Sample - Notebook

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

Section 1 : Code to setup the report. This section includes the parameters you used to run the pipeline, loading necessary packages, data, etc.

Section 2 : Typical sample level T cell repertoire statistics. Each plot has a description about the statistic shown and formulas used to calculate the metric.

Section 3 : TCR convergence

Section 4 : TCR V gene family usage. The x-axis shows the timepoint collected for each individual, and the y-axis shows the proportion of TCRs that use each V gene family.

Section 5 : Tcrdist. Grouping of T-cell receptors (TCRs) based on CDR sequence similarity.

Section 6 : Olga. Likelihood that specific T-cell receptor (TCR) sequences are produced during the random V(D)J recombination process

Section 7 : Phenotype prediction: Innate, CD8, Treg, Mem

1 Report Setup

▶ Code

```
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:         2025-08-11 23:38:15.114882
```

2 Sample level statistics

Below are plots showing basic T cell repertoire statistics. Each plot has a description about the statistic shown and formulas used to calculate the metric when applicable. Specific biological interpretation of each plot is left to the user.

2.1 Number of clones

The number of unique TCR clones in a sample, known as **repertoire richness**, is a fundamental measure of clonal diversity. **This metric provides critical insight into the state of the immune system**. A repertoire with

high richness is considered **polyclonal** and represents a **healthy, vigilant immune system** with a broad capacity to recognize a vast array of potential threats. In contrast, **low richness indicates an oligoclonal repertoire**, where a few T-cell clones have massively expanded and dominate the immune landscape. This is a **hallmark of a highly focused and active immune response**, such as during an acute infection or a targeted anti-tumor response. Tracking changes in richness over time is therefore crucial for understanding immune dynamics, as a drop in diversity often signals the onset of a response, while a subsequent increase can indicate a return to immune homeostasis.

▶ Code

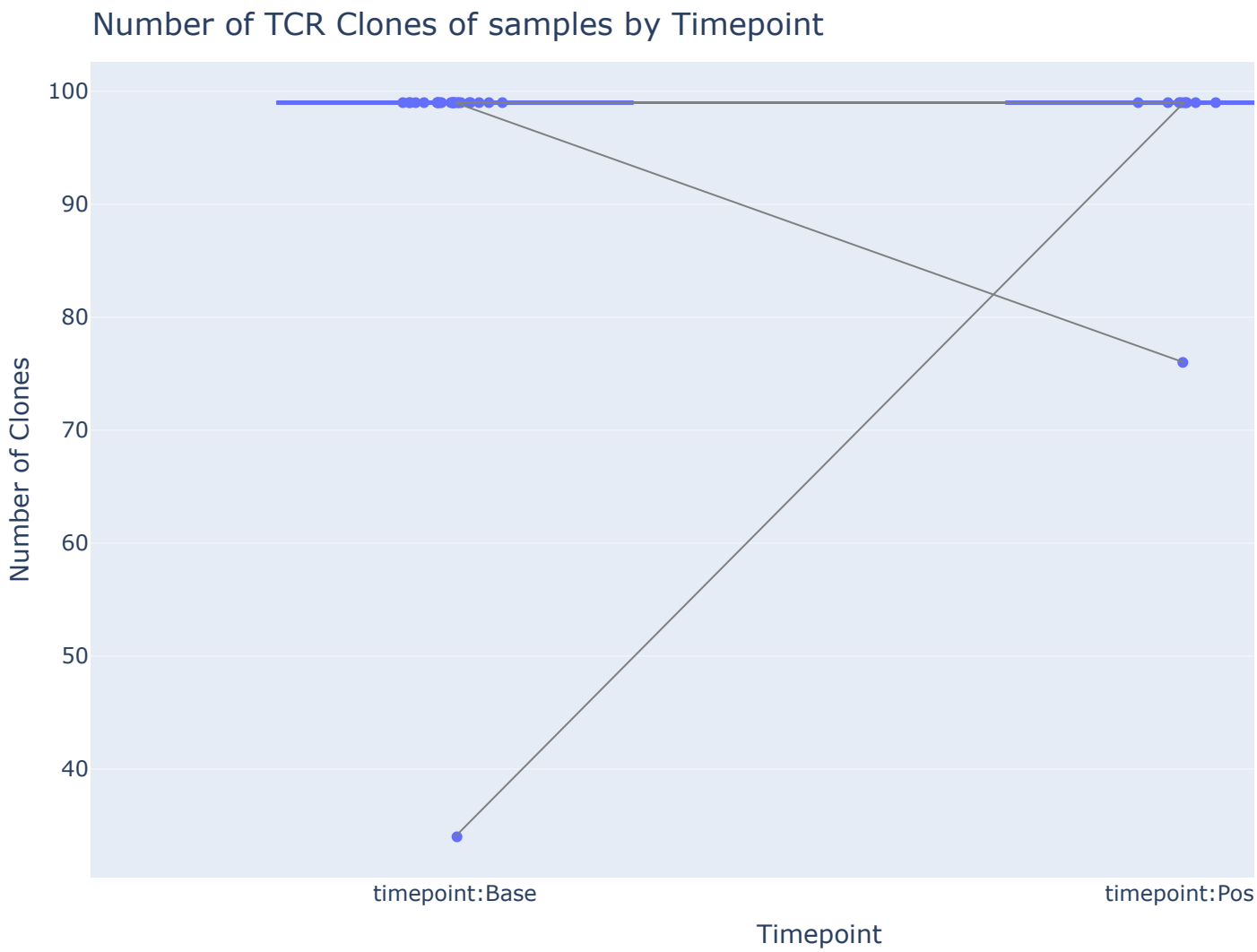


Figure 1. Number of clones across timepoints. A clone is defined as a T cell with a unique CDR3 amino acid sequence. The number of clones is shown on the y-axis and 'origin_timepoint' is shown on the x-axis.

The boxplot allows you to visualize and compare the distribution of repertoire richness across your entire patient cohort at each timepoint. By looking at the boxplot, you assess the overall trend. For instance, if the **median** line of the box for the **"Post" timepoint** is **lower than the "Pre" timepoint, it suggests that the patient group**, **mounted a focused immune response that reduced overall TCR diversity**. The plot also reveals the variability in response across patients and can highlight outliers—individuals whose repertoire diversity changed much more or less dramatically than their peers.

2.2 Clonality

TCR clonality is a measure of the extent to which a T-cell repertoire is dominated by a few highly expanded T-cell clones. It is essentially the opposite of TCR diversity.

What TCR Clonality Tells You About Your Sample Clonality provides a snapshot of the immune system's current focus. Think of your entire T-cell repertoire as an army of soldiers, where each clone is a unique type of soldier trained for a specific target.

Low Clonality (High Diversity): This is like an army with soldiers of every possible type, all standing by in relatively equal numbers.

High Clonality (Low Diversity): This is like an army where a huge portion of the soldiers are of one specific type, mobilized to fight a single, major battle.

▶ Code

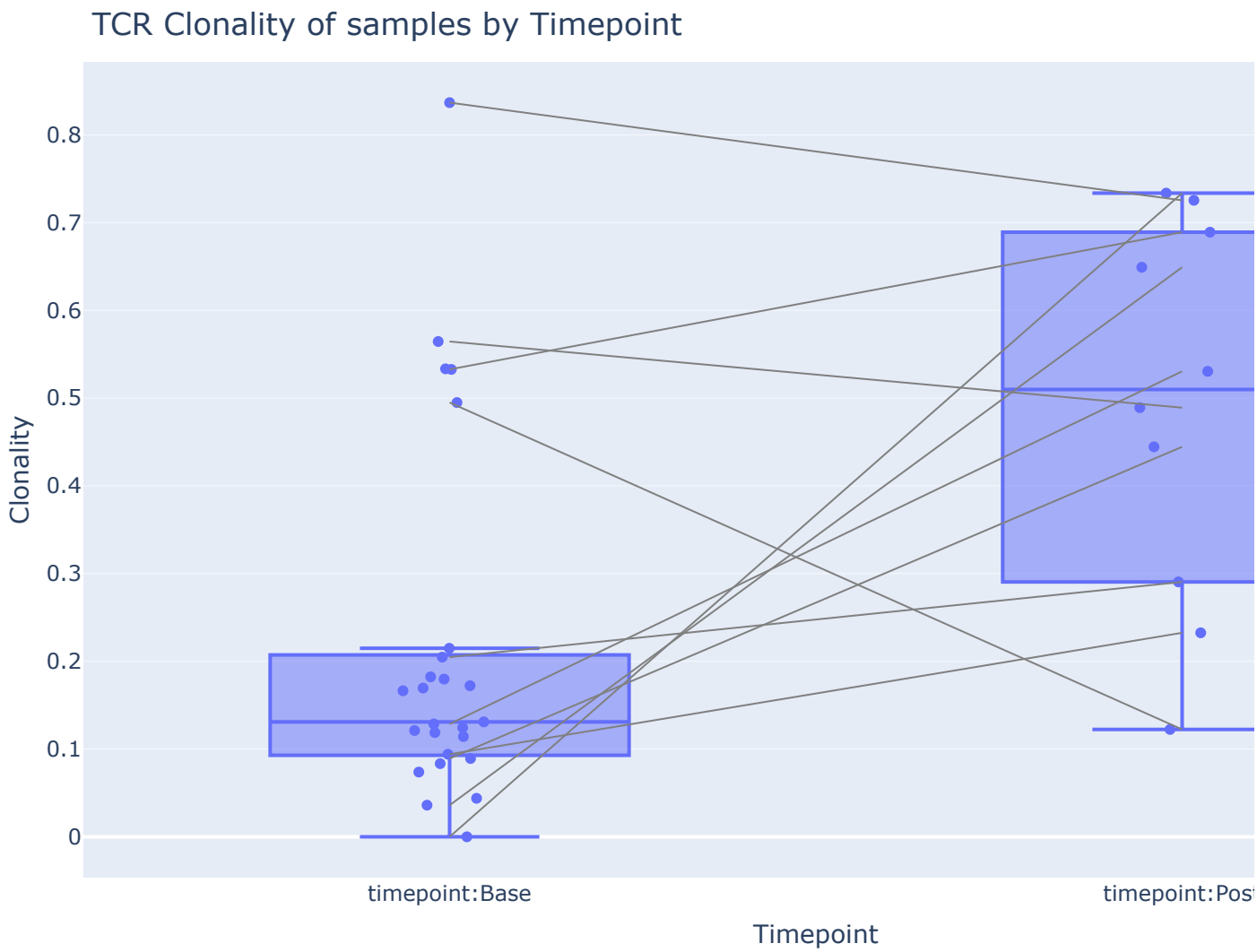


Figure 2. The clonality of samples across timepoints. Clonality is a measure of T cell clonal expansion and reflects the degree to which the sample is dominated by 1 or more T cell clones. Clonality is calculated via:

$$Clonality = \frac{1 - H}{\log_2 N} \quad , \quad H = -\sum_{i=1}^{N} p_i \log_2 p_i$$

where H is the Shannon entropy of a given sample, N is the number of unique TCRs in the sample, and p_i is the frequency of the i th unique TCR in the sample.

