TCR Sharing - Notebook

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

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/
```

2 Analysis

3 TCR Sharing / TCR Publicity

TCR publicity refers to how widely a specific T-cell receptor (TCR) sequence is shared among different individuals. In simple terms, it's a **measure of whether a TCR is unique to one person or common across a population.**

Private vs. Public TCRs 👥

Private TCRs: These receptor sequences are found in only one individual. The vast majority of a person's TCR repertoire is private, reflecting their unique genetic background and history of antigen exposure.

Public TCRs: These are identical TCR sequences found in multiple, unrelated individuals. **While much rarer than private TCRs, these shared receptors are often of great biological importance.**

▶ Code

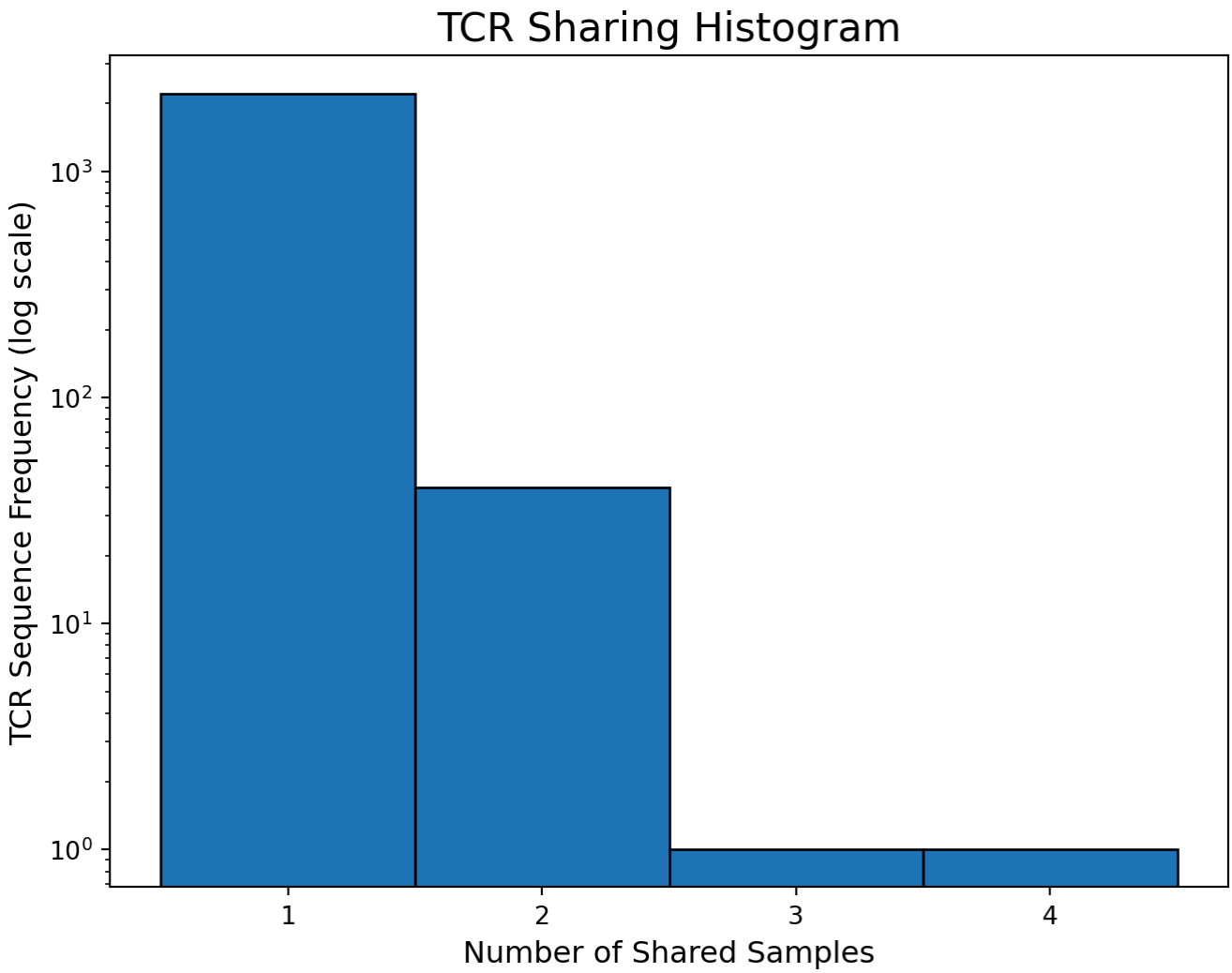


Figure 1. TCR sharing histogram. Number of samples where a TCR has an exact match on the aminoacid level.

This histogram gives you an immediate sense of the immune landscape. A cohort exposed to a common antigen (like a widespread virus or a shared environmental factor) may show a larger-than-expected public repertoire (higher bars for x > 1). **It helps answer the general question: "How similar are the T-cell repertoires among the individuals in this study?"**

▶ Code

Top 10 Most Shared TCR Clones

CDR3b Sequence	# Shared Samples	Generation Probability
CASSYVGNTGELFF	4	2.71e-08
CASSSQETQYF	3	1.16e-06
CASSLAQGANSPLHF	2	1.84e-08
CSAIPQGSYEQYF	2	4.82e-09
CASSYGTDYEQYF	2	7.85e-08
CASSFERGGDTQYF	2	4.64e-09

CASSQDYGQGVGNTIYF	2	1.01e-10
CASSLGDEQYF	2	2.10e-06
CASSLILLGNTEAFF	2	1.59e-09
CAIPGQGAYEQYF	2	2.67e-08

Table 1. Top 10 most shared TCR clones Public TCRs order by number of samples they are seen.

