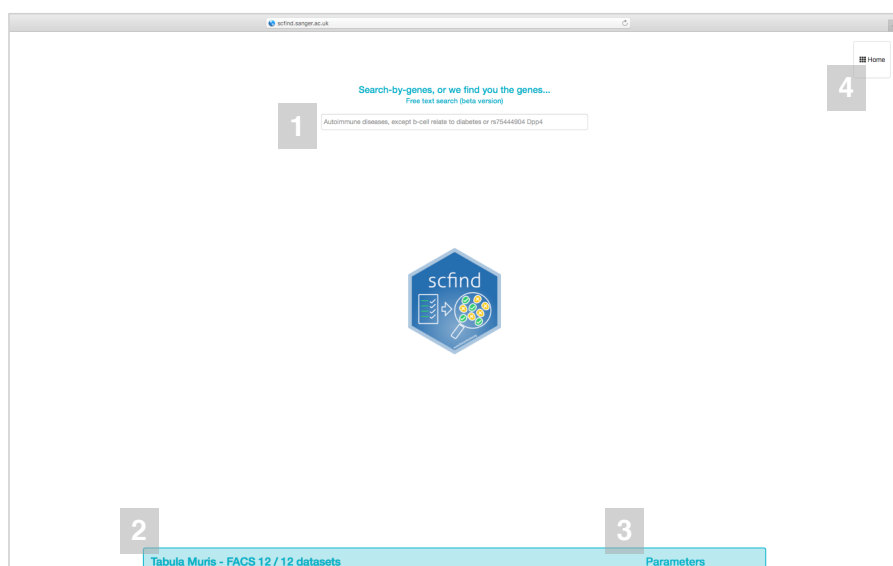




*Search engine for genes in
large single-cell sequencing collections*
<https://scfind.sanger.ac.uk/>
(User Guide)

The User interface of Scfind



- 1 Search box:**
Users can input a list of genes, SNPs ID, MeSH ID or medical/biological/chemical terms
- 2 Dataset check box:**
Where the user can find all the datasets in the index and limit the search result by dataset / tissue
- 3 Parameters:**
Where the user can adjust the parameters for each search
- 4 Home button:**
To the home page of our index collection

4 Steps : From your a query to your target cell type

I

Search-by-genes, or we find you the genes...
Free text search (beta version)

endurance capacity for marathon col5a1 rs12722 or MESH:D001177

II

Search-by-genes, or we find you the genes...
Free text search (beta version)

endurance capacity for marathon col5a1 rs12722 or MESH:D001177

Query Summary

Cells Co-express: AND

Selected genes: 1474

Optimized queries for AND genes:

Query	TF-IDF	Cell No.	Cell types
1. Col5a1	27276	9793	0
2. Col5a1	17118	8198	0
3. Col5a1	10589	4480	0
4. Col5a1	10008	8890	0
5. Col5a1	9802.8	9029	0

Showing 1 to 5 of 54 entries

Exporting specific cell types?
Select above recommended query!

Tabula Muris - FACS 12 / 12 datasets

Parameters

Step I: Input your query

- The bar chart **1** is a summary of the number of cells in the index expressing the gene

- User can always amend the initial gene list by clicking on ☒

Step II: select one of our recommended sub-query on table **2** according to TF-IDF score

- subqueries are ranked by TF-IDF score (*Highest = likely the best*)
- click on **TF-IDF** to sort TF-IDF score in order

III

Search-by-genes, or we find you the genes...
Free text search (beta version)

endurance capacity for marathon col5a1 rs12722 or MESH:D001177

Query Summary

Cells Co-express: AND

Selected genes: 1474

Optimized queries for AND genes:

Query	TF-IDF	Cell No.	Cell types
1. Col5a1	27276	9793	0
2. Col5a1	17118	8198	0
3. Col5a1	10589	4480	0
4. Col5a1	10008	8890	0
5. Col5a1	9802.8	9029	0

Showing 1 to 5 of 54 entries

Selected found 20 cell types for your query

Cell types

Cell type	Cell No.	Cell types
Endothelial cell	985	1196
Myeloid cell	287	887
Neuronal cell	141	280
Stromal cell	380	1111

Showing 1 to 4 of 20 entries

Tabula Muris - FACS 12 / 12 datasets

Parameters

Step III: Select the enriched cell types on table **3**

- Cells expressing all genes in the query are grouped by cell type. The user can rank them based on enrichment by clicking on **p-value**

- User can limit the dataset by clicking on ☒ in **4**

- User can adjust the degree of greediness in **4** to get the best gene set for each term in the query. The higher the degree, the more gene (but less relevancy) will return.

IV

Search-by-genes, or we find you the genes...
Free text search (beta version)

endurance capacity for marathon col5a1 rs12722 or MESH:D001177

Query Summary

Cells Co-express: AND

Selected genes: 1474

Optimized queries for AND genes:

Query	TF-IDF	Cell No.	Cell types
1. Col5a1	27276	9793	0
2. Col5a1	17118	8198	0
3. Col5a1	10589	4480	0
4. Col5a1	10008	8890	0
5. Col5a1	9802.8	9029	0

Showing 1 to 5 of 54 entries

Selected found 20 cell types for your query

Cell types

Cell type	Cell No.	Cell types
Endothelial cell	985	1196
Myeloid cell	287	887
Neuronal cell	141	280
Stromal cell	380	1111

Showing 1 to 4 of 20 entries

Marker evaluation of Col5a1 in 1 cell type

UMAP

Word Cloud

Tabula Muris - FACS 12 / 12 datasets

Parameters

Step IV: Check the best representing genes

- Click on any cell type you are interested in **3**
- 5** allows you to view the cells in a UMAP, compare *precision / recall / f1* scores of each gene corresponding to each cell type and generate a *Word Cloud* relevant to your query.

- User can click on **Copy** **CSV** **Excel** to export the tables as raw text, csv file or .xlsx file