



Publications

Best practices in high-dimensional data analysis

A paper discussing common caveats and best practices in the analysis of high-dimensional cytometry data

[SEE PAPER](#)

Planning your high-dimensional experiments

An open access publication where we provide considerations for new and experienced users to design and carry out high-dimensional experiments to maximize quality data collection.

[SEE PAPER](#)

Overview

In this brief tutorial, we seek to provide an overview of clustering and dimensionality reduction, and how to best to approach leveraging these tools in cytometry analysis. There are also some very helpful websites that discuss the functionality of [tSNE](#) and [UMAP](#) in great deal, with interactive plots.

On this page:

- [Publications](#)
- [Overview](#)
- [1. High-dimensional cytometry data](#)
- [2. Manual analysis: 'gating'](#)
- [3. Computational analysis: clustering and dimensionality reduction](#)
 - [3a. Dimensionality reduction](#)
 - [3b. Clustering](#)
 - [3c. Using these approaches together](#)
- [4. Analysing multiple samples](#)
- [5. The choice of tool can influence throughput and accuracy](#)
- [6. Other computational approaches](#)
- [7. How do I think about each approach?](#)
- [So how do I utilise these 'discovery' approaches?](#)

1. High-dimensional cytometry data

Single cell biology & high-dimensional cytometry technologies

In medical science, we use **cytometry** to measure the properties of individual cells, most commonly immune cells. Every cell in an individual's body contains (approximately) identical **DNA**. As cells develop or respond to stimulus (e.g. inflammation), specific elements of the DNA are 'transcribed' into an intermediate type of molecule called **mRNA**, which is then 'translated' into **protein** that may enact a specific function, with or without post-translational modifications. We typically refer to these proteins as '**markers**'. In cytometry, these markers are often divided into two groups: cellular *identity* or cellular *state* markers. Identity markers are proteins that are expressed at stable levels on specific types of cells (e.g. the '**CD3**' protein is expressed on **T cells**), and state markers denote response to a particular stimulus (e.g. if an individual gets an infection in their arm, a set of signals will lead immune cells in the blood to upregulate cell surface proteins that will encourage them to migrate into the inflamed tissue – elements of DNA will be selected, turned into RNA, which will then be turned into protein). Certain 'markers' are also used as an indicator of disease, such as the presence of proteins that indicate the progression to neoplasia (cancer).

In cytometry, we typically label **cellular proteins** with an **antibody** that is specific for that protein. The antibodies can be bound to a fluorescent molecule that will emit light when hit with a laser, allowing us to measure the signal of each fluorescent molecular on each cell. This is called '**flow cytometry**'. Alternatively, antibodies can be bound to lanthanide metals, which can be measured using a mass spectrometry system. This is called '**mass cytometry**'. Currently, this allows us to measure between ~30 (flow) and ~40 (mass) proteins per cell, at rates of ~400 (mass) to 10,000 (flow) cells/per second in a cost-effective manner. As such, each dataset (and sometimes each sample) may consist of millions of cells

 Cytometry = cell (cyto) measurement (metry).

Single-cell RNA sequencing

The other way we may measure single cells, is using **single-cell RNA sequencing (scRNASeq)**. Here the individual RNA molecules can be sequenced in each cell (similar to how we 'sequence' DNA), allowing us to potentially see *every* RNA molecule, or 'transcript' (up to 10,000 transcripts that might be present). This technology is amazing, but is generally much slower, and more expensive. Moreover, not every RNA transcript is ultimately converted to protein (very irritating), and so the picture is not complete, but the ability to measure many RNA transcripts allows to examine 'types' of response, not just individual molecules. There are alternative approaches to scRNASeq, including approaches that reduce cost and increase sensitivity by measuring only selected important RNA transcripts (e.g. the 400 most important transcripts).

 The measurement of all of the RNA transcripts produced by a cell is called 'transcriptomics'.

What does the data look like and how do we examine it?

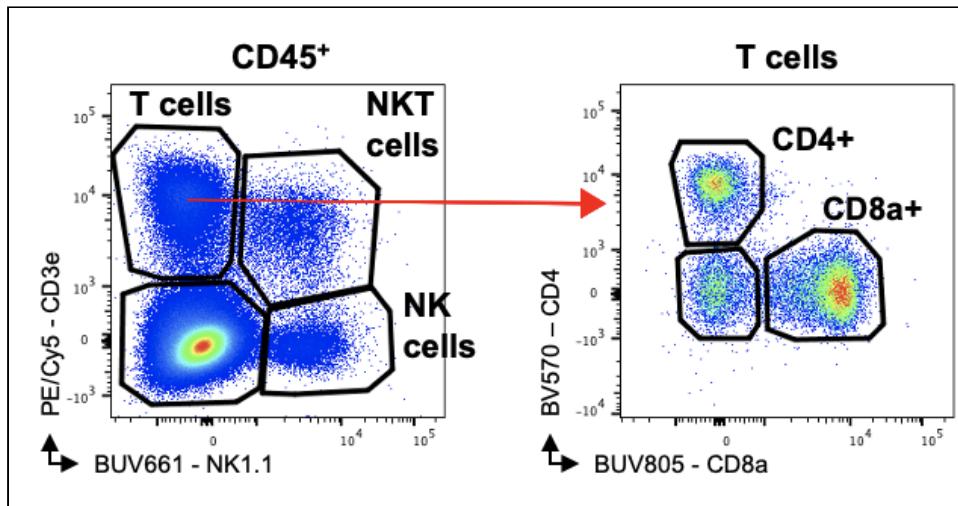
Fundamentally the data is structured as a table of cells (rows) x features/markers (columns) (see Table). The actual *values* at this point are relative measures of expression (but for those not familiar with cellular biology, you could think of it as representing the number of molecules per cell). All cells will exhibit a level of 'background' signal, due to cellular autofluorescence, non-specific sticking of antibodies to cells, or electronic noise from the cytometry platform, which we can determine experimentally (i.e. a cell doesn't have to read 0 to indicate that there is no protein there). As immunologists, we usually divide the raw data values into 'positive' (i.e. the protein IS expressed) or 'negative' (the protein is NOT expressed, or is below the limit of detection). In the table below, the entries are coloured by what we might consider 'positive'.

	Level of 'CD3' protein	Level of 'CD4' protein	Level of 'CD8' protein	Level of 'NK1.1' protein	Summary of positive markers	What immune population is this?
Cell #1	100,000	850,094	534	346	CD3+CD4+	<i>CD4+ T cell</i>
Cell #2	900,000	1424	991,242	128	CD3+CD8+	<i>CD8+ T cell</i>
Cell #3	860,523	849	420	242	CD3+	<i>Double negative T cell</i>
Cell #4	872,049	125	235	952,284	CD3+NK1.1+	<i>NKT cell</i>
Cell #5	457	157	312	892,401	NK1.1+	<i>NK cell</i>