2.3 Simpson Index

The Simpson Index (D) tells you about dominance. The Simpson Index answers the question: "If I draw one TCR, then put it back, and draw a second one, what is the probability that I will draw the same clone twice?"

A high probability (value close to 1) means the repertoire is dominated by one or two clones. This signifies **low diversity**.

A low probability (value close to 0) means the clones are more evenly distributed, making it unlikely you'll pick the same color twice. **This signifies high diversity**.

▶ Code

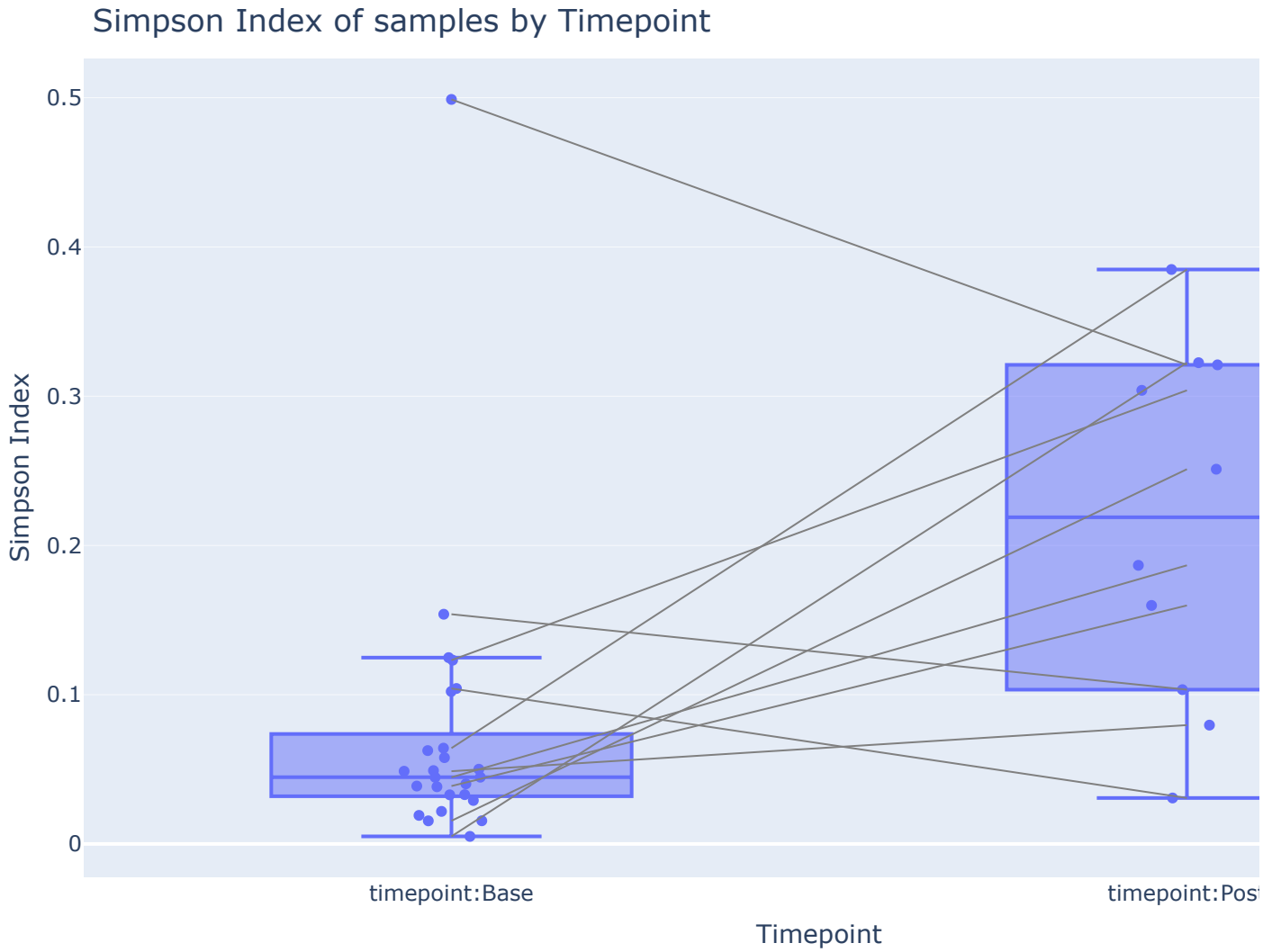


Figure 3. Corrected Simpson Index. The Simpson Index is a measure of diversity that takes into account the number of clones and the relative abundance of each clone in a sample. The corrected Simpson Index, D, is calculated as:

$$D = \sum_{i=1}^{N} \frac{p_i(p_i - 1)}{T(T-1)} \quad , \quad T = \sum_{i=1}^{N} p_i$$

Where N is the number of unique TCRs in the sample, p_i is the frequency of the i th unique TCR in the sample, and T is the total number of T Cells counted in the sample.

2.3.1 How to Use It to Learn About Your Data 🔍

- **Quantify Clonal Dominance:** The Simpson Index tells you how oligoclonal (dominated by a few clones, low diversity) or polyclonal (many clones with even abundance, high diversity) your sample is.
- **Compare Between Groups:** You can use the index to compare diversity between different experimental groups. For example, you could test if a group of patients receiving a certain treatment shows a lower T-cell diversity (indicating a targeted immune response) compared to a placebo group.
- **Track Changes Over Time:** By calculating the index for samples taken at different timepoints, you can track how the structure of a community changes. For instance, in a patient responding to immunotherapy, you might expect to see T-cell diversity decrease as specific anti-tumor clones expand. If the patient recovers and the response contracts, you might see diversity increase again as the repertoire returns to a more balanced, homeostatic state.

2.4 Percent of productive TCRs

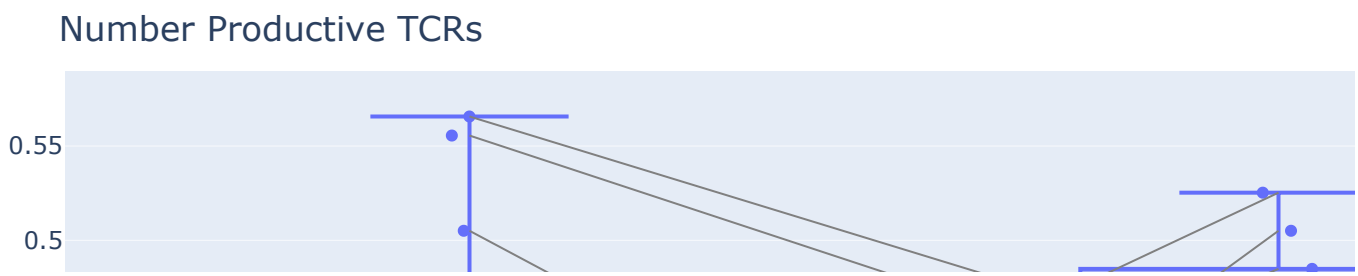
The percent of productive TCRs is one of the first metrics you should check to validate your data's reliability before proceeding to any biological analysis.

As a Quality Filter: You should set a QC threshold (e.g., >75% productive reads). Samples that fail to meet this threshold should be flagged for review or potentially excluded from the analysis. Drawing conclusions about T-cell diversity or clonality from a sample with low productivity is unreliable, as the data is likely noisy and not a true representation of the functional T-cell repertoire.

To Troubleshoot Experiments: If you find that an entire batch of samples has a low productive percentage, it points to a systematic issue in your experimental protocol, most commonly an ineffective gDNA removal step or problems with RNA integrity.

To Ensure Confidence in Results: By confirming that your samples have a high percentage of productive TCRs, you can be confident that the clonotypes you identify and analyze represent real, functional T-cells that are actively participating in the immune response.

▶ Code



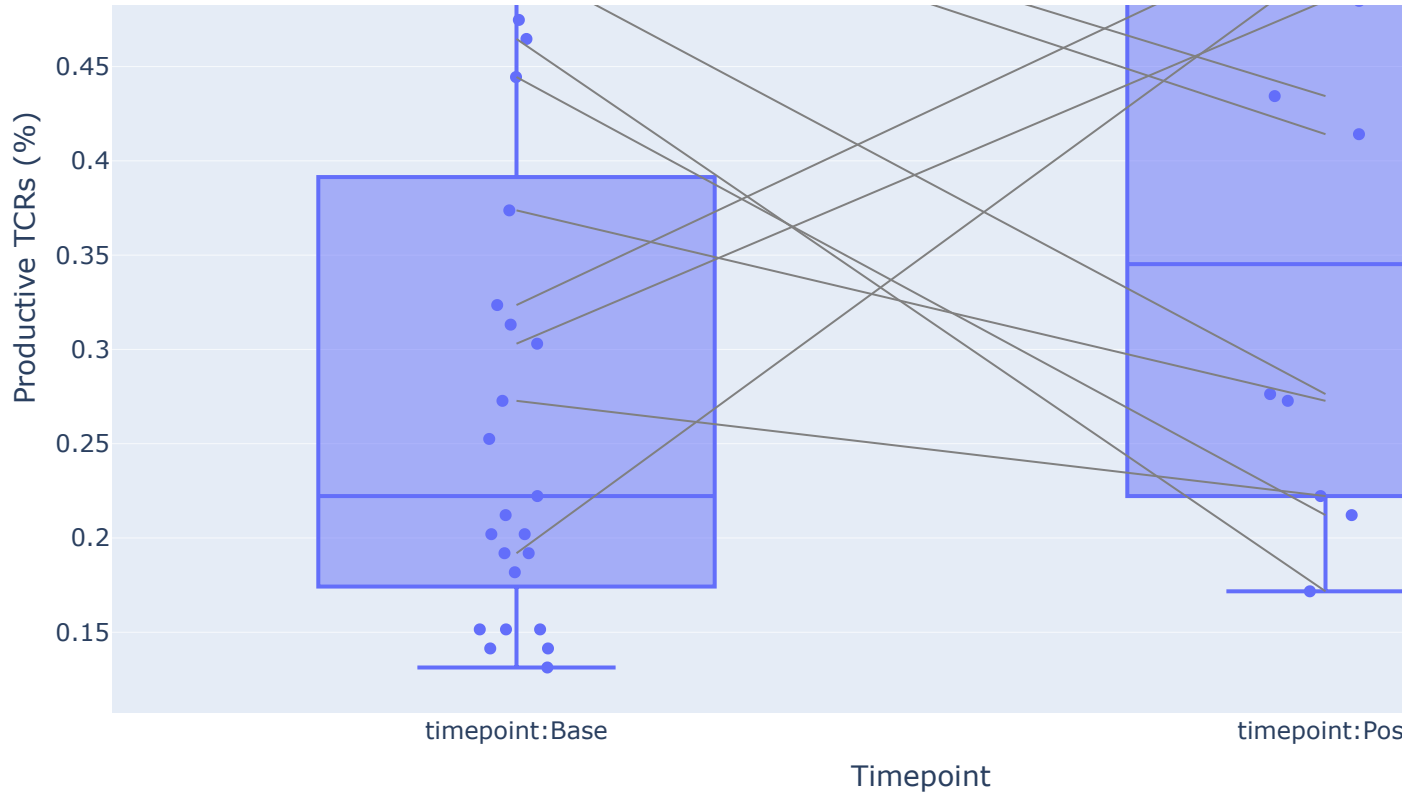


Figure 4. Percent of productive TCRs. A productive TCR is a DNA/RNA sequence that can be translated into a protein sequence, i.e. it does not contain a premature stop codon or an out of frame rearrangement. The percent of productive TCRs is calculated as:

$$Percent\ productive\ TCRs = \frac{P}{N}$$

where P is the number of productive TCRs and N is the total number of TCRs in a given sample.

