HLA-enrichment of Clusters in Mtb-associated TCR sequence repertoires

Flora Zhi-Hui Pang

Supervisors: Prof. Benny Chain

Programme: BSc Immunology and Infection 23/24

Word Count: 3488

1. Capsule Summary

What was already known on this topic?

Meta-clones, a group of non-identical TCR sequences of similar biochemistry and shared antigen specificity, are key to harnessing TCR repertoires as public biomarkers; however, the process of generating meta-clone biomarkers still requires refinement.

What question did this study address?

Can we use an HLA-restriction pipeline to distinguish meta-clones from clusters of TCR sequences lacking shared specificity, and if so, what patterns differentiate meta-clones from antigen challenged and unchallenged TCR sequence repertoires.

What does this study add to our knowledge?

The HLA-restriction pipeline identified TST-derived meta-clones showing enhanced cluster connectivity, adhered to expected HLA Class II restrictions, and encompassed multiple individuals within a single meta-clone, all characteristics distinct from meta-clones of unchallenged repertoires.

What are the implications of this study?

Meta-clone research holds promise to better understand population-wide TCR repertoire dynamics, and how we can harness those dynamics as biomarkers to transform vaccine development, patient risk stratification, and prognosis.

2. Summary

T cell receptor (TCR) repertoires provide insights into individuals' immune history and future response. However, their vast diversity limits their utility as population-wide biomarkers. Meta-clones, clusters of non-identical but biochemically similar TCR sequences with shared antigen specificity, addresses this limitation. Leveraging the *tcrdist3* algorithm, this study applies the meta-clone concept to Tuberculin Skin Test (TST) TCR alpha chain sequences to generate TST meta-clones.

We propose that if meta-clone TCR sequences share antigen specificity they would also have a restriction against a specific Human Leucocyte Antigen (HLA) allele. Our first hypothesis suggests that our HLA-restriction pipeline can distinguish meta-clones from clusters of TCR sequences without a shared antigen specificity. Our second hypothesis suggests that meta-clones derived from TCR sequences of antigen challenged T cell repertoires differ from those of unchallenged repertoires.

The results demonstrate support for both our hypotheses. Our HLA-restriction pipeline identified TST meta-clones with distinct quantitative and qualitative patterns compare to meta-clones derived from PCR negative TCR sequence samples. While *tcrdist3*'s clustering of biochemically similar TCR sequences failed to reflect clusters' presence or absence of a shared antigen specificity, our pipeline showed TST clusters were more significantly and frequently HLA-restricted meta-clones. Unlike PCR negative meta-clones, TST meta-clones aligned with the expected *Mycobacterium tuberculosis* CD4+ T cell-driven response, showing preference for Class II HLA restriction and exhibited greater public representation. TST meta-clones indicated their repertoires' antigen challenge through a distinct convergence of its

TCR sequence towards similar biochemistry and shared antigen specificity across multiple individuals.

This study suggests that meta-clones allow private TCR sequences to become generalisable motifs. Future research should focus on refining meta-clone discovery, exploring paired alpha and beta TCR sequences, and categorising meta-clones as predictors for more specific outcomes. We hope meta-clones will contribute as public biomarkers to improve assessment of infection, disease severity, and vaccine protection.

3. Introduction

3.1 Background

T cells of our immune adaptive system contribute highly specific and protective responses upon recognition of virtually any pathogenic antigen[1]. This recognition is attributed to both genetic variabilities and stochastic recombination of V(D)J gene segments during T cell receptor (TCR) formation, generating up to 10^7 unique structures for the receptor's alpha chain alone[2]. Analysing an individual's TCR repertoire and identifying clonally expanded TCRs, we gain insights into their immunological history[3]. Furthermore, identifying those TCR sequences that drive potent immune responses aids epitope discovery for vaccine development, patient risk stratification, and prognosis[4,5].

For TCR sequences to be generalizable biomarkers, they must be public clonotypes. However, mathematical models[6] predicted and empirical data demonstrated[7] that only 10-15% of TCR alpha or beta chains are shared between at least two randomly selected individuals, with even fewer identical paired sequences. Consequently, focus has shifted towards identifying public meta-clonotypes [8,9,10].