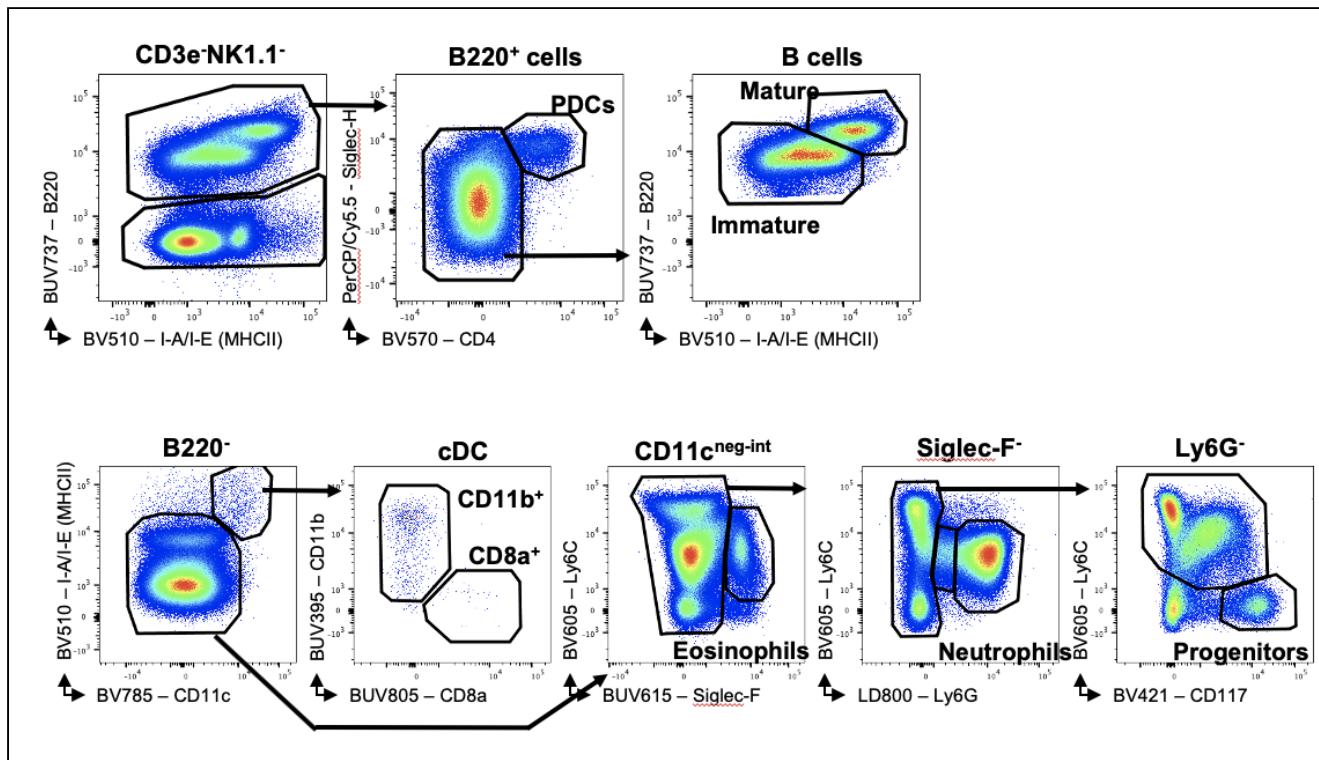
Normally this table would contain thousands to millions of cells, but in this example table we have just included one cell per expression pattern.

2. Manual analysis: 'gating'

This type of data can be plotted on a bi-axial dot plot, where we can perform 'gating' – selecting cells by drawing around the cell groupings. Our approach to gating is dictated by the 'rules' of cellular expression as we understand them. E.g. All T cells express CD3, all NK cells express NK1.1, and NKT cells express both. We know that there are two subset of T cells (CD4+ and CD8+), so we can select the CD3+NK1.1- cells and then subdivide these into CD4+ and CD8+ groups.

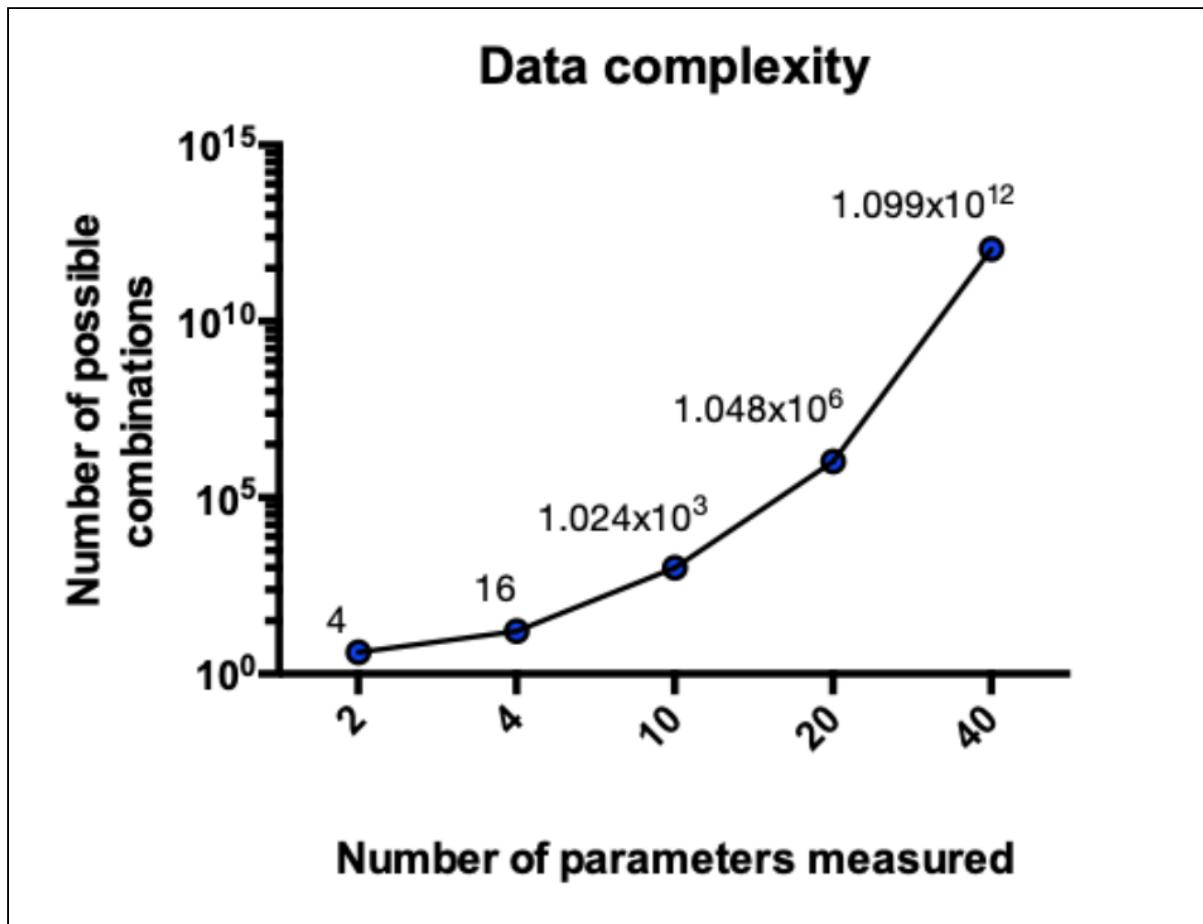


Using this approach, we develop gating 'trees' or 'hierarchies' that allow us to identify all the subsets of interest.

