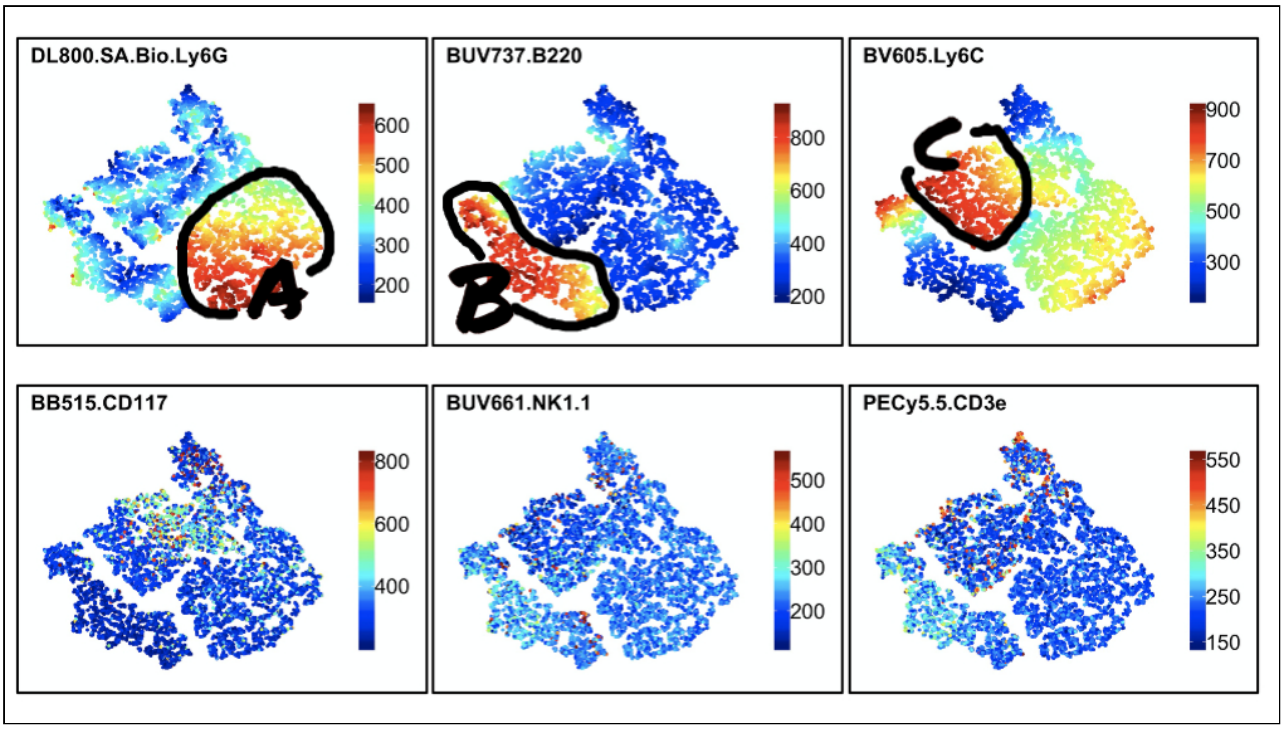
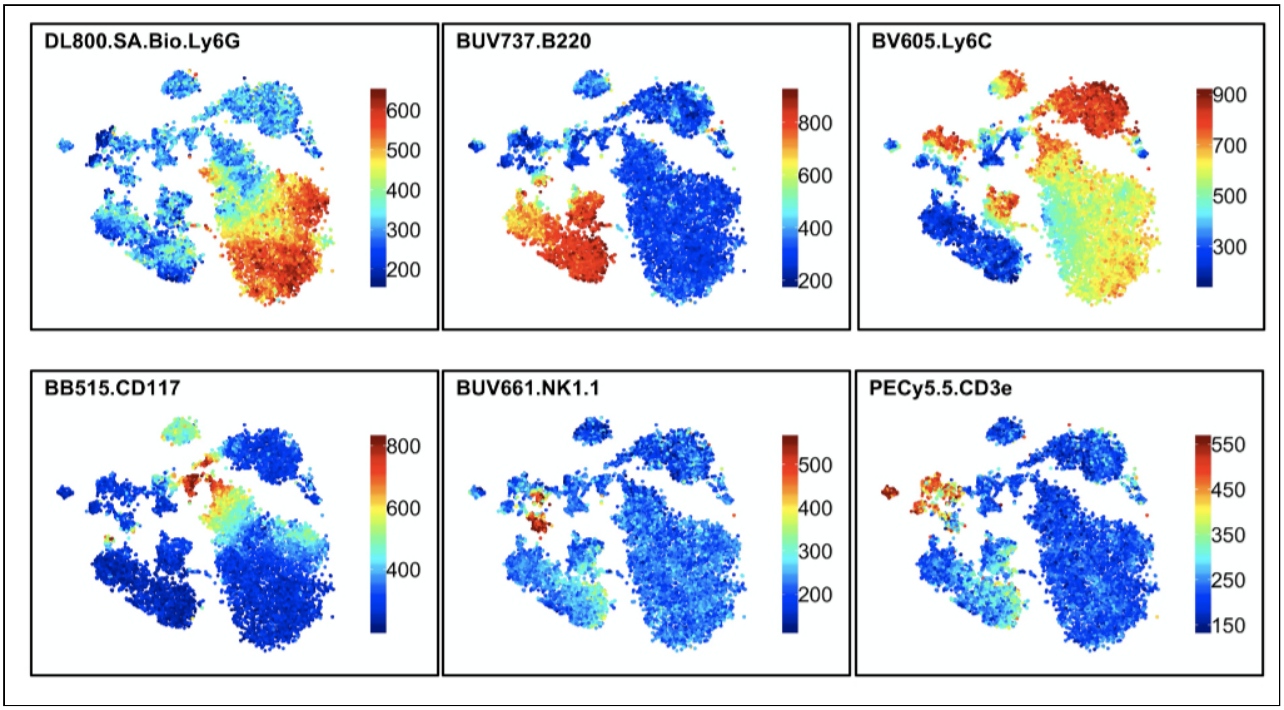
The markers that are included to generate clusters/DR dimensions influence the results

