Supplementary Information 1: SingleR overview

Contents

Introduction

Recent advances in single cell RNA-seq (scRNA-seq) have enabled an unprecedented level of granularity in characterizing gene expression changes in disease models. Multiple single cell analysis methodologies have been developed to detect gene expression changes and to cluster cells by similarity of gene expression. However, the classification of clusters by cell type relies heavily on known marker genes, and the annotation of clusters is performed manually. This strategy suffers from subjectivity and limits adequate differentiation of closely related cell subsets. Here, we present *SingleR*, a computational method for unbiased cell type recognition of scRNA-seq. *SingleR* leverages reference transcriptomic datasets of pure cell types to infer the cell of origin of each of the single cells independently. *SingleR*'s annotations combined with Seurat, a processing and analysis package designed for scRNA-seq, provide a powerful tool for the investigation of scRNA-seq data. We developed an R package to generate annotated scRNA-seq objects that can then use the *SingleR* web tool for visualization and further analysis of the data – http://comphealth.ucsf.edu/SingleR.

Here we explain in details the *SingleR* pipeline and present examples of applying SingleR on publicly available mouse and human scRNA-seq datasets.

SingleR specifications

Reference set: A comprehensive transcriptomic dataset (microarray or RNA-seq) of pure cell types, preferably with multiple samples per cell type.

- **Mouse:** We processed and annotated two reference mouse datasets:
- Immunological Genome Project (ImmGen): a collection of 830 microarray samples, which we classified to 20 main cell types and further annotated to 253 subtypes[1].
- A dataset of 358 mouse RNA-seq samples annotated to 28 cell types[2]. This data set is especially useful for brain-related samples.
- **Human:** For human datasets we use the following reference datasets:
- Human Primary Cell Atlas (HPCA): a collection of Gene Expression Omnibus (GEO datasets), which contains 713 microarray samples classified to 38 main cell types and further annotated to 169 subtypes[3].
- Blueprint+Encode: Blueprint Epigenomics, 144 RNA-seq pure immune samples annotated to 28 cell types[4]. Encode: 115 RNA-seq pure stroma and immune samples annotated to 17 cell types[5]. Altogether, 259 samples classified to 43 cell types.
- For specific applications, smaller datasets can be applicable. *SingleR* is flexible and can be used with any reference dataset.

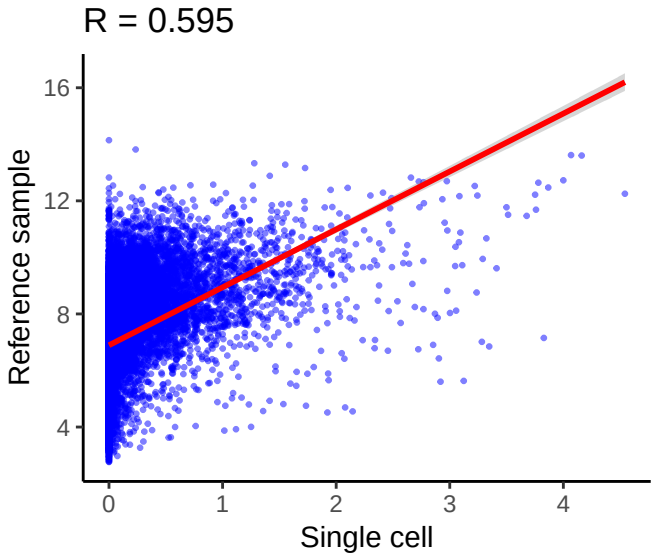
- Samples in each dataset are annotated to highly granular cell states, and also to broad categories, which we term 'main cell types' (i.e. all macrophage subtypes are annotated together). It is suggested to start the analysis by using the 'main types' mode before diving deeper to all cell states.
- All above reference datasets are available with the *SingleR* R package. With loading the *SingleR* library the following objects are observable: immgen, mouse.rna.seq, blueprint_encode, hpca. Each list contains a matrix of the gene expression, the annotations, and the differentially expressed genes between every two cell types.

Single-cell set: Single-cell RNA-seq dataset. It is a good practice to filter-out cells with non-sufficient genes identified and genes with non-sufficient expression across cells. As default we filter-out cells containing less than 500 genes and consider genes which were found in at least one sample. The effect of the gene number threshold on accuracy is discussed in Supplementary Information 2.