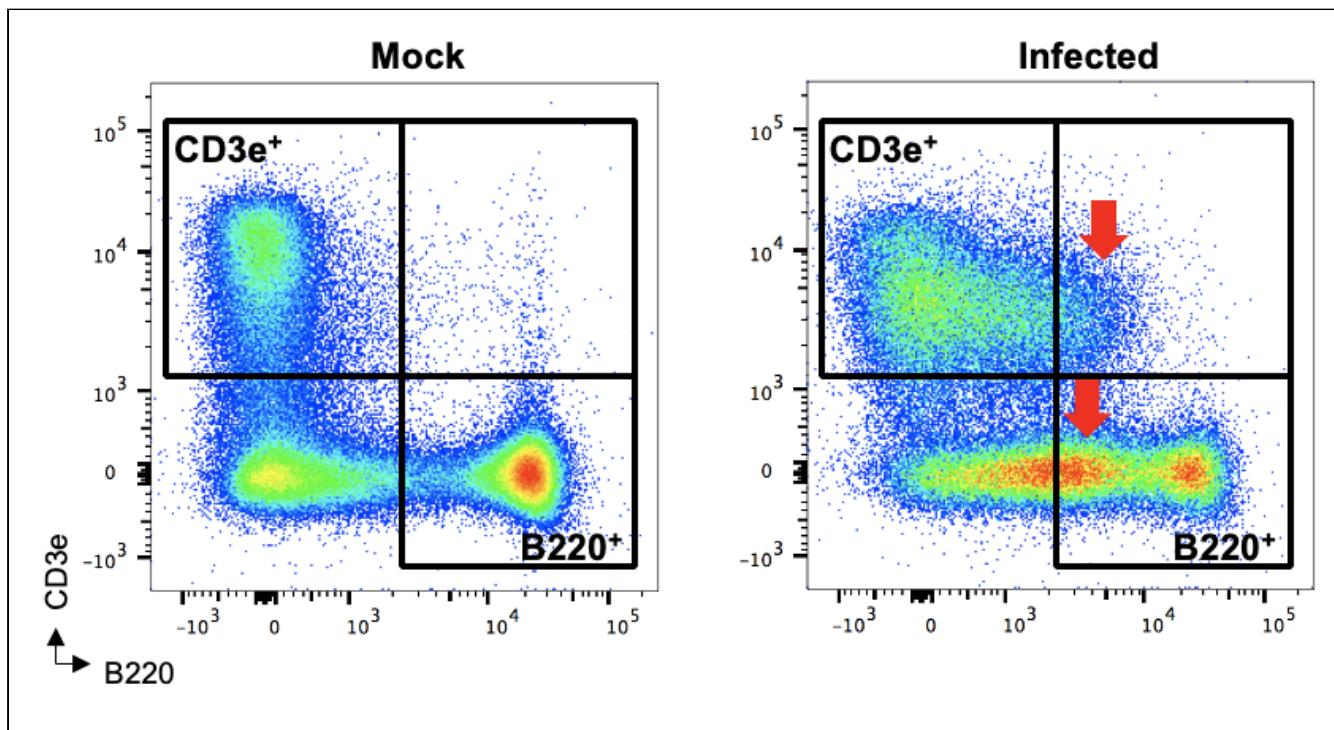


Limitations of manual gating

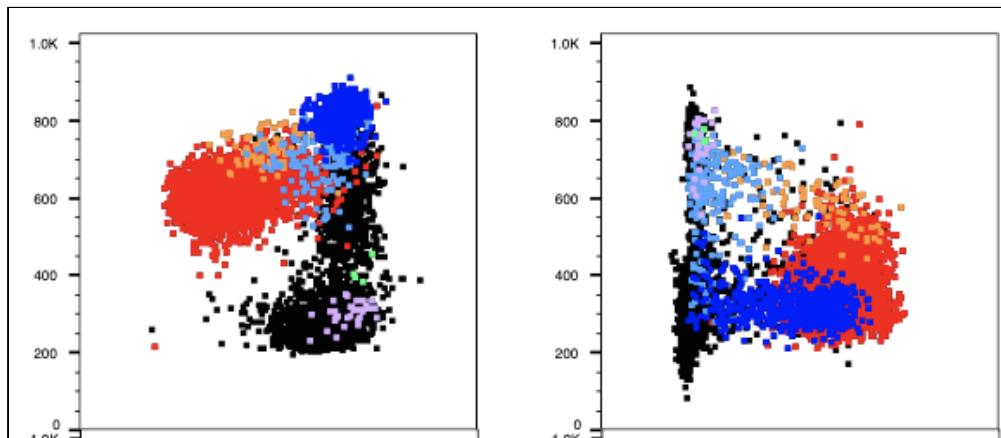
We have approached data analysis this way historically because it depends on our *knowledge* of the cell types being investigated. However, this approach has some significant limitations. One is obviously the laborious manual nature of 'gating', but more significantly, there is no high-/system-level view of the data, meaning that there are many populations that we might miss entirely, depending on how we approach the data. For example, if we consider that each marker could have a positive/negative expression pattern, then measuring 2x markers could result in 4x possible populations (e.g. CD4+CD8-, CD4+CD8+, CD4-CD8+, CD4-CD8-). This can be calculated as 2^n , where n = the number of markers being measured. For 40 markers (2^{40}), this results in 1,099,511,627,776 possible combinations with a +/- phenotype.



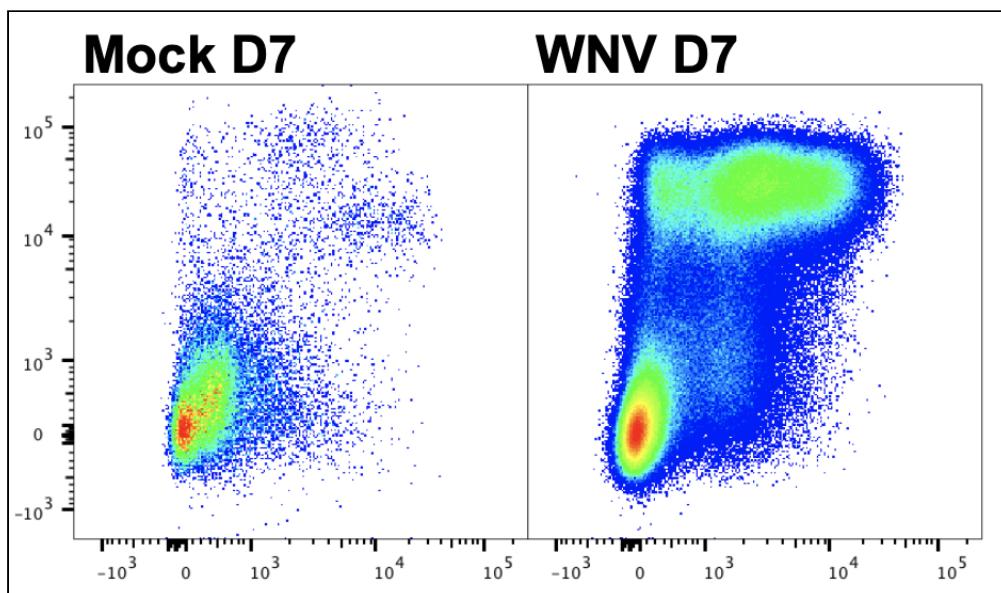
Another limitation is that these gating trees obscure the possible phenotypes of cells that might be present in the dataset, by assuming certain staining patterns. For example, B220 is expressed on B cells, but not on T cells (which are CD3e+) – as a result, B220 would not typically be assessed on CD3e+ cells. However, B220 is able to be upregulated on CD3e+ T cells during exhaustion – a staining pattern that would be missed using traditional gating strategies.



The more markers are measured on single cells, the more assumptions are made to generating gating trees, and the more potential subsets are potentially missed. Moreover, in situations where staining patterns do not generate a neat +/- (binary) staining pattern, such as in the myeloid lineage (especially monocytes, macrophages, and dendritic cells)...



... or cellular infiltration into the mouse CNS during viral infection...



..manual gating does not adequately capture the cellular landscape. This is further complicated when investigating tissues where the cell types present have complex phenotypes, that do not necessarily follow the presumed 'rules'.

3. Computational analysis: clustering and dimensionality reduction

To address the limitations of 'manual gating' analysis, we can leverage computational approaches, including **clustering** and **dimensionality reduction**. These computational approaches can be divided into two broad categories: *discovery* analysis (geared towards understanding the structure of the data, and finding novel subsets/changes) or *replicative* analysis (geared towards an automated 'replication' of an analysis framework – either through automated gating, or some form of cell classification system). **Clustering** and **dimensionality reduction** are typically used in a *discovery* context.