An important consideration is which markers are used to generate clusters or to run tSNE etc. Importantly, not include some markers in the tSNE analysis does not *remove* them from the dataset, but rather simply does not use the information from that marker to inform the generation of a tSNE plot.

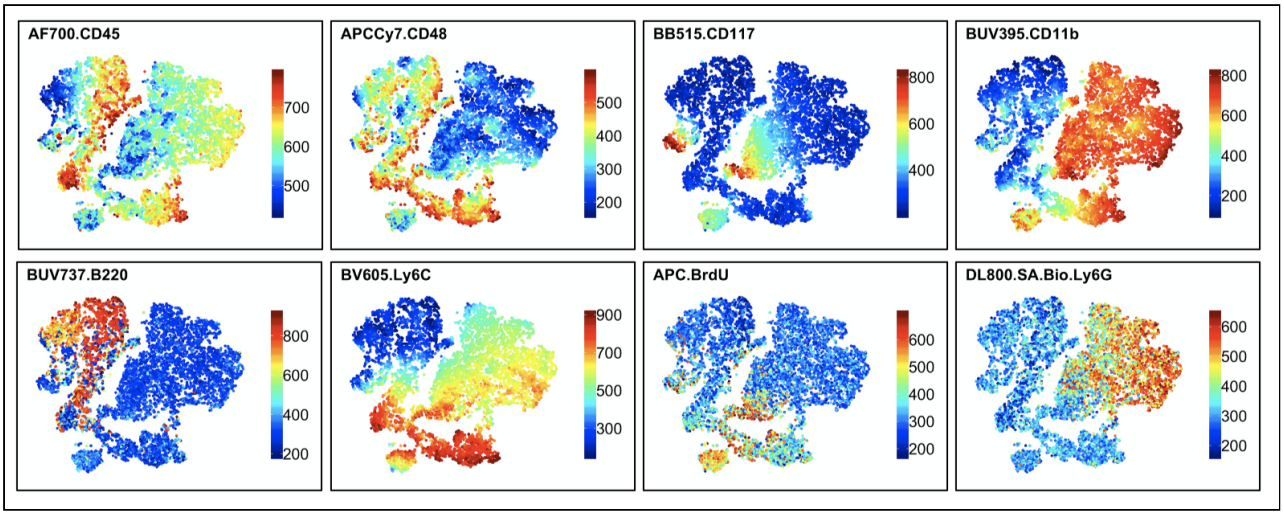
For example, if we run tSNE on mouse bone marrow data using only B220 (B cells), Ly6G (neutrophils), and Ly6C (monocytes), then this will results in 3 rough islands – one for neutrophils (A), one for B cells (B), and another for monocytes (C). When we look at the other markers, such as CD117 (stem cells), NK1.1 (NK cells), and CD3e (T cells), these are distributed throughout the map, as the tSNE algorithm is not able to see them using the information (Ly6G, B220, and Ly6C) provided to it.



