CellKb Immune: a manually curated database of hematopoietic marker gene sets from 7 species for rapid cell type identification

Ajay Patil[1], Ashwini Patil[1*]

[1]Combinatics, Ichikawa-shi, Chiba, Japan.

*** Correspondence:**
Ashwini Patil
ashwini@combinatics.com

Keywords: single-cell, cell types, marker genes, immune cells, annotation, database, cell type prediction

Abstract

Using single-cell RNA sequencing, it is now possible to get the detailed gene expression pattern in each cell for millions of cells in a single experiment under varying conditions. Annotating cells by assigning a type or state to them is important for correct interpretation of results. In spite of current advances in technology, assigning cell types in single-cell datasets remains a bottleneck due to the lack of a comprehensive reference database and an accurate search method in a single tool. CellKb Immune is a knowledgebase of manually collected, curated and annotated marker gene sets from cell types in the immune response in human, mouse and 5 other model organisms. It allows users to find matching cell types in literature given a list of genes within a species and across multiple species. We evaluated the contents and search methodology of CellKb Immune using a leave-one-out approach. We further used CellKb Immune to annotate previously defined marker gene sets from a bulk RNA-seq dataset to confirm its accuracy and coverage. CellKb Immune provides an easy-to-use online tool with a reliable method to find matching cell types and annotate cells in single-cell experiments. It is available at https://www.cellkb.com/immune.

1 Introduction

Single-cell RNA-seq is widely used to study transcriptional patterns of genes in individual cells. Various computational tools have been developed for the annotation of cell types from single-cell experiments, given a reference database and a gene list (1,2). Several single cell reference datasets or atlases have also been prepared (3–5), or compiled into aggregate reference databases of cell types and marker gene sets (6–9). There are currently two primary methods used for annotating cells from single-cell experiments. 1) Manually searching through scientific literature to find marker genes associated with cell types being studied. This requires prior knowledge of the cell types potentially present in the sample and fails to identify rare cell types without established markers. 2) Selecting a single cell reference database/ dataset from literature, downloading and processing the data, identifying and installing a computational method to assign cell types, and finally, writing a program in R or Python to perform the cell type assignment. This method usually limits a researcher to using a single, or a few, datasets as a reference in a single species. Both these methods require significant time and effort but fail to use the vast amount of cell type marker gene sets previously published. Reliable annotations require a resource that contains a comprehensive reference database of cell types

and marker gene sets from literature along with a search method that allows users to look for matching cell types.

CellKb Immune addresses these issues by collecting and aggregating a large amount of immune-related cell type signatures across 7 species from literature and providing an online interface to find matching annotated cell types easily, quickly and accurately.

2 Materials and Methods

2.1 Data collection and curation

The CellKb Immune database consists of marker gene sets with associated cell type annotations collected from literature. The marker genes and their cell type annotations are taken primarily from published single-cell RNA-seq experiments. All publications related to single-cell experiments are taken from PubMed and manually screened to select those identifying cell type specific gene expression patterns.

Cell type specific marker genes are defined as those identified by authors based on their significant change in expression within a group of cells as compared to all other cells in the experiment. The marker gene sets are extracted from tables, figures, supplementary materials and raw, or processed, data of publications describing these experiments. If calculated based on the raw/processed expression data, cell type specific markers are identified using the methods described by the authors in the publication, along with cell type annotations given for the single cells.

Marker gene sets are selected from published experimental studies if they satisfy the following criteria: 1) Raw data related to the experiment is deposited in public databases, 2) The marker gene sets are publicly available for download, 3) The experimental methods and conditions are adequately described in the manuscript to allow detailed annotation, 4) The number of cells and samples studied is clearly specified, 5) The computational methods used to normalize, filter and cluster cell types, along with identification of cluster-specific genes is clearly described, 6) Associated values such as average expression, fold change, statistical significance, raw UMI or read counts, are available for marker genes (marker gene sets specified only as ranked gene lists without associated values are included in some cases), 7) A majority of the gene identifiers in the marker gene set are valid and can be mapped to the latest version of the Ensembl database (14). Single-cell experiments without cell type annotations are not used to calculate cell type specific marker gene sets.