3a. Dimensionality reduction

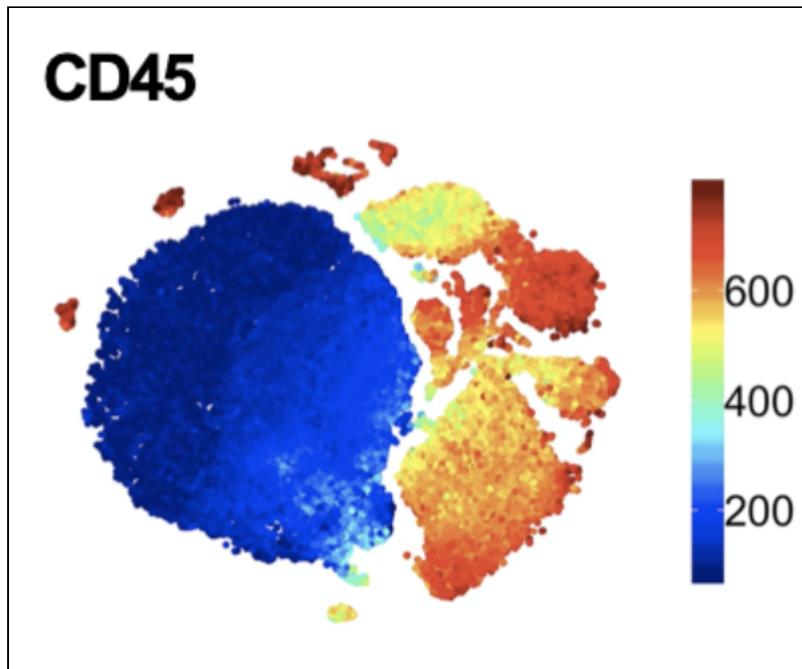
Dimensionality reduction (DR) is a computational approach that seeks to reduce data in n-dimensional space into 2-dimension space, so that every cell can be visualised in a single 2D plot, in such a way that the arrangement of the cells represents the relationship of these cells in high-dimensional space. Critically, the elements of the data structure define how the data is distributed in this 2D space – i.e. as a general rule cells that are similar to each other will group close together on the plot

(more on that later). Cells can be coloured by the level of expression of a marker, or by some other factor (sample, cluster, population, etc). The most well-known example of this is principle component analysis (PCA), a non-linear DR approach, but now includes a range of non-linear approaches, including t-distributed stochastic neighbour embedding (tSNE), and uniform manifold projection (UMAP).

Linear dimensionality reduction (e.g. PCA): ...

IMAGE TBC

Non-linear dimensionality reduction (e.g. tSNE, UMAP): A common example of non-linear dimensionality reduction in cytometry analysis is t-SNE.



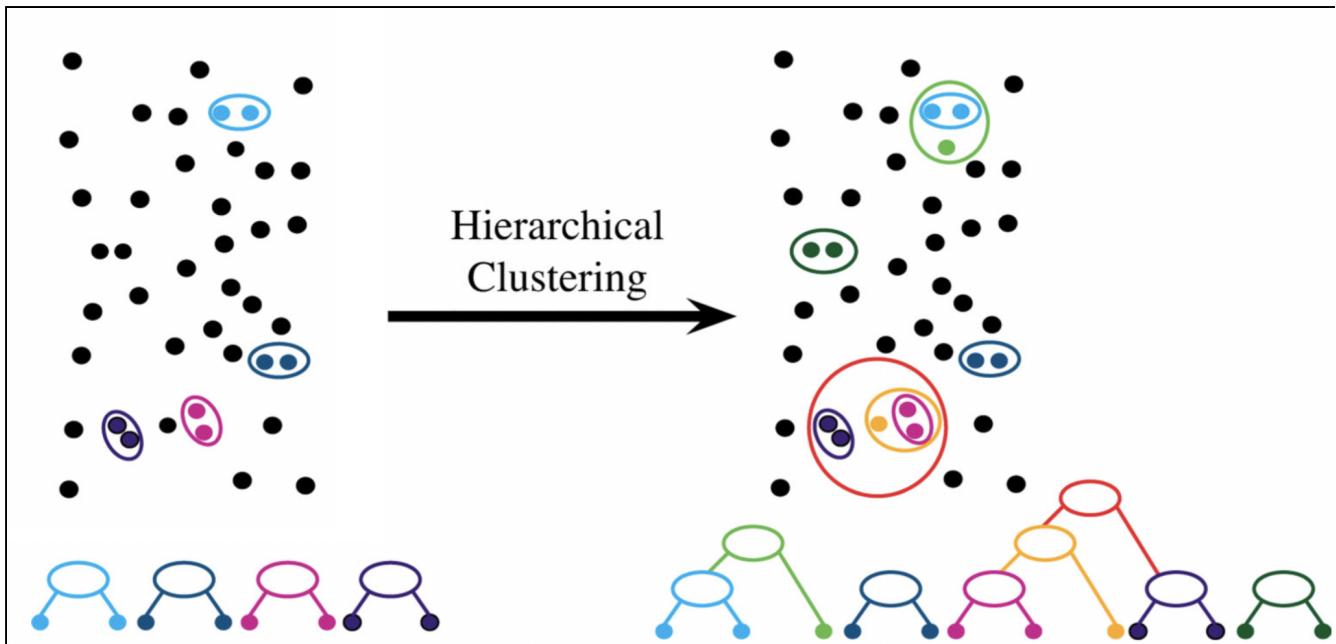
When DR is run, the coordinates of each cell on the tSNE plot can be added to the dataset. For example:

	Level of 'CD3' protein	Level of 'CD4' protein	Level of 'CD8' protein	Level of 'NK1.1' protein	tSNE X	tSNE Y
Cell #1	100,000	850,094	534	346	-2	20
Cell #2	900,000	1424	991,242	128	5	-8
Cell #3	860,523	849	420	242	0	14
Cell #4	872,049	125	235	952,284	10	-12
-Cell #5	457	157	312	892,401	-15	3

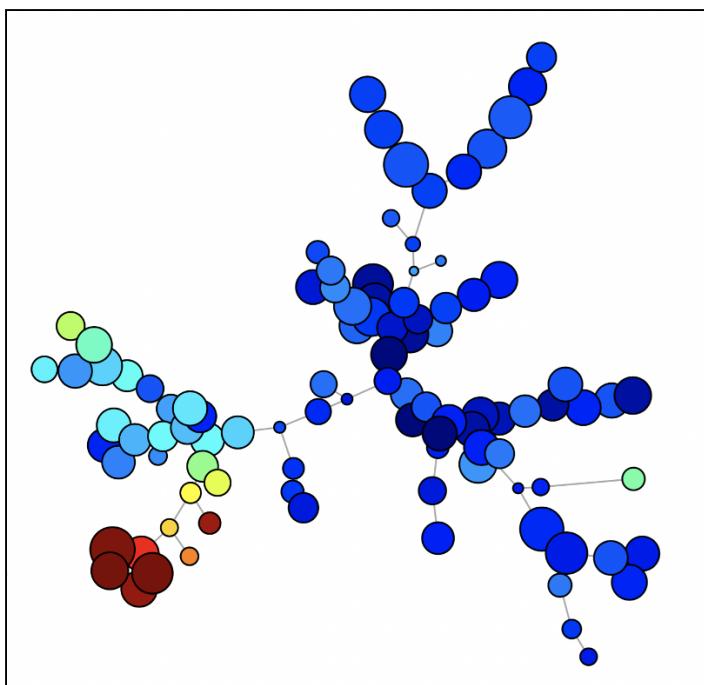
3b. Clustering

Clustering is a computational analysis approach that seeks to group cells together based on their similarity in high-dimensional space. Put simply, clustering attempts to group cells together in an automated and data-driven way. Unlike gating, the user typically cannot influence which cells end up in each cluster (at least not directly). Each 'cluster' will then

contain a number of cells, that have been deemed to be phenotypically similar. A well known clustering approach in the mass cytometry field is spanning-tree analysis of density normalised events (SPADE), but newer tools such as flow self-organising maps (FlowSOM) are used widely. Some implementations of these tools still struggle with large datasets and restricted functionality due to the use of fixed data formats.



