Q&A from
Reconstruction and analysis of B cell lineage trees from single cell data using Immcantation

Speakers: Kenneth Hoehn and Susanna Marquez
Broadcast date: November 9, 2021

Q1 Are there tutorials using the immcantation tools for B cell lineage construction for bulk BCR sequences

You can find tutorials in the Examples and Vignettes sections of Dowser (https://dowser.readthedocs.io) and Change-O (https://changeo.readthedocs.io).

Q2 If we are starting with GEX data that hasn't yet been annotated by cell type, how could we generate these annotations?

We recommend to visit the Vignettes section in the Seurat website, where there are multiple tutorials available. Example: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Q3 I am using 10X single cell Genomics and later using the FASTA and annotation file to run the IgBLAST sequencing alignment, but I know that the CellRanger pipeline doesn't arrange all of the VDJ BCR sequences correctly. So how do we rearrange it so it is correct and what would you recommend to do quality control on the BCR-seq data before classifying clonotypes?

This 10x tutorials in the Immcatation website details how to align the sequences using IgBlast: https://immcantation.readthedocs.io/en/stable/tutorials/10x_tutorial.html

For QC, we usually look for standard things like a high frequency of mitochondrial genes, as well as things like cells with two heavy chains (must be excluded for clonal clustering), overlapping of barcodes between samples with the same sequences (possible contamination), and that cells annotated as B cells by gene expression data have a matching BCR.

Q4 Do the BCR sequences have to be with the same length within a clone ?

Many methods to identify groups of clonally related sequences, start by first splitting sequences into groups that share the same V and J gene assignments, and that have the same junction (or, equivalently CDR3) length. This is based on the assumption that members of a clone will necessarily share all of these properties, because they originate from a common V(D)J recombined ancestor. One limitation of this method is that by requiring the same junction length, sequences that have accumulated insertions or deletions in the junction will not be assigned the same clone. There are other methods (not implemented in Immcantation) that use different types of grouping, that don’t require sharing the same junction length.

Q5 So the threshold is ALWAYS automatically picked?

Sometimes the method cannot automatically find a threshold. You have a few options:

- Select a threshold by visually inspecting the plot

- Include data from an external group (see https://shazam.readthedocs.io/en/stable/vignettes/DistToNearest-Vignette/#calculating-nearest-neighbor-distances-across-groups-rather-than-within-a-groups)
- If the distance-to-nearest distribution is not bimodal, use the spectral clustering method in SCOPer (https://scoper.readthedocs.io/en/stable/topics/spectralClones/), which uses an adaptive threshold to determine the local sequence neighborhood.

Q6 How to incorporate light chain details to cluster B-cells?

SCOPer (https://scoper.readthedocs.io/) can incorporate light chain information in the identification of clones, when using single cell data. The light chain V and J genes can be used to further split the clonal groups identified using the heavy chain information.

Q7 Can we select heavy and light chain at once to compute the threshold ?

The distance threshold is decided using heavy chain data. Heavy chains are sufficient to determine most B cell clonal relationships (Zhou JQ and Kleinstein SH, 2019). The light chain information can be used later to split clonal groups that have differing light chain V and J genes.

Q8 How much improvement do we get on the B-cell clonal clustering by adding the light chain?

Not many studies have looked at this, but a recent one showed that ~80% of heavy chain defined clones have consistent light chains. This means they would not be changed by incorporating the light chain in the way we currently do. See *Ig H Chains Are Sufficient to Determine Most B Cell Clonal Relationships* (Zhou JQ and Kleinstein SH, 2019).

Q9 Will it be possible to use paired heavy and light chains sequences to run IgPhyML and build a tree based on both sequences?

Yes, we're actively working on this within Dowser and IgPhyML. Email me (kenneth.hoehn@yale.edu) if you want to try out these functions early and don't mind using code under development :)

Q10 Can the clonal lineages account for correct class switching events (e.g. IgA -> IgG and not other way) in the lineages constructed for BCR?

No, the lineage trees only use mutation information from the BCR sequences. Class switching is a separate process, so it's not clear how to incorporate that information into the tree building process. However, we do have ways of visualizing isotypes on trees and predicting the simplest sequence of class switching events along a tree: https://www.biorxiv.org/content/10.1101/2020.05.30.124446v1

Q11 Should we rely more on the heavy chain because it has more diversity then the light chain or use both. Lastly, when we define clonotypes if these clonotypes are real should they be confined to a specific cluster of cells on a UMAP for example?

Most analyses use the heavy chain, partially for historical reasons because many pipelines are set up around bulk BCR data. The light chain is important for binding as well, though, so best to include it where possible. It's common for clones to contain cells from different UMAP clusters, so long as they correspond to B cell subtypes. Clones are defined by BCR sequence information, while UMAP clusters are defined by gene expression information.

Q12 Is it possible to generate lineage tree based on somatic hypermutation from two dataset (e.g. longitudinal/different timepoints data) if the CDR3 sequences are not identical, or the CDR3 sequences must be identical to generate lineage tree?

The CDR3 sequences do not need to be identical to group sequences into the same clone. They are allowed to vary based on the Hamming distance threshold found using the Shazam package. A threshold of 0.1, for instance, would group cells that have the same heavy chain V gene, J gene, CDR3 length, and 90% CDR3 nucleotide similarity. Clones can contain sequences from different timepoints – just include data from all timepoints (from the same subject) while assigning clones. We assigned clones across timepoints for this paper: https://elifesciences.org/articles/70873

Q13 Can correlationTest be used between 2 remote timepoints (years) ?

Yes, there's no restriction on time, so long as you find clones that span at least two of the timepoints. We used it for some lineages spanning 14 years in this paper: https://elifesciences.org/articles/70873

Q14 How important is Polymorphism detection and genotyping in lineage tree construction?

Good question. Uncorrected novel alleles will likely increase the length of the "trunk" branch of your trees leading from the germline to the most recent common ancestor of your lineage. Genotyping should increase the size of your clones by restricting the number of V gene alleles your sequences can possibly match to. These should improve the accuracy of lineage tree construction (so long as they're done properly) but I think it remains to be determined how much. Our package for novel allele detection and genotyping is here: https://tigger.readthedocs.io/en/stable/. There's also RAbHIT, which allows for Ig haplotyping: https://yaarilab.bitbucket.io/RAbHIT/.