Annotation: *SingleR* runs in two modes: (1) Single cell: the annotation is performed for each single cell independently. (2) Cluster: the annotation is performed on predefined clusters, where the expression of a cluster is the sum expression of all cells in the cluster.

Step 1: Spearman correlations

Spearman coefficient is calculated for single cell expression with each of the samples in the reference dataset. The correlation analysis is performed only on variable genes in the reference dataset. The example below shows a correlation between the expression of a single cell (x-axis) and a reference sample (y-axis). Each dot in this scatter plot is a gene:

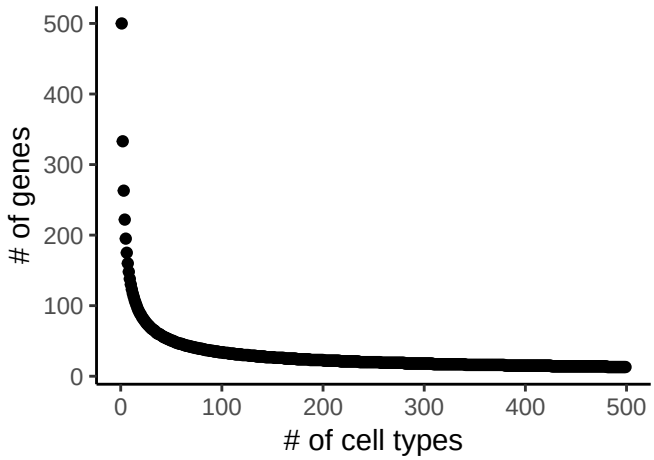


- **Variable genes:** *SingleR* supports two modes for choosing the variable genes in the reference dataset.

1. 'sd' - genes with a standard deviation across all samples in the reference dataset over a threshold. We choose thresholds such that we start with 3000-4000 genes.
2. 'de' - top N genes that have a higher median expression in a cell type compared to each other cell type. We use a varying N, depending on the number of cell types used in the analysis (K):

$N = round(333 * \frac{2}{3}^{log_2 K})$

This function creates the distribution below. The y-axis shows the number of differentially expressed genes used for each pair of cell types used in the analysis. In the first round, when there are many cell types to scores, SingleR uses only a small number of genes, but in the last rounds of the fine-tuning, this number increases up to 500 in the last iteration:



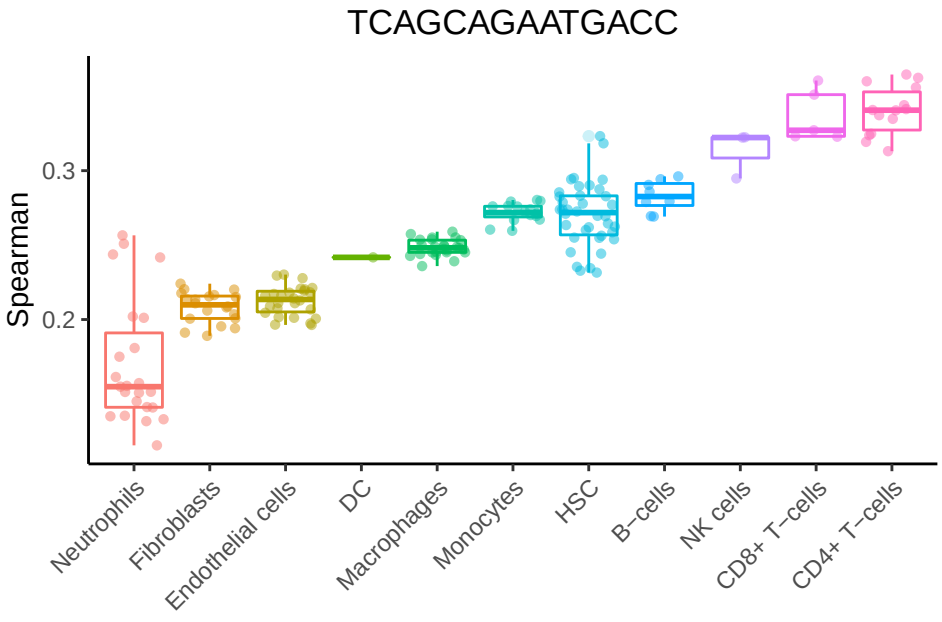
More details can be found in the *SingleR* code. This mode is the default mode and is used in all the studies presented in the manuscript.