Detailed sample and treatment conditions, including sample counts and cell counts are also stored in CellKb Immune. Information about disease state of the samples and the cell types enriched in disease samples are also noted. All data in CellKb Immune is referenced and linked to the PubMed identifier of the source publication.

CellKb Immune also contains marker genes extracted from select bulk RNA-seq and microarray experiments from public databases. These include selected gene signatures from Human Protein Altas (10), MSig-db (11) and previously published bulk RNA-seq datasets of immune cell types (1,12). Cell-type specific interaction networks are created using marker gene sets of all cell types and protein-protein interactions with reliability scores from the HitPredict database (13).

2.2 Data harmonization and annotation

Data harmonization is essential for rapid search and retrieval of matching cell types, within and across species. Given that there is no standard format currently in place for sharing cell type specific marker gene sets, CellKb Immune performs a number of operations to harmonize the data collected. Marker gene sets are extensively curated to identify valid genes and cell types. Genes given in the published marker gene sets are mapped to valid Ensembl identifiers using all possible names, synonyms and identifiers. Assigning a common gene identifier to the marker genes is an important part of data harmonization in CellKb Immune. It also helps identify orthologous genes when using a gene list from one species to search for cell types in another.

Cell types, tissue names and disease conditions are assigned standardized ontology terms using Cell Ontology (15), Uberon Ontology (16) and Disease Ontology (17). Associated values given by authors with each signature are also stored in CellKb Immune. These include, but are not limited to, rank, score, average expression, log fold change and corrected/uncorrected p-values. Marker gene sets that are published in the form of ranked gene lists without associated values are also included in the database if they have been accurately annotated. Unranked gene lists without associated values are not included in CellKb Immune.

Each cell type is assigned a reliability score that gives an indication of how similar it is to other cell types of the same ontology. The reliability score is calculated as the ratio of the average Spearman's correlation coefficient of a marker gene set with others of the same cell type to that of marker gene sets of different cell types. The higher the reliability score of a marker gene set, the closer it is to other marker gene sets of the same cell type and, hence, the more likely it is to be accurate. This helps identify marker gene sets that are incorrect, calculated from erroneous cluster assignments by authors, or that have been calculated from mixtures of cell types.

Finally, all marker genes and their annotations are stored in a uniform format with indexed identifiers enabling rapid search and retrieval of the data.

2.3 Data searching

In CellKb Immune, users can search for cell type specific marker gene sets by giving a list of ranked query genes. Query genes given by the user are compared with every marker gene set in the database and the best match is identified using a match score. The match score is based on the number of common genes between the query and cell type, their ranks, their rank differences and the total number of significant genes in the cell type. This results in cell types sharing highly ranked genes with the query being assigned higher match scores. This method also accounts for differences in the sizes of gene lists between the query and various cell types, such that cell types with fewer marker genes are not disregarded. To search cell types in a species different from that of the query gene list, CellKb Immune identifies the orthologous genes in the species of interest using Ensembl Compara (18) before searching for matching cell types.

3 Results

3.1 Database contents

CellKb Immune contains author-defined immune cell type marker gene sets manually collected from publications describing mainly single-cell, and selected bulk RNA-seq or microarray experiments (1,10,11) (Figure 1). Cell type marker gene sets are assigned a Cell Ontology term (15), where available. Tissues, organs or body parts of the experimental sample are assigned Uberon ontology terms (16). Disease ontology terms are provided where the marker gene sets are obtained from a diseased sample (17). All marker gene sets reference their source publication with a link to its PubMed identifier.

