TCR Overlap - Notebook

Thank you for using TCRtoolkit! This report is generated from sample data and metadata you provided. The report is divided into the following sections:

1 Report Setup

▶ Code

```
Pipeline information and parameters:

Project Name:          TCRtoolkit-Bulk
Workflow command:      nextflow run main.nf --data_dir test_data/minimal-example     --samplesheet test_data/minimal-
example/samplesheet.csv     --output out-minimal-dev     --max_memory 10GB --max_cpus 4
Pipeline Directory:    /Users/kmlanderos/Documents/Johns_Hopkins/Karchin_Lab/Projects/TCRtoolkit/
Date and time:         2025-08-12 18:21:37.943204
```

2 Analysis

2.1 Pair-wise Repertoire Similarity: Morisita-Horn

The **Morisita-Horn Index** is a similarity metric often used to compare the similarity **between two samples**. Unlike simpler metrics that only check if a TCR is present or absent (e.g. Jaccard index), the Morisita-Horn index is **sensitive to clonal abundance**. This is critical for TCR data because it properly weights the contribution of highly expanded T-cell clones that are likely driving an immune response.

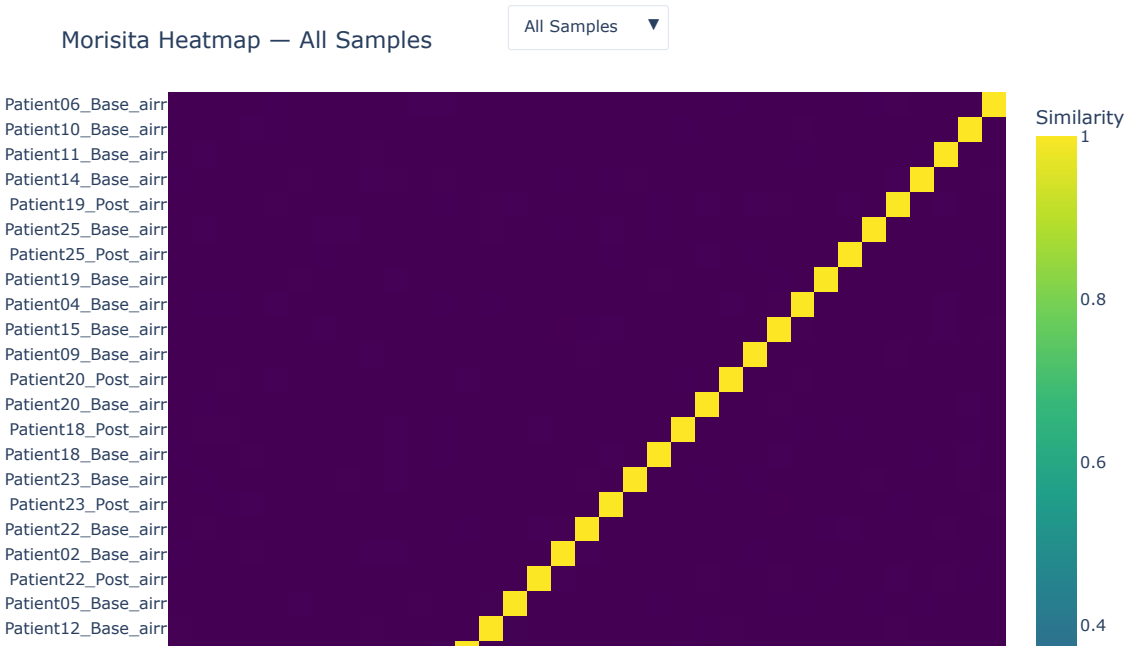
The index is calculated based on the probability that two TCRs, randomly selected from each of the two samples, will belong to the same clonotype. The index metric for 2 given samples is calculated as follows:

