CellKb Immune: a manually curated database of mammalian immune marker gene sets optimized for rapid cell type identification

Ajay Patil and Ashwini Patil*

Combinatics Inc., Shirokanedai, Minato-ku, Tokyo, Japan.

*Corresponding email: ashwini@combinatics.com

Abstract

Single-cell RNA-seq is widely used to study transcriptional patterns of genes in individual cells. 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 a fast 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 mammalian immune response. It finds matching cell types in literature given a list of genes using a novel rank-based algorithm optimized for rapid searching across marker gene lists of differing lengths. We evaluated the contents and search algorithm of CellKb Immune using a leave-one-out approach. We further used CellKb Immune to annotate previously defined marker gene sets from Immgen to confirm its accuracy and coverage. CellKb Immune provides an easy to use database with a fast and reliable method to find matching cell types and annotate cells in single-cell experiments in a single tool. It is available at https://www.cellkb.com/immune.

1. Introduction

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[1]. Annotating cells by assigning a type or state to them is important for correct interpretation of results. This requires a single tool 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 a matching cell types within the database. On the one hand, various computational tools have been developed for the identification of matching cell types given a reference database and a gene list[2,3]. On the other hand, several single cell reference datasets or atlases have been prepared[1,4,5] or compiled into aggregate reference databases of cell types and marker gene sets[6–9]. However, it takes considerable effort from the user to download the reference database/dataset and programmatically integrate the search method with it for cell type annotation.

CellKb Immune addresses all these issues by collecting and aggregating mammalian immune-related cell type signatures from literature and providing an online interface to find matching annotated cell types easily, quickly and accurately.

2. Results

2.1 Database content

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[2,10] in *Mus musculus*.

Version 1 of CellKb Immune contains (Table S1):

- 1,521 marker gene sets from mouse
- 124 unique annotated hematopoietic cell types (Cell Ontology CL:0000988 and all its descendants)
- 130 publications selected from approximately 7000 studies published between 2013 and 2020
- 80 diseases

Marker gene sets are selected from published experimental studies if they satisfy multiple sanity checks such as 1) Deposition of raw data in public databases, 2) Availability of data for download, 3) Type of experimental method used, 4) Number of cells studied, 5) Computational methods used to normalize, filter and cluster cell types, along with identification of cluster-specific genes, 6) Availability of associated values (e.g. average expression, fold change, statistical significance), 7) Number of valid gene identifiers in the marker gene set as mapped to the latest version of the Ensembl database[11].

Marker gene sets are extensively curated to include valid genes and cell types only. Cell types, tissue names and disease conditions are assigned standardized ontology terms[12–14]. 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. The cell type specific marker genes are either directly taken as defined by the authors, or they are calculated based on the associated values provided. 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. Finally, all marker genes and their annotations are stored in a uniform format that enables rapid search and retrieval of the data.

2.2 Functionality

The CellKb Immune web interface allows users to search or browse cell types and their marker gene sets.

1) Search matching cell types by gene list

CellKb Immune can be used to annotate cells given one or more ranked lists of cluster-specific genes from single-cell experiments. Associated values such as log fold changes and p-values are also accepted. Genes in the user query are compared with every marker gene set in the database and a match score is computed, 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 (See Supplementary Materials). This results in cell types sharing highly ranked genes with the query being assigned higher match scores. The match score also accounts for differences in the sizes of gene lists between the query and various cell types, such that cell types with fewer significant genes are not disregarded. Other statistics such as the Fisher's exact test, Pearson's correlation coefficient and Jaccard index are also calculated for each cell type match. Ranks of common genes in the user gene list and matching cell types are provided.

The search result shows the matching cell types along with the annotations extracted from their publication (tissue, experimental condition, disease, Pubmed ID, ontology terms), in addition to a heatmap of the match score assigned to each gene for each cell type. Search results are available for download.