CellKb Immune contains:

- 6,604 marker gene sets
- 560 unique annotated hematopoietic cell types (Cell Ontology CL:0000988 and all its descendants).
- 439 publications selected from approximately 10,000 studies published between 2013 and 2021.
- 852,305 protein-protein interactions from HitPredict (13)
- 7 species - *H. sapiens, M. musculus, D. rerio, R. norvegicus, M. fascicularis, M. mulatta, D. melanogaster*

Figure 2 shows the number of marker gene sets for each species in CellKb Immune.

3.2 Functionality

The CellKb Immune web interface allows users to search matching cell types and their marker gene sets within and across species. It can be used to annotate cells from single-cell RNA-seq data, given a list of statistically significant cluster-specific genes ranked by decreasing log fold change. Associated values such as log fold changes and p-values are also accepted as input by CellKb Immune. Genes in the user query are compared with every marker gene set in the database using a rank-based search method and a match score is computed. To identify matching cell types in another species, CellKb Immune identifies the orthologous genes between the query and the target species before searching.

CellKb Immune is limited to marker gene sets of hematopoietic cell types. However, the user is notified if the query gene list matches non-hematopoietic cell types in 5 out of top 10 cell type matches.

3.3 Use case: Assigning cell types in a single-cell RNA-seq experiment

To demonstrate the functionality of CellKb Immune, we used it to assign cell types to a single-cell RNA-seq dataset of human peripheral blood mononuclear cells from the Single Cell Portal (19).

1) To query CellKb Immune for matching cell types, we used the cell type specific genes, along with their associated log fold changes and p-values calculated in the study for primary lineage clusters defined by the authors. Cell-type specific genes for each cluster were limited to those that were upregulated at least two-fold and had an adjusted p-value less than or equal to 0.05. The filtered query genes were ranked in descending order of their log fold change before submission to CellKb Immune (Table S1).

2) Users have an option to search for matching cell types taken single-cell and/or bulk experiments, and filter the search based on cell types, organs/tissues, diseases and other conditions. No filters were set for this use case.

3) The CellKb Immune results display a heatmap with the rank-based score of the top 100 matching cell types (Figure 3).

4) Details of the experiment containing the matching cell type, such as PubMed ID, condition, number of cells, cell type ontology, etc. are shown in tabular format for the top 10 hits. Other statistics such as the statistical significance by Fisher's exact test, Pearson's correlation coefficient and Jaccard index are also calculated for each cell type match. These results are available for download.

5) Users can also view the top 10 overlapping genes between the query and the marker gene set of the matching cell type.

6) Cell type specific protein-protein interaction network containing the user query genes and the top 10 overlapping genes are also displayed.

CellKb Immune was able to correctly identify 7 of the 8 cell types assigned by the authors. One cell type which was assigned as “Ambiguous/Potential Doublets”, was matched to Plasmablasts by CellKb Immune. Table S1 gives the top 10 matching cell types identified by CellKb Immune for each of the 8 lineage clusters of cells.

The cells in this dataset were also classified into 16 clusters using k-means clustering. We used the cluster specific genes for these 16 clusters to find cell type annotations with CellKb Immune (Table S1). Cell types predicted by CellKb Immune are identical to the author-defined annotations. The top 10 hits for each gene list given to CellKb Immune is shown in Table S1.

3.4 Evaluation

To check the database quality and cell type matching performance of the CellKb Immune search method, we used the leave-one-out approach. We matched each of the 2,223 mouse marker gene sets with cell ontologies in CellKb Immune with all others in the database. 1,568 (70.5%) of gene sets were correctly matched to another gene set with the exact same ontology or its parent, while for 427 (19.2%) gene sets, the ontology of the first hit was a sibling i.e. a descendant of the same parent ontology (Table 1). Thus, 89.7% of marker gene sets were matched to another cell type with the same ontology, or its parent or sibling (Table S2).