Potential Biological Interpretations 🤔

If a sample has passed all other QC checks but still shows a consistently lower or higher percentage of productive TCRs compared to others, it might reflect certain biological states.

1. Thymic Selection and Output The process of T-cell development in the thymus, known as thymic selection, is designed to eliminate T-cells that fail to make a productive TCR rearrangement. A subtle shift in the productive percentage could theoretically reflect changes in this process. For instance, a condition that alters thymic function or leads to an increased export of immature T-cells might result in a slightly different ratio of productive to non-productive transcripts detected in the blood.

2. Gene Dysregulation in Disease Certain disease states, particularly T-cell malignancies like lymphomas, involve significant dysregulation of normal cellular processes. It's hypothesized that a malignant T-cell clone might aberrantly transcribe its non-productively rearranged allele at a higher rate than a normal T-cell. If this malignant clone dominates the sample, it could lead to an overall decrease in the productive percentage. Furthermore, defects in cellular machinery like the nonsense-mediated decay (NMD) pathway, which normally degrades transcripts with premature stop codons, could allow more non-productive sequences to persist and be detected.

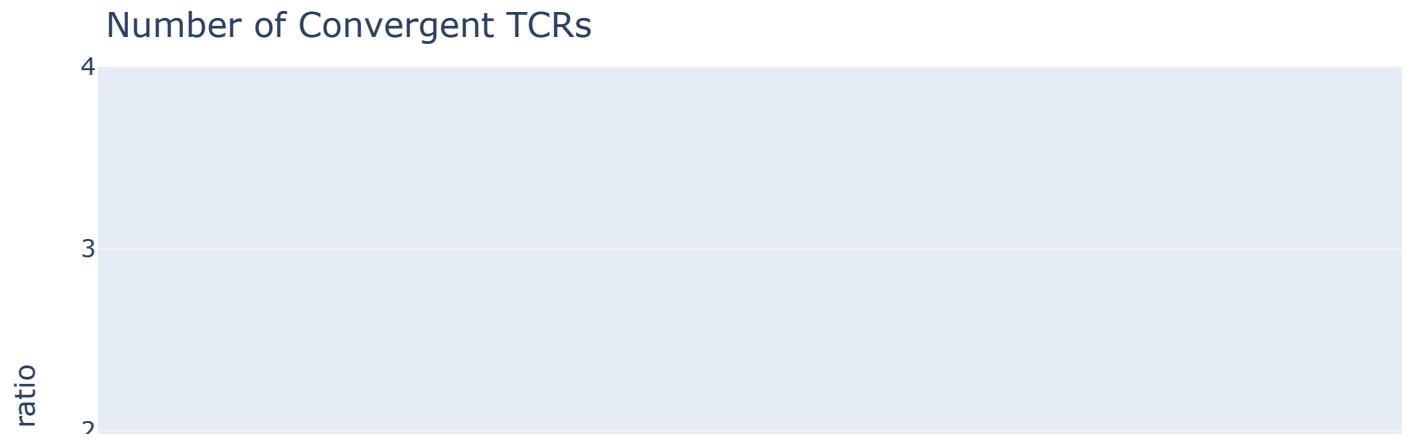
2.5 Average CDR3 Length

▶ Code

Figure 5. Average CDR3 Length The average length of the CDR3 region of the TCR. The CDR3 region is the most variable region of the TCR and is the region that determines antigen specificity.

3 TCR Convergence

▶ Code



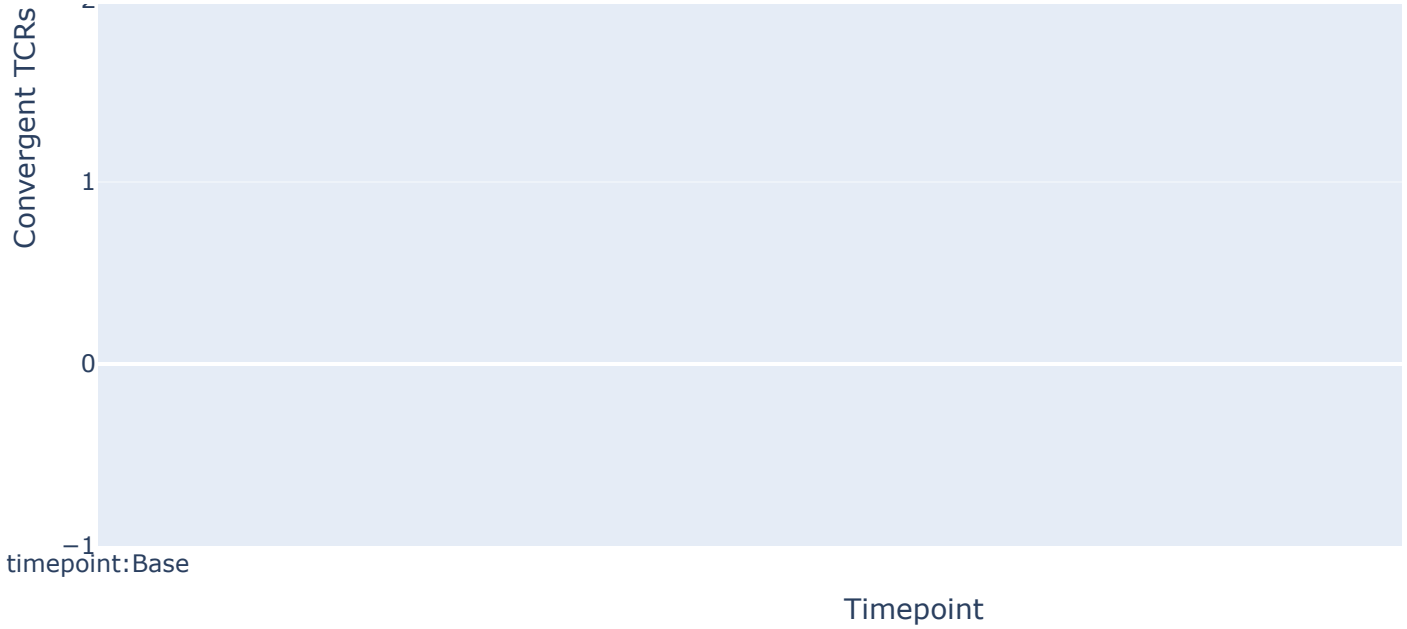


Figure 6. TCR Convergence The ratio of convergent TCRs to total TCRs. A convergent TCR is a TCR that is generated via 2 or more unique nucleotide sequences via codon degeneracy.

3.1 Clonal Expansion

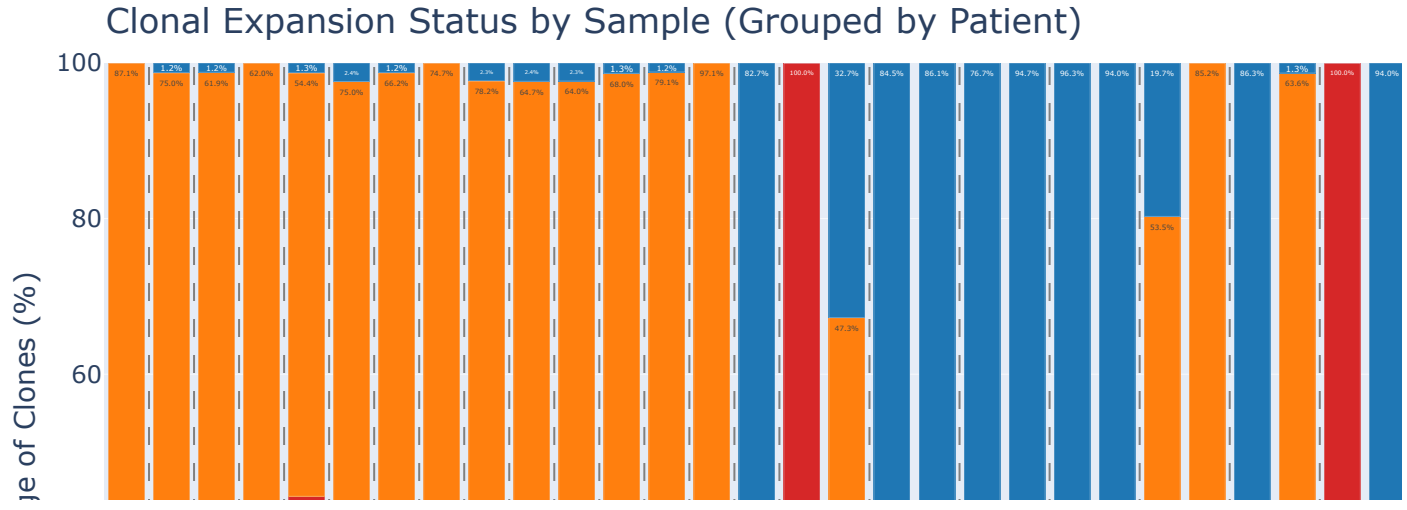
This stacked bar plot provides a snapshot of your T-cell repertoire's structure, revealing the balance between dominant, expanded clones and the diverse pool of rare clones. By categorizing each TCR based on its frequency, we can infer the state of the immune system.

