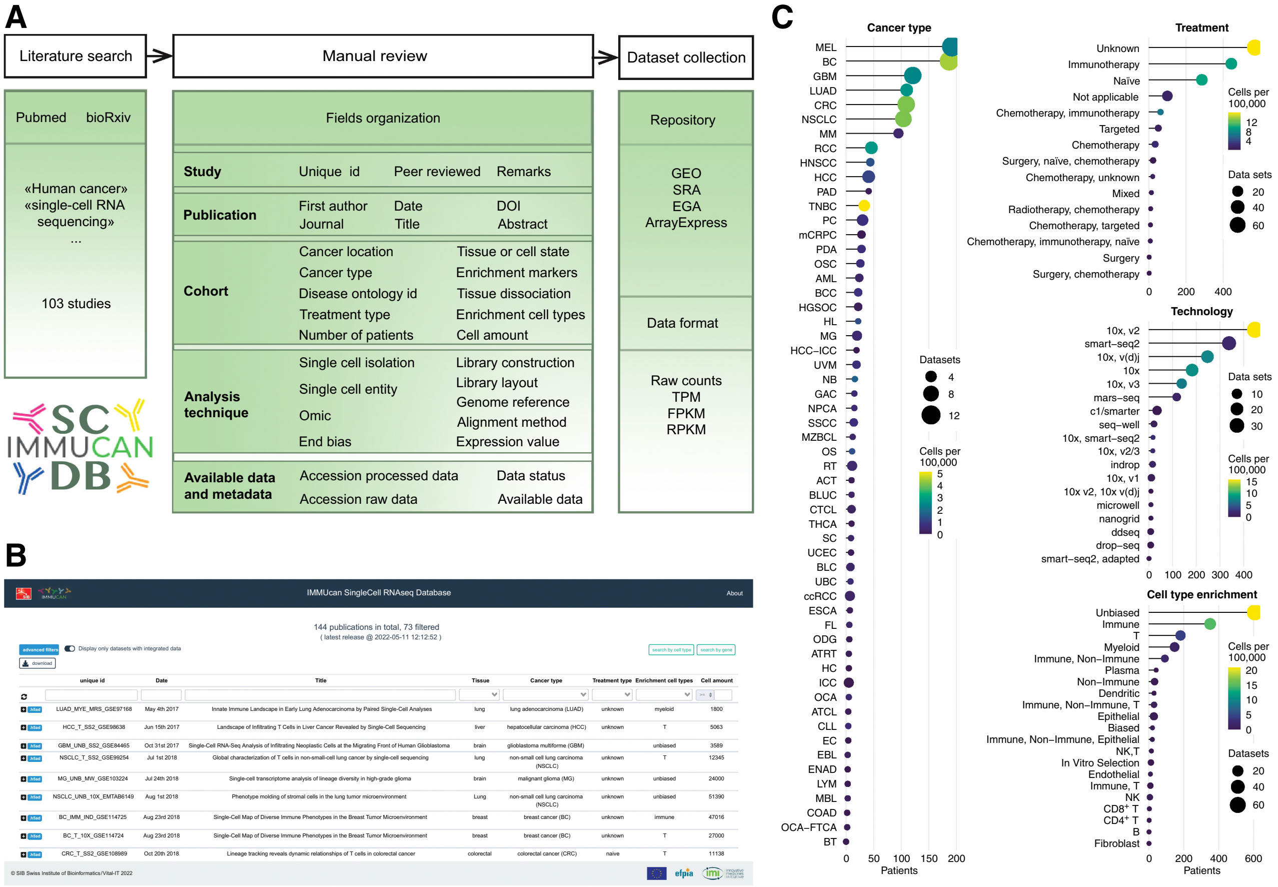
Results

Literature-based creation of the IMMUcan scRNA-seq database

The IMMUcan scRNA-seq database (scDB) was created through four main steps: (i) an exhaustive literature search for human cancer scRNA-seq studies; (ii) a manual review and curation of each relevant article; (iii) the collection of the corresponding datasets through web repositories or by contacting the authors; (iv) the processing and integration of the datasets and all associated metadata to the IMMUcan scDB (**Fig. 1A**).

All available annotations, data, and metadata were integrated into the IMMUcan scDB and can be searched through a user-friendly interface (immucanscdb.vital-it.ch) that enables users to query datasets based on the annotated metadata such as cancer type, treatment type, or the presence of a given cell type (**Fig. 1B**). In addition, the user can rank datasets for the specific expression of a given gene of interest in a subset of clusters or cell types. Once a dataset is selected, the user can mine the data and visualize the contained clusters and cell types as well as plot the expression of multiple genes across the clusters or against each other. The user can also download the normalized (batch corrected) and standardized datasets.

A total of 103 publications and corresponding metadata were successfully integrated into our database, corresponding to 144 datasets (**Fig. 1B**). Fifty-six cancer types were included. The most frequent



IMMUcan SingleCell RNAseq Database

144 publications in total, 73 filtered
(latest release @ 2022-05-11 12:12:52)

Display only datasets with integrated data

unique id	Date	Title	Tissue	Cancer type	Treatment type	Enrichment cell types	Cell amount
LUAD_MYE_MRS_GSE97168	May 4th 2017	Innate Immune Landscape in Early Lung Adenocarcinoma by Paired Single-Cell Analyses	lung	lung adenocarcinoma (LUAD)	unknown	myeloid	1800
HCC_T_SS2_GSE98638	Jun 15th 2017	Landscape of Infiltrating T Cells in Liver Cancer Revealed by Single-Cell Sequencing	liver	hepatocellular carcinoma (HCC)	unknown	T	5063
GBM_UNB_SS2_GSE84465	Oct 31st 2017	Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma	brain	glioblastoma multiforme (GBM)		unbiased	3589
NSCLC_T_SS2_GSE99254	Jul 1st 2018	Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing	lung	non-small cell lung carcinoma (NSCLC)	unknown	T	12345
MG_UNB_MW_GSE103224	Jul 24th 2018	Single-cell transcriptome analysis of lineage diversity in high-grade glioma	brain	malignant glioma (MG)	unknown	unbiased	24000
NSCLC_UNB_10X_EMTAB6149	Aug 1st 2018	Phenotype molding of stromal cells in the lung tumor microenvironment	Lung	non-small cell lung carcinoma (NSCLC)	unknown	unbiased	51390
BC_IMM_IND_GSE114725	Aug 23rd 2018	Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment	breast	breast cancer (BC)	unknown	immune	47016
BC_T_10X_GSE114724	Aug 23rd 2018	Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment	breast	breast cancer (BC)	unknown	T	27000
CRC_T_SS2_GSE108989	Oct 20th 2018	Lineage tracking reveals dynamic relationships of T cells in colorectal cancer	colorectal	colorectal cancer (CRC)	naive	T	11138

© SIB Swiss Institute of Bioinformatics/Vital-IT 2022

Figure 1.