Step 2: Aggregation of scores by cell types

Multiple correlation coefficients per cell type are aggregated according to the named annotations of the reference dataset to provide a single value per cell type per single cell. As described above, the samples are aggregated by broad cell types ('main') or with higher granularity of cell subsets. The default is the 80th percentile of the correlation values per cell type.

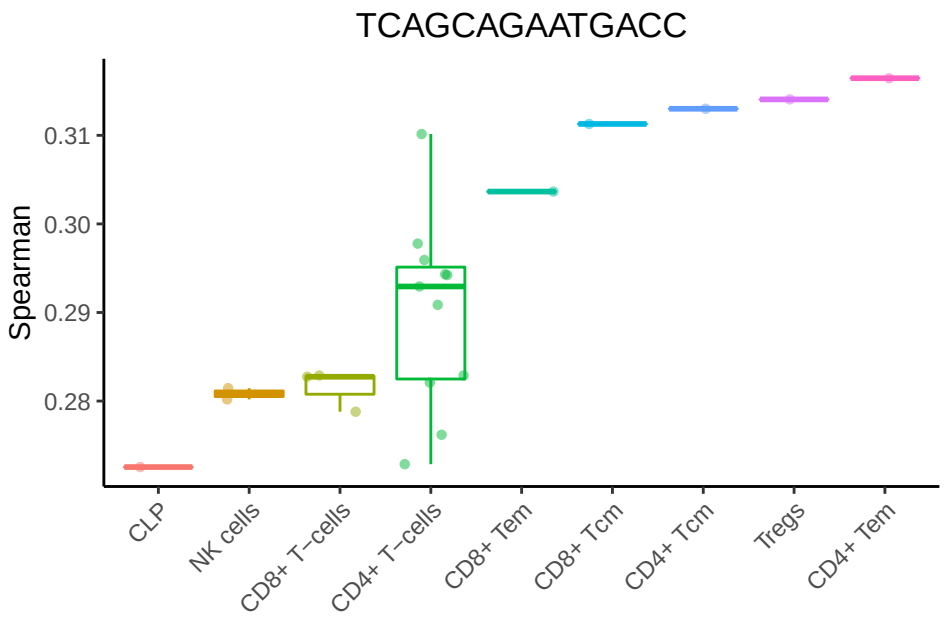
Below is an example for the annotation process for one human single cell whose 10x barcode is TCAGCAGAATGACC. Here the dots are Spearman coefficients of all reference samples (using the Blueprint +Encode reference) with one single cell. The Spearman coefficients were aggregated by cell type (here, a reduced set of the main cell types for simplicity). The SingleR score for each cell type is the 80 percentile in each of the boxplots. This cell type is clearly a T-cell or an NK cell, but its not clear what type exactly.

```
## [1] "Number of DE genes:2375"
## [1] "Number of cells: 2"
```



The analysis above grouped together cell subsets and states to main cell types. *SingleR* allows a more granular view of cell types (showing only top scoring cell types):

```
## [1] "Number of DE genes:3782"
## [1] "Number of cells: 2"
```