Each bar is a complete TCR repertoire from one sample, broken down into three key functional categories:
- **Highly Expanded (or Hyperexpanded):** Clones constituting > 1% of the total productive TCR reads.
- **Expanded (or Large):** Clones with a frequency between 0.1% and 1%.
- **Non-expanded (or Small/Rare):** Clones with a frequency < 0.1%.

Clonal frquency ia calculated as:

$$f_i = \frac{\text{Read count for clone } i}{\text{Total reads for all productive TCR clones}} \times 100\%$$

▶ Code



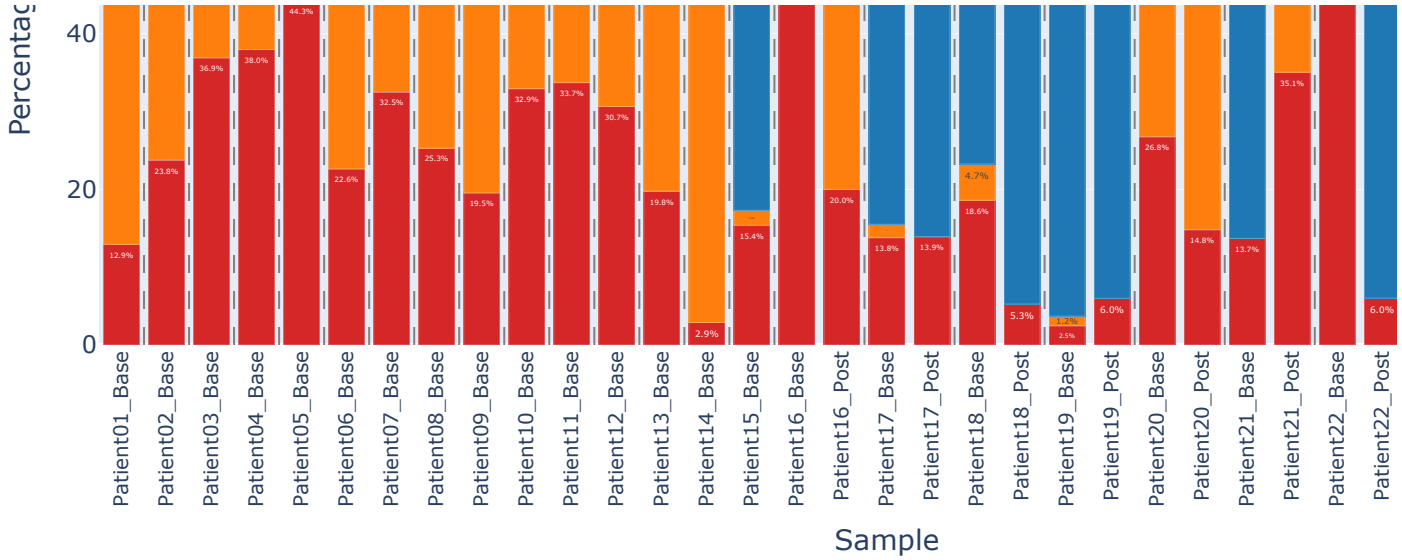


Figure 7. Sample's expansion composition Stacked barplot showing the percentage (y-axis) of "Non-Expanded", "Expanded" and "Highly Expanded" clones per each sample (x-axis). Samples derived from the same patient are clubbed together.

Understanding Each Category

Highly Expanded Clones (> 1%)

- What they are: These are the dominant "fighters" of the immune system. A single clone making up more than 1% of the entire repertoire is exceptionally frequent and almost certainly represents a T-cell that has undergone massive proliferation in response to a specific antigen.
- Biological Insight: A large "Highly Expanded" portion indicates a highly focused, or oligoclonal, immune response. This is characteristic of:
    - An acute viral infection (e.g., flu, COVID-19).
    - A strong response against cancer antigens (especially with immunotherapy).
    - Certain autoimmune diseases where a specific self-antigen is targeted.

Expanded Clones (0.1% - 1%)

- What they are: These are moderately expanded clones. They are participating in an immune response but are not as dominant as the hyperexpanded clones. They can represent secondary responders or clones from a previous infection that have settled into a memory state.
- Biological Insight: This category provides a picture of the active, but not overwhelming, immune activity. A significant portion of the repertoire in this category suggests a broad, active response without being dominated by a single specificity.

Non-expanded Clones (< 0.1%)

- What they are: This category represents the vast diversity of the TCR repertoire. It includes the massive pool of naïve T-cells waiting for an antigen and the wide variety of memory cells from past immune encounters.

- Biological Insight: A large “Non-expanded” portion indicates a highly diverse, or polyclonal, repertoire. This is the hallmark of a healthy, resting immune system with the potential to respond to a vast array of future threats. It is your immunological “reservoir.”

Comparing timepoint samples within Patients

The real power of this plot comes from comparing samples, such as before (Pre) and after (Post) a treatment or vaccination.

Look for a shift towards “Highly Expanded”: If the blue “Highly Expanded” bar grows significantly from the “Pre” to the “Post” sample, it strongly suggests a successful immune response was mounted. The specific T-cells recognizing the target antigen (e.g., from a vaccine or tumor) have proliferated dramatically.

Look for a return to diversity: If a sample starts with a large “Highly Expanded” portion (e.g., during an infection) and then shifts towards a larger “Non-expanded” portion at a later timepoint, it can signify the contraction phase of an immune response. The dominant clones recede after clearing the antigen, and the repertoire returns to a more diverse, homeostatic state.

By analyzing the changing proportions of these categories, you can build a powerful narrative about how the immune system is responding to disease, vaccination, or therapy.

3.2 Clonal Expansion

This table is useful because it reveals the identities of the dominant clones driving the immune response in each sample. By listing the most highly expanded T-cell receptors, **you can pinpoint the specific “effector” cells that have proliferated most, likely in response to a key antigen from a tumor or pathogen.** Having their precise sequences allows you to track these key clones over time, compare them across different patients to find shared public clones, and provides the necessary information for downstream functional studies to determine their antigen specificity.

▶ Code

Top 15 Expanded Clones

CDR3b	TRBV	TRBJ	Co
CASSYVGNTGELFF	TRBV6-5*01	TRBJ2-2*01	61
CASSLELNTEAFF	TRBV5-1*01	TRBJ1-1*01	34
CASSWTGIGTSPNYGYTF	TRBV5-1*01	TRBJ1-2*01	33
CASSQRGLAFF	TRBV4-1*01	TRBJ1-1*01	12
CASSQVFRGGVGDTQYF	null	TRBJ2-3*01	73
CASSLRAPYIMNTEAFF	TRBV5-6*01	TRBJ1-1*01	62
CASSLFTGAYTEAFF	TRBV14*01	TRBJ1-1*01	53
CASSILYPGAGELFF	null	TRBJ2-2*01	50
CASSPSGGGNTEAFF	TRBV9*01	TRBJ1-1*01	39
CASGNEKLFF	null	TRBJ1-4*01	35
CATSDSYEQYF	null	TRBJ2-7*01	33

CASSPGLAGLEQYF	TRBV5-6*01	TRBJ2-7*01	30
CASSEEKWTGSTEAFF	TRBV5-1*01	TRBJ1-1*01	30
CASRATGKANVLTF	TRBV6-1*01	TRBJ2-6*01	29
CASSAGLMNTEAFF	TRBV5-6*01	TRBJ1-1*01	28

Table 1. Top 15 most expanded TCR clones TCRs are rearranged decreasingly based on thei “Counts” column.

Visualizing the clone size distribution provides an intuitive snapshot of the repertoire’s entire structure. This plot reveals the balance between the vast number of rare clones (the “tail”) and the few highly expanded clones (the “head”) that often drive an immune response. A distribution heavily skewed to the left with a short tail signifies a diverse, polyclonal repertoire with many different rare T-cells. In contrast, a distribution with a long, heavy tail extending to the right is a clear visual indicator of an oligoclonal repertoire, dominated by large, expanded clones. This allows for a quick, qualitative assessment of a sample’s clonality and diversity, complementing more quantitative summary metrics. **By comparing all distributions, you can visually assess the overall impact of your experimental condition across the entire patient cohort.**

▶ Code

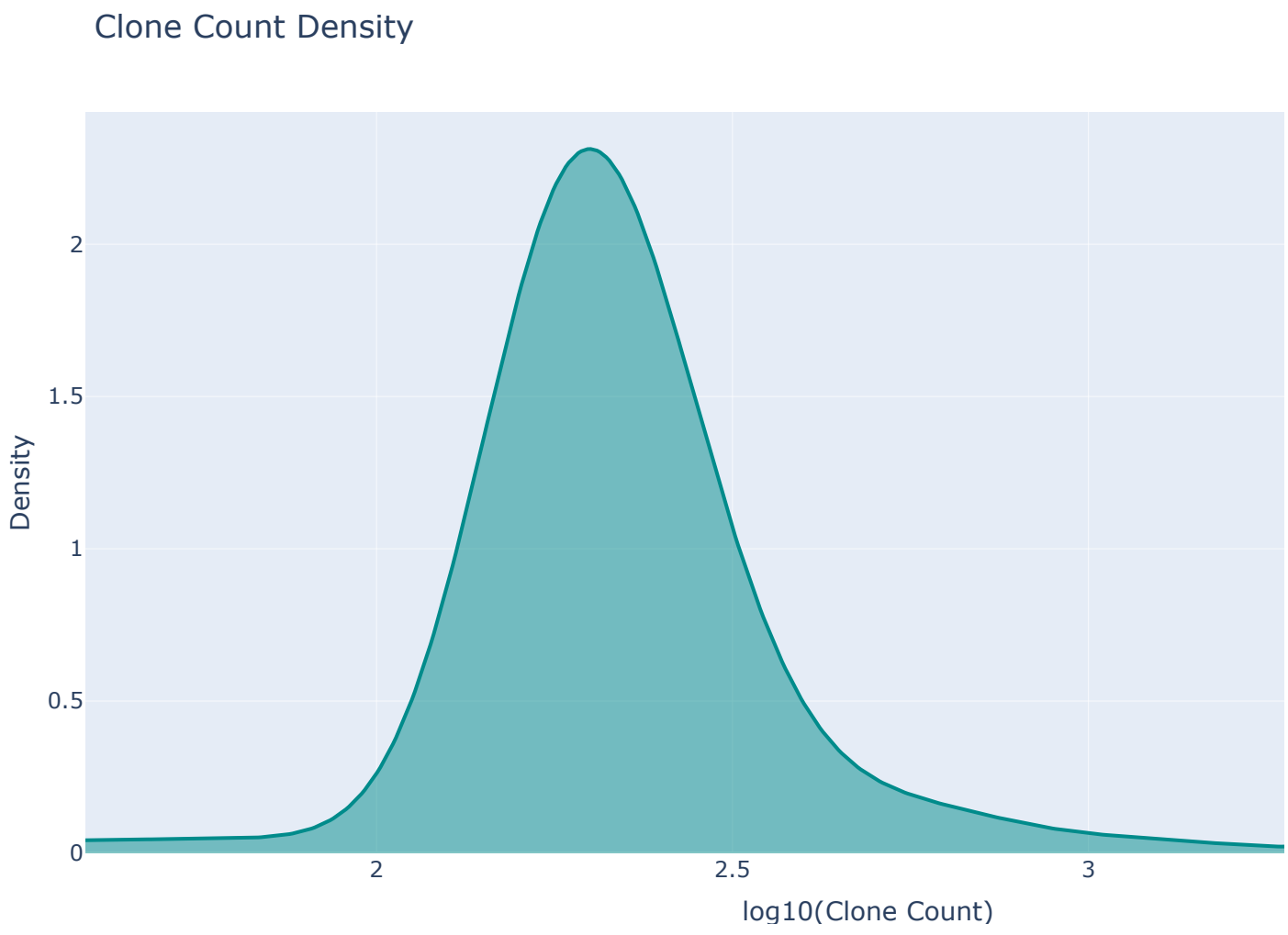


Figure 8. Clone Density Density plot showing the number of sequencing counts (x-axis) and the number of TCRs (y-axis) having them. A value around 0 in the x-axis, represets clones that are non-expanded.