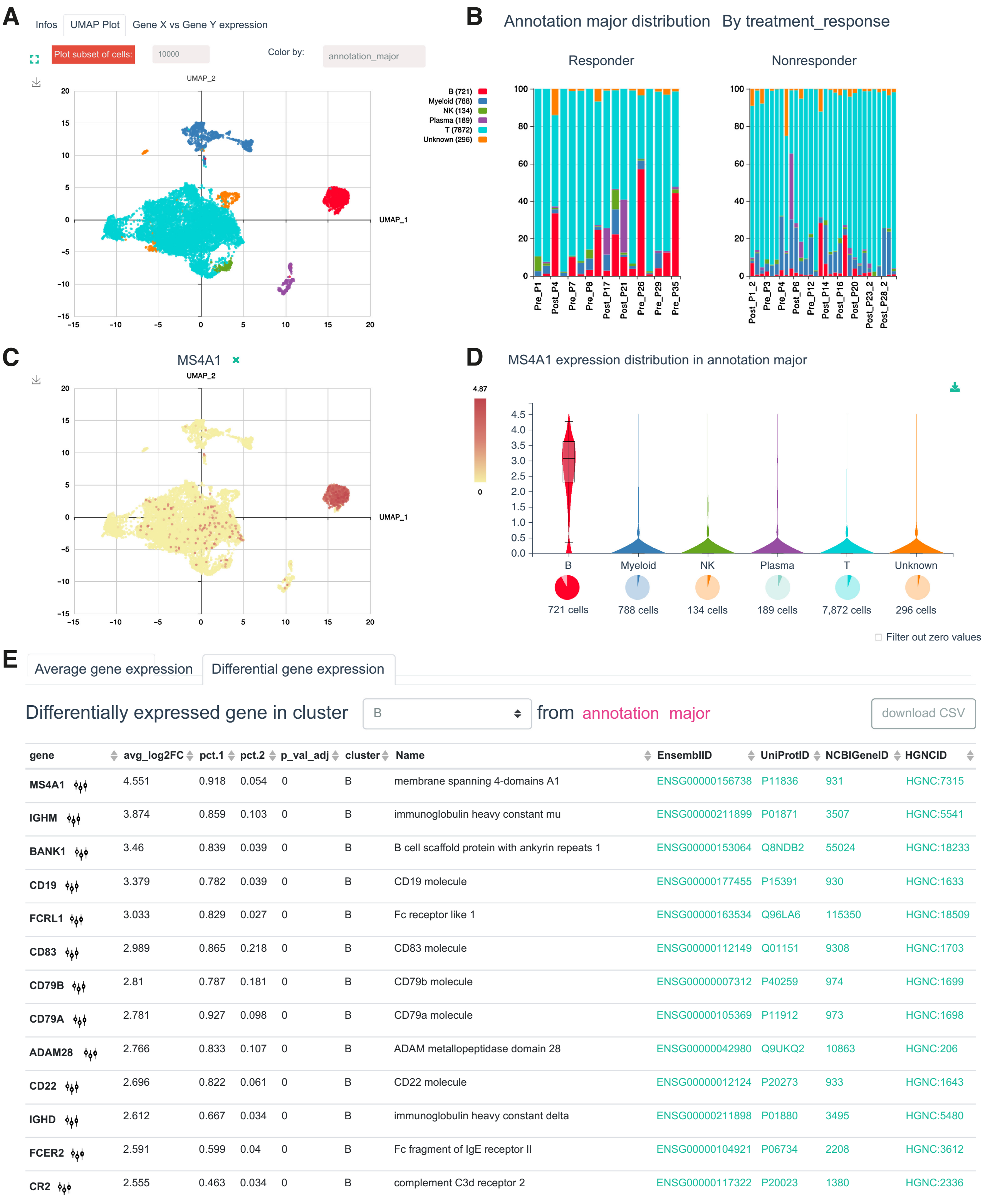
scRNA-seq database workflow. **A,** Strategy used to create the IMMUcan scRNA-seq database (scDB). **B,** Overview of the home page of the database web interface. **C,** Statistics of the database content represented as lollipop plot. The information (cancer type, cell type enrichment, treatment, and technology) is shown on the *y*-axis while the related number of datasets is shown on the *x*-axis. Point size correspond to the number of patients and the color-gradient represents the number of cells per 100,000.

cancer types were melanoma (13 datasets with a total of 192 patients with melanoma), followed by breast cancer (12 datasets, 187 patients), and glioblastoma (10 datasets, 121 patients; **Fig. 1C**). Less frequent tumor types included acute T-cell leukemia, renal cell carcinoma, and certain childhood tumors like medulloblastoma. The majority of the datasets were generated from single-cell suspensions with no prior enrichment (unbiased; 61 datasets), followed by immune cell selection through CD45 enrichment (23 datasets), and T-cell enrichment (15 datasets; **Fig. 1C**). Overall, 21 different types of enrichment protocols were applied across the different studies. Patient treatment was known and described for 61% of the patients, corresponding to 56% of the datasets. This information allows for specific analyses such as identifying cell type and transcriptomic changes specific to certain cancer treatments. Lastly, the database contains data generated via eleven different single-cell sequencing technologies with most studies having employed SMART-seq2 or 10X Genomics single-cell sequencing (**Fig. 1C**).