2) Search matching cell types by keyword

CellKb Immune can also be used as a reference database to find previously published datasets related to a specific experimental condition, cell type, gene, tissue or disease via the keyword search. For each cell type identifier, CellKb Immune displays the top 10 marker genes along with associated values from the original publication.

3) Browse cell types

The Browse functionality in CellKb Immune is an index using which users can drill down through categories and find the cell type of interest and their marker gene sets, experimental annotations and publication information.

2.3 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 compared each of the 1,510 marker gene sets with cell ontologies in CellKb Immune with all others in the database. 1,111 (73.6%) of gene sets were correctly matched to another gene set with the exact same ontology or its parent, while for 268 (17.8%) gene sets, the ontology of the first hit was a sibling i.e. a descendant of the same parent ontology (Table 1). Thus, 91.4% of marker gene sets were matched to another cell type with the same ontology, or its parent or its sibling (Table S2).

We further evaluated the performance of the CellKb Immune using marker gene lists from Immgen[15] (which is not included in CellKb Immune). Using the most significantly upregulated genes for each cell type in Immgen, 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 Immgen. Of the 276 cell type signatures representing 77 distinct cell types in Immgen, CellKb Immune was able to find the exact matching ontology or its parent as the first hit in 143 instances (51.8%). In another 41 cell types (14.9%), the ontology of the first hit was a sibling i.e. a descendant of the same parent ontology as that in Immgen (Table 1). Thus, a total of 66.7% of cell types in Immgen had their first hit with the same ontology, its parent or its sibling (Table S3). This performance is expected to improve as data in CellKb Immune increases.

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 to some extent, from Immgen.

3. Discussion

CellKb Immune addresses several issues in existing single-cell databases. 1) Reanalyzing the raw expression patterns using a common pipeline, as is done by some reference databases, ignores the cluster definitions given in the original study. User-defined marker gene sets in literature 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. 2) 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. 3) 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. 4) 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 search method used by CellKb Immune enables searching through marker gene sets of differing size, 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.

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

4. Methods

4.1 Novel rank-based match score

CellKb Immune can be searched using a list of genes to find matching cell type signatures published in literature. CellKb Immune uses a rank-based score to identify the cell type marker gene sets matching the user's gene list.

Match score for the jth cell type in CellKb = $\sum_{i=1}^{n} m_i * \left(f_1(r_i) * f_2(d_i) * f_3(c_j)\right)$

Where n = number of genes in the user list

m_i = 1 if the ith gene is present in the jth cell type, 0 otherwise

r_i = min (r_{iq}, r_{ij})

r_{iq} = rank of the ith gene in the user list

r_{ij} = rank of the ith gene in the jth cell type

d_i = difference in ranks of the ith gene in the user query and cell type

c_j = number of marker genes in the jth cell type

$f_1(r_i),\ f_2(d_i),\ f_3(c_j)$ are bounded functions, inversely related to $log_{10}\ (r_i),\ log_2\ (d_i)$ and $log_{10}\ (c_j)$, respectively

The match score also accounts for differences in the sizes of gene lists between the query and various cell types, such that cell types with fewer significant genes are not disregarded.

4.2 Cell type reliability score

A reliability score is calculated for each cell type marker gene set is calculated as the ratio of its average rank correlation with marker gene sets sharing its cell type ontology and its average rank correlation with marker gene sets belonging to all other cell type ontologies.

Data availability: All data in CellKb Immune is collected from publicly available datasets and is appropriately referenced at https://www.cellkb.com/immune.