4 Gene Family Usage

4.1 V gene family usage

V (Variable), D (Diversity), and J (Joining) genes are the genetic building blocks a T-cell uses to construct a unique T-cell receptor (TCR) through a "mix-and-match" process called V(D)J recombination. To create the immense diversity needed to recognize countless potential threats, each developing T-cell randomly selects and permanently stitches together one V, one D (for the beta chain only), and one J gene segment from a large genetic library. This process creates a single, unique TCR gene for that cell, with the most critical, hypervariable region known as the CDR3 being formed at the junction where these segments meet, which is the part that directly binds to antigens. 🧬

The V gene family usage of the TCRs in each sample is shown in the plots below. The x-axis shows the timepoint collected for each individual, and the y-axis shows the proportion of TCRs that use each V gene family. The V gene usage proportion, V_k, is calculated via:

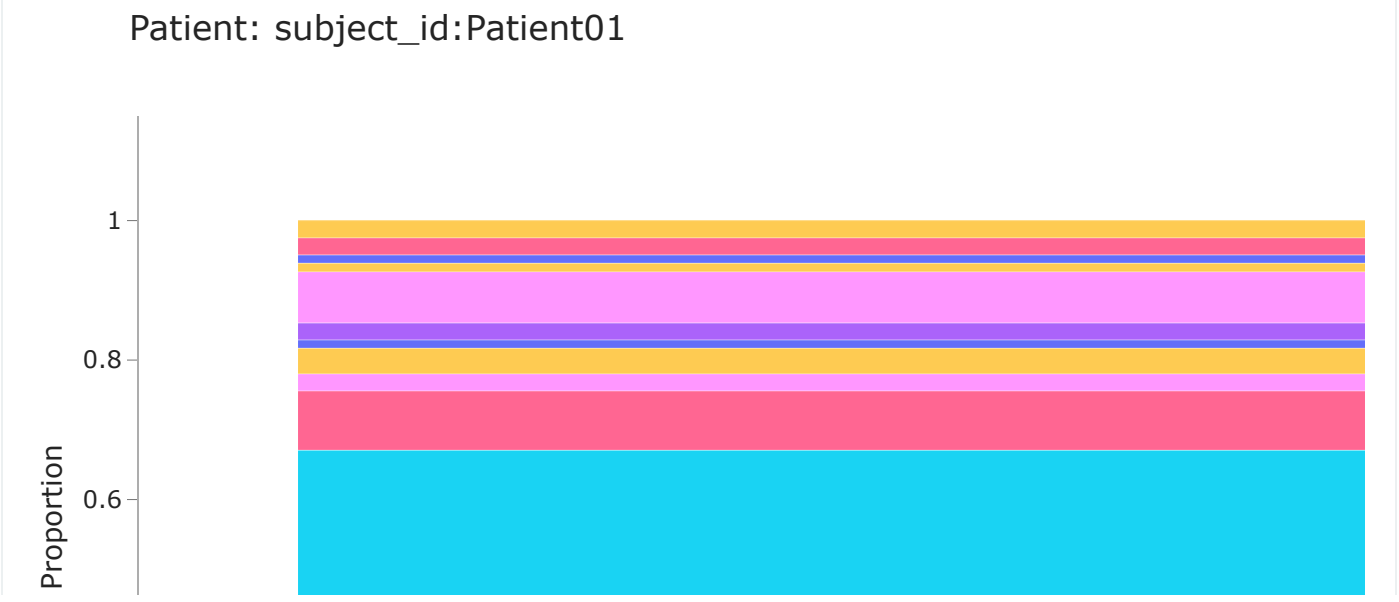
$$V_k = \frac{N_k}{T} \quad , \quad T = \sum_{i=1}^{N} p_i$$

where N_k is the number of TCRs that use the k th V gene, and T is the total number of TCRs in the sample.

▶ Code

▶ Code

Single Timepoint | Multiple Timepoints



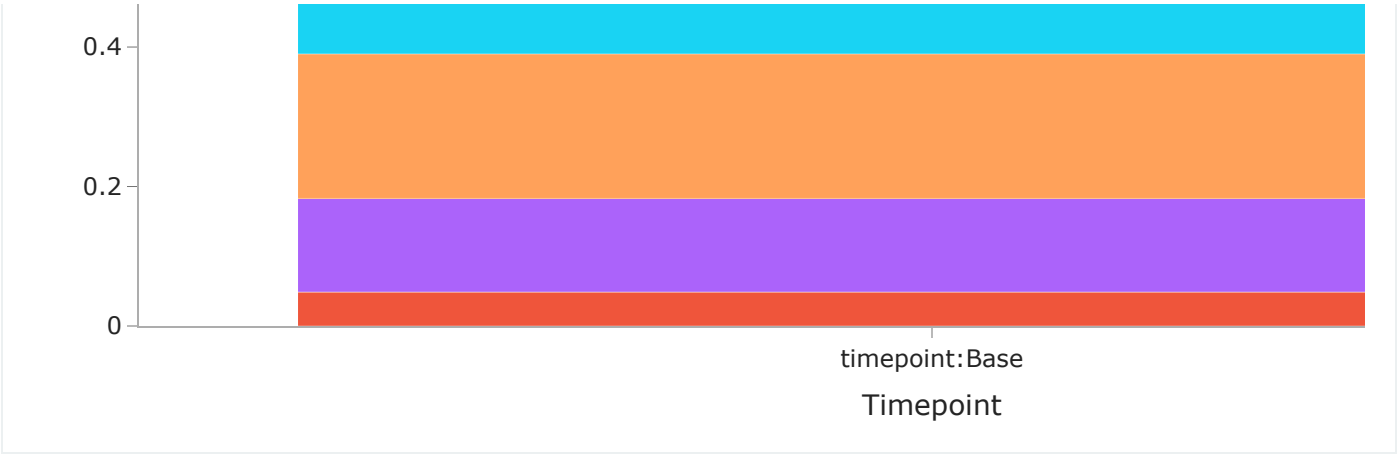


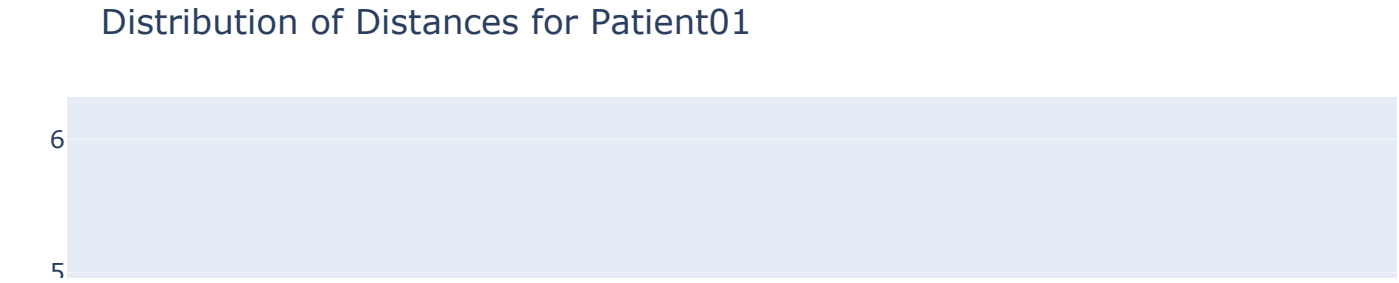
Figure 9. V/J gene usage Stacked barplot showing gene proportion (y-axis) at a given sample (x-axis). Samples cointaining multiple timepointes are shown side-by-side to facilitate comparisons.

Use this visualization to detect biased gene usage, which is a strong indicator of a significant, antigen-driven immune response. When T-cells responding to a specific antigen proliferate, the V and J genes they are built from will naturally increase in proportion to all other genes in the repertoire. Comparing the usage profiles between two timepoints is the most powerful application of this chart. Look for significant changes: A bar that dramatically increases in height from one timepoint to the next indicates that an immune response involving T-cells from that specific gene family has been initiated or has expanded.
Example scenario:
If you observe that the usage of the TRBV7-9 gene increases from 4% in a pre-treatment sample to 35% in a post-treatment sample, this provides strong evidence of a treatment-induced response driven by T-cells that utilize the TRBV7-9 gene segment.

5 TCRdist3

TCRdist is a **metric that quantifies the similarity between any two T-cell receptors (TCRs) by calculating a biochemically-aware distance based on their amino acid sequences**. Instead of just checking if two TCR sequences are identical, **TCRdist focuses on the key regions that bind to antigens: Complementarity-Determining Regions (CDRs).** It compares the amino acid sequences of the CDR1, CDR2 (obtained from the V gene), and especially the hypervariable CDR3 loops of two TCRs. The “distance” is calculated using a substitution matrix (like BLOSUM62) that scores how similar two different amino acids are in their biochemical properties. Swapping two functionally similar amino acids results in a small distance penalty, while swapping two very different ones results in a large penalty. The total distance is a weighted sum of these scores, with the CDR3 distance contributing the most.