Meta-clonotypes, or meta-clones, are clusters of TCRs whose sequences are biochemically similar in some way and recognize a common peptide-major histocompatibility complex (p-MHC) complex[11]. This concept poses that different individuals may not share identical TCR sequences but may share biochemically similar TCR sequences with a shared antigen specificity[12]. In relaxing the requirement for identity between TCRs, meta-clones attempt to find public signatures for antigen specificity from otherwise private TCR sequences. The *tcrdist3* algorithm

is a leading framework towards generating meta-clones[11]. *Tcridst3* uses a centroid TCR sequence and a pre-set biochemical distance threshold to determine whether other TCR sequences are sufficiently similar to be grouped together (Figure 1). *Tcridst3*'s original study demonstrated that private TCR sequences could become, in some measure, generalizable across the individuals composing the meta-clone, enhancing the utility of TCR sequences in biomarker development.

In this study, we identified meta-clones against *Mycobacterium tuberculosis* (Mtb), the causative agent of tuberculosis disease (TB). In 2022, 7.5 million new TB cases and 1.3 million deaths were reported globally[13]. *M. bovis* bacilli Calmette-Guérin is currently the only licensed TB vaccine, but fails to break the cycle of transmission, showing incomplete and inconsistent efficacy in preventing pulmonary TB in adults[14,15]. Developing new TB vaccines has thus been a priority. Notably, T cells play a key role in TB disease protection and immunopathogenesis, where functional impairment and depletion of CD4+ T cells in HIV+ patients cause greater susceptibility to disease[16]. Despite this, correlates of disease or protection in the T cell immune response remain unknown. Previous research has attempted to identify TB-associated meta-clones using a *tcdist3*-like algorithm called GLIPH2 to group TCR sequences in Mtb immune responses[17,18].

Here we further previous meta-clone research in several ways. First, TCR sequences isolated from Tuberculin Skin Test (TST) biopsies were used instead of peripheral blood samples, as in earlier studies. These biopsies are believed to be highly enriched for TB-specific T cells, increasing confidence that our TST-derived meta-clones are TB-specific. Second, this study seeks to distinguish between

clusters of TCR sequences that share only biochemistry and meta-clones composed of TCR sequences with also shared antigen specificity. To achieve this, we examined human leukocyte antigens (HLAs) of the individuals contributing TCRs to each cluster. A meta-clone's TCR sequences with shared target specificity likely recognize the same HLA and would also be observed from individuals with a common HLA. A cluster lacking HLA restriction could suggest TCR sequences' biochemical similarity without shared specificity. GLIPH2 includes HLA-restriction analyses but does not incorporate non-CDR3 regions' contribution to TCR antigen recognition. Contrastingly, *tcrdist3* considers all three CDR regions but lacks built-in HLA-restriction analyses. *Tcrdist3*'s original paper attempted an HLA-restriction method but was constrained by missing participants' genotyped data, requiring computational Individual-HLA pair predictions[11]. In this study, we address these limitations, combining *tcrdist3*'s comprehensive CDR consideration with an HLA-restricted pipeline supported by exact genotyped data.

Ultimately, this study poses two hypotheses. First, we hypothesise some clusters consist of biochemically similar TCR sequences from different individuals who share an HLA allele. We further suggest, but do not experimentally verify, clusters' HLA-enrichment reflect a shared antigen specificity. Second, existing research underscore that following antigenic challenge, selective expansion of T cells with high affinity for the target antigen transforms the T cell repertoire's clonal composition[19,20,5]. We hypothesise that TST meta-clones exhibit distinct quantitative and qualitative patterns reflecting the Mtb challenged-TCR repertoire landscape; patterns absent in unchallenged repertoire-derived meta-clones. Putatively, meta-clones represent a population level convergence of specific, similar TCR sequences.

3.2. Aims and Implications

This project aims to:

1. Develop an HLA-restriction pipeline to putatively distinguish meta-clones from clusters of similar TCR sequences lacking shared antigen specificity.
2. Elucidate discernible quantitative and qualitative patterns specific to meta-clones derived from antigen challenged TCR sequences.