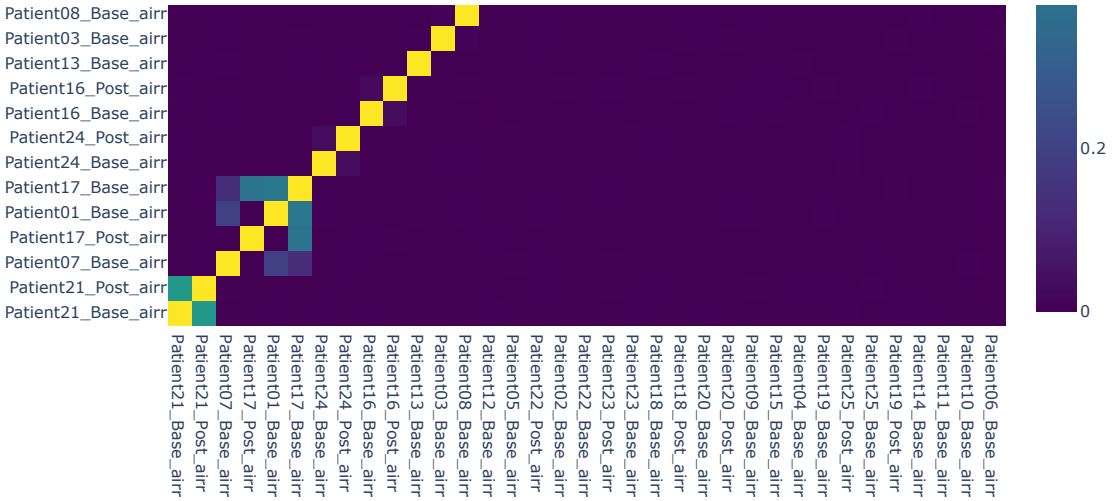
$$M(A,B) = \frac{2\sum_{i=1}^{S} a_i b_i}{(D_a + D_b)AB}; D_a = \frac{\sum_{i=1}^{S} a_i^2}{A^2}, \quad D_b = \frac{\sum_{i=1}^{S} b_i^2}{B^2}$$

Where:

- A and B are the sets of unique CDR3 amino acid sequences (TCRs) in samples A and B.
- a_i (b_i) is the number of times TCR i is represented in the total A (B) from one sample.
- S is the total number of unique TCRs in the two samples.
- D_a and D_b are the Simpson Index values for samples A and B, respectively.

▶ Code





2.1.1 Interpreting the Index Value

It answers to the question: *How similar are the structures of two TCR repertoires, considering which clones are shared and how frequent those shared clones are in each sample?*

M(A,B)=0 (No Overlap): The two TCR repertoires are completely different. They do not share any of the same TCR clonotypes.

M(A,B)=1 (Identical): The two TCR repertoires are identical in structure. They consist of the same TCR clonotypes at the exact same relative frequencies.

0<M(A,B)<1 (Partial Overlap): This indicates some degree of similarity. A value closer to 1 means the repertoires are very similar, sharing many of the same clones, especially the dominant ones. A value closer to 0 means they share very few clones, or the clones they do share are rare in one or both samples.

A key strength of this index is that two samples can share a TCR, but if that TCR is highly abundant in one sample and extremely rare in the other, the index will be low, correctly reflecting the different immunological states.

2.2 Persistent and Unique Clones within patients

To understand how the TCR repertoire changes over time, we need to identify which T-cell clones are persistent, which are new, and which have disappeared. Analyzing the TCR overlap for patients with multiple timepoint can be an ideal way of understanding changes in the repertoir over time.

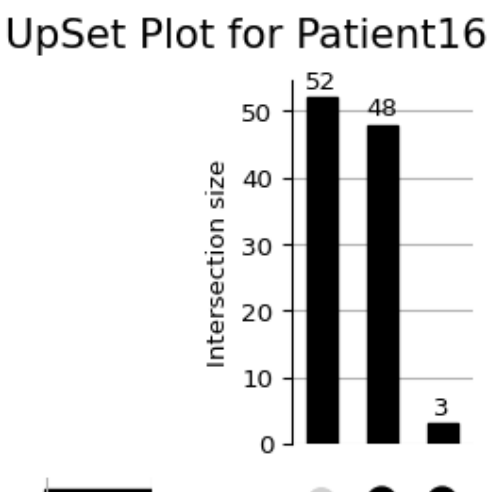
Persistent Clones (The Intersection):
The height of this bar tells you the number of T-cell clones that persisted between the two timepoints. These could be long-lived memory T-cells or clones responding to a chronic antigen or tumor. A large persistent set suggests a stable immune state.