▶ Code



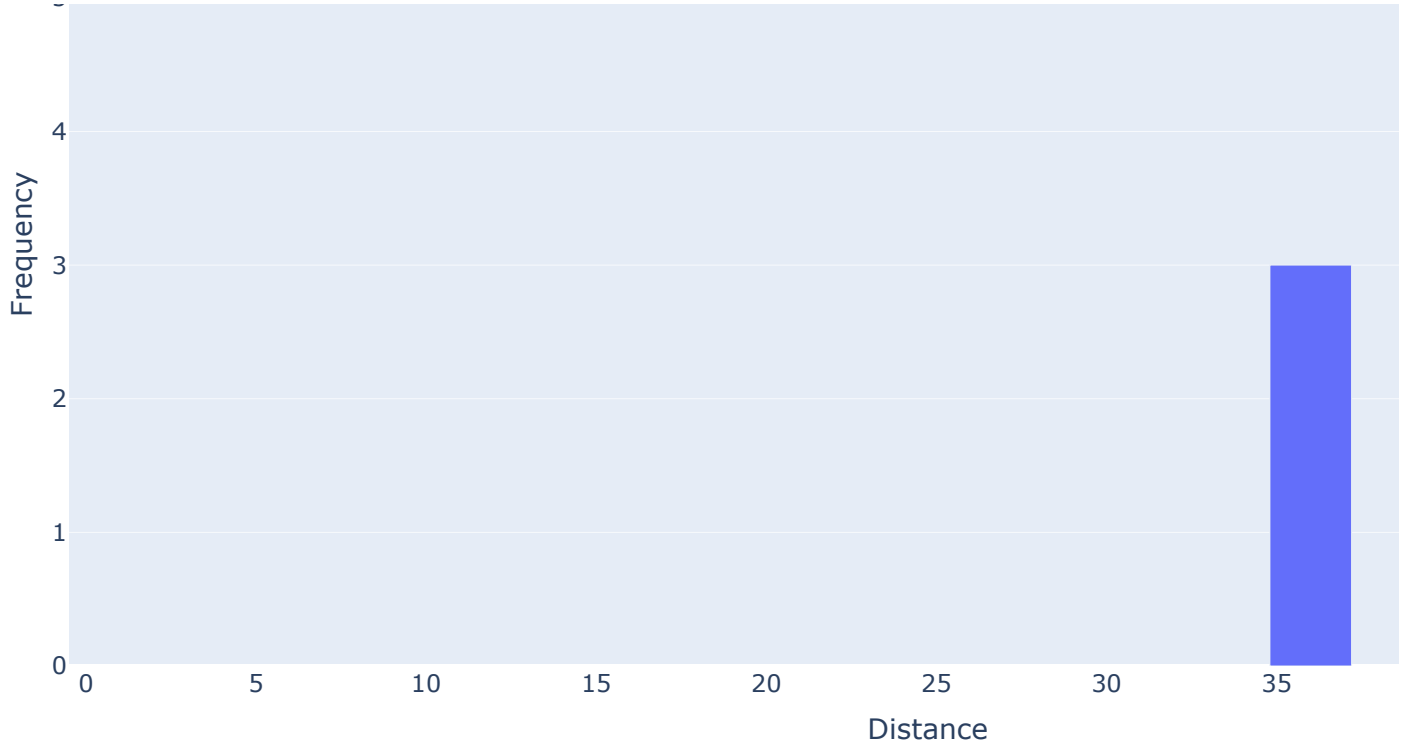


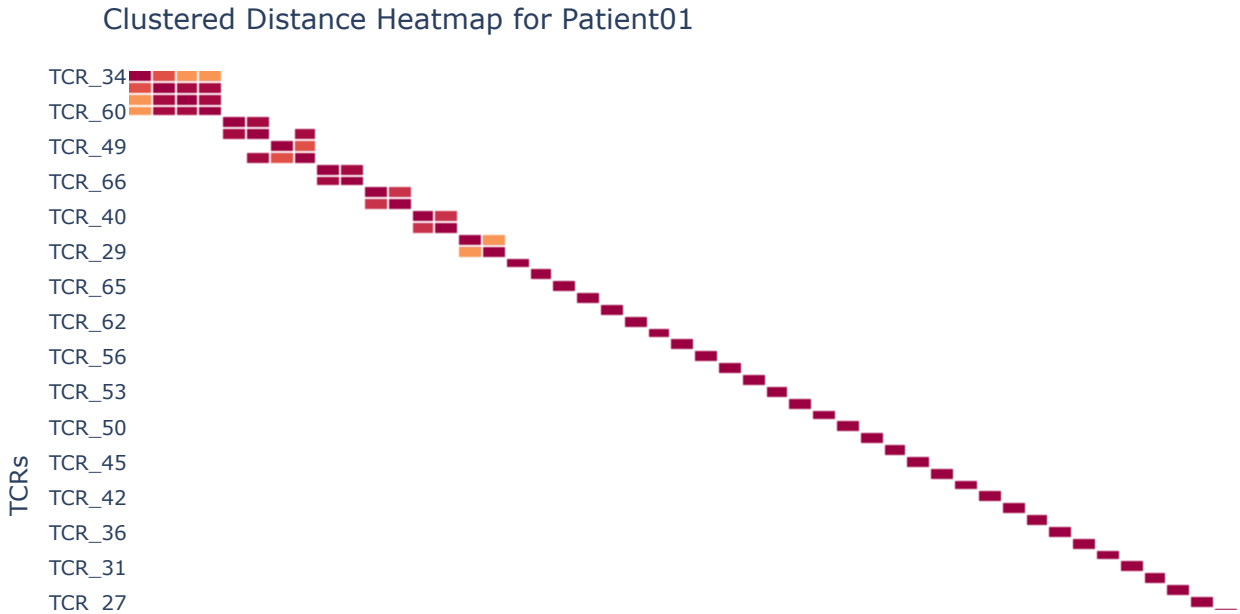
Figure 10. Sample Distances Barplot showing biochemical distances between sample's TCRs

Why This Metric Is Helpful? 🌐

The primary benefit of TCRdist is its ability to identify groups of T-cells that are likely to recognize the same antigen, even if their TCRs are not identical. This phenomenon is known as cross-reactivity. By calculating the distance between all TCRs in your sample, you can move beyond analyzing single, identical clones and instead identify functional "neighborhoods" of similar TCRs. This provides a much more comprehensive and biologically accurate picture of an immune response. It allows you to group related T-cells together to see the full breadth of the response to a specific antigen, rather than just the single most expanded clone.

▶ Code

```
Skipping Patient16: No valid distance data.
Skipping Patient25: No valid distance data.
```



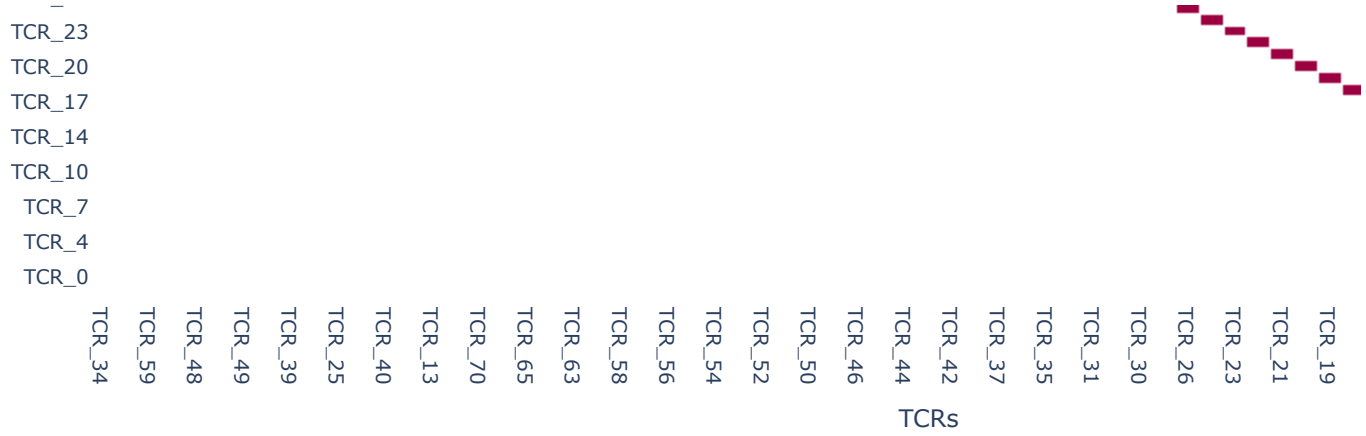


Figure 11. Distances Heatmap showing distance between all sample's TCRs to help identify related cones within a sample.

This heatmap visualizes the pairwise TCRdist for every T-cell receptor (TCR) within a sample, creating a comprehensive map of the repertoire's similarity landscape. Darker colors represent a smaller distance (higher similarity), and **the key features to look for are the square blocks** of dark color off the main diagonal. **These blocks identify "neighborhoods" or clusters of TCRs that are biochemically similar and likely recognize the same antige**n, providing a more complete picture of the immune response than analyzing single clones in isolation. It allows you to visually confirm the presence and composition of these functionally related TCR groups within your data.