On the other hand, if we run tSNE on mouse bone marrow data using all markers then we can see a far more granular distribution of islands, representing all the major populations in the bone marrow.



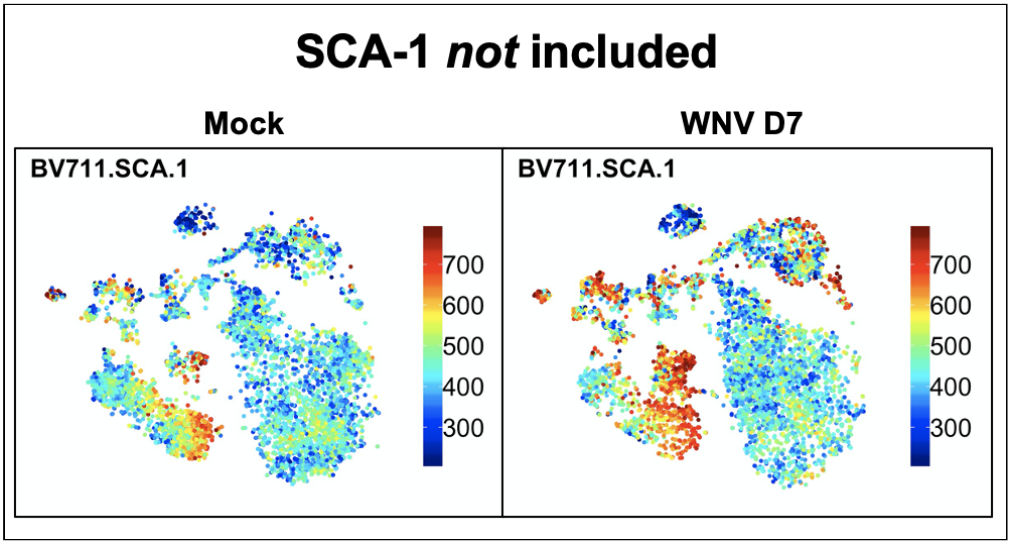
However, a minimal use of markers can still be used strategically. If we run tSNE using only SSC, CD117, CD45, CD48, Ly6C, CD11b, you can see that neutrophils and B cells have roughly been arranged into islands, despite the lack of use of markers that are specific for these populations (Ly6G and B220). This is because combinations of other markers (SSC, CD117, CD48, Ly6C, CD11b etc) have specific patterns on these cell types.



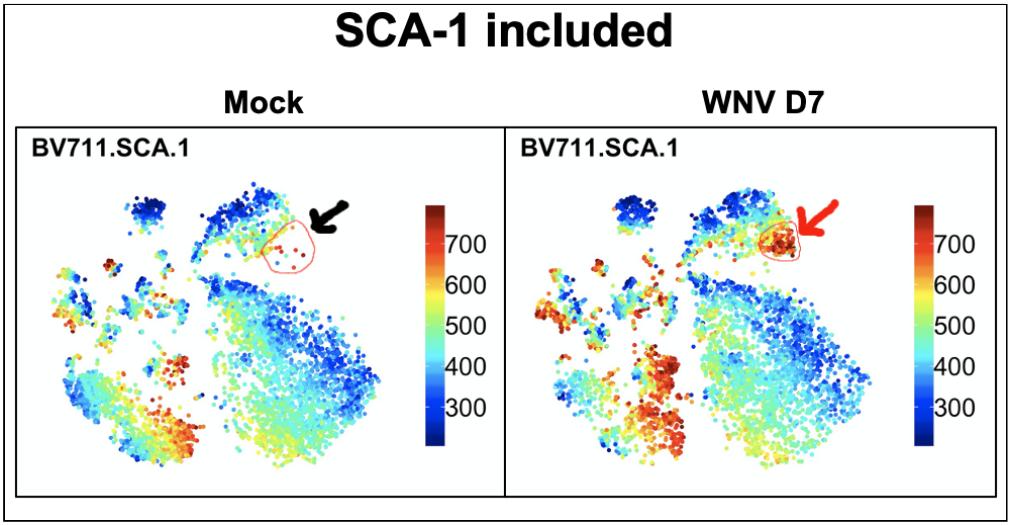
Using cell 'identity' and 'state' markers