We compare TCR sequences sampled from:

a) Biopsies from TST study participants as the main antigen challenged dataset.
b) Peripheral blood mononuclear cells from COVIDSortium PCR negative participants as an unchallenged, control dataset.

(Methods and Materials 6.3 for TCR sequence data filtering details preceding *tcrdist3*)

4. Results

4.1 Optimizing a distance threshold for meta-clone identification

Meta-clone identification begins with *tcrdsit3* quantifying biochemical distances between two TCR sequences. Using a predetermined *tcridst*-unit distance threshold, pairs of TCR sequences with distances below this threshold are grouped into clusters as biochemically similar sequences (Materials and Methods 6.4 and Figure 1).

The subsequent step in meta-clone identification assesses whether the grouped TCR sequences are from individuals with a shared HLA allele through our HLA-restriction pipeline (Materials and Methods 6.5 and Figure 2). First, each TCR sequence in a cluster was mapped to the individual from whom it was observed in, linking them to their specific HLA alleles through a pre-established Individual-HLA Python dictionary. Secondly, we identified the most prevalent HLA allele within each cluster and compared its enrichment against the overall dataset using Fisher's Exact test. Lastly, to validate this enrichment, we tested on the clusters again but with a shuffled Individual-HLA dictionary, wherein dictionary keys and values were randomly re-paired to destroy any HLA enrichment pattern resulting from TCR sequences' distance-based grouping. Rather than using a p-value threshold of 0.05 to identify the meta-clones, which could compromise the study's statistical robustness due to multiple testing, we employed the Benjamini-Hochberg Procedure for a stricter significance criterion. Meta-clones were identified with a 1% false discovery rate when comparing correctly and shuffled HLA-restricted clusters.

Seeing how positive meta-clones disappeared on shuffling, we sought to use this to test different *tcrdsit*-unit distance thresholds and optimize meta-clone identification. Defining which TCR sequences to group is crucial for meta-clone generation; the ideal distance threshold should include biochemically similar TCR sequences with shared antigen specificity and exclude those with differing specificity. To do this, we proposed maximizing the number of resultant significant HLA-restricted meta-clones (Materials and Methods 6.6). 16 *tcrdist*-units gave the greatest number of meta-clones (45) and was the distance threshold for our subsequent analyses (Figure 3).

4.2 No quantitative differences between the number of TCR clusters detected in TST and control TCR repertoires.

After biochemically similar TCR sequences were grouped into sets of undirected graphs, clusters, we observed a similar number of clusters derived from TST (2540) and PCR negative (2488) datasets. For clusters greater than 4, both datasets had the same median size (7 TCR sequences) and comparable mean sizes (8.6 and 10.5) (Figure 4a). The cluster connectivity, measured by the edges-nodes ratio, was also similar (0.669 and 0.666). These patterns remained consistent when considering all clusters (Figure 4b).

The quantitative similarity between the distance-based grouping of TST and control repertoires suggests that clusters of biochemically similar TCR sequences cannot reflect meta-clones' shared antigen specificity. We thus explored our first hypothesis that, based on the assumption of HLA-restriction between TCR sequences sharing antigen specificity, identifying clusters with specific HLA enrichments could identify putative meta-clones.

4.3 Quantitative Differences between HLA-associated meta-clones in TST and control repertories

After testing clusters for HLA-enrichment, we found more meta-clones in the TST repertories (45, 1.77% of all TST clusters) than in PCR negative control repertories (6, 0.241%) (Figure 5). The size of TST meta-clones were also consistently matched or exceeded that of control meta-clones (Figure 6a). TST meta-clones also exhibited a higher median (15.0) and mean (17.3) number of component TCR sequences (control meta-clones: 6.00, 6.33). Additionally, cluster connectivity revealed notable difference between TST (4.11) and control (1.45) meta-clones. These results supported our first hypothesis, that our HLA pipeline could identify putative meta-clones from *tcrdist3* TCR sequence clusters.

We can attempt to explain the identification of meta-clones in control clusters (Figure 5b). This may be due to the natural production of biochemically similar and HLA-restricted TCR sequences, even in the absence of antigenic challenge. Nonetheless, the p-value threshold set for the 1% false discovery rate for control meta-clones (4.33×10^{-4}) was much greater than for TST meta-clones (1.94×10^{-8}). This suggests a lower confidence of identifying meta-clones in PCR negative data based on HLA enrichment.