Why TCR Publicity is Important?

The primary reason public TCRs are important is convergent selection. The generation of TCRs is a random process, and the potential diversity is astronomical. Therefore, **if the exact same TCR sequence appears in many different people, it's highly unlikely to be by chance. Instead, it strongly implies that different immune systems, when faced with the same threat** (like a virus or cancer cell), have independently "converged" on the same optimal TCR as an effective weapon.

▶ Code

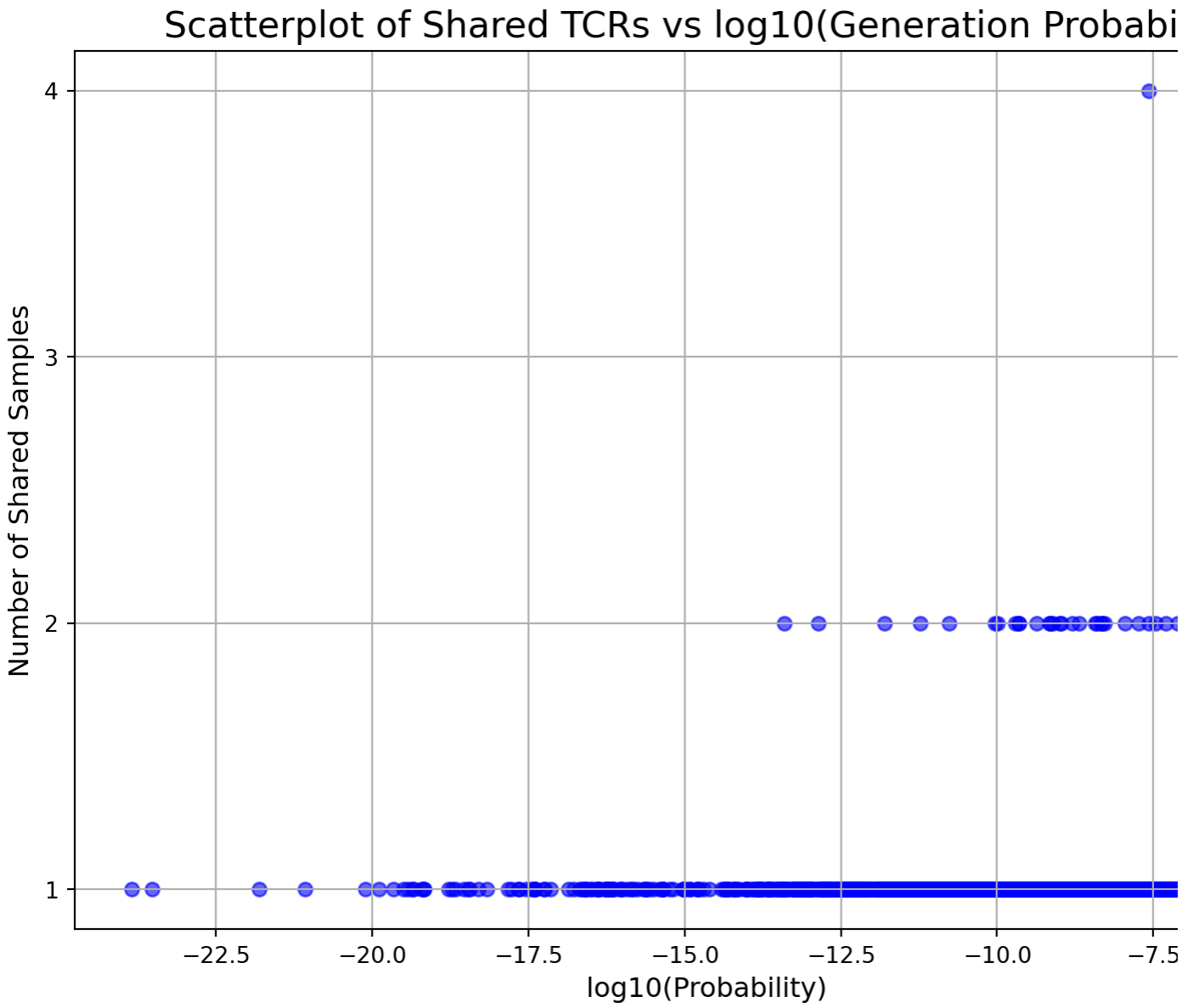


Figure 3. TCR sharing and generation probability correlation Relationship between the likelihood of a T-cell receptor (TCR) being generated (generation probability, Pgen) and the number of samples it is shared across.

How to Interpret the scatter plot?

The General Trend: You will notice that points on the right (higher Pgen, "easier-to-make" TCRs) tend to have higher y-values (more sharing). This is expected; **if a TCR sequence is easy to generate, it's more likely to appear in multiple people just by chance.**

The Key Insight (Upper-Left Quadrant): The most biologically significant points are those that defy this trend. Pay close attention to the TCRs in the upper-left area of the plot. These are **TCRs with a very low generation probability (they are rare and hard to make) that are nevertheless shared across multiple samples.**

3.1 GLIPH2: TCRB CDR3 motif clusters visualization

GLIPH2 (Grouping of Lymphocyte Interactions by Paratope Hotspots, version 2) is a computational tool used in TCR sequence analysis to **identify clusters of TCRs that likely recognize the same or similar antigens**.

By grouping TCRs with shared sequence motifs, particularly in the CDR3 regions (the main antigen-binding region), GLIPH2 helps infer functional relationships and **predict which TCRs might target the same or related epitopes**.

For the current stage of analysis, we are focusing solely on the **CDR3 region of the TCR beta chain (TCRB)**. a network was created to facilitate result visualization.