We further evaluated the performance of the CellKb Immune by checking the number of correct cell types assigned to gene lists obtained from a bulk RNA-seq dataset of mouse immune cell types (12) from Immgen (21). Using the most significantly upregulated genes for each cell type in the query dataset, we searched the CellKb Immune database for matching cell types. We then compared the cell ontology of the top 10 hits with that of the ontology assigned to the cell types in the query dataset. Of the 98 cell type signatures representing 40 distinct cell types, CellKb Immune was able to find the exact matching ontology or its parent as the first hit in 61 instances (62.2%). In another 18 cell types (18.4%), the ontology of the first hit was a sibling i.e. a descendant of the same parent ontology as that in the bulk RNA-seq dataset (Table 1). Thus, a total of 79.5% of cell types in the query dataset had their first hit with the same ontology, its parent or sibling (Table S3).

These results show that 1) the CellKb Immune database includes high quality cell type signatures from single-cell experiments, and 2) the search methodology used by CellKb Immune to find matching cell types is able to recapitulate the original cell types from within its own database and other unseen bulk and single-cell datasets.

4 Discussion

CellKb Immune has been designed as a resource to assign cell types to cells in single-cell RNA-seq experiments. CellKb Immune addresses several issues in existing single-cell reference databases and tools used for this purpose.

CellKb Immune consists of user-defined marker gene sets from literature. These carry significant knowledge since the cell types in the form of cell clusters are often selected by authors based on biological information. CellKb Immune is able to capture and aggregate this biological information. On the other hand, several reference databases reanalyze the raw expression patterns using a common pipeline while ignoring the methodology and cluster definitions used by the authors in the original study, thus losing valuable biological information.

CellKb Immune provides deep annotations, in the form of extensive cell type information and description, curated and validated marker genes ranked by significance, along with fold change and significance values associated with gene expression. Additionally, CellKb Immune provides a web-based interface to find matching cell types in published data sets given a user gene list, independent of the experimental platform, analysis methods and insufficient marker gene sets. Thus, users do not need to spend time integrating data and search methods programmatically.

Unlike other methods which require the presence of associated expression or fold change values and the same number of genes in all target cell types, the rank-based search method used by CellKb Immune enables searching through marker gene sets of differing sizes, in the absence of expression fold changes, and independent of experimental platform and pre-processing methods. Searching of cell types happens by direct comparison of the query gene list with the marker genes of every cell type in CellKb Immune using a fast and reliable method without any type of score aggregation or machine learning. As a result, CellKb Immune provides information about the matching cell type in the form of the publication, experimental details and the overlapping genes between the query and matching marker gene set. This allows users to further explore the study to which their cell type matches.

Since CellKb Immune contains marker gene sets from multiple publications for a single a cell type, it can also be used to evaluate the validity of known marker genes for the cell type of interest. The reliability score assigned to each cell type marker gene set can also be used to identify those that are incorrectly annotated in publications. Additionally, comparing the query gene list with marker gene sets from multiple publications helps assess the accuracy of the matching cell types found by CellKb Immune.

Thus, CellKb Immune provides an easy-to-use reference database with a fast and reliable method to find matching cell types and marker gene sets to help annotate cells in single-cell experiments in a single tool.

5 Conflict of Interest

Ashwini is CEO at Combinatics Inc., a privately-held company. Ajay and Ashwini are stock-holders in Combinatics Inc.

6 Author Contributions

Ashwini conceived of the project and collected the data. Ajay and Ashwini designed and developed the database and web application, and prepared the manuscript.

7 References

1. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, Chak S, Naikawadi RP, Wolters PJ, Abate AR, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* (2019) doi:10.1038/s41590-018-0276-y

2. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* (2019) doi:10.1038/s41592-019-0535-3

3. Rozenblatt-Rosen O, Stubbington MJT, Regev A, Teichmann SA. The Human Cell Atlas: From vision to reality. *Nature* (2017) doi:10.1038/550451a

4. Han X, Wang R, Zhou Y, Fei L, Sun H, Lai S, Saadatpour A, Zhou Z, Chen H, Ye F, et al. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* (2018) doi:10.1016/j.cell.2018.02.001