The quantitative differences between meta-clones from Mtb antigen challenged and unchallenged TCR repertoires also positively encouraged our second hypothesis. This prompted us to explore further distinct qualitative patterns of the TST meta-

clones. Our hypothesis entailed that while control meta-clones may be identified through HLA-restriction, they would lack those qualitative patterns.

4.4 Preferential Class II HLA Enrichment in TST meta-clones

Considering that the TST dataset contains TCR sequences expanded in response to Mtb challenge and CD4 T cells preferentially drive Mtb immune responses, we anticipated that TST meta-clones would predominantly be enriched for Class II HLA alleles. Our results aligned with this predication. 86.7% of TST meta-clones were restricted to Class II HLA alleles (Figure 6b), with DRB1*15 being the most enriched allele (Table 1, 62.2% of TST meta-clones). Furthermore, TST meta-clones enriched for Class I HLA alleles had larger p-values, indicating their lower confidence as TST meta-clones (Table 1). Meta-clones from unchallenged control repertoires were not biased towards either HLA class (Figure 6c).

4.5 TST meta-clones incorporate TCRs from various individuals proving representation in a population-wide coverage

A key objective of public biomarker development through meta-clones is to identify shared antigen specificity among the population. Our analysis demonstrated that a single TST meta-clone encompassed up to 34 unique individuals, whereas control meta-clones had a maximum of 6 (Figure 7). Notably, 93.5% of individuals from the original TST TCR sequence dataset were represented by at least one TST meta-clone. Control meta-clones only covered 40% of its total individuals. Both the single meta-clone and spread level indicate that TST meta-clones represent collections of individuals, rather than merely sets of similar TCR sequences from a single individual.

4.6 TST Meta-clone logo motifs show conserved residues in TCR CDR3 animo acid sequences

To simplify future identification of meta-clone TCR sequences within individual repertoires for patient risk stratification and prognosis, we translated component TCR sequences into sequence logos and regular expressions (Regex) (Figure 8). TST meta-clones consistently had identical V and J-genes across their respective sequences, differences then specifically relied on the CDR3 region. Sequence logos revealed that within TST meta-clones, CDR3 amino acid sequences only differed at two or three positions and typically centrally located. The logo plots suggest critical residues conserved among component TCR sequences to allow for the shared specificity of its meta-clone.

5. Discussion

The results of this study robustly support both our hypotheses. First, while *tcrdist3* calculated TCR sequences' biochemical similarities, our HLA-restriction pipeline provided the basis to putatively determine shared antigen specificity. Identifying TST meta-clones validated our pipeline to distinguish them from clusters of TCR sequences sharing only similar biochemistry. Additionally, the pipeline estimated varying contributions of different HLA alleles to participants' T cell responses against Mtb. We found that DR, DQ, DP, A, and C alleles were all significantly enriched in TST meta-clones, but with a preference for Class II HLAs and a dominance of DRB1*15 enrichment. These results are consistent a prior study on HLA alleles' role in Mtb CD4 T cell responses, which also highlight predominant restriction by DR alleles[21]. These findings not only bolster confidence that our TST meta-clones likely comprise of TB-specific TCR sequences expanded in response to Mtb challenge but also affirm the pipeline's validity in aligning with biological expectations. Methodologically, optimised distance threshold and the Benjamini-Hochberg Procedure emphasise this study's attention towards mitigating biases and statistical inaccuracies during data handling. Thereby enhancing our confidence in our HLA-restriction pipeline towards meta-clone identification.