Once cells have been clustered, they can be plotted using a minimum spanning tree or similar format, where each circle represents a cluster, that may contain a certain number of cells, and these can be coloured by the average expression of a specific marker in that cluster:

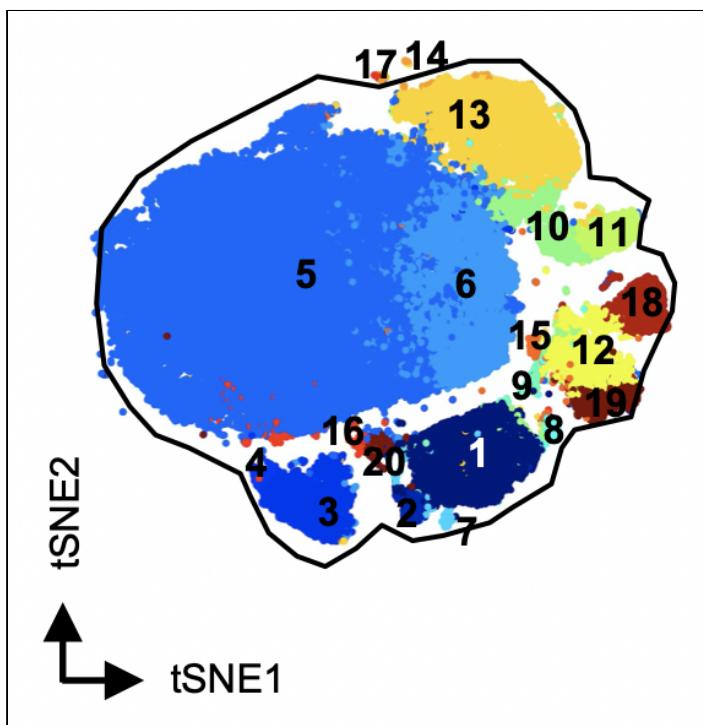


When clustering is run, the cluster assignment of each cell can be added to the dataset. For example:

	Level of 'CD3' protein	Level of 'CD4' protein	Level of 'CD8' protein	Level of 'NK1.1' protein	Cluster
Cell #1	100,000	850,094	534	346	1
Cell #2	900,000	1424	991,242	128	2
Cell #3	860,523	849	420	242	3
Cell #4	872,049	125	235	952,284	4
-Cell #5	457	157	312	892,401	5

3c. Using these approaches together

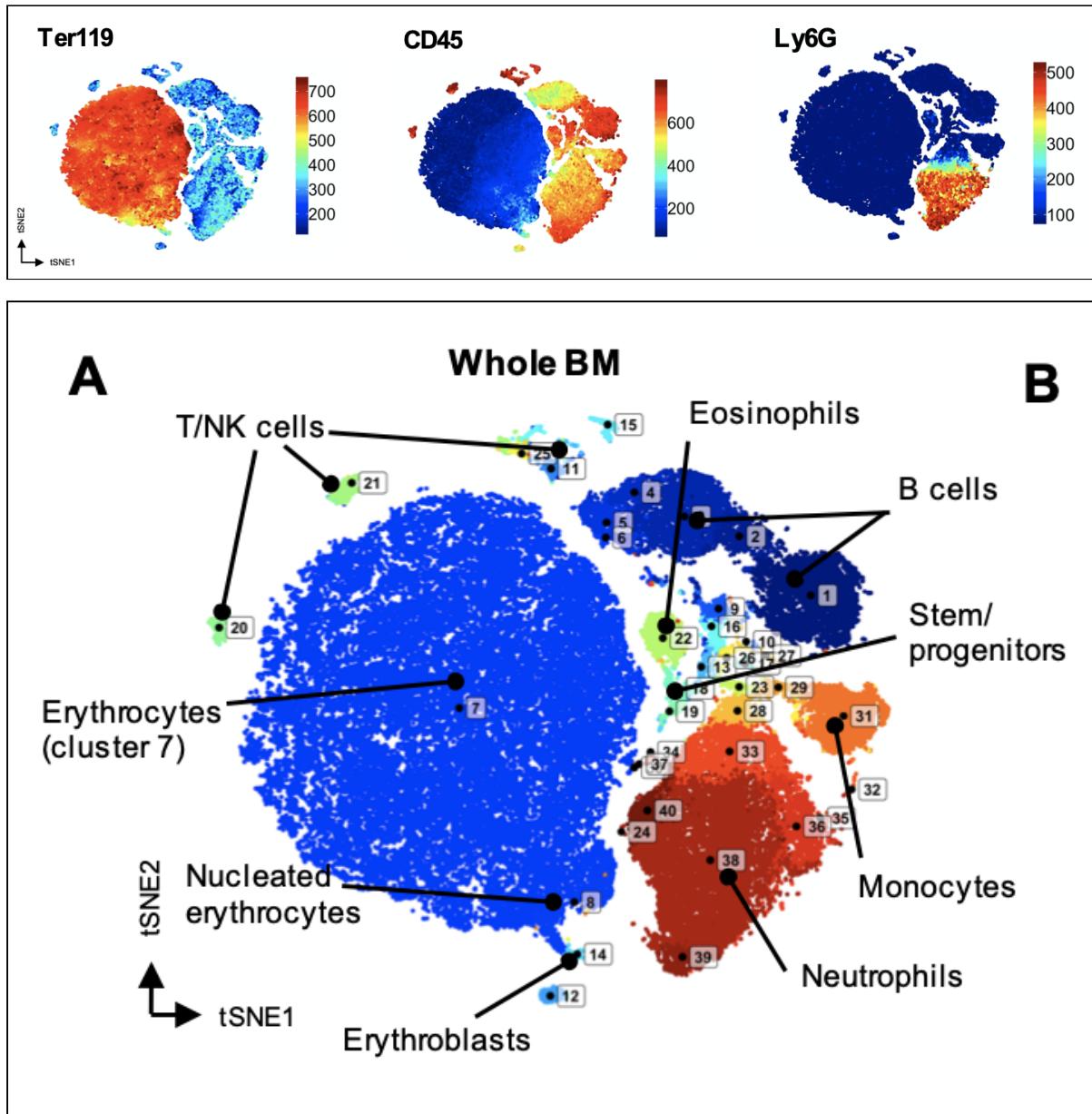
Clustering will group cells into 'clusters' or 'nodes'. However, what kind of cells are in each cluster? Are they truly as identical as possible, or could multiple subsets have been inadvertently captured within each cluster? Because we typically look at clustering using some form of cluster plot (i.e. each circle is a cluster) we do not get to see which cells are contained within each cluster. By using dimensionality reduction, we can visualise each individual cell in the dataset, and then colour each cell by which cluster it belongs to. Many discovery approaches (including Spectre) use a combination of clustering and dimensionality reduction to cluster and visualise all cells in the dataset through a data-driven approach. Typically these are executed by clustering on the cellular markers, and the resulting cluster identities are plotted using a tSNE plot (i.e. clustering is not performed on the tSNE parameters themselves).



Now we have the DR and clustering assignments added to the dataset.

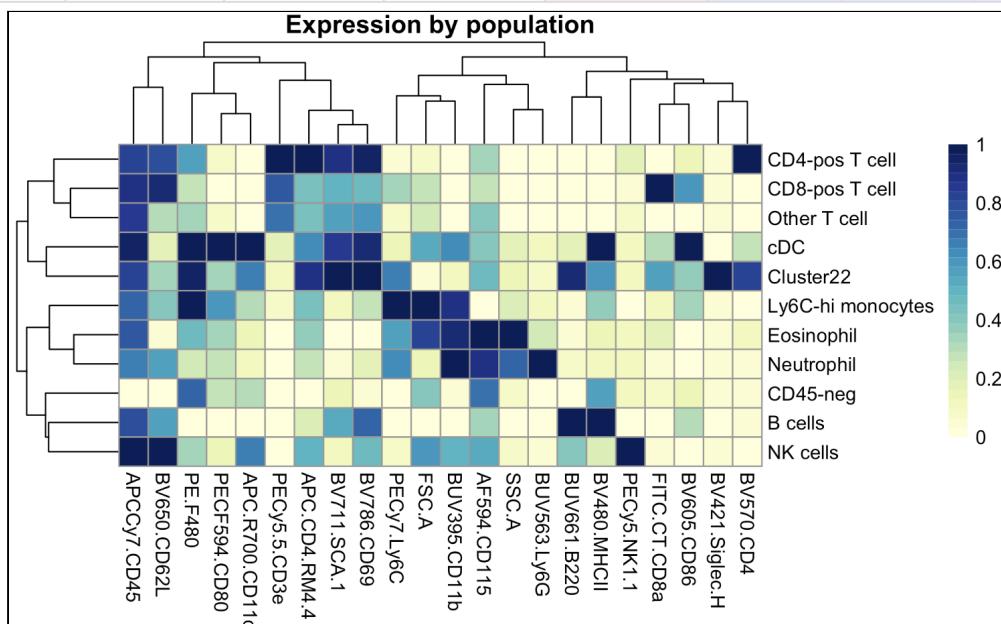
	Level of 'CD3' protein	Level of 'CD4' protein	Level of 'CD8' protein	Level of 'NK1.1' protein	tSNE X	tSNE Y	Cluster
Cell #1	100,000	850,094	534	346	-2	20	1
Cell #2	900,000	1424	991,242	128	5	-8	2
Cell #3	860,523	849	420	242	0	14	3
Cell #4	872,049	125	235	952,284	10	-12	4
-Cell #5	457	157	312	892,401	-15	3	5

In this case, we are able to see what kind of cells are contained within each cluster. We are also able to compare this to the level of expression of various markers.