Cell type–based exploration of the IMMUcan scDB

To demonstrate the usefulness of the IMMUcan scDB we first focused on the cell type–specific use case of identifying cell types overrepresented in ICI treatment responders versus nonresponders. To this end, we searched in the scDB interface for datasets from immunotherapy-treated patients and then selected the melanoma dataset MEL_IMM_SS2_GSE120575 because it contains a comprehensive set of around 17,000 TME cells from patients before and after anti–PD-1 therapy.

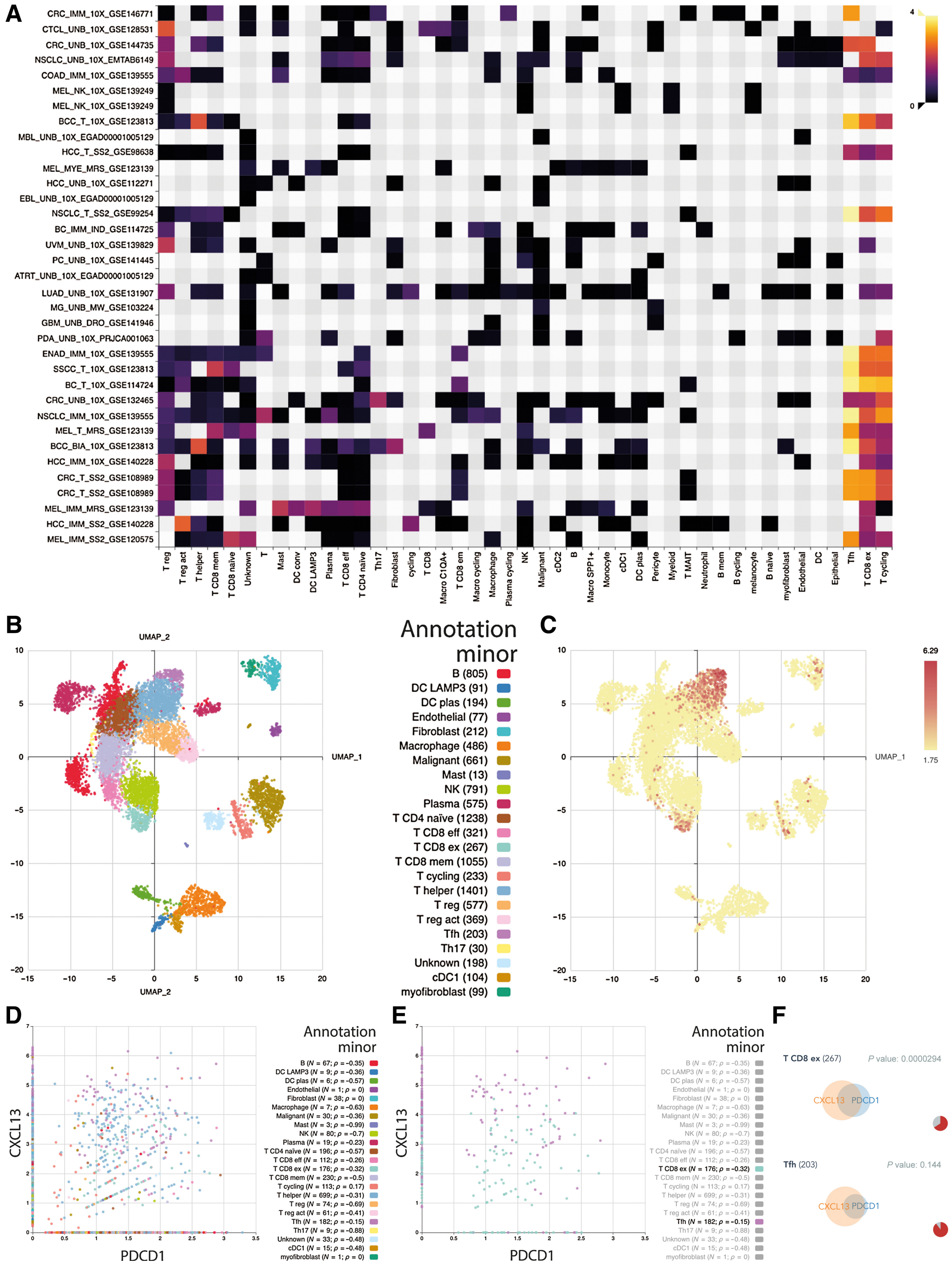
Selecting a dataset directly opens a panel showing an UMAP visualization of the data. In this plot, cells can be colored according to various levels of annotation such as the automatic CHETAH cell type assignments, their tissue of origin, or patient treatment (**Fig. 2A**). The interactive legend allows to select and deselect cell groups and displays group names and sizes. To increase plotting speed, a pulldown menu next to the UMAP plot allows to limit the visualization to a random selection of 10,000 cells. Next to the UMAP plot, for a given annotation such as CHETAH cell type assignments, a stacked bar chart visualizes the corresponding cell type composition of each sample in the study. By selecting a clinical annotation on top of the bar chart such as “treatment response,” multiple plots are created that allow for comparing the cell type composition between responders and nonresponders (**Fig. 2B**). Our bar chart visualization shows that B cells were increased in melanoma patients responding to anti–PD-1 therapy.