Consistent with our second hypothesis, our analysis revealed quantitative and qualitative distinctions between meta-clones from antigen challenged and unchallenged TCR repertoires. Again, the preference for an HLA class was only observed in TST meta-clones. Despite starting with similar cluster numbers, the HLA-restriction process resulted in more TST meta-clones. The number of TCR sequences in TST meta-clones also equalled or exceeded those of PCR negative

control meta-clones and exhibited greater interconnectivity. This suggests that antigen challenge and the proliferation of high-affinity TCR sequences drive convergence towards similarity in biochemistry and antigen specificity. Unlike control meta-clones, TST meta-clones provided a repertoire coverage sufficient to use as population-wide biomarkers. Sequence logos reveal conservation of amino acids in CDR3 regions of meta-clones' TCR sequences, even across different individuals. While further research is needed to understand the functional implications those amino acids with variation, these findings indicate that TCR sequences expanded in response to antigen exposure are shared, to a similar degree, across the population. These logos could help identify public motifs for the shared antigen recognition in meta-clones. In contrast, the preservation of a polyclonal TCR population during the absence of challenge, leads to PCR negative control meta-clones lacking these distinct characteristics. TCR antigen recognition critically shapes the TCR repertoire, and our TST meta-clones exhibit distinct quantitative and qualitative patterns reflective of their associated history as Mtb antigen challenged repertoires.

This study also adds to published literature. We are the first to identify TB-associated meta-clones with *tcrdist3* and more importantly develop a process of HLA-restriction for *tcrdist3* based on HLA genotyping. In the specific contest of Mtb meat-clones, Musvosvi's paper[18] currently provides the most closely related and comprehensive to ours. They used GLIPH2 and identified TB specific HLA-restricted meta-clones from TST data, observing that 67% of GLIPH2-identified clusters were also identified by *tcrdist3.* Although we did not side-by-side compare GLIPH2 and *tcrdist3*, it is worth noting GLIPH2's disregard for CDR1 and CDR2 TCR regions despite their critical

roles in TCR-p-MHC cooperative binding[22]. It will therefore be interesting to explore whether our pipeline identifies meta-clones undetected by GLIPH2.

Our study has several limitations that future work could be considered. Firstly, we focused solely only alpha chains, overlooking the importance of paired alpha and beta chains in TCR-p-MHC interactions[23,24]. Understanding pairings is necessary enhance the accuracy of grouping non-identical T cells based on a shared antigen specificity. Secondly, our meta-clone discovery is based on a single cohort. Future studies should test the generalisability of our findings in diverse, independent validation cohorts. We could also consider an HLA-independent method for meta-clone identification when HLA genotyping is unavailable. A potential approach could involve topological analyses of *tcrdist3* clusters. For instance, upon visualising the *tcrdist3* clusters as in Figure 7., the connectivity of the graph (e.g. counting the number of triangles formed by a cluster's edges) could differentiate meta-clones from clusters of TCR sequences lacking shared antigen specificity. Investigating a connectivity threshold represented by triangle count could provide valuable insights into meta-clone identification.

In conclusion, this study supports our hypothesis that clusters of biochemically similar TCR sequences derived from different individuals with statistically significant enhanced sharing of an HLA allele are present in TST repertoires. This finding is consistent with the hypothesis that such meta-clones not only reflect TCR sequences' similar biochemistry but also a putative shared antigen specificity in the local tissue following a standardized antigen challenge. The number and size of these HLA-restricted meta-clones is much higher in the TST repertoires compared to

control blood repertoires from healthy individuals, likely reflecting the recent and strong antigen exposure. Our results indicate that meta-clone identification can increase the effective publicity of a TCR signature from a group of unrelated individuals targeting the same p-MHC complex. However, identifying meta-clones is only the first step in biomarker development, biomarkers predictive of infection, disease severity, or vaccine protection will require correlating specific meta-clones with robust clinical data (such as disease, protection etc.) collected from prospective studies of Mtb infection. By doing this, we hope to further understand the immune system’s response to disease, bringing us closer to effective biological tools, so that more lives can be saved.

6. Materials and Methods

6.1 Data Science Tools

All analyses used the Python programming language[25]. *Tcrdist3*, a package for multi-CDR TCR repertoire analysis and visualization, computed distances between TCR sequences[11]. The Networkx package handled complex graph networks and analysed meta-clones[26]. The Pickle module deserialised data[25]. The Matplotlib library visualized data[27]. The Logomaker package generated sequence logos[28]. NumPy and Pandas packages managed calculations, data storage, and retrieval.[29,30].