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

5. References

1. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* (2018) doi:10.1038/s41586-018-0590-4.
2. Aran, D. *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.
3. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* (2019) doi:10.1038/s41592-019-0535-3.
4. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: From vision to reality. *Nature* (2017) doi:10.1038/550451a.
5. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* (2018) doi:10.1016/j.cell.2018.02.001.
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. *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, L. M. & Björkegren, J. L. M. 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. *et al.* Expression Atlas update: From tissues to single cells. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkz947.
10. Subramanian, A. *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.
11. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* (2020) doi:10.1093/nar/gkz966.
12. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome Biol.* (2005) doi:10.1186/gb-2005-6-2-r21.
13. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome Biol.* (2012) doi:10.1186/gb-2012-13-1-r5.
14. Schriml, L. M. *et al.* Disease ontology: A backbone for disease semantic integration. *Nucleic Acids Res.* (2012) doi:10.1093/nar/gkr972.
15. Aguilar, S. V. *et al.* ImmGen at 15. *Nature Immunology* (2020) doi:10.1038/s41590-020-0687-4.

6. Tables

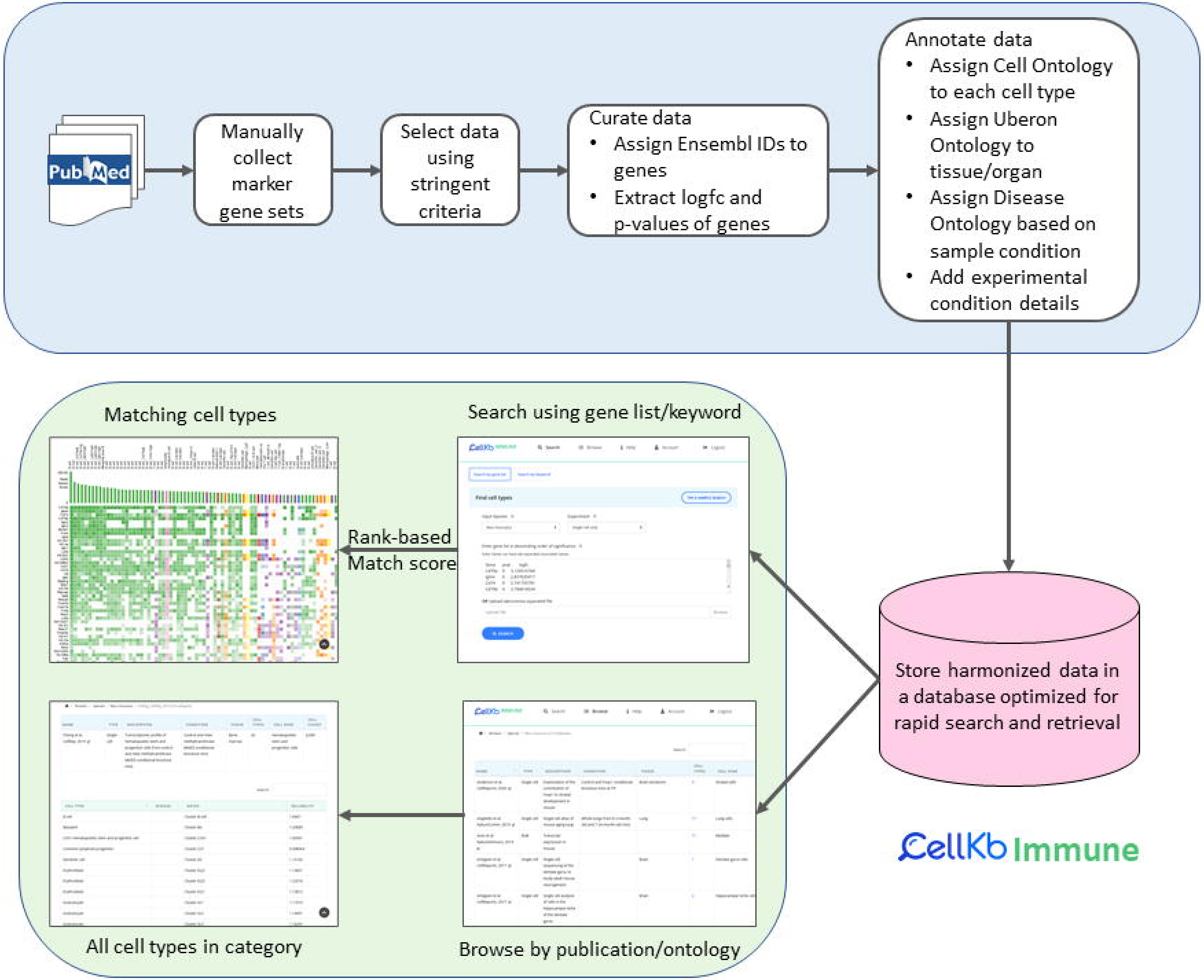
Table 1. Performance evaluation of CellKb Immune