▶ Code

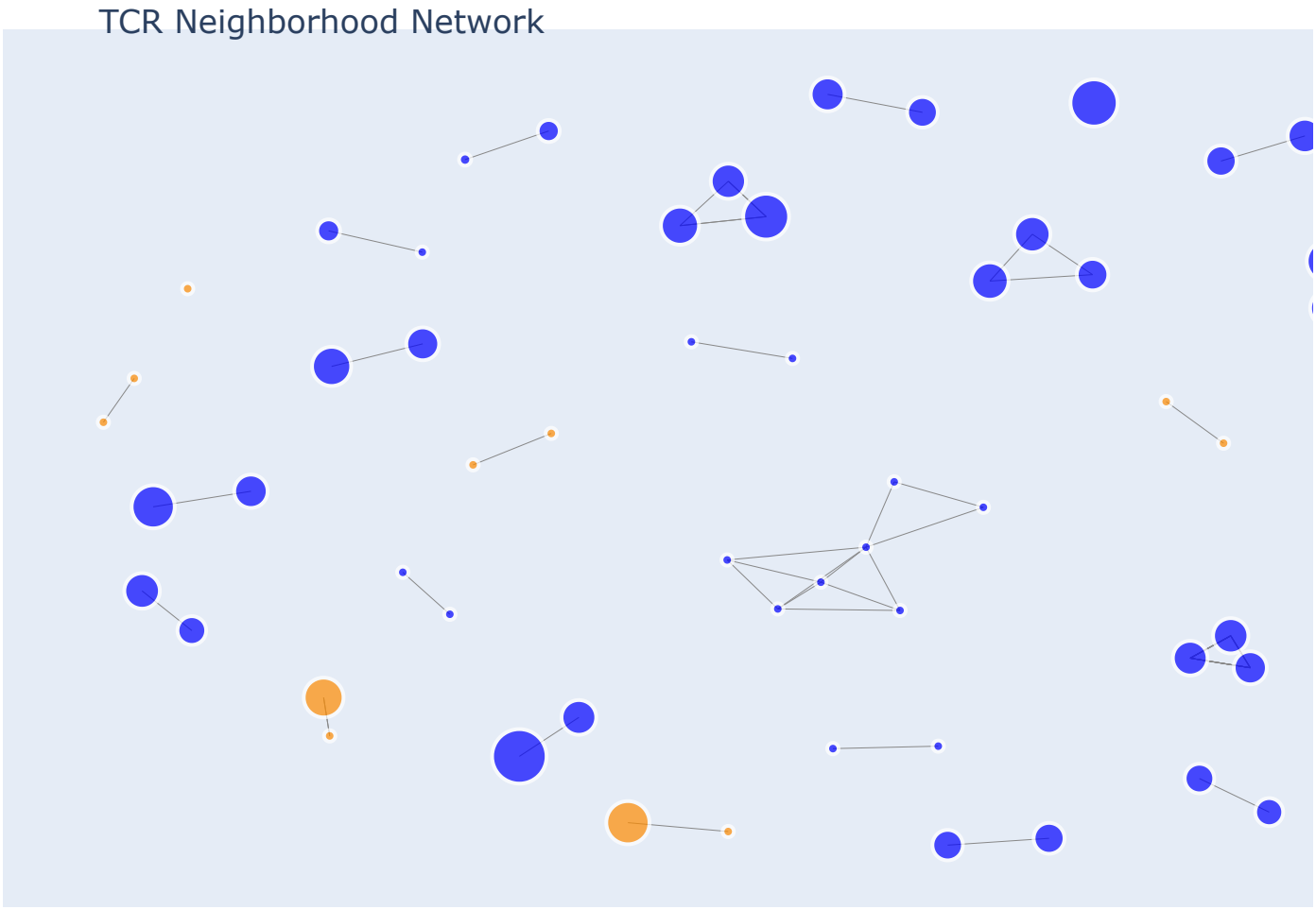


Figure 12. TCRdist Network Graph showing all samples for which a proper distance was calculated across all samples. Node size is proportioanl to clone “Counts” and color varies depending on wether the clone is present in “Pre-”, “Post” or both conditions.

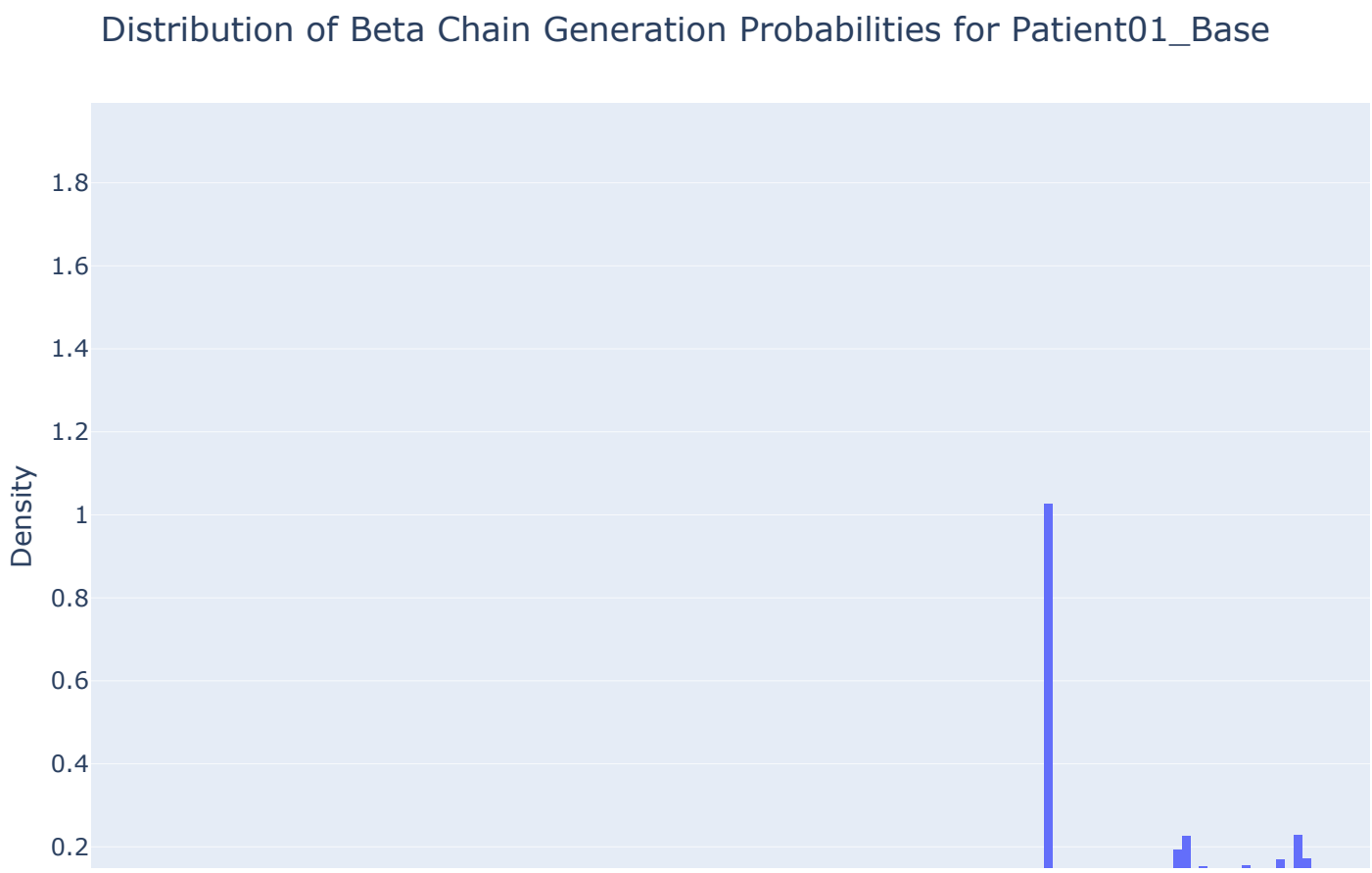
This network graph provides a **global view of TCR relationships across all samples, integrating multiple layers of data.** Each node is a unique TCR clone, its size reflects its abundance, and its color indicates the sample condition where it was found. A line, or edge, connects any two TCRs that are biochemically similar (i.e., have a TCRdist below a set threshold), forming connected clusters.

The primary utility of this plot is to visualize these “TCR neighborhoods” across the entire dataset, allowing you to see how functionally related clones are shared between conditions or form a response to a treatment, such as a large cluster of clones appearing specifically in the “Post” condition.

6 TCR generation probabilities

Calculating these probabilities in bulk TCR-seq data is valuable because **it provides a baseline understanding of the “expected” frequency of each TCR**. By comparing observed TCR frequencies against their generation probabilities, it is easier to identify TCRs that are “overrepresented” (clonally expanded) due to an immune response (e.g., to infection or cancer) rather than just being common due to biases in the generation process itself. This helps distinguish antigen-driven selection from inherent generation biases, offering clearer insights into immune repertoire dynamics and responses.

▶ Code



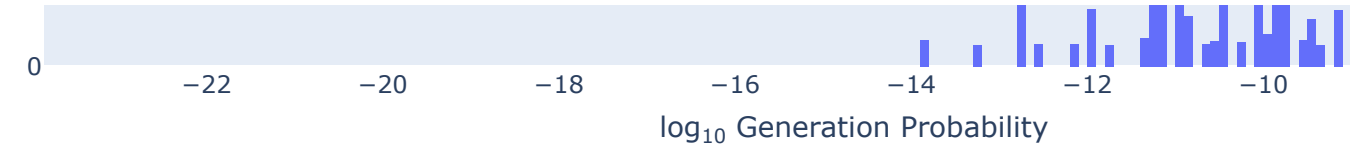


Figure. 13 TCR Generation Probabilities Probabilities shown above are weighted, accounting for the counts per clone within a repertoire.

7 TCRPheno: From TCR sequence to Phenotype

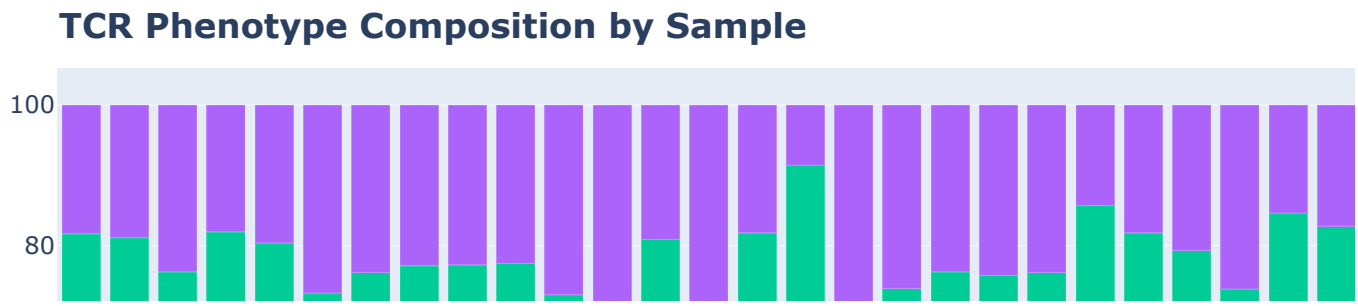
TCRpheno is a machine learning model that predicts the likely functional phenotype of a T-cell based solely on its TCR sequence. This is powerful because it allows you to infer the roles of different T-cells (e.g., killer, regulator, memory) from your bulk sequencing data without needing corresponding cell surface marker information.

How to Interpret the Scores? 🔍
The output from TCRpheno is a set of scores, not a definitive classification. For each TCR, a higher score in a category indicates a higher probability that the T-cell belongs to that functional group.

