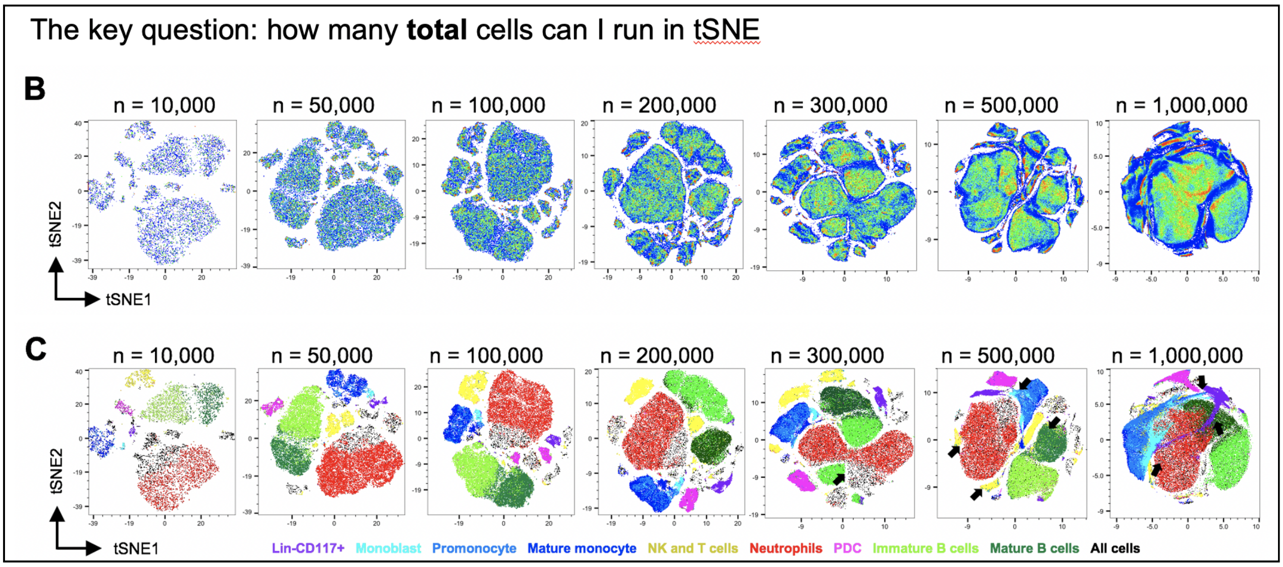
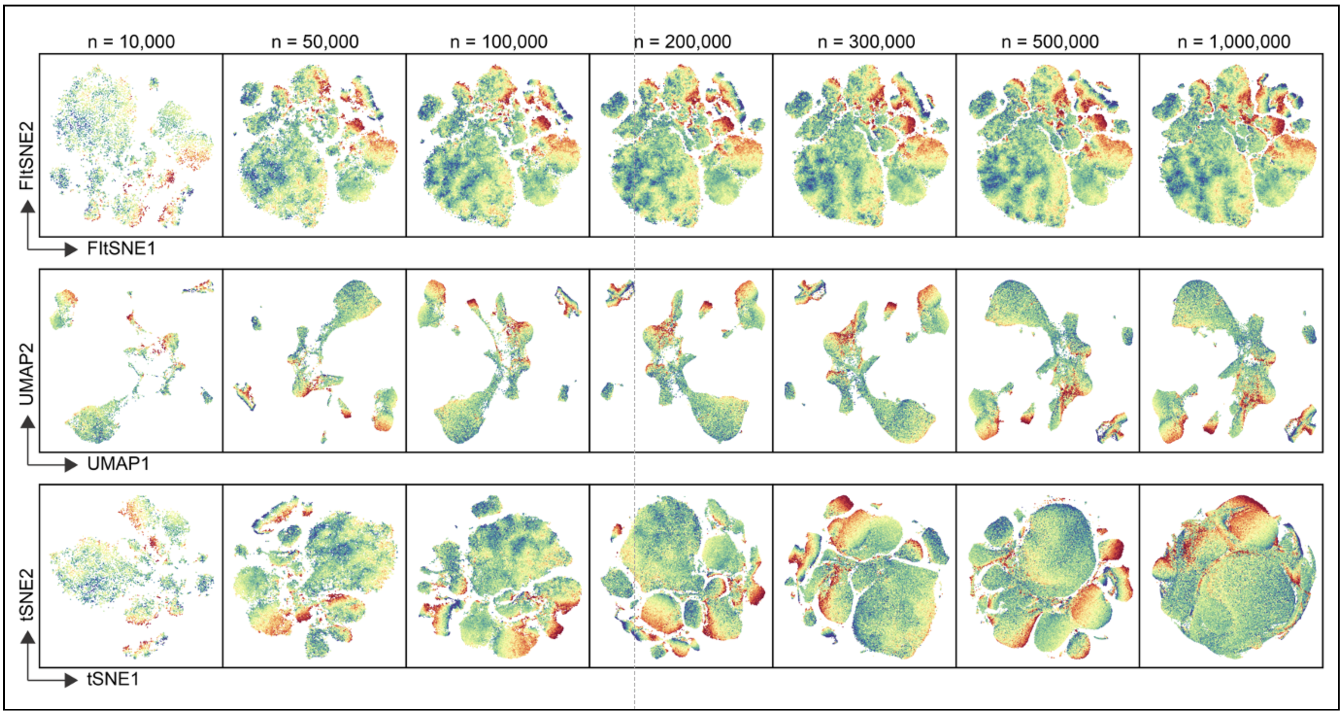
How many cells can you include?