GLIPH2 classifies specificity groups in two categories. The color of each cluster tells you whether the shared motif is **local** (motif-based), meaning that motifs' position within CDR3 are restricted within 3 amino acids. Or **global**, where member CDR3s need to be of the same length, and differ at the same position.

▶ Code

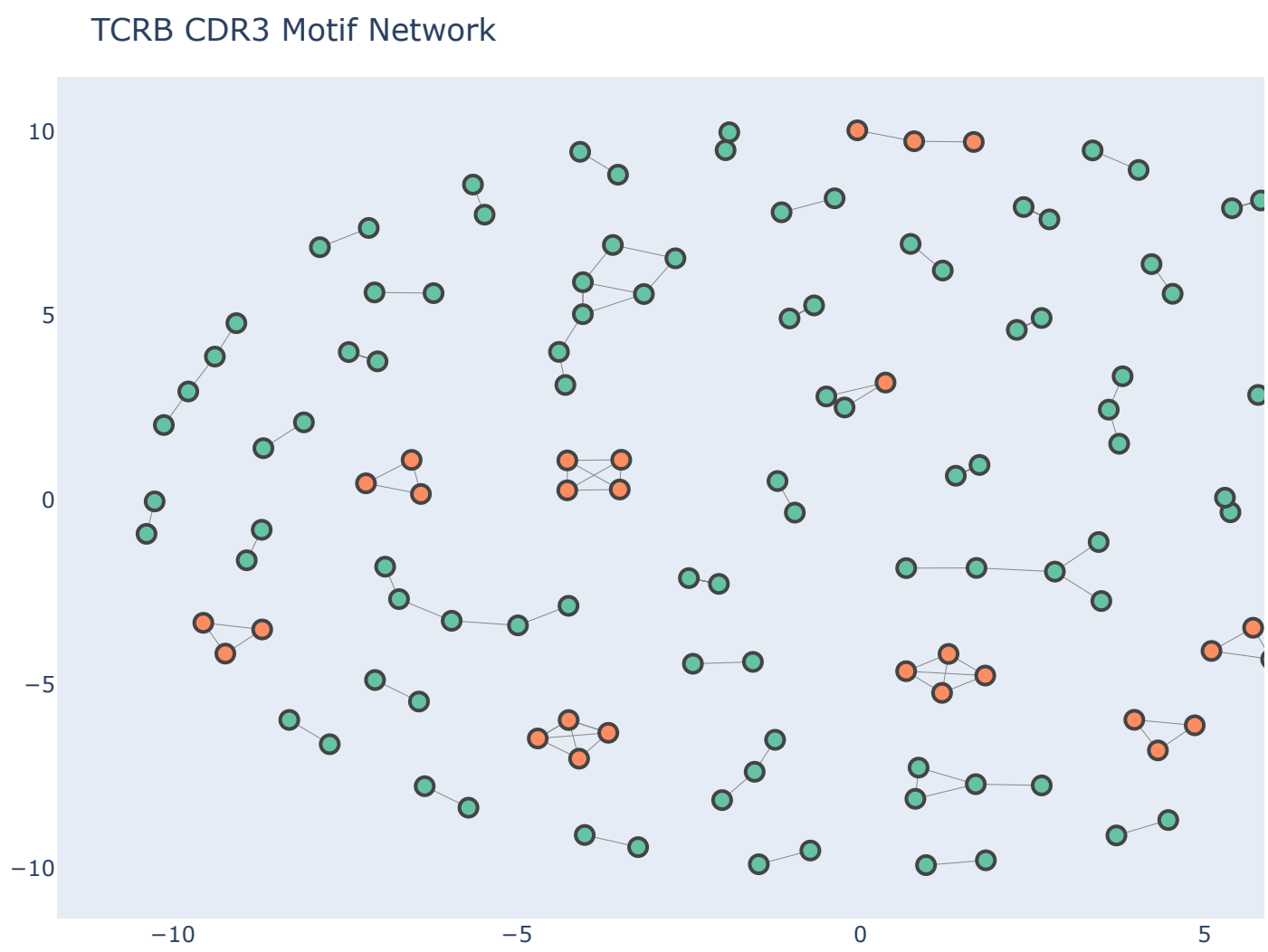


Figure 4. Specificity groups across patients Each circle (node) represents a unique TCR clonotype across all data. Nodes are connectionected if they share a common sequence motif, placing them into the same specificity group.

How to Extract Biological Knowledge from the Figure?

1. Are there large, dense clusters? A large cluster containing many connected nodes signifies a convergent immune response. This means that multiple different T-cell lineages have independently evolved to recognize the same target, implying a strong and focused immune pressure against a specific antigen.