5. Schaum N, Karkanias J, Neff NF, May AP, Quake SR, Wyss-Coray T, Darmanis S, Batson J, Botvinnik O, Chen MB, et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* (2018) doi:10.1038/s41586-018-0590-4

6. 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) doi:10.4049/jimmunol.1700272

7. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, Luo T, Xu L, Liao G, Yan M, et al. CellMarker: A manually curated resource of cell markers in human and mouse. *Nucleic Acids Res* (2019) doi:10.1093/nar/gky900

8. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database* (2019) doi:10.1093/database/baz046

9. Papatheodorou I, Moreno P, Manning J, Fuentes AMP, George N, Fexova S, Fonseca NA, Füllgrabe A, Green M, Huang N, et al. Expression Atlas update: From tissues to single cells. *Nucleic Acids Res* (2020) doi:10.1093/nar/gkz947

10. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z,

Edfors F, et al. A pathology atlas of the human cancer transcriptome. *Science (80-)* (2017) doi:10.1126/science.aan2507

11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* (2005) doi:10.1073/pnas.0506580102

12. H Y, CA L, RN R, SA R, B M, A W, F D, A C, A M, C D, et al. The cis-Regulatory Atlas of the Mouse Immune System. *Cell* (2019) **176**:897-912.e20. doi:10.1016/J.CELL.2018.12.036

13. López Y, Nakai K, Patil A. HitPredict version 4: Comprehensive reliability scoring of physical protein-protein interactions from more than 100 species. *Database* (2015) **2015**: doi:10.1093/database/bav117

14. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, Azov AG, Bennett R, Bhai J, et al. Ensembl 2021. *Nucleic Acids Res* (2021) **49**:D884–D891. doi:10.1093/NAR/GKAA942

15. Bard J, Rhee SY, Ashburner M. An ontology for cell types. *Genome Biol* (2005) doi:10.1186/gb-2005-6-2-r21

16. Mungall CJ, Torniai C, Gkoutos G V., Lewis SE, Haendel MA. Uberon, an integrative multi-species anatomy ontology. *Genome Biol* (2012) doi:10.1186/gb-2012-13-1-r5

17. Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease ontology: A backbone for disease semantic integration. *Nucleic Acids Res* (2012) doi:10.1093/nar/gkr972

18. Herrero J, Muffato M, Beal K, Fitzgerald S, Gordon L, Pignatelli M, Vilella AJ, Searle SMJ, Amode R, Brent S, et al. Ensembl comparative genomics resources. *Database* (2016) **2016**:96. doi:10.1093/DATABASE/BAV096

19. Single Cell Portal. Available at: https://singlecell.broadinstitute.org/single_cell/study/SCP345/ica-blood-mononuclear-cells-2-donors-2-sites?scpbr=immune-cell-atlas [Accessed October 25, 2021]

20. Hurskainen M, Mižíková I, Cook DP, Andersson N, Cyr-Depauw C, Lesage F, Helle E, Renesme L, Jankov RP, Heikinheimo M, et al. Single cell transcriptomic analysis of murine lung development on hyperoxia-induced damage. *Nat Commun* (2021) **12**:1565. doi:10.1038/s41467-021-21865-2

21. Aguilar SV, Aguilar O, Allan R, Amir EAD, Angeli V, Artyomov MN, Asinovski N, Astarita J, Austen KF, Bajpai G, et al. ImmGen at 15. *Nat Immunol* (2020) doi:10.1038/s41590-020-0687-4

8 Figures

Figure 1. CellKb Immune workflow – data collection, harmonization and search interface.

Figure 2. Number of marker gene sets for each species in CellKb Immune.

Figure 3. CellKb Immune user interface used to search matching cell types given a gene list.