Evaluation database	**Marker gene sets**	**CellKb Immune performance**	
		Correct ontology match	**Parent ontology match**
CellKb Immune (leave-one-out)	1,510	1,111 (73.6%)	268 (17.8%)
Immgen (all-against-all)	276	143 (51.8%)	41 (14.9%)

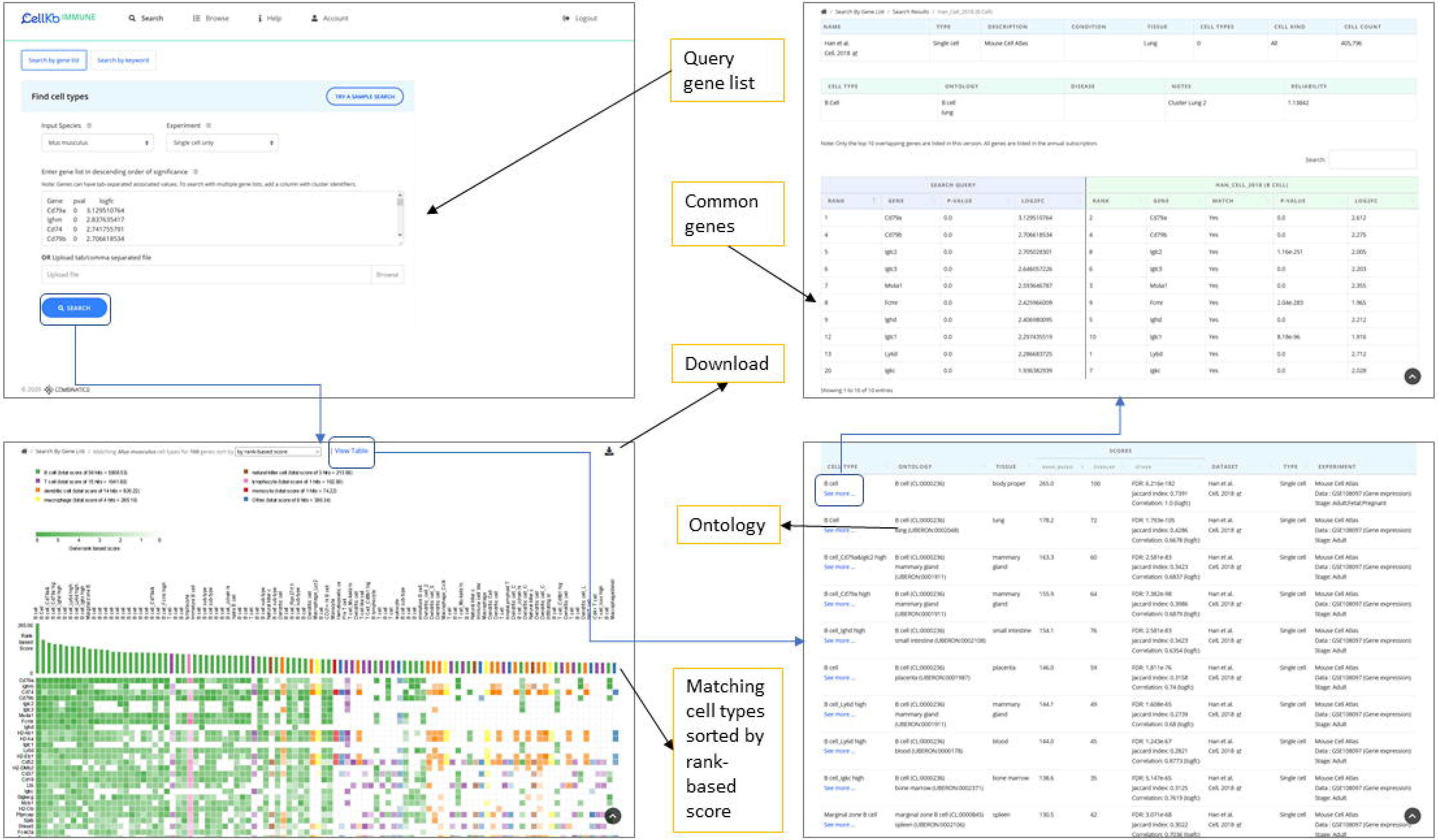
7. Figure Legends

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

Figure 2. CellKb Immune user interface used to search matching cell types given a gene list.


PubMed
Manually collect marker gene sets
Select data using stringent criteria
Curate data
• Assign Ensembl IDs to genes
• Extract logfc and p-values of genes
Annotate data
• Assign Cell Ontology to each cell type
• Assign Uberon Ontology to tissue/organ
• Assign Disease Ontology based on sample condition
• Add experimental condition details
Store harmonized data in a database optimized for rapid search and retrieval
CellKb Immune
Search using gene list/keyword
Rank-based Match score
Matching cell types
Browse by publication/ontology
All cell types in category

Search by ranked gene list



CellKb IMMUNE Search Browse Help Account Logout

Search by gene list | Search by keyword

Find cell types TRY A SAMPLE SEARCH

Input Species: Mus musculus

Experiment: Single cell only

Enter gene list in descending order of significance

Note: Genes can have tab-separated associated values. To search with multiple gene lists, add a column with cluster identifiers.