2. Are the most prominent clusters driven by Local or Global motifs?

3. What are the specific motifs in the key clusters? Hover over the nodes within a large, local (orange) cluster to identify the shared sequence motif (e.g., CASSINQPQHF). You can take this motif sequence and:

▶ Code

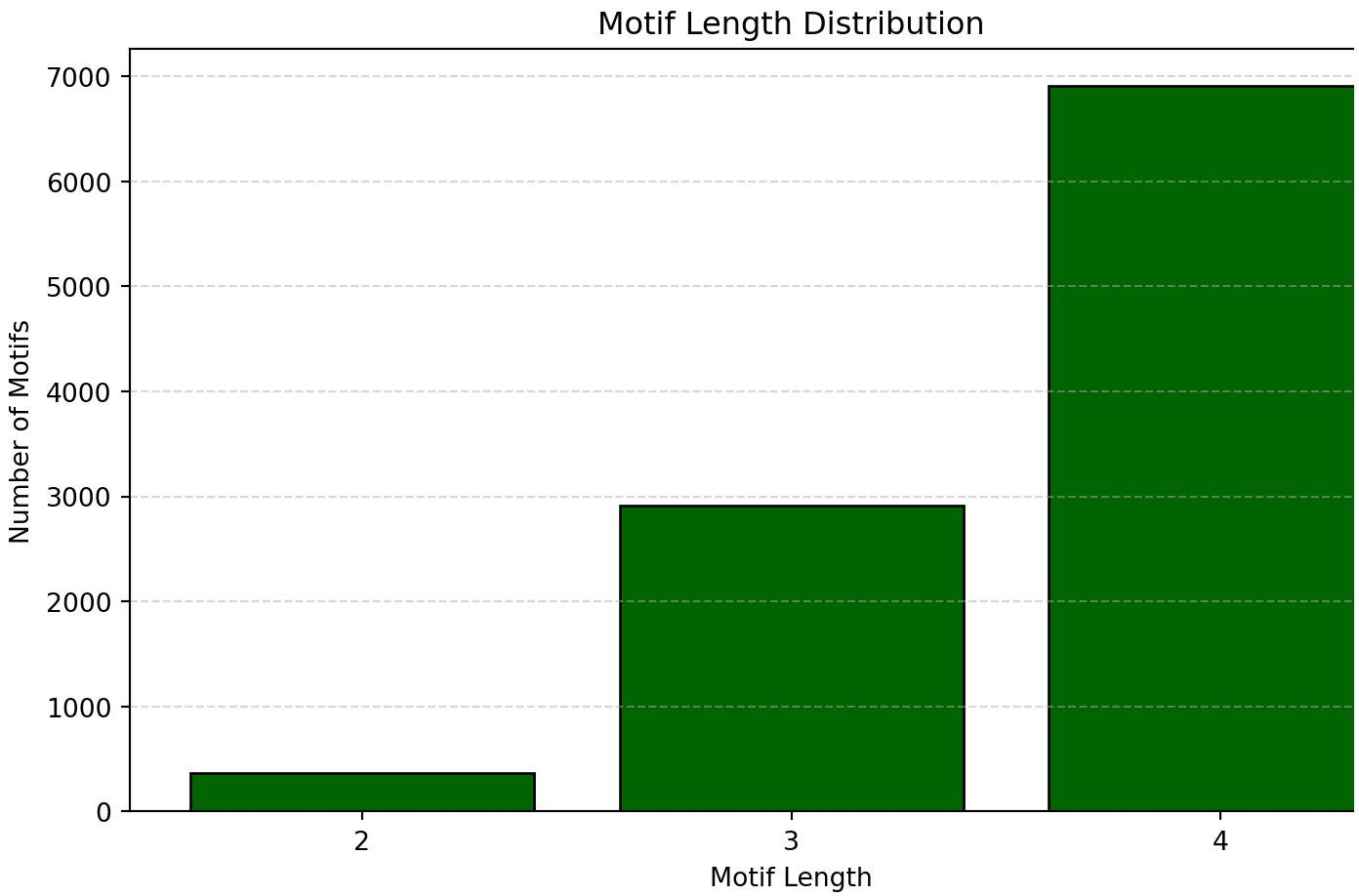


Figure 5. TCRB CDR3 motif lengths

Knowing the length of the amino acid motifs is useful because it provides insight into both the biological plausibility of the findings and the likely functional relationship between the clustered T-cells.

Reflecting the Biophysics of Binding

From a biological standpoint, **the interaction between a TCR and its target peptide-MHC is often driven by a small number of critical "contact points" or "hotspots" within the CDR3 loop.** These hotspots typically consist of a short stretch of 2-5 amino acids. When GLIPH2 identifies a large number of **motifs with a length of 3 or 4**, it serves as a crucial sanity check. It suggests the algorithm **is not just finding random statistical noise, but is successfully identifying patterns that are biologically plausible** and reflect the known mechanics of T-cell recognition.

Short Motifs (2-3 amino acids): These are like a broad. They are **more likely to occur by chance** and may define more "promiscuous" groups of TCRs that share a common structural feature but might still bind to a range of similar peptides.

Long Motifs (4-5 amino acids): These are like a specific. They are statistically much rarer and impose a much stronger constraint on the TCR's structure. Therefore, a group of TCRs defined **by a longer motif is very** likely to be highly specific for the exact same antigen.