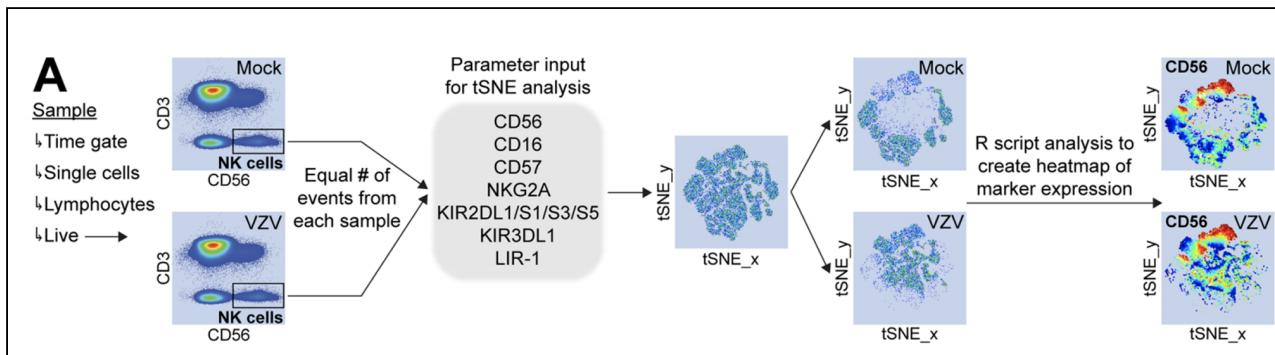
As a result we can see the clusters, determine what cluster represents which population, and label them (e.g. cluster 1 represents CD4+ T cells, cluster 2 represents NK cells etc). We can then quantify changes in sample numbers or expression patterns between samples, etc.

	Level of 'CD3' protein	Level of 'CD4' protein	Level of 'CD8' protein	Level of 'NK1.1' protein	tsNE X	tsNE Y	Cluster	Annotation
Cell #1	100,000	850,094	534	346	-2	20	1	CD4+ T cell
Cell #2	900,000	1424	991,242	128	5	-8	2	CD8+ T cell
Cell #3	860,523	849	420	242	0	14	3	Double negative T cell
Cell #4	872,049	125	235	952,284	10	-12	4	NKT cell
-Cell #5	457	157	312	892,401	-15	3	5	NK cell



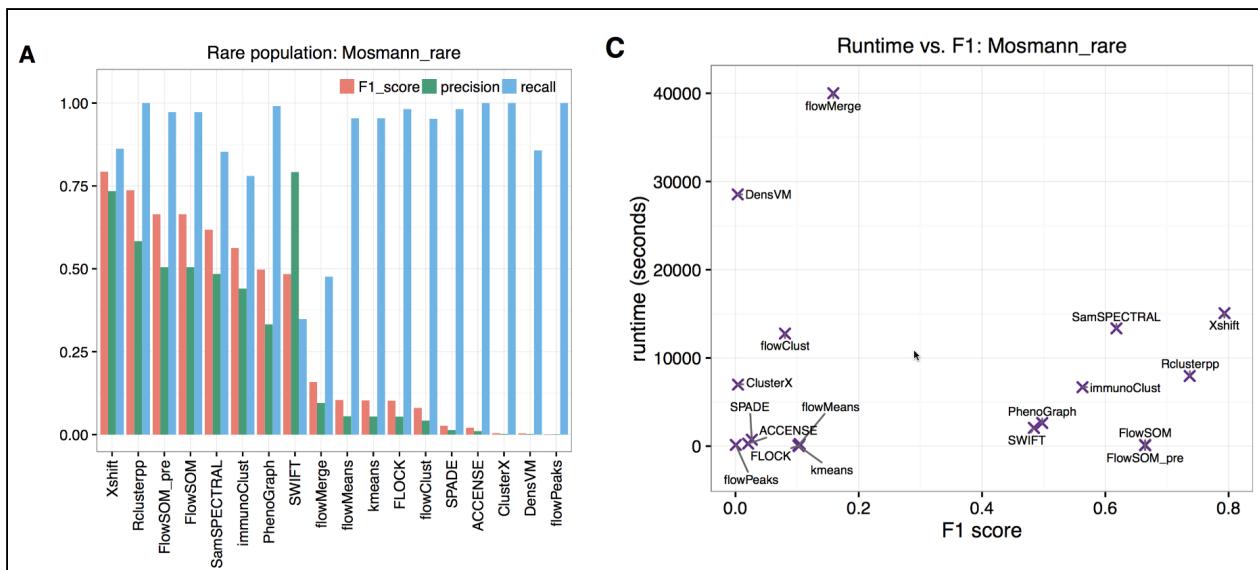
4. Analysing multiple samples

Typically, the objective of a clustering + dimensionality reduction approach is to merge multiple samples from an experiment, generate clusters that represent groups of cells in the dataset, plot these using dimensionality reduction, and quantitatively compare these changes between experimental groups. For this analysis to work, we merge all the samples that are to be analysed into a single dataset. As such, clustering and dim. reduction are performed on the *whole* dataset, not on each sample individually. This way, differences between samples can be examined and compared directly.



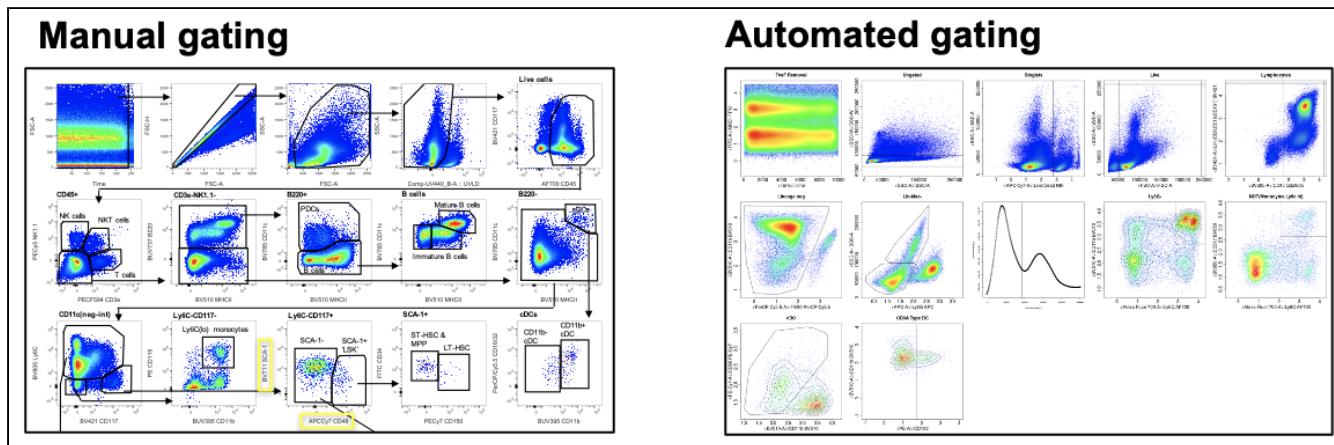
5. The choice of tool can influence throughput and accuracy

Run-time is one of the major limitations of various tools. Tools that run faster can be repeated, modified, and optimised based on the required conditions. Tools that take a long time to run encourage the user to not experiment with settings and approaches, as the run time delays the work. The clustering tool FlowSOM appears to have fast run times, and performs with reasonable accuracy. As a result, it is a favoured approach by many in the field.