gene	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	Name	EnsemblID	UniProtID	NCBIGeneID	HGNCID
MS4A1	4.551	0.918	0.054	0	B	membrane spanning 4-domains A1	ENSG00000156738	P11836	931	HGNC:7315
IGHM	3.874	0.859	0.103	0	B	immunoglobulin heavy constant mu	ENSG00000211899	P01871	3507	HGNC:5541
BANK1	3.46	0.839	0.039	0	B	B cell scaffold protein with ankyrin repeats 1	ENSG00000153064	Q8NDB2	55024	HGNC:18233
CD19	3.379	0.782	0.039	0	B	CD19 molecule	ENSG00000177455	P15391	930	HGNC:1633
FCRL1	3.033	0.829	0.027	0	B	Fc receptor like 1	ENSG00000163534	Q96LA6	115350	HGNC:18509
CD83	2.989	0.865	0.218	0	B	CD83 molecule	ENSG00000112149	Q01151	9308	HGNC:1703
CD79B	2.81	0.787	0.181	0	B	CD79b molecule	ENSG00000007312	P40259	974	HGNC:1699
CD79A	2.781	0.927	0.098	0	B	CD79a molecule	ENSG00000105369	P11912	973	HGNC:1698
ADAM28	2.766	0.833	0.107	0	B	ADAM metallopeptidase domain 28	ENSG00000042980	Q9UKQ2	10863	HGNC:206
CD22	2.696	0.822	0.061	0	B	CD22 molecule	ENSG00000012124	P20273	933	HGNC:1643
IGHD	2.612	0.667	0.034	0	B	immunoglobulin heavy constant delta	ENSG00000211898	P01880	3495	HGNC:5480
FCER2	2.591	0.599	0.04	0	B	Fc fragment of IgE receptor II	ENSG00000104921	P06734	2208	HGNC:3612
CR2	2.555	0.463	0.034	0	B	complement C3d receptor 2	ENSG00000117322	P20023	1380	HGNC:2336

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, visualized 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 colors 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 $\log_2$-fold change compared with the rest of the cells in the dataset.


A
CRC_IMM_10X_GSE146771
CTCL_UNB_10X_GSE128531
CRC_UNB_10X_GSE144735
NSCLC_UNB_10X_EMTAB6149
COAD_IMM_10X_GSE139555
MEL_NK_10X_GSE139249
MEL_NK_10X_GSE139249
BCC_T_10X_GSE123813
MBL_UNB_10X_EGAD00001005129
HCC_T_SS2_GSE98638
MEL_MYE_MRS_GSE123139
HCC_UNB_10X_GSE112271
EBL_UNB_10X_EGAD00001005129
NSCLC_T_SS2_GSE99254
BC_IMM_IND_GSE114725
UVM_UNB_10X_GSE139829
PC_UNB_10X_GSE141445
ATRT_UNB_10X_EGAD00001005129
LUAD_UNB_10X_GSE131907
MG_UNB_MW_GSE103224
GBM_UNB_DRO_GSE141946
PDA_UNB_10X_PRJCA001063
ENAD_IMM_10X_GSE139555
SSCC_T_10X_GSE123813
BC_T_10X_GSE114724
CRC_UNB_10X_GSE132465
NSCLC_IMM_10X_GSE139555
MEL_T_MRS_GSE123139
BCC_BIA_10X_GSE123813
HCC_IMM_10X_GSE140228
CRC_T_SS2_GSE108989
CRC_T_SS2_GSE108989
MEL_IMM_MRS_GSE123139
HCC_IMM_SS2_GSE140228
MEL_IMM_SS2_GSE120575
T reg
T reg act
T helper
T CD8 mem
T CD8 naive
Unknown
T
Mast
DC conv
DC LAMP3
Plasma
T CD8 eff
T CD4 naive
Th17
Fibroblast
cycling
T CD8
Macro C1QA+
T CD8 em
Macro cycling
Macrophage
Plasma cycling
NK
Malignant
cDC2
B
Macro SPP1+
Monocyte
cDC1
DC plas
Pericyte
Myeloid
T MAIT
Neutrophil
B mem
B cycling
melanocyte
B naive
myofibroblast
Endothelial
DC
Epithelial
Tfh
T CD8 ex
T cycling
4
0
B
UMAP_2
10
5
0
−5
−10
−15
−20
UMAP_1
−15
−10
−5
0
5
10
15
20
UMAP_2
Annotation
minor
B (805)
DC LAMP3 (91)
DC plas (194)
Endothelial (77)
Fibroblast (212)
Macrophage (486)
Malignant (661)
Mast (13)
NK (791)
Plasma (575)
T CD4 naïve (1238)
T CD8 eff (321)
T CD8 ex (267)
T CD8 mem (1055)
T cycling (233)
T helper (1401)
T reg (577)
T reg act (369)
Tfh (203)
Th17 (30)
Unknown (198)
cDC1 (104)
myofibroblast (99)
C
10
5
0
−5
−10
−15
−20
UMAP_1
−15
−10
−5
0
5
10
15
20
6.29
1.75
D
CXCL13
7
6
5
4
3
2
1
0
0
0.5
1
1.5
2
2.5
3
3.5
PDCD1
Annotation
minor
B (N = 67; ρ = −0.35)
DC LAMP3 (N = 9; ρ = −0.36)
DC plas (N = 6; ρ = −0.57)
Endothelial (N = 1; ρ = 0)
Fibroblast (N = 38; ρ = 0)
Macrophage (N = 7; ρ = −0.63)
Malignant (N = 30; ρ = −0.36)
Mast (N = 3; ρ = −0.99)
NK (N = 80; ρ = −0.7)
Plasma (N = 19; ρ = −0.23)
T CD4 naïve (N = 196; ρ = −0.57)
T CD8 eff (N = 112; ρ = −0.26)
T CD8 ex (N = 176; ρ = −0.32)
T CD8 mem (N = 230; ρ = −0.5)
T cycling (N = 113; ρ = 0.17)
T helper (N = 699; ρ = −0.31)
T reg (N = 74; ρ = −0.69)
T reg act (N = 61; ρ = −0.41)
Tfh (N = 182; ρ = −0.15)
Th17 (N = 9; ρ = −0.88)
Unknown (N = 33; ρ = −0.48)
cDC1 (N = 15; ρ = −0.48)
myofibroblast (N = 1; ρ = 0)
E
CXCL13
7
6
5
4
3
2
1
0
0
0.5
1
1.5
2
2.5
3
3.5
PDCD1
Annotation
minor
B (N = 67; ρ = −0.35)
DC LAMP3 (N = 9; ρ = −0.36)
DC plas (N = 6; ρ = −0.57)
Endothelial (N = 1; ρ = 0)
Fibroblast (N = 38; ρ = 0)
Macrophage (N = 7; ρ = −0.63)
Malignant (N = 30; ρ = −0.36)
Mast (N = 3; ρ = −0.99)
NK (N = 80; ρ = −0.7)
Plasma (N = 19; ρ = −0.23)
T CD4 naïve (N = 196; ρ = −0.57)
T CD8 eff (N = 112; ρ = −0.26)
T CD8 ex (N = 176; ρ = −0.32)
T CD8 mem (N = 230; ρ = −0.5)
T cycling (N = 113; ρ = 0.17)
T helper (N = 699; ρ = −0.31)
T reg (N = 74; ρ = −0.69)
T reg act (N = 61; ρ = −0.41)
Tfh (N = 182; ρ = −0.15)
Th17 (N = 9; ρ = −0.88)
Unknown (N = 33; ρ = −0.48)
cDC1 (N = 15; ρ = −0.48)
myofibroblast (N = 1; ρ = 0)
F
T CD8 ex (267)
P value: 0.0000294
CXCL13
PDCD1
Tfh (203)
P value: 0.144
CXCL13
PDCD1