▶ Code

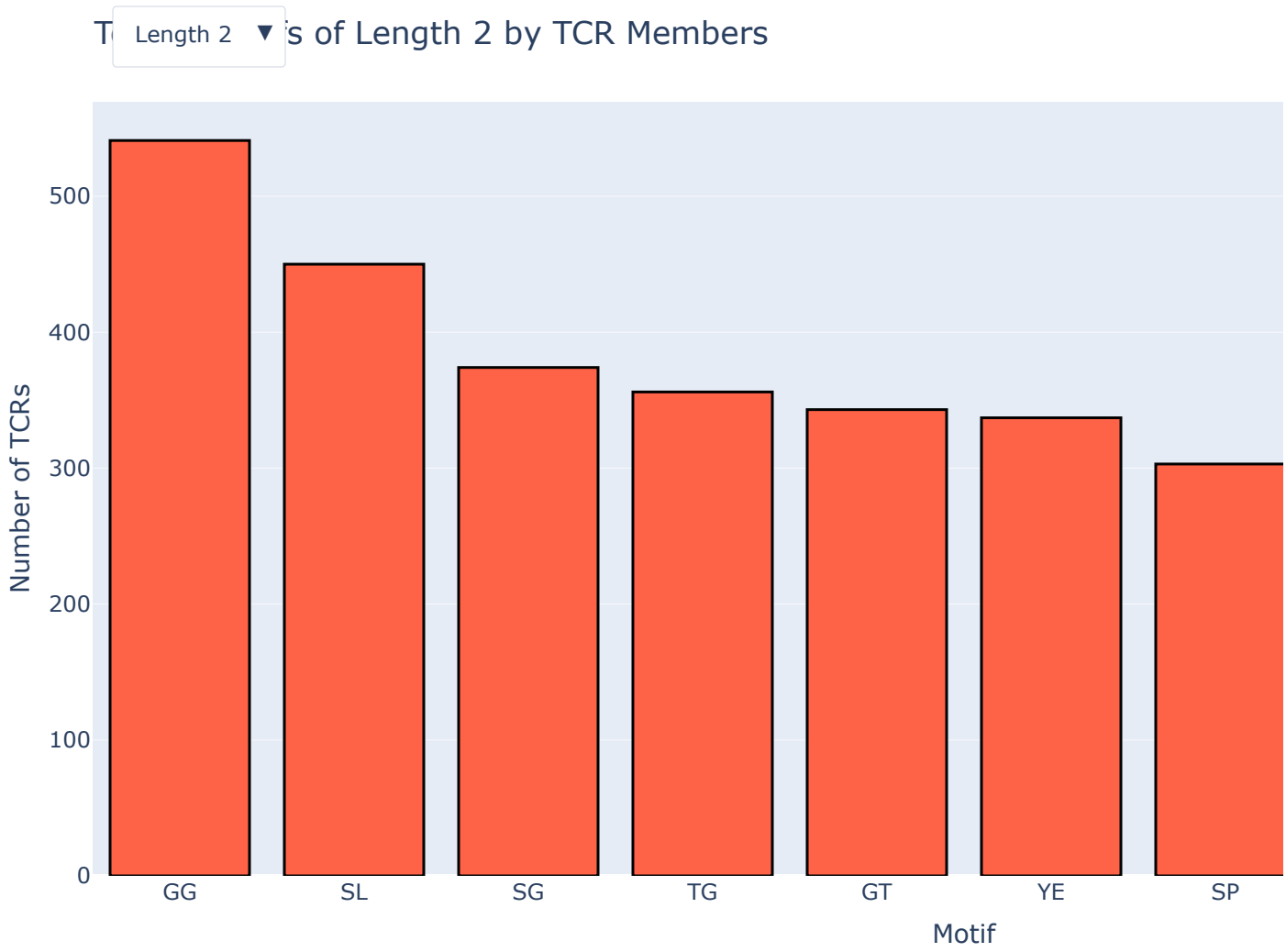


Figure 1

Figure 6. Top 10 TCR specificity motifs ranked by member count. Bar chart displaying the most prevalent amino acid motifs identified by the GLIPH2 algorithm, ranked by the number of unique T-cell receptors (TCRs) that share each motif.

This bar chart ranks the top conserved amino acid motifs found by GLIPH2 based on the number of unique T-cell receptors (TCRs) that belong to each specificity group. Each bar represents a specific motif (e.g., 'GG'), and its height indicates the size of that "TCR neighborhood"—the total number of different TCRs that share this core binding pattern.

Identifying the motifs shared by the largest number of unique TCRs helps **pinpoint the most immunodominant recognition strategies used by the T-cells** in your dataset; you are effectively highlighting the **most common solutions the immune system has evolved to target key antigens**, thus providing a data-driven way to prioritize the most significant TCR communities for further study. However, it is crucial to interpret these findings with care, as **the most frequent motifs are often short and may not always define the most functionally specific T-cell groups compared to smaller groups with longer, more complex motifs.**

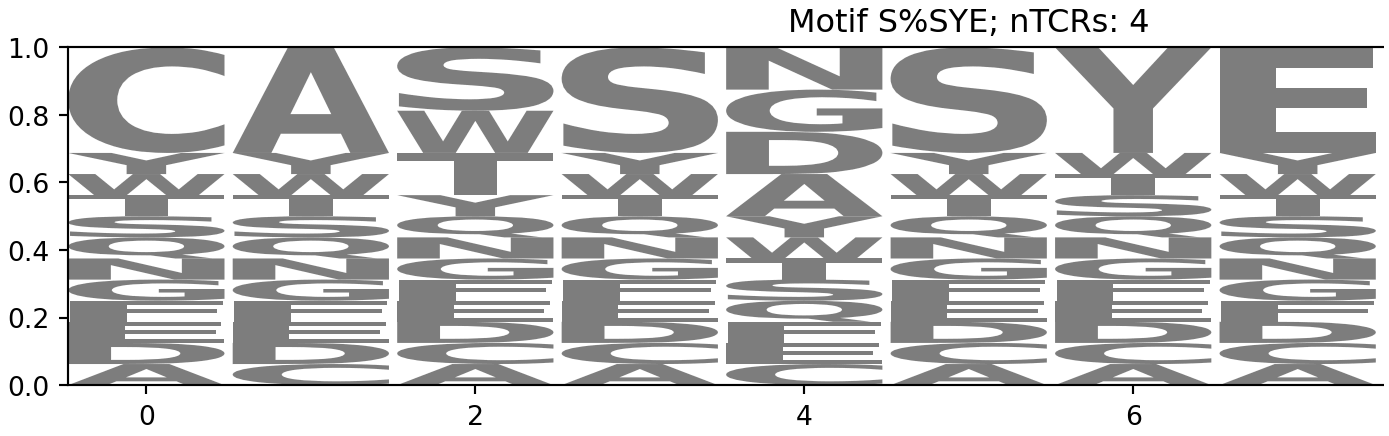
3.2 GLIPH2 global Motifs

GLIPH2 identifies TCRB CDR3 clusters based on either local motifs (short, position-restricted patterns within 3 amino acids) or **global similarity (requiring identical CDR3 length and differences at the same positions).** Visualizing global motifs as sequence logos reveals the conserved amino acid preferences at each position, and can provide insights into the sequence features that potentially drive shared antigen recognition.