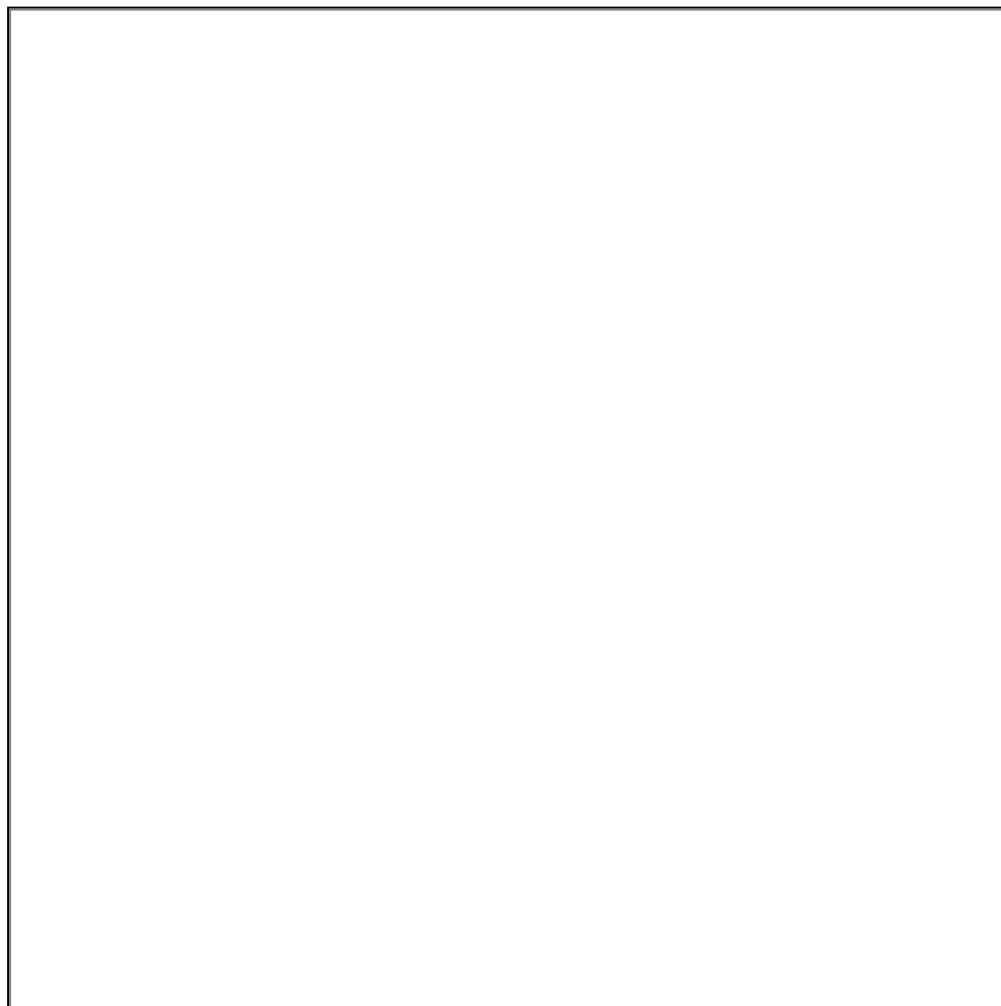


6. Other computational approaches

Manual gating is also a form of discovery analysis, and gates can be created/manipulated experimentally to explore the dataset. However, the manual nature of this 2D approach makes it challenging as a complete discovery approach for high-dimensional datasets. *Automated gating* is an approach which seeks to address this limitation, and is geared towards the identification of subsets of cells (similar to clustering) by automatically generating a gating tree.



Another use of 'computational' analysis approaches is in *replicative* analysis. This is essentially a process of 'replicating' cellular labels (derived from manual gating or clustering) or new samples or datasets. This could be done by 'replicating' gates on multiple samples in one of the automated gating approaches (with some form of automated adjustment for slight variations in marker expression), or through the use of an automated cell classifier, such as a Random Forest classifier.

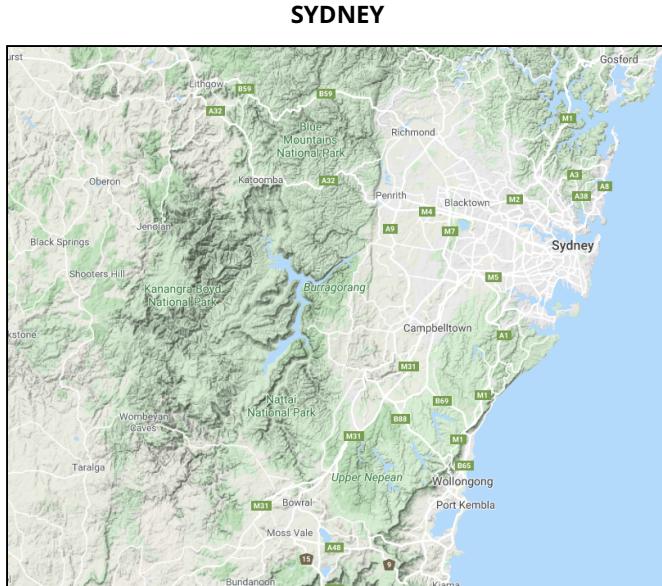


7. How do I think about each approach?

A thematic overview

We can think of 'discovery' and 'replicative' approaches like two different types of maps.

Discovery (clustering/dimensionality reduction): topographical map



What does each map provide

A topographical map allows us to assess many things about an area (elevation, terrain type, water, distance). In other words, it allows us to assess the *structure* of the landscape, whether it be a city, or a previously unexplored area of wilderness.

What does each map let us do

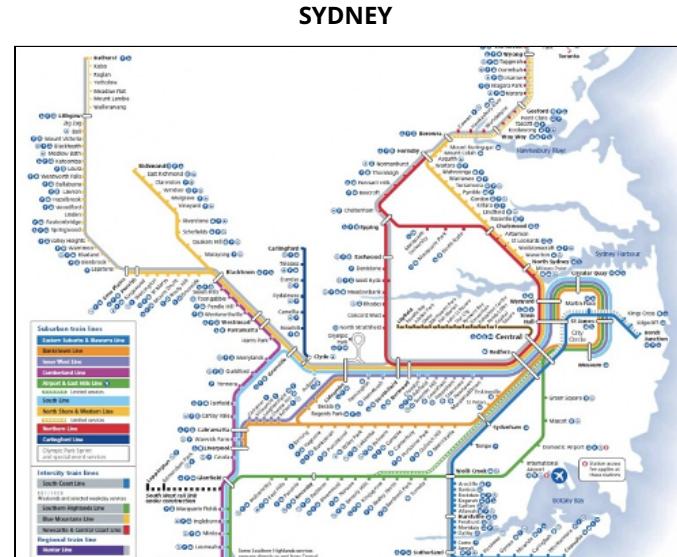
A topographical map allows us to learn about the environment, and if we need, determine how best to navigate through the area, whether previously navigated or not. This is especially helpful if we are trying to plan a hike through the national parks in the area.

What each map does not do

A topographical map does not give us a clear selection of important sites in the city, nor provide an easy to interpret method for getting a quick train from home to the city.

How would each map manage in a wilderness area

Replicative: railway network map



A railway map allows us to locate the important locations (stations) in the city we are visiting. These locations have already been deemed as important – either for population density reasons (major suburban areas), or functional reasons (city centre, etc).

A railway map allows us to determine the easiest and quickest way to get to and from pre-determined important locations (e.g. your suburb to the city).

A railway map doesn't give us any information about the nature of the landscape, and would not help with navigating the area on foot, bike, or car. It also doesn't give us an *accurate* representation of distances, as the map is warped in such a way to fit all the stations onto the map.

MFANGANO ISLAND, KENYA

MFANGANO ISLAND, KENYA



In a wilderness area, a topographical map can be generated by satellite imaging, and provide us with information about an area, even if it is previously unexplored.

How do these map concepts translate to approaches to cytometry analysis

Discovery analysis (much like a topographical map) allows us to analyse *any* dataset, whether well studied, or never before examined. It is driven by the *landscape* of the data, not by our pre-selected important populations. This allows us to study the data and make decisions about how we are to analyse and interpret it.

A railway map does not exist for previously unexplored areas, until they have been developed.