9 Tables

Table 1. Performance evaluation of CellKb Immune

Evaluation database	**Marker gene sets**	**CellKb Immune performance**		
		Correct ontology match	**Parent ontology match**	**Total ontologies matched**
CellKb Immune Mouse (leave-one-out)	2,223	1,568 (70.5%)	427 (19.2%)	**1,995 (89.7%)**
Yoshida et al. (12) **Mouse (all-against-all)**	98	61 (62.2%)	18 (18.4%)	**79 (80.6%)**

10 Supplementary Material

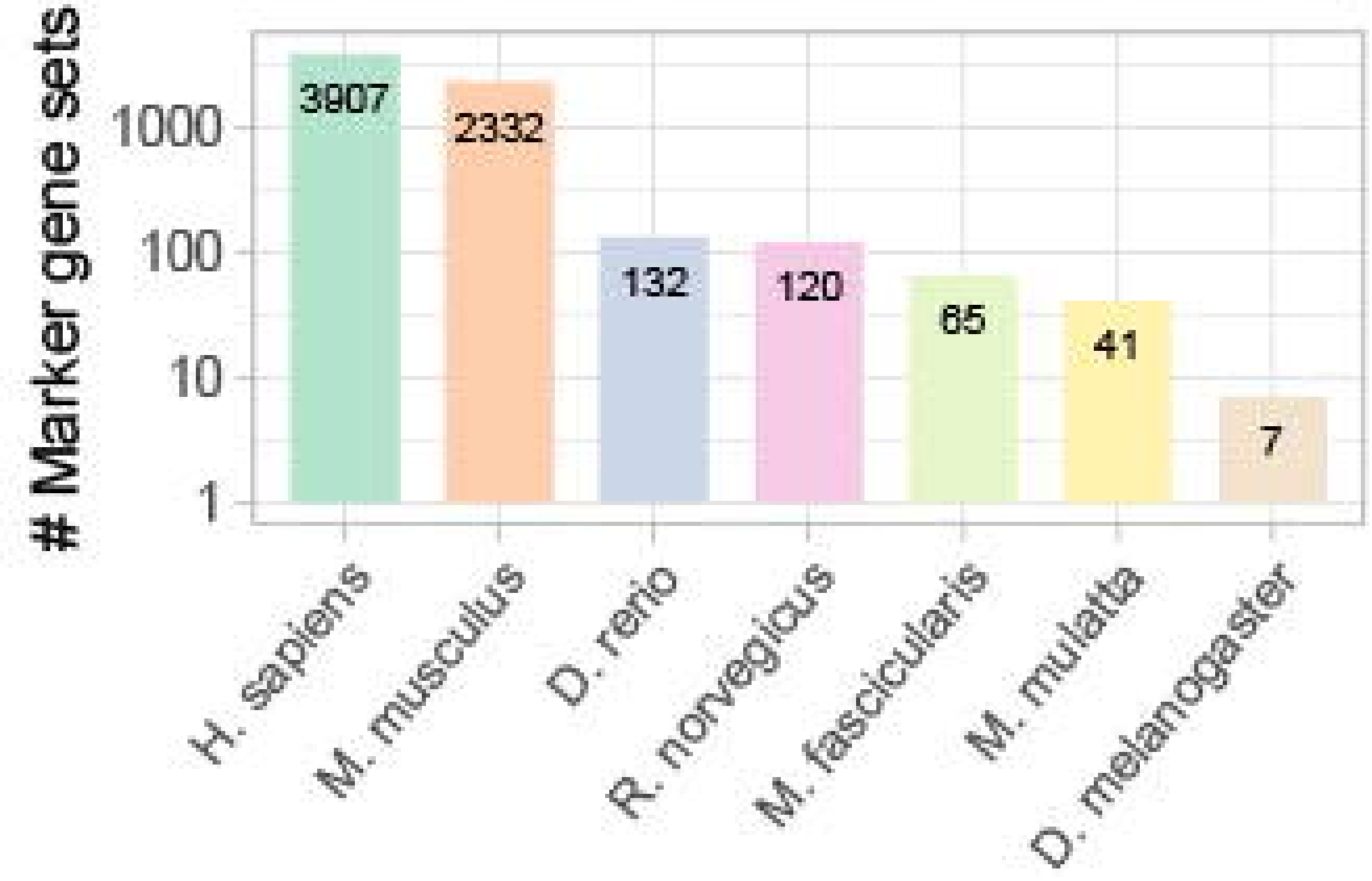
Table S1. List of cluster-specific genes for 8 and 16 pre-defined clusters in human peripheral blood mononuclear cells used as input to CellKb Immune. Top 10 results of matching cell types obtained from CellKb Immune.