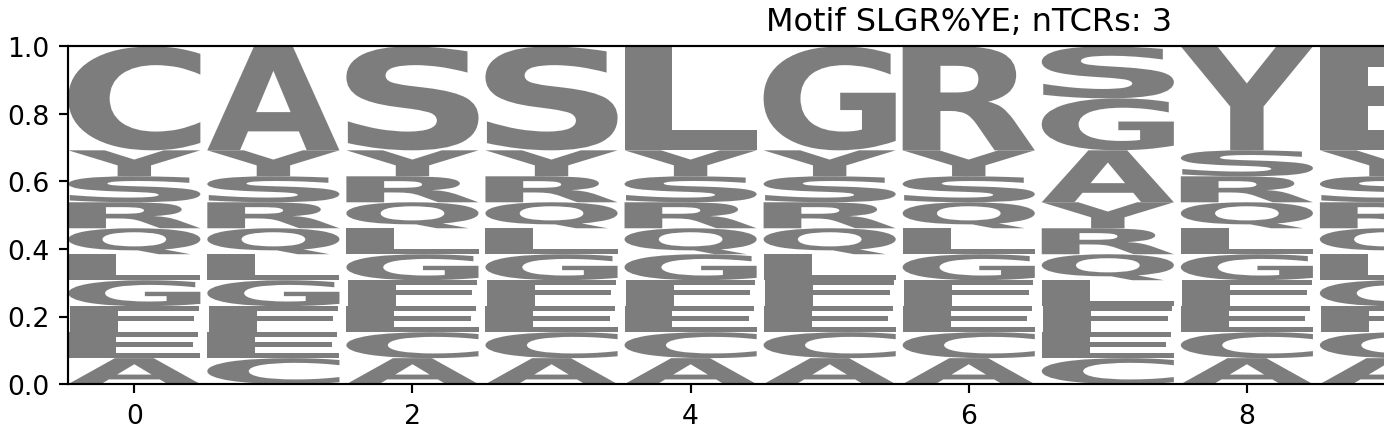
Logo visualization of the top 10 global motifs based on no. TCR members

▶ Code

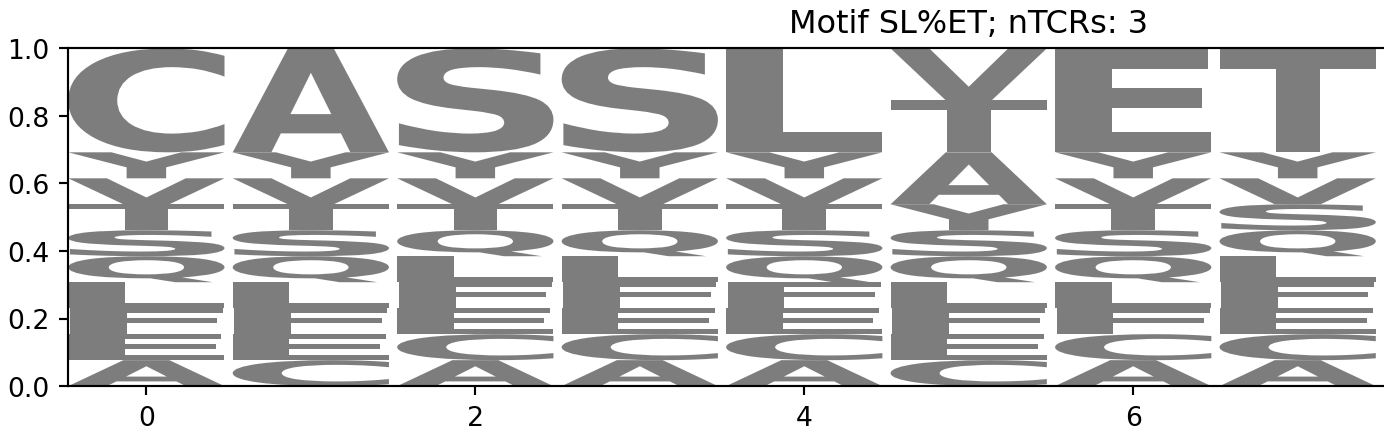
```
<Figure size 960x288 with 0 Axes>
```



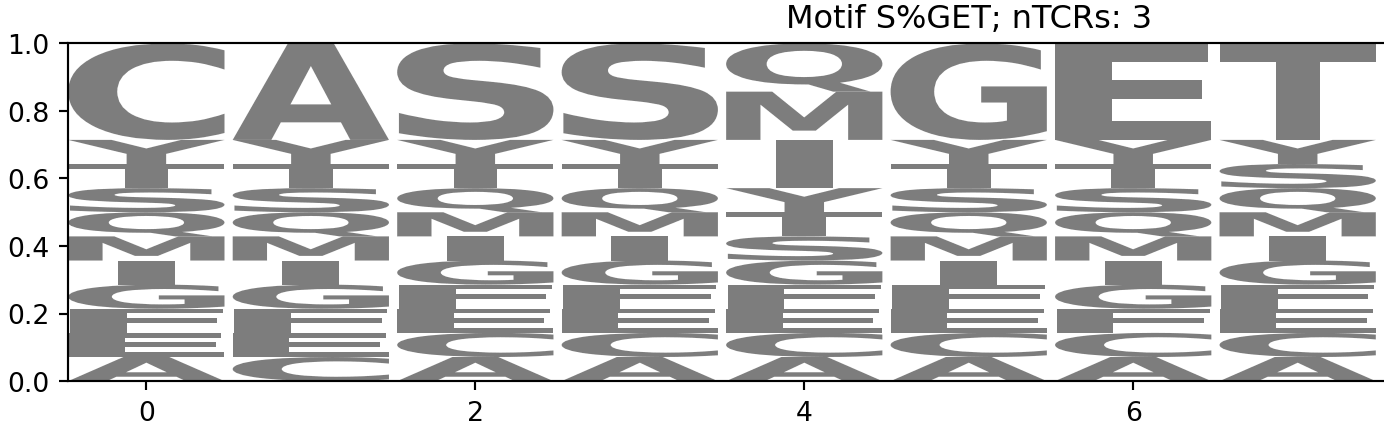

```
<Figure size 960x288 with 0 Axes>