Two gene tables are automatically loaded on the bottom of the page that show the average expression of each gene per cell type as well as the level of differential expression of each gene between a user selected cell type and all other cells in the dataset. The expression of each gene can be visualized on an additional UMAP plot (**Fig. 2C**) and as a violin plot (**Fig. 2D**). To improve the interpretation of these plots, the absolute cell number is represented as a pie chart below the violin plot and the percentage of cells expressing the gene appears in a mouseover (**Fig. 2C**). In addition, the expression of the selected gene is also displayed in a separate UMAP plot. We visualized the expression of the top marker gene of B cells showing here that *MS4A1* (CD20) is the most selective gene expression marker for B cells (**Fig. 2E**).

Gene-based searching of the IMMUcan scDB

Recently, a study of multiple bulk transcriptomic cancer datasets has shown that *CXCL13* and *CXCL9* could be used as predictive biomarkers for checkpoint immunotherapy response (16). As use case for a gene centric search, we employed IMMUcan scDB to identify the cell types expressing these two genes across different cancer types.

Upon entering a gene in the corresponding search field at the top right side of the entry page, the IMMUcan scDB displays a heatmap of the gene's average expression in each cell type in every dataset in which the gene is expressed. We searched for *CXCL13* and using "annotation minor" as cell type resolution observed that it is most highly expressed in T follicular helper cells (Tfh) and exhausted CD8$^+$ T cells (CD8$^+$ T ex) in most cancer indications including basal cell carcinoma (BCC), melanoma (MEL), and non–small cell lung cancer (NSCLC; **Fig. 3A–C**). On the other hand, and in line with recent publications, *CXCL9* was found to be expressed across myeloid cells with highest levels in LAMP3 positive dendritic cells and macrophages (17) from various indications including hepatocellular carcinoma (HCC), NSCLC, and melanoma.

Evaluating gene coexpression in the IMMUcan scDB

The IMMUcan scDB also makes it possible to quantify the coexpression of two genes. To this end, after selecting a dataset on the entry page, the user can select the "Gene X vs. Gene Y expression" panel and enter the names of two genes. A scatter plot is then created with one point per individual cell. Cells are colored on the basis of the selected cell type resolution level. The legend lists all cell types expressing both genes together with the corresponding number of cells and associated Pearson correlation coefficients. In addition, Venn diagrams are displayed that visualize for each cell type the number of cells coexpressing both genes and a *P* value for the significance of the overlap. To demonstrate the coexpression features we selected the BCC study BCC_BIA_10X_GSE123813 based on the high expression of *CXCL13* in exhausted CD8$^+$ T cells within this dataset. We investigated the coexpression of *CXCL13* with *PD1* (*PDCD1*), another well-known marker for T-cell exhaustion and observed highly significant coexpression of the two genes in exhausted CD8 T cells with an overlap *P* value of 2.9×10^{-5} (**Fig. 3D–F**).

In line with the results from the above gene-based search, *CXCL9* and *CXCL13* were not coexpressed in any of the cell types from the BCC_BIA_10X_GSE123813 dataset (Supplementary Fig. S3) as expression of *CXCL9* was restricted to cells of myeloid origin. Accordingly, the Venn diagrams show no overlap between the *CXCL9* and *CXCL13* positive cells and the scatter plot shows no cells expressing both genes.

Identifying common changes in cell type frequencies associated with malignant transformation

We next used the opportunity of having a large collection of harmonized single-cell datasets to identify changes to the T-cell and macrophage compartments observed consistently across the TME of different cancer types. To this end we selected all 25 datasets from the IMMUcan scDB containing both tumor and normal tissue and then compared both changes in macrophage and T-cell subtype frequencies and corresponding changes in gene expression patterns associated with malignant transformation.

