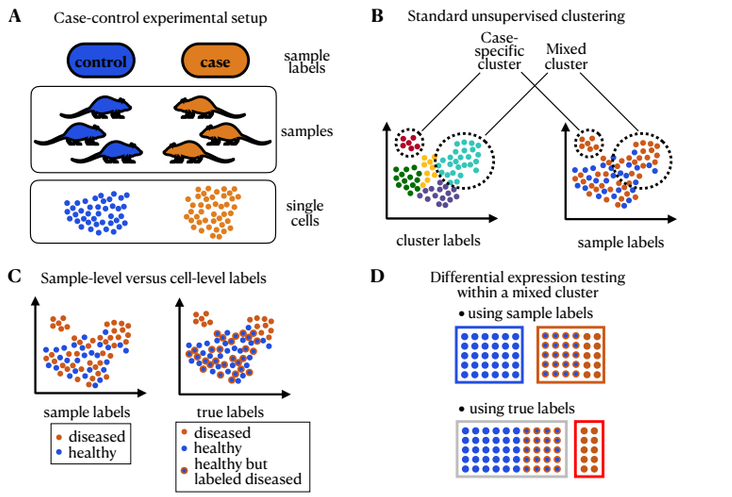
Improved marker detection through label refinement in case-control single-cell RNA-seq studies

Aleksandrina Goeva[1], Michael-John Dolan[1], and Evan Macosko[1,2]

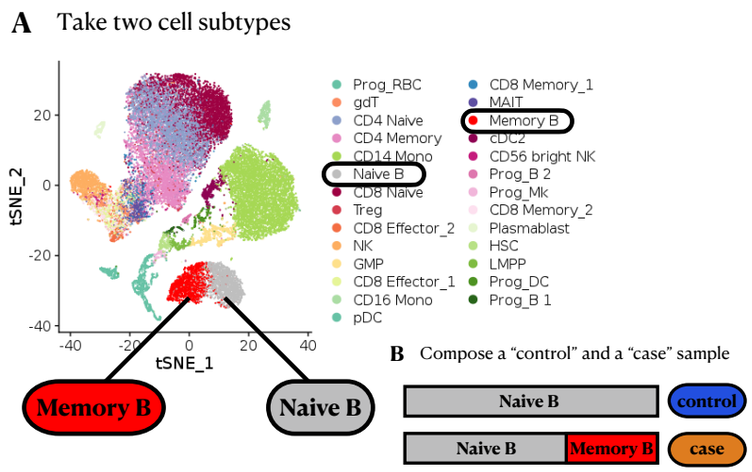
[1] Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA
[2] Massachusetts General Hospital, Department of Psychiatry, Boston, MA, USA

1. Not all cells in case samples are different from cells in control samples and this can obscure subtle signals

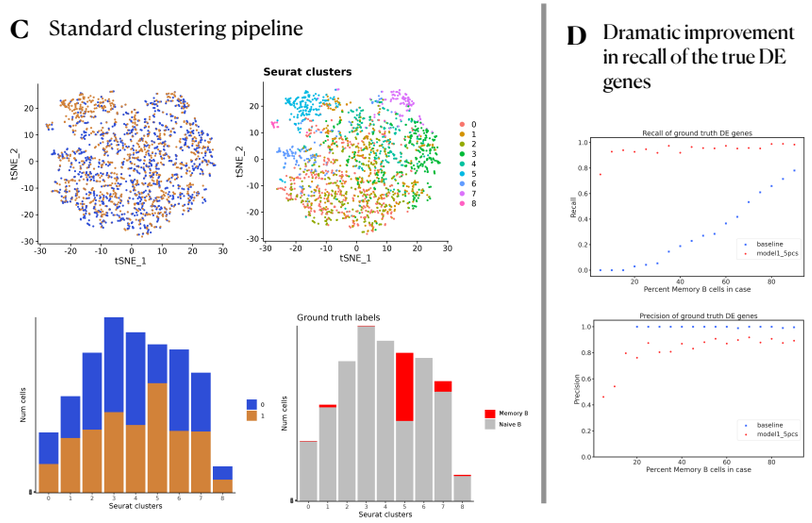


2. Consider wanting to 'discover' Memory B cells in background of Naive B cells

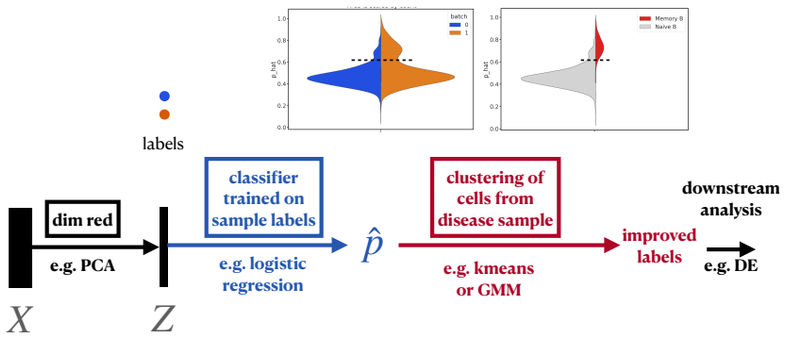
Mix the two related PBMC[1] subtypes in the following way



3. The standard approach fails to detect the Memory B cells and their markers so we developed a method that does

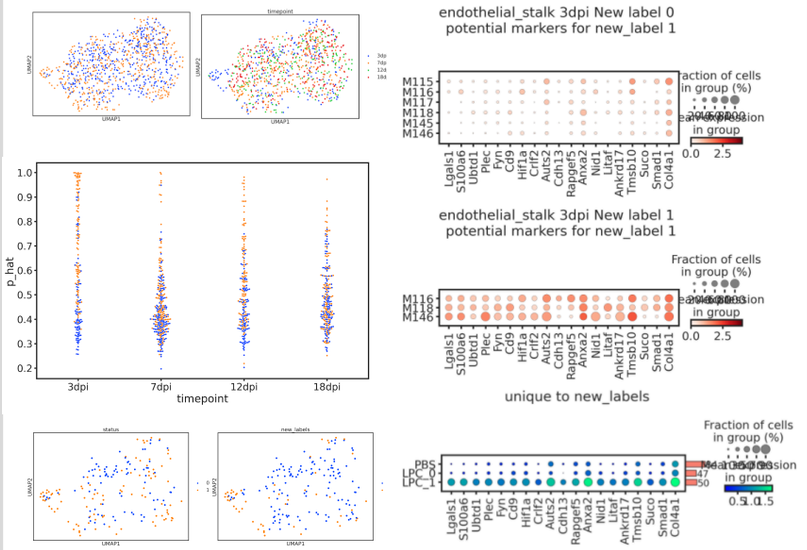


4. Our method combines partially correct sample labels with gene expression profiles to refine the labels



The classifier outputs predicted probability that a cell is affected by the disease. We can either directly use these 'soft' labels, or further cluster them to redefine the healthy and disease binary annotation.

5. Applied to endothelial cells from a demyelination experiment, we uncover early changes to the blood-brain barrier



6. A tale of two mice: summary and remarks

- We have developed a method that generates hypotheses about **which are the truly affected cells** in a case-control experiment.
- The **refined labels improve downstream tasks**, such as DE.
- Applications **beyond the health-deasese context**: e.g. sex differences, perturbation experiments.
- Our method outputs **soft labels** and can capture a **gradient** of the effect of the condition.

References

[1] Stuart*, Butler* et al, Cell (2019)

agoeva@broadinstitute.org