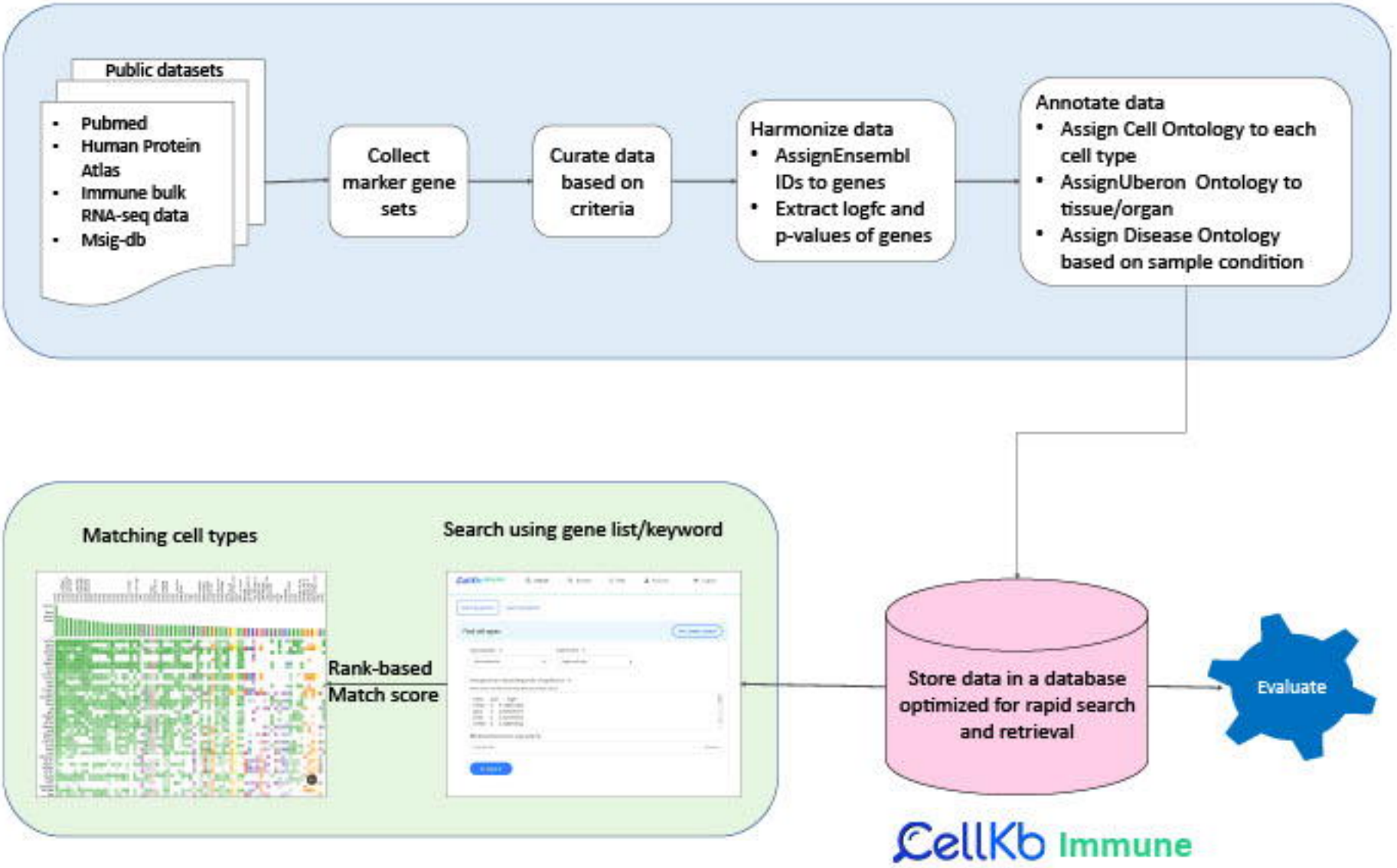
Table S2. Results of the leave-one-out evaluation in CellKb Immune shows the first hit of each mouse cell type marker gene set compared to all other cell types.