A key limitation of many of these tools is how many cells can be included in a run. FlowSOM scales extremely well to large datasets, and we have successfully run datasets with 40 million cells. In this case the more cells are included, the longer the run takes, but this relationship is linear, so there is no penalty for additional cells.

DR tools are different. While the run-time for tSNE is linear for additional cells, the way it is plotted is not. Using the default settings in tSNE, running 10,000 - 50,000 cells results in reasonable distributions. However, when using 100,000 cells or higher, the *crowding effect* starts to cause the cells to mash together, resulting in difficulties in interpretation.



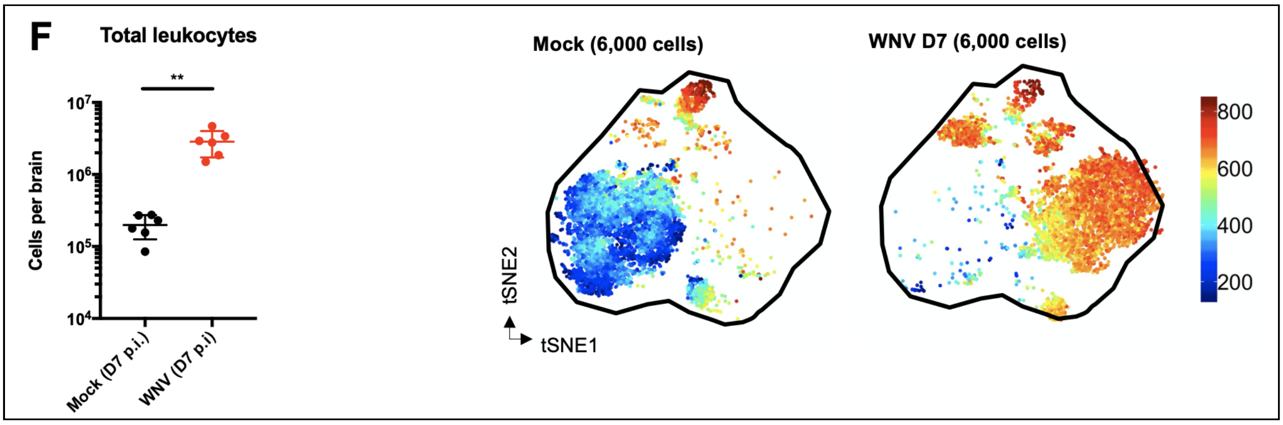
Variations of tSNE (opt-SNE, FIT-SNE), along with other tools such as UMAP, are able to effectively distribute millions of points on a 2D plot, and are better suited to large datasets.