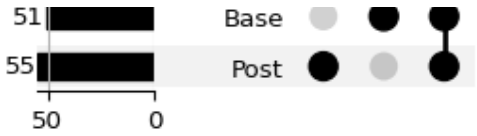
Unique Clones:
- ***Waning Clones:*** Look for the bar with a solid dot for Time 1 only. These are clones that were present initially but disappeared or contracted by Time 2. This could represent the resolution of an acute infection.
- ***Emerging Clones:*** Look for the bar with a solid dot for Time 2 only. These are new T-cell clones that expanded between the timepoints. This is a strong indicator of a new immune response, perhaps due to a new infection, vaccination, or a response to treatment.

▶ Code

▶ Code

Patient16 ▼





Understanding the Upset plot:

1. Top Bar Chart: Intersection Size

Each bar represents the number of TCR clones found in a specific combination of samples. The taller the bar, the more clones there are in that specific group. You can quickly see which combinations (e.g., clones unique to Time 1) are the largest.

2. Dot Matrix: Set Membership

Below the bar chart, a matrix of dots tells you which samples are included in each intersection. A solid dot (●) means that the sample is part of the combination for the bar directly above it. A faint gray dot (○) means the sample is not part of that combination. A line connecting two or more solid dots (●–●) indicates that the bar above represents clones shared between those connected samples.

3. Left Bar Chart: Total Set Size

This horizontal bar chart on the left shows the total number of unique clones in each original sample. This gives you a quick overview of the repertoire diversity for each timepoint before looking at the overlaps.

2.3 Expanding and Contracting clones within patients

To track how the T-cell repertoire changes over time, we first identify clones present at both pre- and post-treatment timepoints. For these **shared clones**, we calculate an **"Expansion Score"**. This score is a powerful metric that quantifies the magnitude and direction of a clone's change in frequency, defined as:

$$Expansion_score = \log_2 \left(\frac{\text{Frequency}_{\text{Post}}}{\text{Frequency}_{\text{Pre}}} \right)$$

- **A positive score indicates expansion** (the clone makes up a larger proportion of the repertoire after treatment).
- **A negative score indicates contraction** (the clone's proportion has decreased).
- A score around 0 represents a stable behavior (the clone's proportion has remained quite similar).

We use **clonal frequency** (the proportion of a specific clone relative to the total number of T-cells sequenced in that sample) **rather than raw counts**. This is crucial because **sequencing depth can vary between samples**. Using frequency normalizes the data, ensuring a fair comparison and preventing a sample with more total reads from appearing to have more expanded clones simply due to deeper sequencing.

From Observations to Significance

While the "Expansion Score" tells us the magnitude of a change, it doesn't tell us if that change is meaningful. The immune system has natural fluctuations (**biological noise**), and the sequencing process itself has variability (**technical noise**). A small change in frequency could easily be a result of this random noise.

To address this, **we perform a statistical test (Fisher's Exact Test) on the raw counts of each shared clone**. This test calculates the probability (the p-value) that an observed change is merely a random fluke. By adjusting these p-values for the thousands of tests run (one for each clone), we get a q-value, which controls the False Discovery Rate. A low q-value gives us high confidence that the observed expansion or contraction is a real, statistically significant event.

The following tables display the T-cell clones that are shared between the pre- and post-treatment timepoints across all patients, focusing on those that have an Expansion Score greater or less than zero. **To highlight the most compelling results, clones that passed our dual-criteria significance test (i.e., have a q-value < 0.05 and an abs(Expansion Score) > 1.0) are shown in bold**.

▶ Code

▶ Code

▶ Code

Table 1: Top Expanding Clonotypes

subject_id	CDR3b	TRBV	TRBJ	Frequency_Pre	Frequency_Post	Expansion_Score	counts_pre
Patient18	**CASSFGRNQPQHF**	**TRBV5-1*01**	**TRBJ1-5*01**	**5.92e-05**	**0.334284**	**12.46**	**2.0**
Patient20	**CASSYSVAGGPYNEQFF**	**TRBV6-8*01**	**TRBJ2-1*01**	**0.0001586**	**0.0116505**	**6.2**	**2.0**
Patient20	CASRDPGTDYGYTF	TRBV7-3*01	TRBJ1-2*01	0.0003172	0.007767	4.61	4.0
Patient17	**CASSPTSGVAGELFF**	**TRBV4-3*01**	**TRBJ2-2*01**	**0.0526844**	**0.0852425**	**0.69**	**6509.0**
Patient17	**CASSCAGTPIGTIYF**	**TRBV21-1*01**	**TRBJ1-3*01**	**0.1342647**	**0.1678802**	**0.32**	**16588.0**
Patient17	**CASSLILLGNTEAFF**	**TRBV7-2*01**	**TRBJ1-1*01**	**0.0949517**	**0.1159721**	**0.29**	**11731.0**