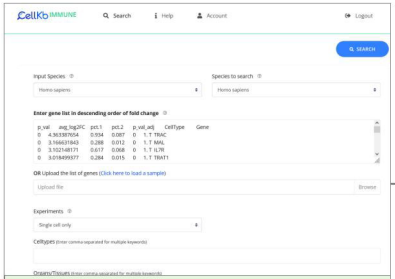
Table S3. First hit of each cell type marker gene set in a bulk RNA-seq dataset of immune cell types compared to all cell types in CellKb Immune mouse.

11 Data Availability Statement

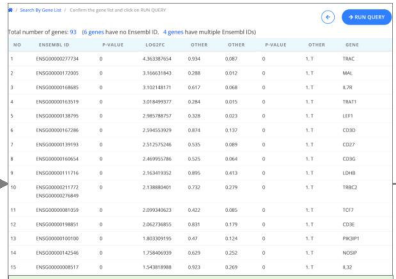
CellKb Immune is available at https://www.cellkb.com/immune. All data used in CellKb Immune is publicly available and appropriately referenced in the web application.


Marker gene sets
1000
100
10
1
3907
2332
132
120
65
41
7
H. sapiens
M. musculus
D. rerio
R. norvegicus
M. fascicularis
M. mulatta
D. melanogaster

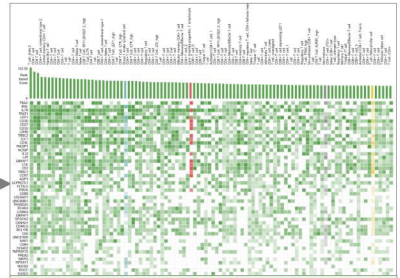

Public datasets
Pubmed
Human Protein Atlas
Immune bulk RNA-seq data
Msig-db
Collect marker gene sets
Curate data based on criteria
Harmonize data
AssignEnsembl IDs to genes
Extract logfc and p-values of genes
Annotate data
Assign Cell Ontology to each cell type
AssignUberon Ontology to tissue/organ
Assign Disease Ontology based on sample condition
Matching cell types
Search using gene list/keyword
Rank-based
Match score
Store data in a database optimized for rapid search and retrieval
Evaluate
CellKb Immune



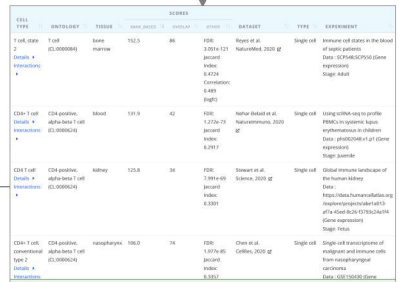
Search cell types using ranked gene list



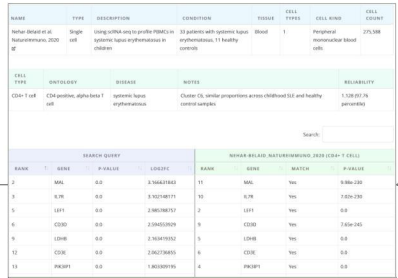
View mapping of gene names to Ensembl IDs



View heatmap of matching cell types with gene scores



View table of matching cell types with source details



View matching marker genes in hit cell type

View cell type specific protein interaction network