MONASH BIOINFORMATICS PLATFORM

MONASH University

Annotating single-cell RNAseq clusters by similarity to reference single-cell datasets

Sarah Williams, Sonika Tyagi, David Powell
Monash Bioinformatics Platform, Monash University

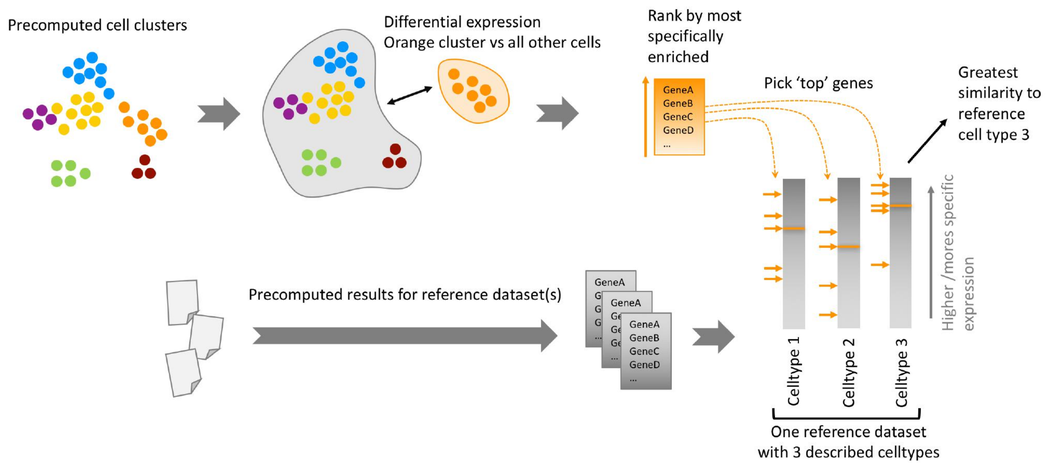
INTRODUCTION

- Single cell RNAseq (scRNAseq) is often used to examine cell types within tissue samples. There are a multitude of methods available for clustering sequenced cells into transcriptionally-similar groups, putatively corresponding to cell type or state [3].
- Those clusters are then labelled via known cell-type marker genes with specific or enriched expression [1,2,5].
- Cluster labelling can be a time-consuming interactive process which requires specialist knowledge of the cell types in the sample and their transcriptomic signature.

Aims:

1. To take pre-computed cell-clusters and **automatically annotate cell type information to each cluster** in a quick, accessible manner on the basis of similarity to cell clusters from publically available scRNAseq datasets.
2. To provide a quick starting point to start characterisation of targeted cells, exclude uninteresting cell-types or uncover unexpected cell-types.

METHODS



Input = Cell clusters + Counts table

1. Clusterwise differential expression
2. Identify 'top' genes
3. Rank on known cell types

Output = Annotated clusters

'Top' gene selection: up to 100 significantly (corrected p-value <0.01) at least 2-fold enriched using SCDE[4]. Genes were ranked by the lower 95% confidence interval of their fold-change to emphasise larger changes, avoiding low-expression genes and dropouts.

RESULTS

Same tissue comparison – brain: To check performance on similar cell types mouse brain cell-clusters [5] were compared to human brain [1] clusters. Fluidigm C1 data.

There was a good match between neuronal cell types, even with differences in the cluster annotations.

Cross-tissue comparison: The same mouse brain dataset was compared to a mouse lacrimal gland dataset [2] (10X data). This tests if common cell types might be identified from different tissues, and provides a negative control of dissimilar cell types.

As anticipated most brain clusters had no obvious homology among lacrimal gland cell types, yet microglia cells show their similarity to macrophages.

Gene-ranking correlation vs 'Top' gene approach:

- The results are similar, but correlation shows higher similarity within studies.
- Focusing on the 'top' genes should minimise the influence of total sample composition on cluster characterisation.
- One challenge of this approach is its directionality; it results in some unintuitive non-reciprocal similarities (Figure 3A). However, reciprocal hits might improve results for clusters with fewer 'top' genes – as bi-directional groupings of similar cell types (e.g. endothelial, microglia/macrophages) emerge in the network visualisation.