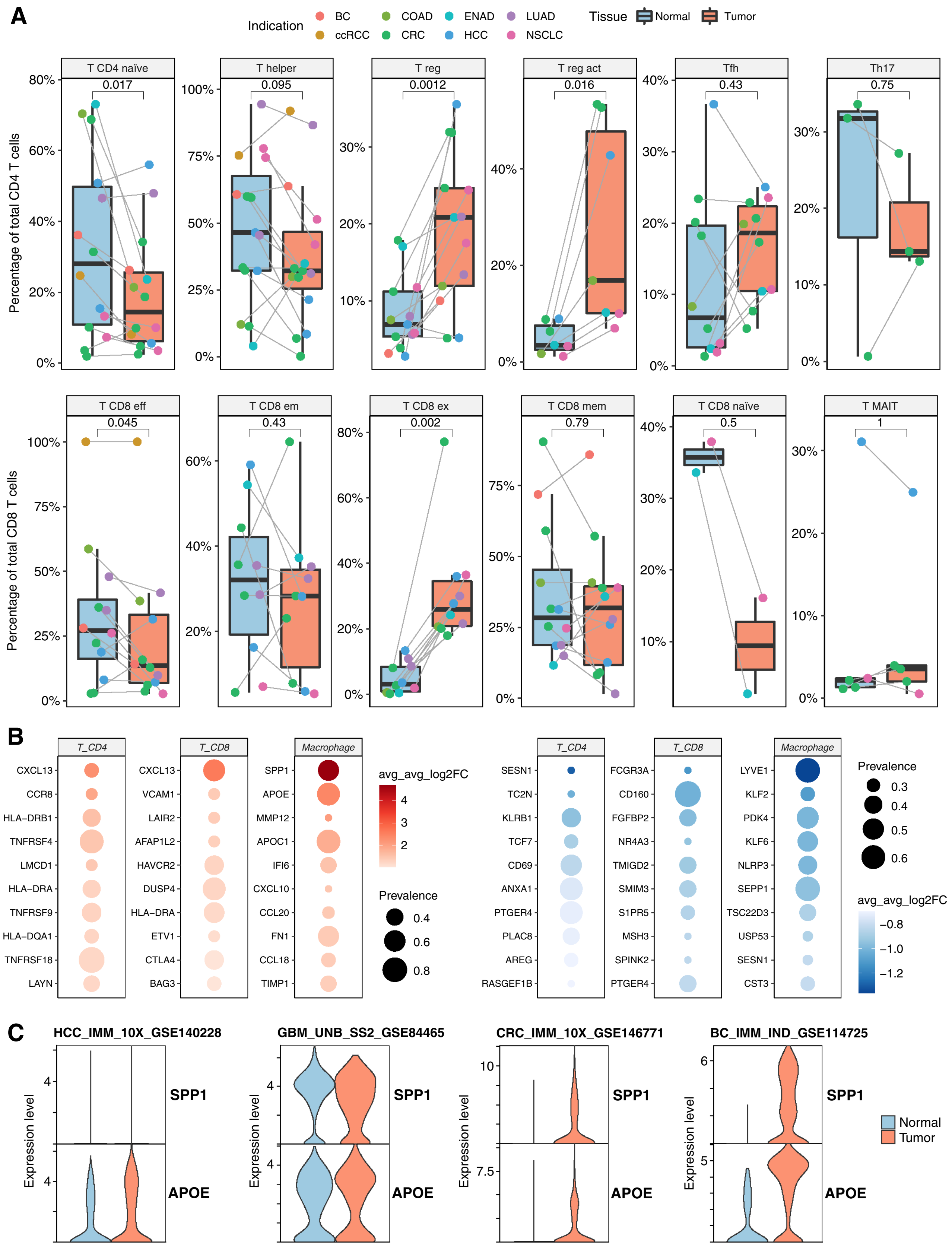
We found a total of 705 upregulated and 611 downregulated genes over 11 cell types with log-fold change differences greater than 1 and a multiple testing corrected *P* value of 0.001. Many of the top DE genes were corresponding to dissociation-artefacts (18) such as heat shock proteins and tissue specific genes such as alveolar and surfactant genes. Therefore, we created a gene blacklist removing all genes all genes associated with these effects from downstream analyses. The gene blacklist contained heatshock proteins, other dissociation-associated genes such as DNAses, FOS and JUN, immunoglobins, mitochondrial genes, tissue-specific genes, ribosomal genes, and ERCC spike-ins. Resulting differentially expressed genes were prioritized by the average fold change over all datasets, excluding genes that were observed in less than 20% of the datasets (**Fig. 4A**).

As shown in **Fig. 4A** we find the TME to be associated with a drastic increase in regulatory T cells (Treg) across multiple cancer types including NSCLC, CRC, and HCC. In particular, activated Tregs appear to be present almost exclusively in the TME while being nearly absent in corresponding normal tissues. In contrast, we see a consistent decrease in the percentage of naïve CD4 and CD8 T cells in the TME compared with normal tissue (**Fig. 4A**). In line with the observed changes in cell type frequencies we observe significantly lower expression of genes associated with naïve T cells such as *TCF7* while genes related to T-cell activation such as *TNFRSF4*, *TNFRSF9*, and *TNFRSF18* and T-cell exhaustion such as *HAVCR2* and *CTLA4* appear higher expressed in the TME (**Fig. 4B**). In addition, next to *CXCL13*, an attractant of B cells, we also find markers for activated Tregs including *CCR8* and *LAYN* as highly tumor specific. In line with this observation, CCR8 has recently been identified as a tumor Treg-specific target, leading to anti-CCR8 antibodies currently being tested in clinical trials for Treg depletion approaches (19).

For macrophages, we find a strong upregulation of *SPP1* and *APOE* in the TME (**Fig. 4B**). Interestingly, this upregulation is absent in tumor indications like HCC and glioblastoma while most other indications such as CRC and breast cancer show drastic upregulations of both genes (**Fig. 4C**). SPP1 is hypothesized as a mediator of proinflammatory pathways and immunotherapy response in NSCLC (20) while APOE is hypothesized to promote immune suppression in

Figure 3.

Gene-based exploration of the IMMUcan scDB using CXCL13, a predictive biomarker for immunotherapy response. **A,** Heatmap 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.


A
Indication
BC
ccRCC
COAD
CRC
ENAD
HCC
LUAD
NSCLC
Tissue
Normal
Tumor
Percentage of total CD4 T cells
T CD4 naïve
0.017
T helper
0.095
T reg
0.0012
T reg act
0.016
Tfh
0.43
Th17
0.75
Percentage of total CD8 T cells
T CD8 eff
0.045
T CD8 em
0.43
T CD8 ex
0.002
T CD8 mem
0.79
T CD8 naïve
0.5
T MAIT
1
B
T_CD4
CXCL13
CCR8
HLA-DRB1
TNFRSF4
LMCD1
HLA-DRA
TNFRSF9
HLA-DQA1
TNFRSF18
LAYN
T_CD8
CXCL13
VCAM1
LAIR2
AFAP1L2
HAVCR2
DUSP4
HLA-DRA
ETV1
CTLA4
BAG3
Macrophage
SPP1
APOE
MMP12
APOC1
IFI6
CXCL10
CCL20
FN1
CCL18
TIMP1
avg_avg_log2FC
4
3
2
Prevalence
0.4
0.6
0.8
T_CD4
SESN1
TC2N
KLRB1
TCF7
CD69
ANXA1
PTGER4
PLAC8
AREG
RASGEF1B
T_CD8
FCGR3A
CD160
FGFBP2
NR4A3
TMIGD2
SMIM3
S1PR5
MSH3
SPINK2
PTGER4
Macrophage
LYVE1
KLF2
PDK4
KLF6
NLRP3
SEPP1
TSC22D3
USP53
SESN1
CST3
Prevalence
0.3
0.4
0.5
0.6
avg_avg_log2FC
−0.8
−1.0
−1.2
C
HCC_IMM_10X_GSE140228
GBM_UNB_SS2_GSE84465
CRC_IMM_10X_GSE146771
BC_IMM_IND_GSE114725
Expression level
SPP1
APOE
Normal
Tumor

pancreatic cancer (21). The fact that we observe these genes to be highly upregulation in tumor-associated macrophages from nearly all datasets suggests that their effect on immune suppression could be more widespread than previously suggested.