```



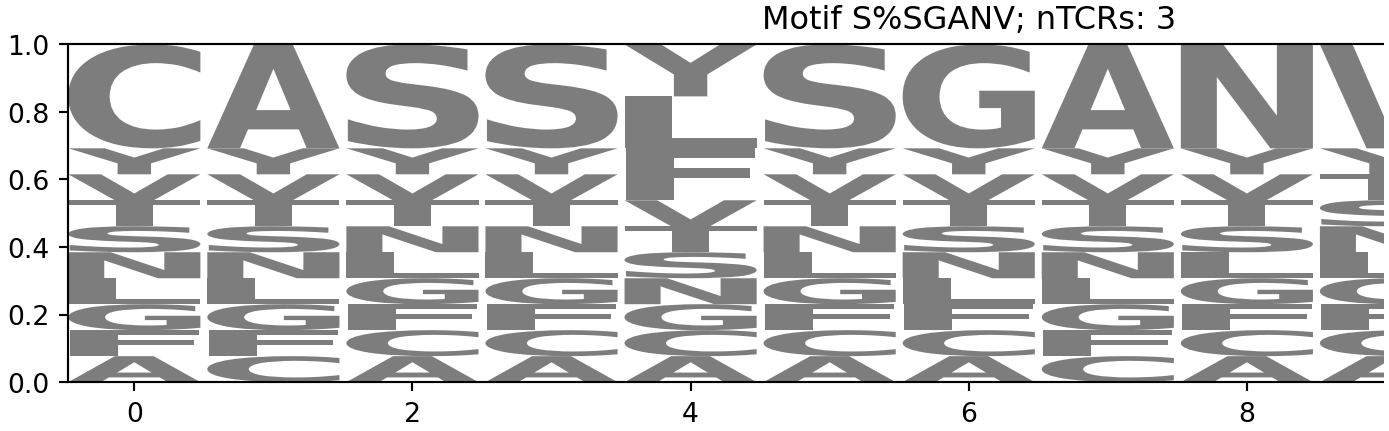
```
<Figure size 960x288 with 0 Axes>
```



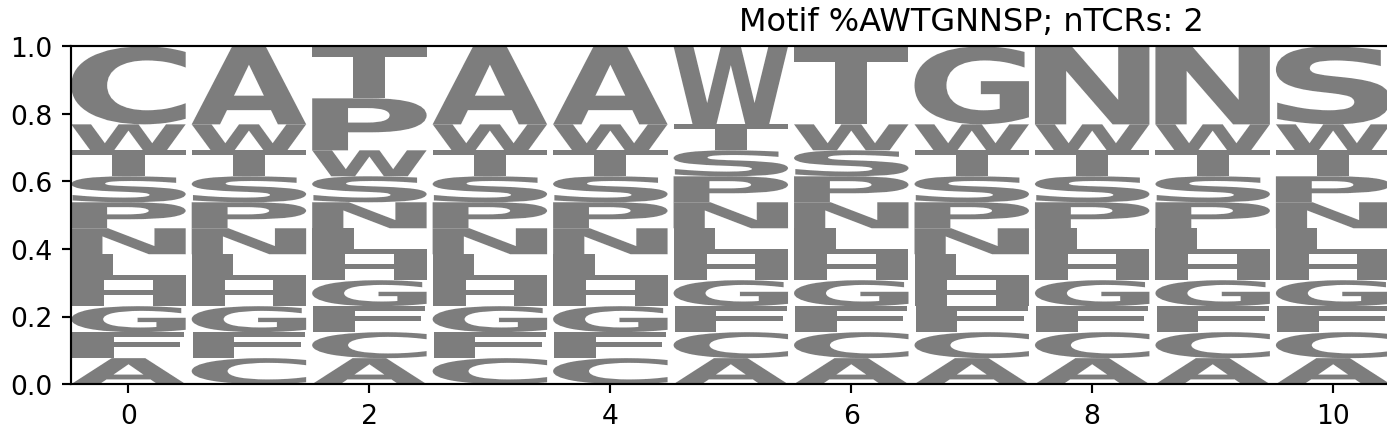

```
<Figure size 960x288 with 0 Axes>
```



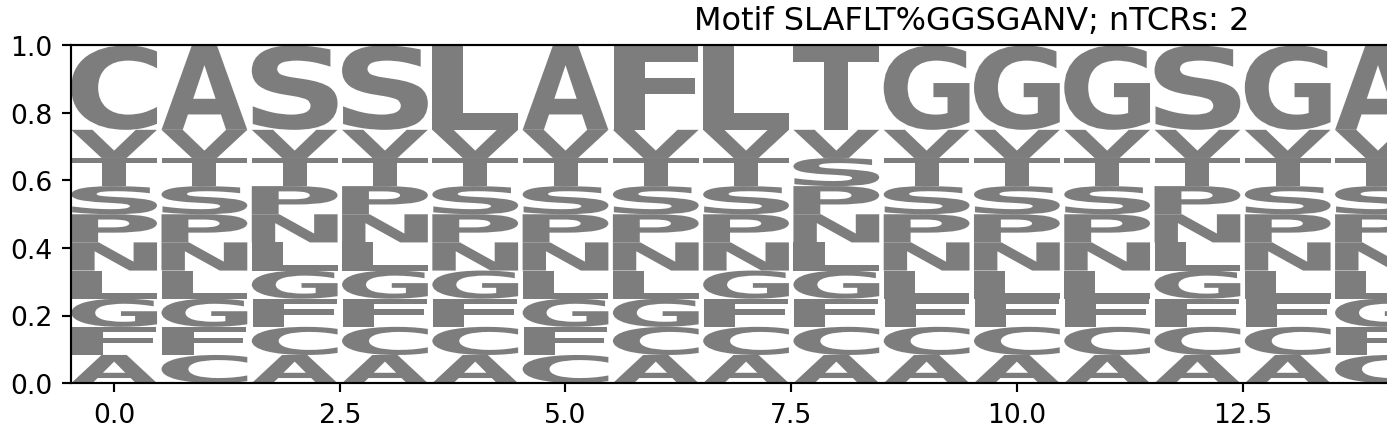

```
<Figure size 960x288 with 0 Axes>
```