```
Gene    pval    logfc
Cd79a   0       3.1295107664
Ighm    0       2.837635417
Cd74    0       2.741755791
Cd79b   0       2.706618534
```

OR Upload tab/comma separated file

Upload file | Browse

SEARCH

© 2020 COMBINATICS

Query gene list

Common genes

Download

Ontology

Matching cell types sorted by rank-based score

Search By Gene List / Search Results / Han_Cell_2018 (B Cell)

NAME	TYPE	DESCRIPTION	CONDITION	TISSUE	CELL TYPES	CELL KIND	CELL COUNT
Han et al. Cell, 2018	Single cell	Mouse Cell Atlas		Lung	0	All	405,796

CELL TYPE	ONTOLOGY	DISEASE	NOTES	RELIABILITY
B Cell	B cell lung		Cluster Lung 2	1.13842

Note: Only the top 10 overlapping genes are listed in this version. All genes are listed in the annual subscription.

Search:

SEARCH QUERY				HAN_CELL_2018 (B CELL)				
RANK	GENE	P-VALUE	LOG2FC	RANK	GENE	MATCH	P-VALUE	LOG2FC
1	Cd79a	0.0	3.129510764	2	Cd79a	Yes	0.0	2.612
4	Cd79b	0.0	2.706618534	4	Cd79b	Yes	0.0	2.275
5	Iglc2	0.0	2.705028301	8	Iglc2	Yes	1.16e-251	2.005
6	Iglc3	0.0	2.646057226	6	Iglc3	Yes	0.0	2.203
7	Ms4a1	0.0	2.593646787	3	Ms4a1	Yes	0.0	2.355
8	Fcmr	0.0	2.425966009	9	Fcmr	Yes	2.04e-283	1.965
9	Ighd	0.0	2.406980095	5	Ighd	Yes	0.0	2.212
12	Iglc1	0.0	2.297435519	10	Iglc1	Yes	8.10e-96	1.916
13	Ly6d	0.0	2.286683725	1	Ly6d	Yes	0.0	2.712
20	Igkc	0.0	1.936382939	7	Igkc	Yes	0.0	2.028