In conclusion, here we present the IMMUcan scDB, a curated database of scRNA-seq studies of the human TME that is easily searchable and explorable. By means of three use cases, we showed that the IMMUcan scDB is an efficient tool to validate observations from literature, to generate new hypotheses and to provide novel biological insights.

Discussion

The number of scRNA-seq studies in human cancer has increased exponentially in recent years. The first studies provided a large-scale description of tumor cells and TME ("atlas" view), extending from common tumor types (melanoma, breast cancer, NSCLC) to rare cancers, such as atypical teratoid rhabdoid tumor (22) or less frequent molecular subtypes, such as triple-negative breast cancer (23). We anticipate that scRNA-seq "atlas" studies will gradually focus on an even broader diversity of tumor types, and include increasing numbers of patients, samples, and cells. Parallel to these descriptive studies, scRNA-seq was applied more recently to identify mechanisms of resistance (24), or response to immune checkpoint inhibitors (25). Such hypothesis-driven studies should also grow in numbers and magnitude, with the diffusion and the increased accessibility of scRNA-seq technologies. Another type of study design includes the comparison of different anatomical sites, such as primary versus metastatic tumor location (26). The number and diversity of scRNA-seq studies justifies a resource that would be fully dedicated to human cancer datasets, to provide a detailed annotation, easy and efficient search functions, as well as multiple implemented methods for meta-analysis. We believe this to be the most optimal way to cope with an anticipated number of several hundred datasets in the coming years. In this respect, we will pay particular attention to the prospective integration of newly published datasets according to the standardized strategy that we have established. Within the IMMUcan consortium we will maintain the database with monthly updates.

Public data repositories offer access to an increasing number of large-scale ("omics") datasets, in particular genomics and transcriptomics. However, clinical annotation is often missing or reduced to a minimal amount of information, such as the tumor type. This greatly limits the possibilities for integration of clinical and biological data in the analysis and interpretation. Single-cell portals, such as UCSC Cell Browser (27), Broad Institute Single Cell Portal (available from: https://singlecell.broadinstitute.org/single_cell) and single-cell expression atlas (28), do not include this level of annotation. Cancer scRNA-seq databases such as CancerSea (5) or TISCH (6), include minimal clinical information, restricted to tumor type, primary or metastatic stage, and treatment type. In our study, we have gone through the manual process of extracting and mapping to reference ontologies detailed clinical features (9 items) associated to each patient cohort and datasets. In comparison with the other resources, IMMUcan scDB is the only database specifically dedicated to human cancer single-cell transcriptomic datasets (**Table 1**). It integrates the information of 144 studies, including 73 datasets. The tumor clinical annotations are one of the most detailed among all existing resources. IMMUcan scDB allows most of functionalities that are offered by other resources and is the only such database with interactive graphs (allowing to display graph according to clinical features of interest, such as splitting graph according to tissue type, treatment, or patients). This should allow biologists and clinicians to focus on datasets corresponding to a particular clinical scenario, and to compare datasets across clinical settings. It should also provide important insight into cell types, cell states and associated signatures.

Different from bulk transcriptomics, scRNA-seq generates data from a large number of cells even in individual samples. Assuming that cell numbers are sufficient, this offers the possibility for robust characterization of cellular clusters and associated gene expression programs in individual patients. In parallel, the aggregated analysis of several datasets fulfilling common conditions is also important to identify unifying patterns associated to a tumor type, a specific anatomical location, or a treatment effect. A recent study has constructed a "pan-cancer blueprint" of stromal cell heterogeneity using original scRNA-seq datasets from four cancer types (29). It revealed shared gene expression programs in infiltrating immune cells. In our IMMUcan scDB, we have implemented robust methodologies to integrate several samples to identify common patterns and increase statistical relevance to a given clinical setting. As a result, users may apply focused strategies on individual patient samples.

The IMMUcan scDB offers large possibilities for biomedical applications. Exploratory analysis allows an early discovery process to generate hypotheses for further validation. For example, comparison of cell type–specific signatures from different clinical settings may reveal interesting mechanisms of immune activation or immune escape, or novel therapeutic targets. Conversely, hypothesis-driven analyses may establish the expression pattern of specific genes or signatures according to different annotation terms. Finally, our database can be used to validate findings established in an independent study. The large and increasing number of scRNA-seq datasets offers unique possibilities for cross-validation of results coming from different technologies, such as proteomics, genomics, or spatial transcriptomics.

Integrating such a large number of scRNA-seq datasets into a single database has potential risks and limitations. As all literature-based resources, sample quality and dataset annotation rely on the quality of the information provided in the original publication. In this respect, we found tremendous heterogeneity in patient cohorts' description, both in the amount and in the quality of the clinical information. An important step forward would be the improvement and generalization of standardized 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. The processing of scRNA-seq datasets generated in different studies, using various tissue dissociation and enrichment

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 upregulated and downregulated 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 normalized expression of SPP1 and APOE between macrophages from matched normal and tumor samples in four selected datasets. CRC, colorectal cancer.