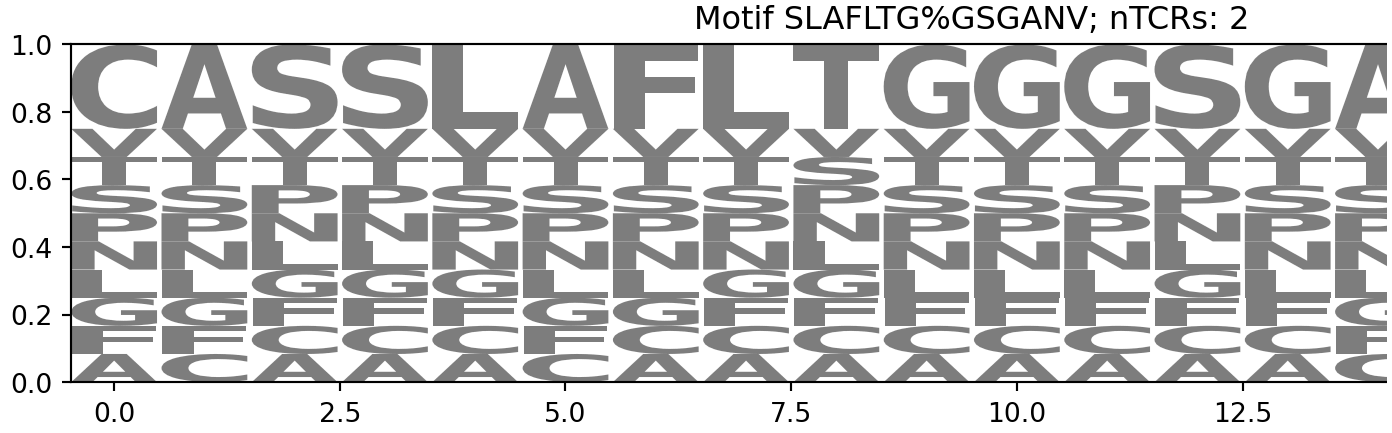

```
<Figure size 960x288 with 0 Axes>
```



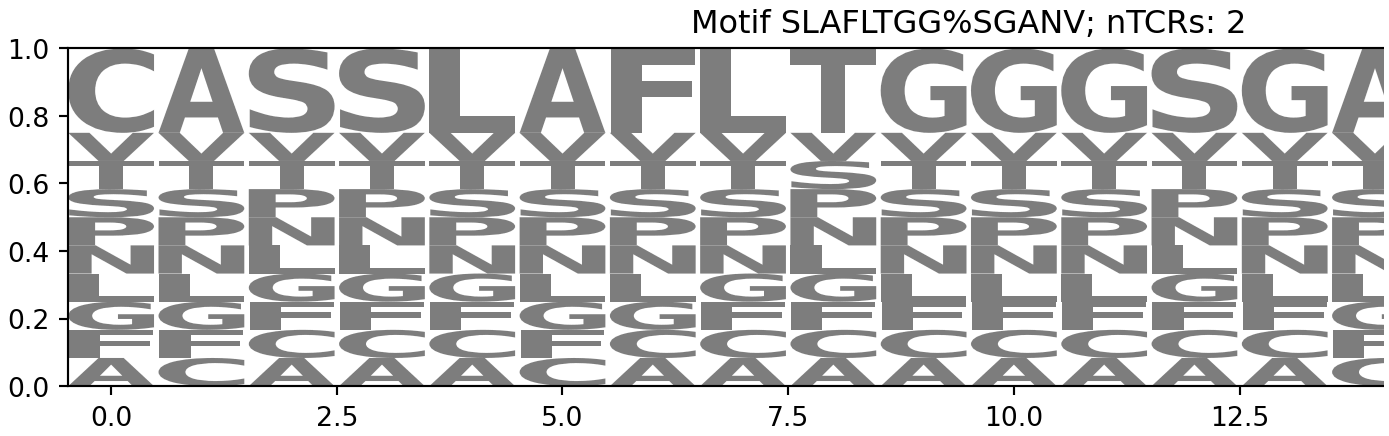

```
<Figure size 960x288 with 0 Axes>
```




```
<Figure size 960x288 with 0 Axes>
```




```
<Figure size 960x288 with 0 Axes>
```




```
<Figure size 960x288 with 0 Axes>
```