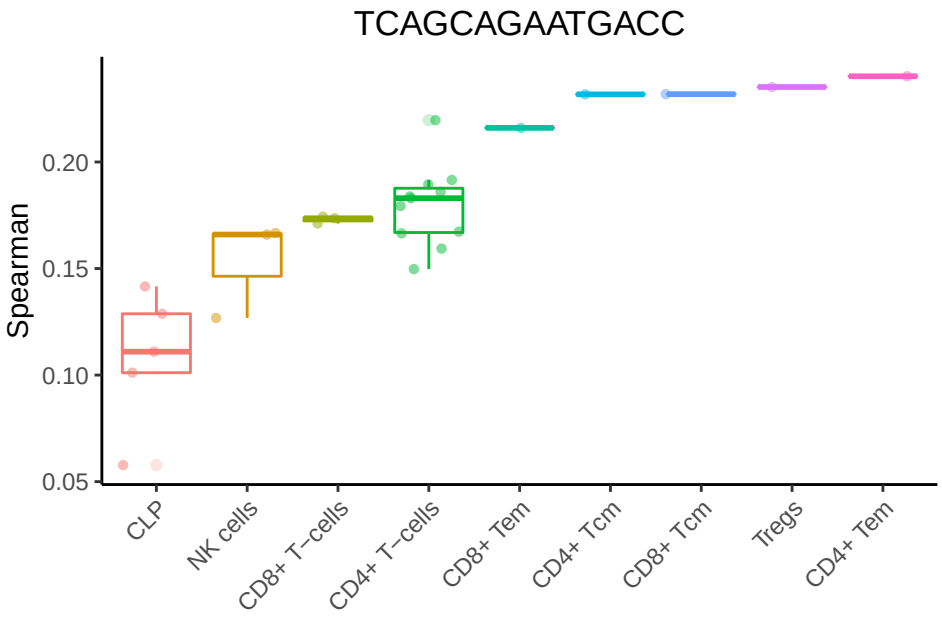


Step 3: Fine-tuning

In this step *SingleR* reruns the correlation analysis, but only for the top cell types from step 2. The analysis is performed only on **variable genes between these cell types**. The lowest value cell type is removed (or a margin of more than 0.05 below top value), and then this step is repeated until only two cell types remain. The cell type corresponding to the top value after the last run is assigned to the single cell.

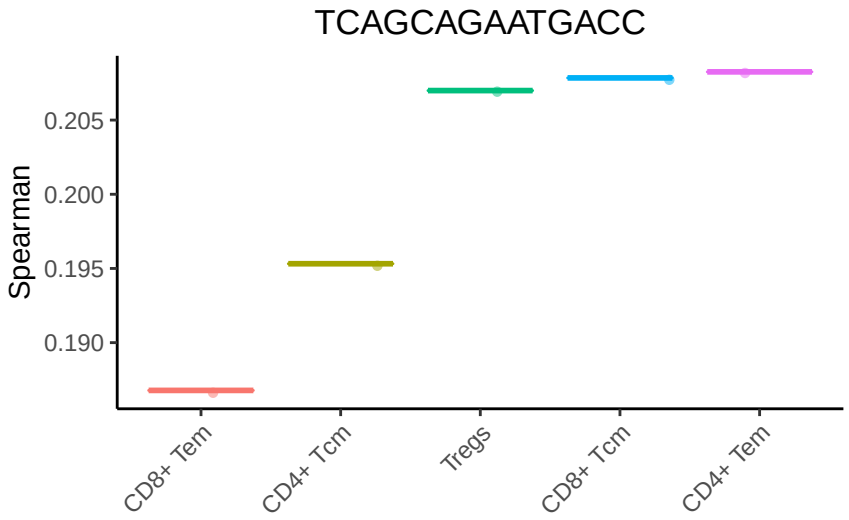
In the example above, *SingleR* is clearly suggesting the single cell is a memory T-cell. However, it is hard to suggest which of these cell subsets best fits it. The fine-tuning step helps to diffentiate closely related cell types. In the first fine-tuning iteration the top cell types (up to 0.05 difference from the CD4+ Tem score) are chosen. The Spearman correlation analysis is then performed, but only using the variable genes between those cells. Before fine-tuning, with all cell types, 3,782 genes were used. In the first fine-tuning iteration only 1,819 genes are used to differentiate 9 cell types.

```
## [1] "Number of DE genes:1819"
## [1] "Number of cells: 2"
```



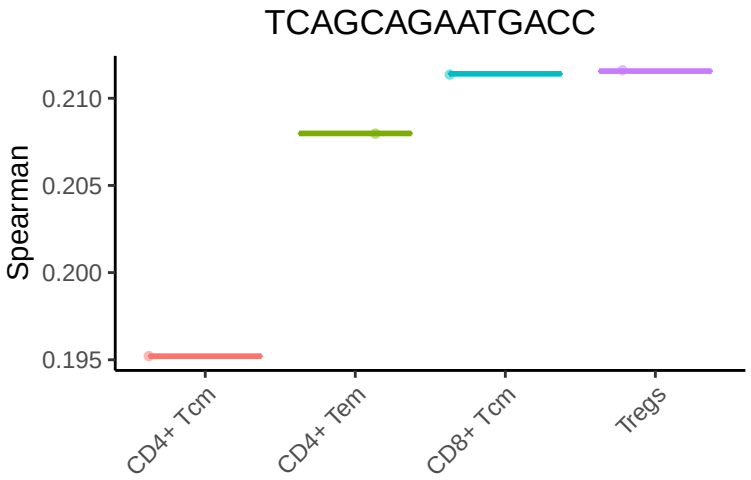
After this iteration, 5 cell types remain.

```
## [1] "Number of DE genes:1287"
## [1] "Number of cells: 2"
```