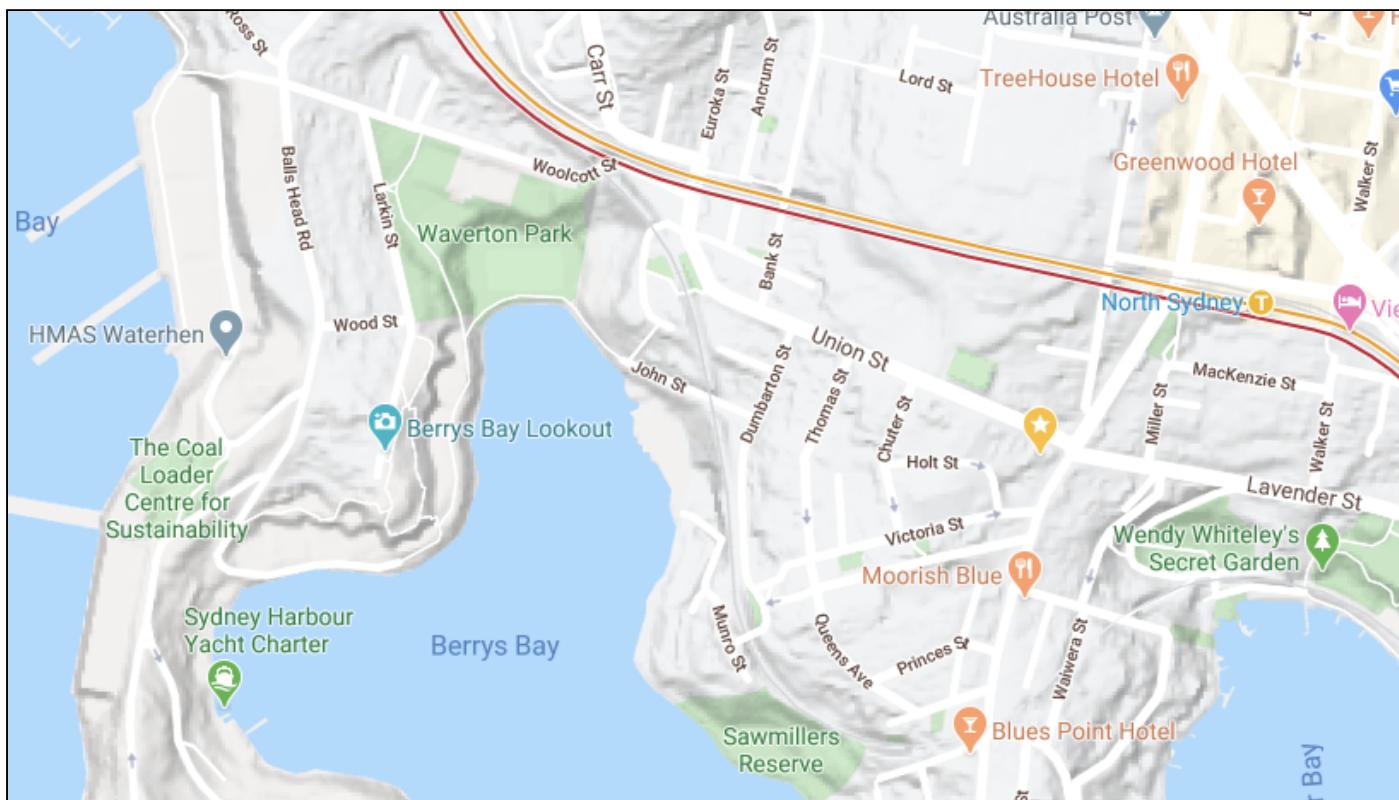
N/A

Replicative analysis (much like a railway map) allows us to quickly replicate major important populations (driven by their importance in the field, and how well they have been studied). This approach assumes that we do not need to know anything about the structure of the data, nor about unexpected phenotypes, but are simply replicating the labelling of populations we deem important.

Side note: the 'cell atlas' – an integrated 'Google Map'

The ideal type of map would be an integrate map, such as Google Maps. These maps integrate topography (landscape structure), water, distance, major train stations, roads, traffic, cafes, restaurants, and a tools for calculating optimal driving, public transport, or walking routes.

NORTH SYDNEY



This is essentially the ultimate aim of projects such as the [Human Cell Atlas](#): to create an integrated atlas of all cell types in the human body, using multiple single-cell technologies, across various ages, backgrounds, and states of inflammation.

So how do I utilise these 'discovery' approaches?

This depends on the objective of your analysis. Here are a few considerations:

These tools are an extension of the scientist, not a replacement. This is perhaps the most important point to be made on this page. Whether performing manual gating, automated gating, clustering, or classification; these are all tools that become an extension of the investigator. This means that any finding: novel populations, changes in cell states or frequencies from one experimental group to another, are only demonstrated to be true through multiple modes of analysis and consistent reproducibility. A cluster generated that appears to be 'novel' must be validated through other means (e.g. cell sorting to identify the 'form' and 'function' of the cell) and through consistent reproducibility by multiple investigators. Ultimately, a meta-assessment by the field (multiple laboratories all being able to reproduce a finding) is necessary for any 'novel' cell type, or for any experimental finding, to be accepted.



You don't just have to pick a single approach. Using multiple methods to investigate a dataset is wise, as it allows you to gain different perspectives on your analysis. For example, after performing clustering we encourage a manual exploration of the clusters in the datasets (using gating) to help utilise the domain knowledge of the researcher in understanding the dataset.

Clustering/dimensionality reduction allows you to understand the structure of the dataset. This approaches is great for analysing cytometry datasets, allowing you to see the mutually-exclusive groups of cells that are present – rather than depending on presumed +/- expression rules. However, each cluster generated does not necessarily represent a specific cell type. Because clusters are generated through a *data-driven* approach, they are biased by strong patterns in the data, and by more frequent populations of cells. If trying to find *known* populations in the dataset, then two potential problems might occur – the clustering might:

1. Over-cluster (splitting known cell types into more than one group) or
2. Under-cluster (grouping multiple cell types together into a single cluster).

Under-clustering is addressed by re-running an analysis to generate a higher number of clusters. Over-clustering is addressed by annotating clusters in a meaningful way (E.g. if both cluster 1 and 2 are neutrophils, then these can be 'merged' by adding the label 'neutrophil' to both). However, these fact that multiple clusters were generated could suggest that there are previously not-considered subsets of this larger group. Alternatively, this may just be to slight variation in the marker expression on these cells. If the cell types are *not* known, then a more careful assessment is required, which brings us to our next point...

When is a cluster a population? This is the key question. A cluster is generated computationally, based on features of the data, not by the knowledge of the scientists, or by other studies that are able to validate the identity and function of a group of cells. A great example of a study that performed a comprehensive characterisation of a 'novel' cell type, using RNA (scRNAseq), protein (flow/sorting), form (cytology), and function (sorting/culture) is demonstrated in [Villani et al 2017](#).

There are situations where clustering/dimensionality reduction is not straight forward. If you are trying to 'replicate' the results of a carefully designed (and validated!) gating strategy, then clustering won't necessarily reproduce exactly the same populations – perhaps a small population that is known to be important (through sorting experiments etc) has weak marker correlation, and so is divided amongst larger groups of cells. Alternatively, cells that exist at the intersection of multiple lineages (e.g. multi-lineage progenitors) might come out as their own cluster, or be divided amongst the clusters containing their progeny. This does not mean that such approaches can't be used, or don't add value, but it does require a considered approach.

What about batch effects? One key area where the clustering/dimensionality reduction approaches on their own will struggle, is in managing batch effects. If samples have been run in multiple batches, then some changes in marker expression might result in clusters being generated by the batches, rather than the populations. This is managed in manual gating because slight shifts in gates are able to be made by examining 2 markers at a time. This increases the workload of the investigator, but allows for a careful adjustment for obvious/evident batch effects – though importantly, not for global or unknown batch effects.

There are excellent batch-alignment and data integration strategies available to allow clustering/dimensionality reduction approaches to be used on cytometry data – you can [read more here](#).