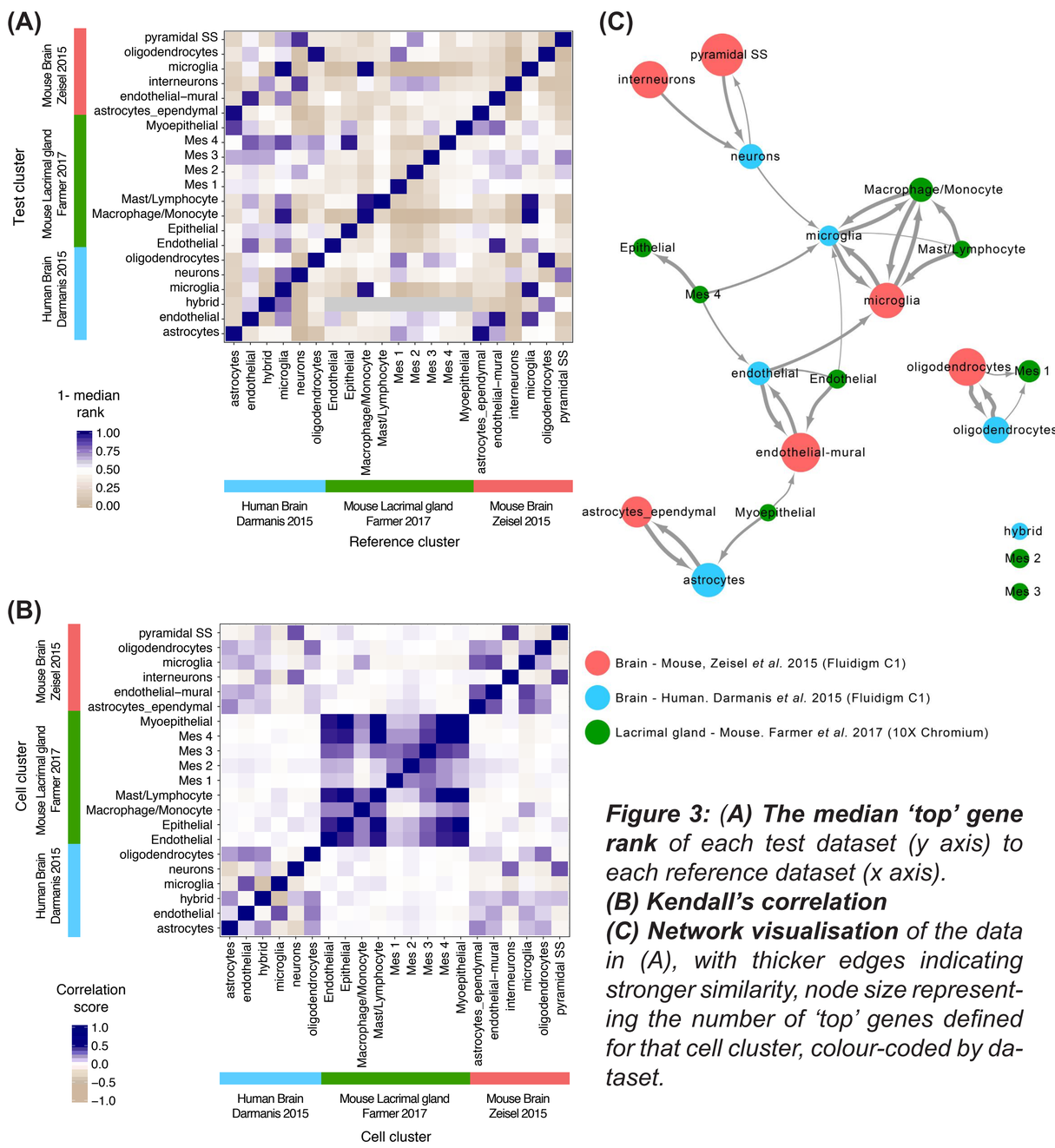


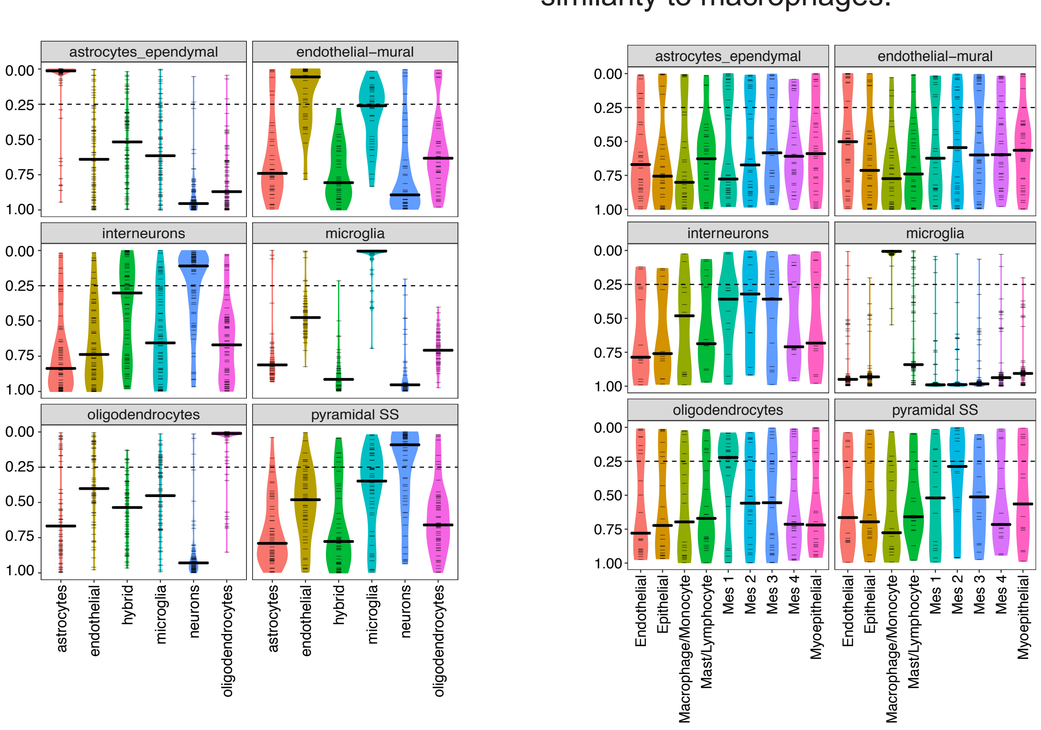
Figure 3:* (A) *The median 'top' gene rank *of each test dataset (y axis) to each reference dataset (x axis).*
(B) Kendall's correlation
(C) Network visualisation *of the data in (A), with thicker edges indicating stronger similarity, node size representing the number of 'top' genes defined for that cell cluster, colour-coded by dataset.*

CONCLUSIONS

This simple 'top' gene based approach is surprisingly effective at matching cell types between different scRNAseq experiments, even across tissue types and platforms, without overmatching unrelated cell types.

The next step will be formally evaluating the detection sensitivity through computational mixing experiments and using those results to define sensible reporting thresholds.

This work is a proof-of-concept, which will be developed into a useful tool for labelling cell clusters from scRNAseq experiments. Such a tool would enable users to more quickly identify cell types. It may also help in selecting an appropriate clustering algorithm for an experiment, by providing an easy way to evaluate different cluster-sets with respect to known cell types.



Distribution of 'top' genes from query clusters (panels) on reference clusters. Reference cluster rankings are ordered from highest to lowest specific expression – as defined by the lower fold-change 95% confidence interval. Higher median rankings (thick line) indicate higher similarity.

REFERENCES

1. *Darmanis, S.et al (2015). A survey of human brain transcriptome diversity at the single cell level. PNAS, 112(23), 201507125.*
2. *Farmer, D. T et al. (2017). Defining epithelial cell dynamics and lineage relationships in the developing lacrimal gland. Development, 144(13), 2517–2528.*
3. *Freytag, S. et al. (2017). Cluster Headache: Comparing Clustering Tools for 10X Single Cell Sequencing Data, (4).*
4. *Kharchenko, P. Vet al (2014). Bayesian approach to single-cell differential expression analysis. Nature Methods, 11(7), 740–742.*
5. *Zeisel, A. et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science, 347(6226), 1138–42.*

Bioinformatics.Platform@monash.edu
platforms.monash.edu/bioinformatics
@MonashBioinfo