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

		scRNASeqDB	SCPortalen	PanglaoDB	JingleBells	CancerSEA	TISCH	SPICA	IMMUcanDB
Data	Number of datasets	38	66	1368	302	74	79	21	144 [73][a]
	Number of human oncology datasets	3	2	10	14	20	75	4	144 [73][a]
	Number of criteria for datasets query	3	5	6	3	2	9	2	7+
Sample type	Cell lines	X	X	X	X	X	–	–	–
	Xenograft	X	X	X	X	X	–	–	–
	Mouse samples	–	X	X	X	X	X	X	–
	Human tissues/blood	X	X	X	X	X	X	X	X
Clinical annotations	Tumor type	–	–	–	–	X	X	X	X
	Tissue site	–	–	X	–	X	X	–	X
	Primary vs metastatic	–	–	–	–	–	X	–	X
	Treatment type	–	–	–	–	–	X	–	X
	Response to treatment	–	–	–	–	–	X	X	X
	Cell enrichement strategy	–	–	–	–	–	X	–	X
Availability of processed data	BAM	–	X	–	X	–	–	–	–
	Average gene expression per cell type	–	X	X	–	X	X	X[b]	X
	Differentially expressed genes	X	–	–	–	–	X	X	X
	Single-cell object	–	–	–			–		X
Gene specific functionality across datasets	Gene expression distribution	X	X	X	–	–	X	X	X
	Signature expression among datasets	–	–	–	–	–	X	–	–
	Dataset filtering	X	–	X	–	–	X	–	X
Visualisation	UMAP	X	X	X	–	–	X	X	X
	Cluster proportion	–	–	–	–	–	X	X	X
	Gene signature expression	–	–	–	–	–	X	–	X
	Interactive graphs	–	–	–	–	–	–	–	X

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

protocols, as well as potentially different technological platforms, is certainly challenging and subject to technical biases. In our processing pipeline, we have implemented robust and validated methodologies at each step. We have selected Harmony as a method to reduce experimental bias during multiple datasets integration. Harmony uses reiterative clustering to remove batch effects between experiments and patients. From recent benchmark studies on scRNA-seq data integration (31–33), Harmony was among the top performers and it is recommended as integration method over methods such as CCA (34), Liger (35) and UMI downsampling (bioRxiv 2021.11.15.468733) for its good performance. In addition, we have seen no substantial differences between harmony and other integration algorithms when we tested this on a selection of IMMUcan scDB datasets (Supplementary Fig. S4). Users should be aware of all limitations and possible biases and may use their own cross-validation methodologies to increase the robustness of their findings. Improving the performance of our data processing will remain a top priority in the coming years. Overall, we believe that the power and possibilities offered by integrating such a large number of datasets largely outweighs the limitations and weaknesses inherent to meta-analysis. We hope that our resource will facilitate the exploitation of publicly available scRNA-seq datasets to address existing and novel challenges in human cancer research.

Authors' Disclosures

Jordi Camps, Mahmoud Ibrahim, and Helge Roider are full-time employees of Bayer AG. F. Noel reports grants from Innovative Medicines Initiative during the conduct of the study. C. Hoffmann reports personal fees from Nanobiotix and personal fees and other support from Owkin outside the submitted work. J. Pollard reports personal fees from Sanofi during the conduct of the study. V. Soumelis reports grants from Sanofi and Roche, and personal fees from Leo and from Menarini outside the submitted work; and is a full-time employee at Owkin since March 2022. No disclosures were reported by the other authors.

Authors' Contributions

J. Camps: Conceptualization, data curation, software, formal analysis, visualization, methodology, writing–original draft. **F. Noël:** Data curation, visualization, writing–original draft. **R. Liechti:** Software, funding acquisition, visualization. **L. Massenet-Regad:** Data curation. **S. Rigade:** Resources, data curation. **L. Götz:** Software. **C. Hoffmann:** Data curation. **E. Amblard:** Data curation. **M. Saichi:** Data curation. **M.M. Ibrahim:** Software, methodology. **J. Pollard:** Conceptualization, supervision. **J. Medvedovic:** Conceptualization, supervision. **H.G. Roider:** Conceptualization, supervision, writing–original draft. **V. Soumelis:** Conceptualization, supervision, funding acquisition, writing–original draft.

Acknowledgments

Swiss Institute of Bioinformatics would like to thank Jorge Molina (Core-IT, Swiss Institute of Bioinformatics) for his expert support in setting up the IT infrastructure of the web portal. The authors would like to thank Cristina Ghirelli for her help in reviewing the manuscript. They also would like to acknowledge Donald Jackson and all their colleagues from IMMUcan consortium, who gave them insightful feedback on the database.

The IMMUcan project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 821558. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. https://IMI.europa.eu. C. Hoffmann was funded by Cancéropole INCA, Fondation de France, and Institut Curie. L. Massenet-Regad was supported by a PhD fellowship from La Ligue Contre le Cancer. M.M. Ibrahim was supported by the Life Science Collaboration (LSC) program. This work of V. Soumelis's team was supported by Ligue nationale contre le cancer (labellisation EL2016.LNCC/VaS).

The publication costs of this article were defrayed in part by the payment of publication fees. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 USC section 1734.

Note

Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Received January 10, 2022; revised August 15, 2022; accepted November 29, 2022; published first December 2, 2022.

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 visualization 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;38:1384–6.
8. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol 2019;15: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:W541–545.
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;37:963–7.
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;16:1289–96.
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:e95.
14. Gao R, Bai S, Henderson YC, Lin Y, Schalck A, Yan Y, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. Nat Biotechnol 2021;39:599–608.
15. Cakir B, Prete M, Huang N, van Dongen S, Pir P, Kiselev VY. Comparison of visualization tools for single-cell RNA-seq data. Nucleic Acids Res 2020;2: lqaa052.
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 sensitization to checkpoint inhibition. Cell 2021;184:596–614.
17. Park MK, Amichay D, Love P, Wick E, Liao F, Grinberg A, et al. The CXC chemokine murine monokine 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: 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: 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;81:2983–94.
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;39:1594–609.
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: 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;51:1702–13.
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;173:879–93.
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;175:984–97.
25. Sade-Feldman M, Yizhak K, Bjorgaard SL, et al. Defining T-cell states associated with response to checkpoint immunotherapy in melanoma. Cell 2018;175:998–1013.
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;171:1611–24.
27. Speir ML, Bhaduri A, Markov NS, Moreno P, Nowakowski TJ, Papatheodorou I, et al. UCSC cell browser: Visualize 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;48:77–83.
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;30:745–62.
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;19:41–50.
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: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;21:12.
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;36:411–20.
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;177:1873–87.