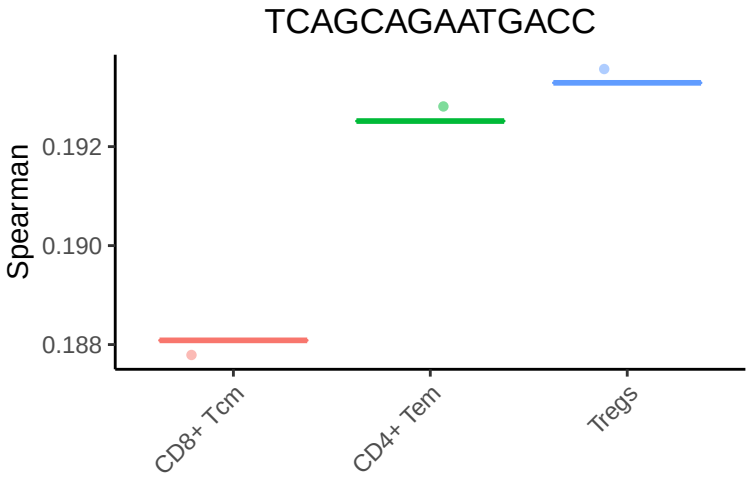


SingleR continues these iterations, each time taking the top cell types or removing the lowest scoring cell types.

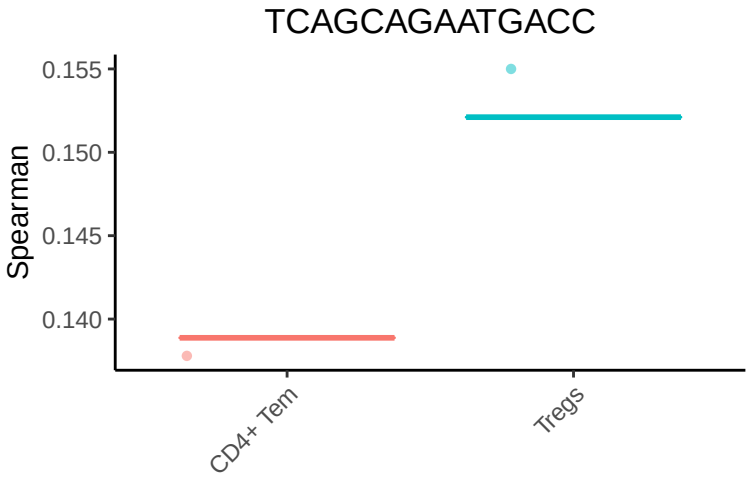
```
## [1] "Number of DE genes:1145"
## [1] "Number of cells: 2"
```




```
## [1] "Number of DE genes:949"
## [1] "Number of cells: 2"
```




```
## [1] "Number of DE genes:666"
## [1] "Number of cells: 2"
```




```
## [1] "Number of iterations: 5"
```

In the end the winning annotation is of a regulatory T-cell (Treg). This cell is in fact a sorted Treg;

however, it does not express known markers such as FOXP3 and CTLA4, making it hard to detect by marker-based approaches.

The open source code and specifications for using *SingleR* are available at the *SingleR* Github repository https://github.com/dviraran/SingleR.

References

1. Heng, T. S. P. *et al.* The Immunological Genome Project: networks of gene expression in immune cells. *Nature Immunology* **9**, 1091–1094 (2008).

2. Benayoun, B. A. *et al.* Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. *bioRxiv* 336172 (2018). doi:10.1101/336172

3. Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C. & Hume, D. A. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC genomics* **14**, 632 (2013).

4. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98**, 1487–9 (2013).

5. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).