- **TCRbeta.CD8:** A high score here suggests the TCR likely belongs to a cytotoxic CD8+ T-cell. These are the **"killers" of the adaptive immune system**, responsible for destroying virus-infected cells and tumor cells. A repertoire with many high-scoring CD8 TCRs indicates a strong, active anti-pathogen or anti-tumor response.
- **TCRbeta.reg:** This score corresponds to regulatory T-cells (Tregs). These are the **"peacekeepers" that suppress other immune cells** to prevent autoimmune reactions and maintain tolerance. A high Treg score for a TCR suggests it plays a role in immune suppression.
- **TCRbeta.mem:** This indicates a likely memory T-cell phenotype. These are the l**ong-lived "veterans" of the immune system that persist after an infection is cleared**, providing rapid protection upon re-exposure to the same pathogen.
- **TCRbeta.innate:** A high score in this category suggests the TCR may belong to an innate-like T-cell, such as a MAIT or NKT cell. These cells bridge the gap between the innate and adaptive immune systems, **acting as rapid first responders to certain types of threats.**

It's important to remember that a single TCR can have moderate scores in multiple categories, reflecting the **potential for cellular plasticity** or shared sequence features between different T-cell types. **The phenotype with the highest score is considered the most likely identity for that TCR.**

▶ Code



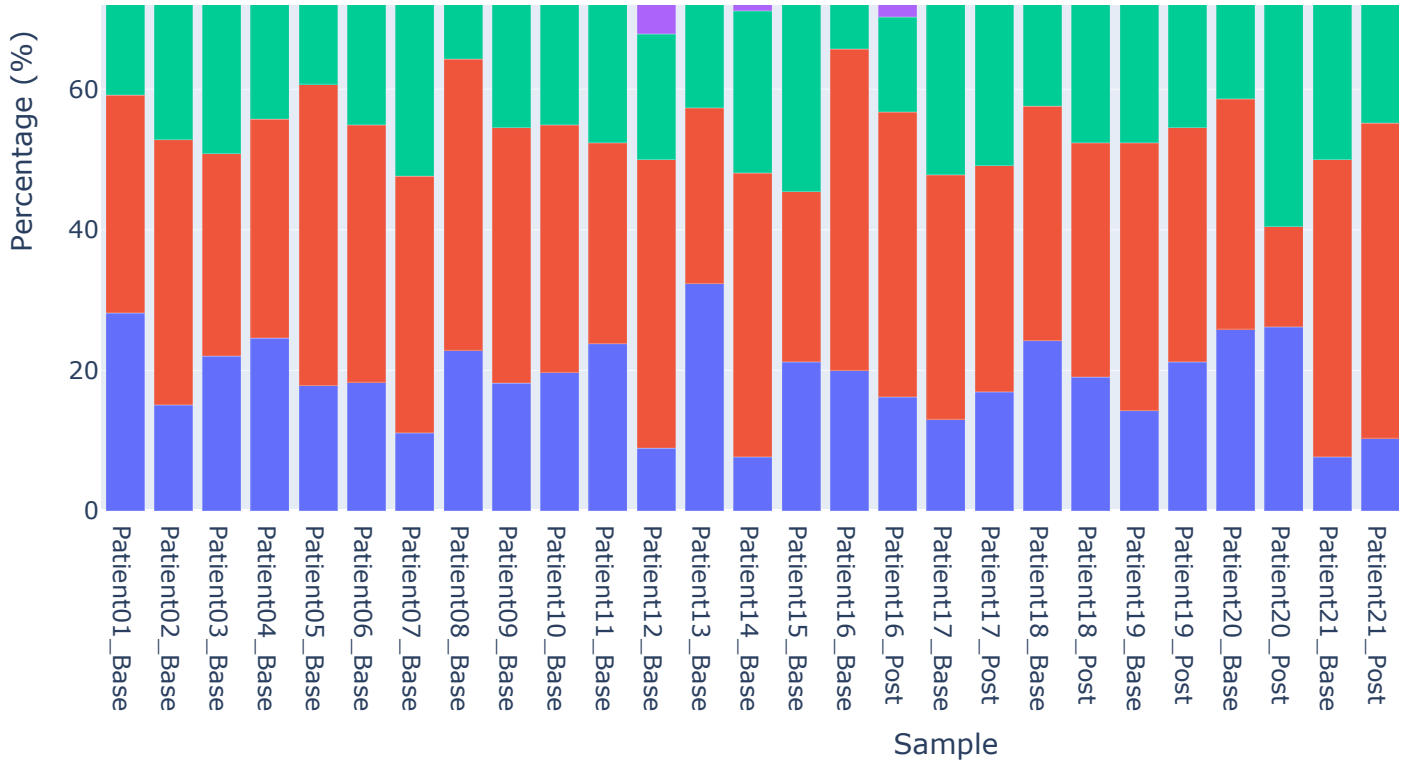
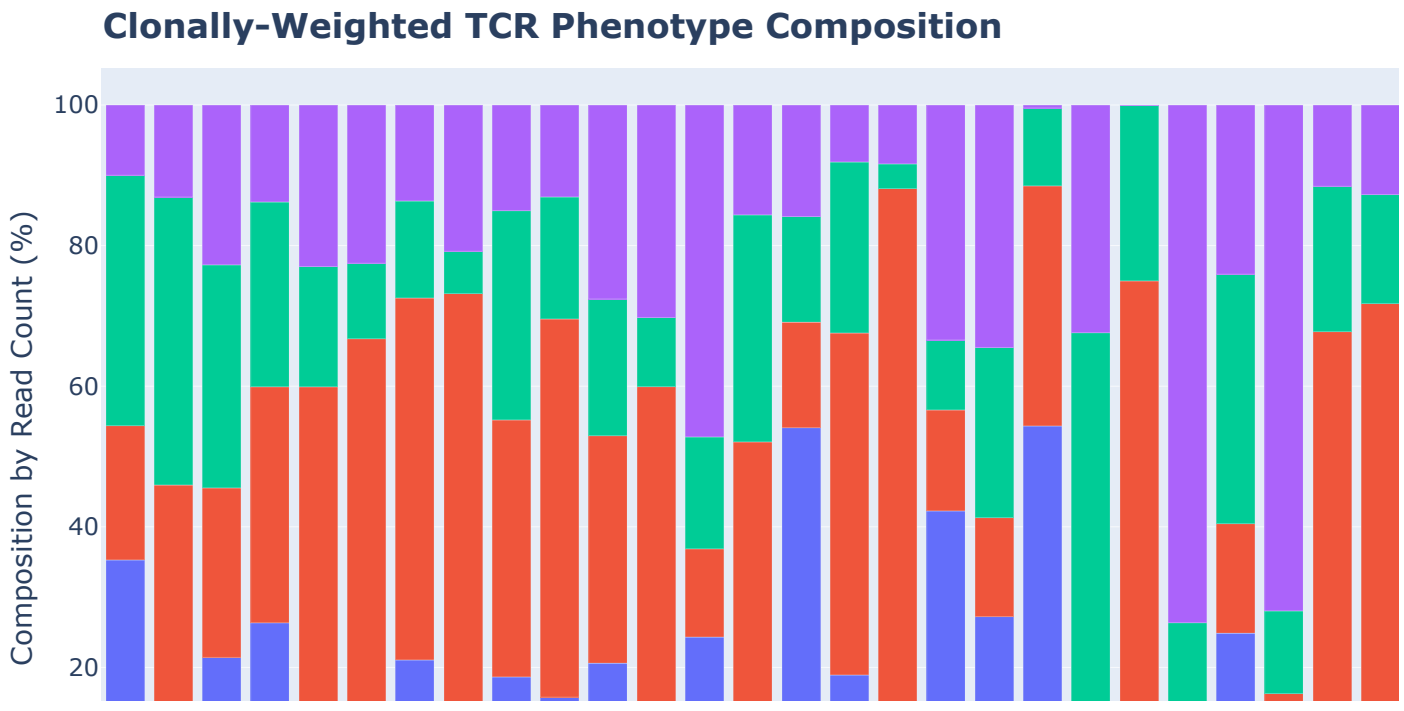


Figure. 14 Composition of Unique TCR Clonotypes by Predicted Phenotype Stacked bar chart showing the proportional diversity of predicted T-cell phenotypes (memory, regulatory, CD8, or innate-like) for each sample, independent of clonal expansion. Each bar indicates the percentage of clones assigned to each category based on their highest TCRpheno score.

This plot answers the question: *"Of all the different types of T-cell soldiers available, what is the breakdown of their specialties?"*

It reflects the underlying potential of the repertoire. Comparing "Base" vs. "Post" samples can reveal if a treatment induced the emergence of many new and different types of T-cells with a certain phenotype.

▶ Code



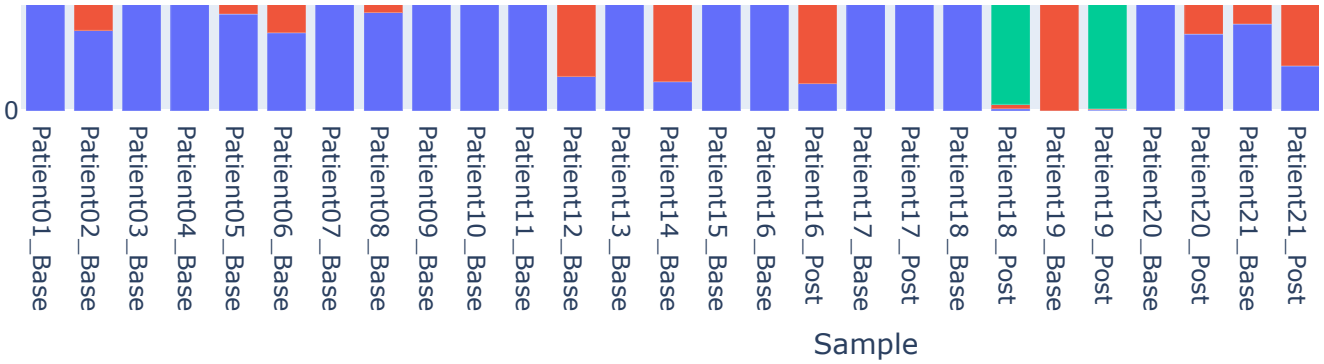


Figure 15: Clonally-Weighted TCR Phenotype Composition Stacked bar chart showing the proportional abundance of predicted T-cell phenotypes for each sample, weighted by clonal size. The size of each segment is determined by summing the read counts of all TCRs assigned to that functional category, thus directly reflecting clonal expansion.

This plot answers the question: *"Of all the T-cell soldiers currently fighting on the battlefield, which specialty is dominating by sheer numbers?"*

Higher Proportion of mem (Memory) T-cells 🛡️
This indicates the individual has a robust history of past immune responses, likely from resolved infections or successful vaccinations. These "veteran" cells are not actively fighting a major battle but are circulating long-term to provide rapid protection if a known pathogen reappears.

HHigher Proportion of reg (Regulatory) T-cells 🕊️
A high proportion of regulatory T-cells (Tregs) points to an immunosuppressive state. These "peacekeeper" cells work by dampening the activity of other immune cells. An abnormally high proportion can be detrimental in other contexts, such as cancer, where Tregs can be co-opted by tumors to protect themselves from being attacked by the patient's own immune system.

HHigher Proportion of CD8 T-cells ⚔️ A repertoire dominated by CD8-predicted T-cells is the clearest sign of an active, ongoing cytotoxic immune response. These "killer" T-cells are the primary soldiers responsible for identifying and destroying virus-infected cells and tumor cells.

▶ Code