Showing 1 to 10 of 10 entries

Search By Gene List / Matching *Mus musculus* cell types for 100 genes sort by: by rank-based score | View Table

B cell (total score of 54 hits = 5868.53)
T cell (total score of 15 hits = 1041.83)
dendritic cell (total score of 14 hits = 630.22)
macrophage (total score of 4 hits = 265.10)
natural killer cell (total score of 3 hits = 213.00)
lymphocyte (total score of 1 hits = 102.80)
monocyte (total score of 1 hits = 74.22)
Other (total score of 8 hits = 380.34)
6 5 4 3 2 1 0
Gene rank based score
265.00
Rank based Score
0
Cd79a
Ighm
Cd74
Cd79b
Iglc2
Iglc3
Ms4a1
Fcmr
Ighd
H2-Ab1
H2-Aa
Iglc1
Ly6d
H2-Eb1
Cd52
H2-DMb2
Cd37
Cd19
Ltb
Igkc
Siglecg
Mzb1
H2-Ob
Pou2af1
Spib
Blnk
Fcer2a

CELL TYPE	ONTOLOGY	TISSUE	SCORES			DATASET	TYPE	EXPERIMENT
			RANK_BASED	OVERLAP	OTHER			
B cell See more ...	B cell (CL:0000236)	body proper	265.0	100	FDR: 6.216e-182 Jaccard index: 0.7391 Correlation: 1.0 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult;Fetal;Pregnant
B Cell See more ...	B cell (CL:0000236) lung (UBERON:0002048)	lung	178.2	72	FDR: 1.793e-105 Jaccard index: 0.4286 Correlation: 0.6678 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult
B cell_Cd79a&Iglc2 high See more ...	B cell (CL:0000236) mammary gland (UBERON:0001911)	mammary gland	163.3	60	FDR: 2.581e-83 Jaccard index: 0.3423 Correlation: 0.6837 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult
B cell_Cd79a high See more ...	B cell (CL:0000236) mammary gland (UBERON:0001911)	mammary gland	155.9	64	FDR: 7.382e-88 Jaccard index: 0.3986 Correlation: 0.6879 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult
B cell_Ighd high See more ...	B cell (CL:0000236) small intestine (UBERON:0002108)	small intestine	154.1	76	FDR: 2.581e-83 Jaccard index: 0.3423 Correlation: 0.6354 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult
B cell See more ...	B cell (CL:0000236) placenta (UBERON:0001987)	placenta	146.0	59	FDR: 1.871e-76 Jaccard index: 0.3158 Correlation: 0.74 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult
B cell_Ly6d high See more ...	B cell (CL:0000236) mammary gland (UBERON:0001911)	mammary gland	144.7	49	FDR: 1.608e-65 Jaccard index: 0.2739 Correlation: 0.68 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult
B cell_Ly6d high See more ...	B cell (CL:0000236) blood (UBERON:0000178)	blood	144.0	45	FDR: 1.243e-67 Jaccard index: 0.2821 Correlation: 0.6773 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult
B cell_Igkc high See more ...	B cell (CL:0000236) bone marrow (UBERON:0002371)	bone marrow	138.6	35	FDR: 5.147e-65 Jaccard index: 0.3125 Correlation: 0.7619 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult
Marginal zone B cell See more ...	marginal zone B cell (CL:0000845) spleen (UBERON:0002106)	spleen	130.5	42	FDR: 3.071e-68 Jaccard index: 0.3022 Correlation: 0.7036 (logfc)	Han et al. Cell, 2018	Single cell	Mouse Cell Atlas Data : GSE108097 (Gene expression) Stage: Adult