**Pre-processing of single cell VDJ sequencing**

**Step 1:** Check all files exist in the right locations

**Step 2:** Run VDJAnnotator\_10X.py

This has four steps:

1. For each barcode, identify IGH/L or TRA/B sequences;
2. Split each sequence type into difference fasta file (for annotation);
3. Perform local VDJ annotation (using Local\_immune\_repertoire\_annotator\_1.0.py) on the fasta files;
4. Read the output of step c and collate this information for each barcode.

**Outputs below for BCR:**

|  |  |
| --- | --- |
| File name (for sample "HC\_10X1") | Description |
| Annotation\_HC\_10X1\_IGH.txt | Annotations for each heavy chain sequence |
| Annotation\_HC\_10X1\_IGL.txt | Annotations for each light chain sequence |
| CDR3\_frequencies\_HC\_10X1\_IGH.txt | CDR3 frequencies for each heavy chain sequence |
| CDR3\_frequencies\_HC\_10X1\_IGL.txt | CDR3 frequencies for each light chain sequence |
| CDR3\_lengths\_HC\_10X1\_IGH.txt | CDR3 lengths for each heavy chain sequence |
| CDR3\_lengths\_HC\_10X1\_IGL.txt | CDR3 lengths for each light chain sequence |
| Cell\_annotation\_HC\_10X1\_IG.txt | Combined heavy and light chain cell annotation (similar to output provided in Seurat object) |
| Nucleotide\_trimmed\_codon\_HC\_10X1\_IGH.fasta | Heavy chain nucleotide sequences trimmed to start of first codon |
| Nucleotide\_trimmed\_codon\_HC\_10X1\_IGL.fasta | Light chain nucleotide sequences trimmed to start of first codon |
| Protein\_translation\_HC\_10X1\_IGH.fasta | Protein translation of heavy chain sequences |
| Protein\_translation\_HC\_10X1\_IGL.fasta | Protein translation of Light chain sequences |
| Sequences\_HC\_10X1\_IG\_IGH.fasta | Heavy chain sequences |
| Sequences\_HC\_10X1\_IG\_IGL.fasta | Light chain sequences |
| Trimmed\_sequences\_HC\_10X1\_IG\_IGH.fasta | Heavy chain nucleotide sequences trimmed to start of first codon and with the constant region removed |
| Trimmed\_sequences\_HC\_10X1\_IG\_IGL.fasta | Light chain nucleotide sequences trimmed to start of first codon and with the constant region removed |
| filtered\_contig\_annotations\_BCR\_HC\_10X1.csv | raw contig file |

**Outputs below for TCR:**

|  |  |
| --- | --- |
| File name (for sample "HC\_10X1") | Description |
| Annotation\_HC\_10X1\_TRA.txt | Annotations for each TRA chain sequence |
| Annotation\_HC\_10X1\_TRB.txt | Annotations for each TRB chain sequence |
| CDR3\_frequencies\_HC\_10X1\_TRA.txt | CDR3 frequencies for each TRA chain sequence |
| CDR3\_frequencies\_HC\_10X1\_TRB.txt | CDR3 frequencies for each TRB chain sequence |
| CDR3\_lengths\_HC\_10X1\_TRA.txt | CDR3 lengths for each TRA chain sequence |
| CDR3\_lengths\_HC\_10X1\_TRB.txt | CDR3 lengths for each TRB chain sequence |
| Cell\_annotation\_HC\_10X1\_IG.txt | Combined TRA and TRB chain cell annotation (similar to output provided in Seurat object) |
| Nucleotide\_trimmed\_codon\_HC\_10X1\_TRA.fasta | TRA chain nucleotide sequences trimmed to start of first codon |
| Nucleotide\_trimmed\_codon\_HC\_10X1\_TRB.fasta | TRB chain nucleotide sequences trimmed to start of first codon |
| Protein\_translation\_HC\_10X1\_TRA.fasta | Protein translation of TRA chain sequences |
| Protein\_translation\_HC\_10X1\_TRB.fasta | Protein translation of TRB chain sequences |
| Sequences\_HC\_10X1\_IG\_TRA.fasta | TRA chain sequences |
| Sequences\_HC\_10X1\_IG\_TRB.fasta | TRB chain sequences |
| Trimmed\_sequences\_HC\_10X1\_IG\_TRA.fasta | TRA chain nucleotide sequences trimmed to start of first codon and with the constant region removed |
| Trimmed\_sequences\_HC\_10X1\_IG\_TRB.fasta | TRB chain nucleotide sequences trimmed to start of first codon and with the constant region removed |
| filtered\_contig\_annotations\_BCR\_HC\_10X1.csv | raw contig file |