Table 2: Top Contracting Clonotypes

subject_id	CDR3b	TRBV	TRBJ	Frequency_Pre	Frequency_Post	Expansion_Score	counts_pre	coun
Patient18	**CASSGRLTEAFF**	**TRBV6-1*01**	**TRBJ1-1*01**	**0.1245707**	**0.0004074**	**-8.26**	**4207.0**	**2.0**
Patient24	CASSEDWVQGPEAFF	TRBV6-1*01	TRBJ1-1*01	0.0153846	8.4e-05	-7.52	2.0	2.0
Patient24	CASSLYSEGAGELFF	TRBV7-9*01	TRBJ2-2*01	0.0153846	8.4e-05	-7.52	2.0	2.0
Patient18	**CASSLAQGANSPLHF**	**TRBV13*01**	**TRBJ1-6*01**	**0.0659718**	**0.0004074**	**-7.34**	**2228.0**	**2.0**
Patient21	**CASSPMKGYEQYF**	**TRBV18*01**	**TRBJ2-7*01**	**0.1354716**	**0.0089**	**-3.93**	**62599.0**	**388**
Patient23	CASSPFNNNEQFF	TRBV9*01	TRBJ2-1*01	0.0235294	0.0031397	-2.91	4.0	2.0
Patient23	CASSVHTGPSYEQYF	TRBV9*01	TRBJ2-7*01	0.0117647	0.0031397	-1.91	2.0	2.0
Patient21	**CASSYSGVSNQPQHF**	**TRBV27*01**	**TRBJ1-5*01**	**0.0463554**	**0.0126541**	**-1.87**	**21420.0**	**552**
Patient21	**CASTHPGGIYGYTF**	**TRBV27*01**	**TRBJ1-2*01**	**0.0447453**	**0.0129014**	**-1.79**	**20676.0**	**563**
Patient21	CASRQGRVWQPQHF	TRBV2*01	TRBJ1-5*01	0.1152241	0.0407185	-1.5	53243.0	1778
Patient21	**CASSPTGSTEAFF**	**TRBV5-4*01**	**TRBJ1-1*01**	**0.1639384**	**0.0686684**	**-1.26**	**75753.0**	**299**
Patient17	**CASSSQETQYF**	**TRBV2*01**	**TRBJ2-5*01**	**0.0005018**	**0.0002539**	**-0.98**	**62.0**	**50.0**
Patient17	CAWNFDREGEQYF	TRBV30*01	TRBJ2-7*01	1.62e-05	1.02e-05	-0.67	2.0	2.0
Patient21	**CASSQDYGQGVGNTIYF**	**TRBV4-1*01**	**TRBJ1-3*01**	**0.1089763**	**0.0685746**	**-0.67**	**50356.0**	**299**
Patient21	**CASSLGSSSYEQYF**	**TRBV7-2*01**	**TRBJ2-7*01**	**0.0444185**	**0.0293829**	**-0.6**	**20525.0**	**128**
Patient17	**CASSEGRGAAEGKLFF**	**TRBV10-1*01**	**TRBJ1-4*01**	**0.1279594**	**0.0910665**	**-0.49**	**15809.0**	**179**
Patient21	**CASKDGTGSYNEQFF**	**TRBV7-9*01**	**TRBJ2-1*01**	**0.1208811**	**0.0996216**	**-0.28**	**55857.0**	**435**

Interpret the Results

It is critical to not interpret the Expansion Score in isolation. A clone's biological importance depends on the interplay between three factors:

- **Expansion Score (Magnitude)**: How large was the change?
- **Statistical Significance (Confidence)**: How likely is it that the change was real? (q-value)
- **Clone Counts (Abundance)**: What is the clone's overall presence in the sample?

For example, a massive Expansion Score resulting from a change of 1 count to 10 might not be statistically significant and could be noise. Conversely, a highly significant q-value for a clone that changes from 2 counts to 8 might be statistically real but biologically unimportant due to its low abundance.

Therefore, the most robust conclusions come from paying attention to clones that demonstrate both a high abs(Expansion Score) and a low q-value, contextualized by their raw clone counts.