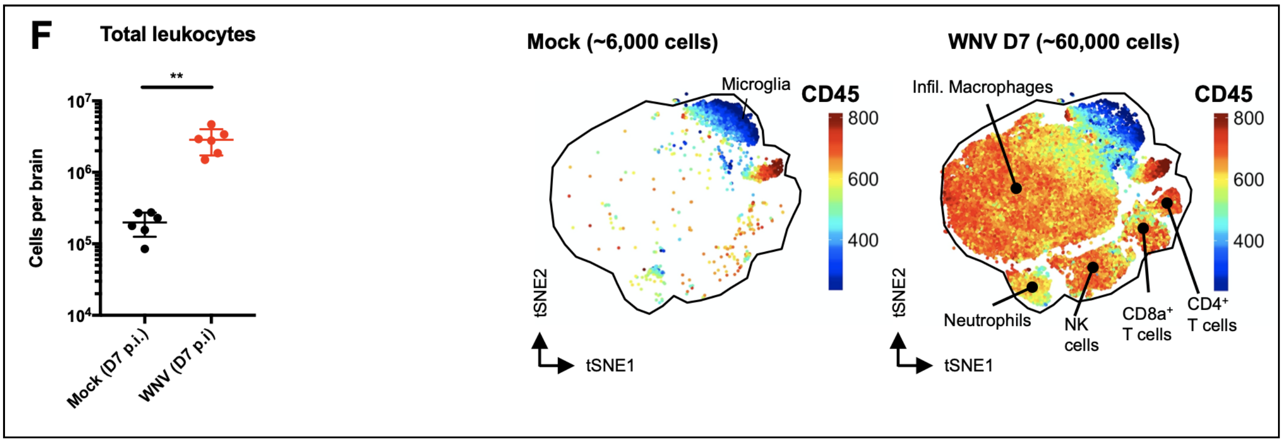
Balancing the cell numbers between groups

In most publications using these tools in a cytometry context, they are being applied to PBMC samples where the total blood counts/volume of blood are not recorded. As such, the comparisons between groups are simply performed as proportions. As such, the number of cells from each group are identical. E.g. if we had 3x 'control' samples and 3x 'test' samples to analyse, we may include 10,000 cells from each sample (resulting in 30,000 cells in each group). However, this approach isn't always suitable.

When examining tissues, the actual number of cells that are present in each tissue sample is critically important. For example, in studying viral infection of mouse brains, the number of leukocytes in a normal brain are around 10^5 (mostly resident microglia), but this increases 10-fold to 10^6 cells following West Nile virus (WNV) infection. In this case, the number of microglia don't change, but the number of cells increases due to the infiltration of leukocytes from the blood. If we take the same number of cells from the mock- and WNV-infected samples, the blue group of cells (microglia) appears to dramatically reduce in the WNV-infected sample. However, this is simply because the *proportion* of microglia has decreased.



If we keep the number of cells included in each group in proportion to the actual cell counts, then the distribution of cells in each plot more faithful represents what is occurring in the sample. In this case, the number of microglia (blue) are consistent, but we find a number of new cell types infiltrating from the blood.


F
Total leukocytes
**
Cells per brain
10^7
10^6
10^5
10^4
Mock (D7 p.i.)
WNV (D7 p.i)
Mock (~6,000 cells)
Microglia
CD45
tSNE2
tSNE1
WNV D7 (~60,000 cells)
Infil. Macrophages
CD45
Neutrophils
NK cells
CD8a+ T cells
CD4+ T cells
tSNE2
tSNE1