**Step 3:** Run sequences through IMGT (for “gold standard” annotation).

This will be very similar to what we got locally, but this will be easier for justifying in manuscripts and aligns with other groups’ pipelines.

This has five steps:

1. Batch to together sequences in groups of <=1,000,000;
2. Upload to IMGT and run;
3. Download output and unpackage unto distinct directories for each batch;
4. Run R code to split up key output files per sample. (lines 184-695)
5. Reruns VDJ annotation using the IMGT output (using VDJAnnotator\_10X\_IMGT.py) and collates this information for each barcode.

**Step 4:** Cluster TCRs/BCRs per group (lines 752-820)

Runs “Network\_generation\_Single\_cell\_grouped\_2.0\_IMGT.py” based on groups of samples (defined by variable Overall\_sample\_group):

e.g.

Overall\_sample\_group sample\_id

[1,] "HC\_10X1" "HC\_10X1"

[2,] "HC\_10X3" "HC\_10X3"

[3,] "HC\_10X4" "HC\_10X4"

[4,] "HC\_10X5" "HC\_10X5"

[5,] "HC1" "HC1"

[6,] "HC2" "HC2"

[7,] "RCC-metP" "01\_Kidney\_met\_panc1\_biopsy\_CD45p1"

[8,] "RCC-metP" "02\_Kidney\_met\_panc1\_biopsy\_CD45p2"

[9,] "RCC-metP" "04\_Kidney\_met\_panc1\_blood\_CD45p1"

*e.g. 1 sample in group "HC\_10X1" but 3 samples in group "RCC-metP"*

This has X steps run per chain:

1. Batch to together fasta sequences from all samples within group for that chain;
2. Group together all identical V(D)J sequences;
3. Cluster together all similar V(D)J sequences (using CD-hit) and perform network generation (max distance = 3bp, due to lower capture of sequences by single cell, this was considered to be a good balance for identifying sequences that were clonally related where full mutational chains were missing);
4. Output cluster information.

**Outputs below for BCR IGH:**

|  |  |
| --- | --- |
| File name (for sample "HC\_10X1" heavy chain clustering) | Description |
| Cluster\_identities\_RCC-metP\_IGH.txt | Cluster information: col 1 = sequence number, col 2 = cluster ID, col 3 = cluster node ID, col 4 = number of sequences in node, col 5 = mapping for raw droplet IDs |
| Edges\_RCC-metP\_IGH.txt | Edge information for network analysis: col 1 = edge number, col 2 = node 1, col 3 = node 2 |
| Reduced\_sequence\_identities\_RCC-metP\_IGH.txt | Cluster mapping for raw droplet IDs to node IDs |
| Reduced\_sequences\_RCC-metP\_IGH.fasta | Sequences for node IDs |
| Sequences\_combined\_RCC-metP\_IGH.fasta | Combined sequences for clustering |
| Tmp\_RCC-metP\_IGH\_ | Temporary file |
| Tmp\_RCC-metP\_IGH\_.bak.clstr | Temporary file |
| Tmp\_RCC-metP\_IGH\_.clstr | Temporary file |
| Tmp\_RCC-metP\_IGH\_\_coclustered | Temporary file |

**Step 5:** Combine annotation (IMGT) and clustering information for both BCR and TCR into single object for further single cell analysis.

From line 823 onwards