Broadly, we can group cellular markers into two categories: markers for *cell types*, and markers for *cell states*. Cell identity markers (also called *phenotypers*) are markers that are stably expressed on specific cell types, which helps us to identify them (e.g. CD3 on T cells, CD19 on B cells etc). Cell state markers indicate some sort of change to the status of a cell, but not specifically it's identity (e.g. CD69 is upregulated on multiple lymphocyte types following 'activation'). An extension of the state markers are things like the presence of viral proteins which indicate the status of 'infected'. Sometimes markers might belong to both categories (e.g. SCA-1 is expressed on mouse stem cells, but is also inducible on multiple cell types following interferon signalling).

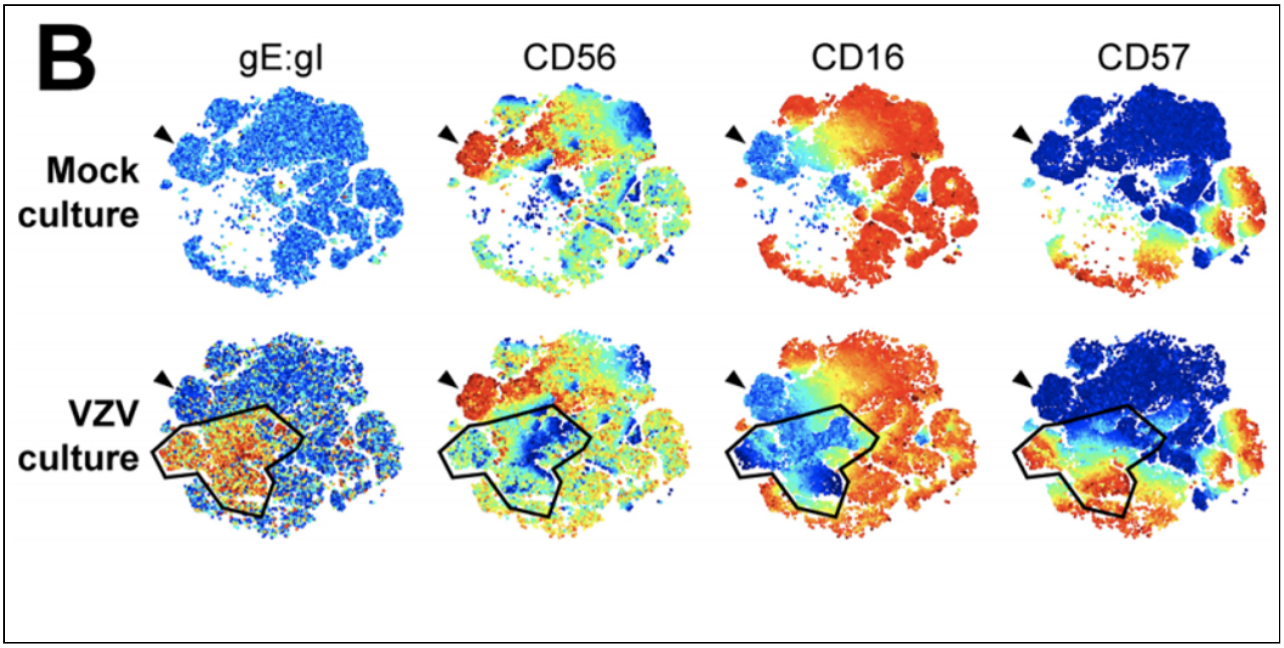
For example, in the example below we ran tSNE analysis on mouse bone marrow, containing cells from both mock- and WNV-infected samples. Importantly we did not include SCA-1 in the markers used to inform tSNE, and so cells that have upregulated SCA-1 following WNV-infection do not change position on the plot.



However, if we include SCA-1 in the markers used to inform tSNE, then cells that upregulate SCA-1 following WNV-infection appear in a new location on the plot.



Sometimes, even if you don't use specific markers to run tSNE, the expression of other markers might correlate highly with those cells, resulting in the cells appearing in a new location on the plot. In the example below, gE:gI is a viral protein that is detected in infected cells. This was not used to run tSNE, but the cells that are infected substantially downregulate CD16 (which is used to run tSNE), resulting in the CD16(low) gE:gI(hi) cells occupying a distinct position on the plot.


B
gE:gI
CD56
CD16
CD57
Mock culture
VZV culture