6.2 Data

896,633 alpha chain TCR sequences from the TST study's 100 participants provided by the Noursadeghi lab. The TST study intradermally injected a Mtb purified protein derivative (PPD), with TCR sequences obtained from TCR sequencing biopsies at the TST site. UK research ethics committees granted research ethics and regulatory approvals under reference numbers 11/LO/1863, 14/LO/0505, and 18/LO/0680, and all participants provided written informed consent.

A control set of 717,479 alpha chain TCR sequences from the peripheral blood of the COVIDSortium study's 50 PCR negative individuals[31]. TCRs were sequenced using the Chain ligation protocol[32] and analysed using the Decombinator software package[31].

These works were done before I started the project.

6.3 Creating a 10,000 TCR sequence dataset

TST and PCR negative TCR sequences were first filtered into *tcrdist3*-appropriate data frames:

1. Filtered for Day 7 TST TCR sequences (PPD-reactive T cell enrichment peaks 7 days after antigenic challenge[33])
2. Selected TCR sequences observed more than once (suggesting clonal expansion).
3. Retained V gene, J gene, CDR3 sequence columns from the Decombinator files for *tcrdist3*[34]
4. Appended '*01' to each entry in the 'v_call' column (*tcrdist3*-requirement)
5. Excluded TCR sequences from HLA-unrestricted[35,36] invariant natural killer T cells (TRAV10 and TRAJ18 gene region pairings) and mucosal associated invariant T cells (TRAV1-2 with TRAJ33, TRAJ12 or TRAJ20 pairings).

Due to *tcrdist3*'s 10,000-sequence limit, we sampled representatively from the datasets. 150 TCR sequences were randomly sampled from each of TST study's 100 individuals, estimating just over 10,000 sequences after filtering. A random sample of 10,000 was then finally taken as the input dataset for *tcrdist3*. We repeated this process for the PCR negative dataset, taking 250 sequences from each of the 50 individuals.

6.4 *Tcrdist3* distance metric

Tcrdist3 calculated pairwise distances between TCR sequences. Its TCRRrep package computes Levenshtein distance as a weighted sum of BLOSUM62 substitution penalties across all CDRs (CDR1, CDR2, CDR2.5 and CDR3). The CDR3 penalty is weighted three times that of the other CDRs (Figure 1b).

6.5 HLA-restriction Pipeline: from *tcrdist3* distance matrix to meta-clones

The *tcrdist3* distance matrix was converted into a set of undirected graphs (clusters) by iteratively connecting all TCR sequences with *tcrdist3* distances below the set threshold. A cluster of TCR sequences was defined as a set of connected TCRs (a connected component of the graph). HLA analysis was restricted to clusters with more than 4 TCR sequences to reduce multiple testing and increase statistical power.

To putatively determine whether a cluster of similar TCR sequences also shared antigen specificity as a meta-clone, each cluster underwent HLA-restriction testing. Each TCR sequence was mapped to the individual they were observed in, listing the individual's carried HLA allele in a non-redundant manner – counting homozygous alleles once. Then, when compiling individuals' unique HLA alleles across the cluster, this list was redundant – allowing allele repeats (Figure 2a). From this collated list, we identified the most common HLA allele in the cluster and used a Fisher's Exact Test to assess its enrichment. We compared the proportion of TCR sequences which individuals carried the most common HLA to the proportion prevalence of that HLA in the original TCR dataset population. The Benjamini-Hochberg Procedure corrected for multiple testing.

As a control, we randomly shuffled the keys and values of the Individual-HLA dictionary and repeated the mapping of clusters' TCR sequences to HLA. By comparing Fisher's Exact Test results from clusters mapped by unshuffled and

shuffled dictionaries, a 1% false discovery rate identified significantly HLA restricted meta-clones in correctly mapped clusters (Figure 2).

6.6 Optimizing the *tcrdist3* distance threshold

After pairwise calculation of biochemical distance, the *tcrdist3* distance threshold determines whether one TCR sequence is sufficiently similar to another. To optimise the threshold, we performed iterative analyses to observe how its variation impacted the resulted meta-clone count. The objective was to find the threshold which maximised the number of meta-clones. A threshold too low loses sensitivity and deems too few TCR sequences similar, resulting in no or few meta-clones. A threshold too high deems too many TCR sequences as sufficiently similar and compromises the meta-clone TCR sequences' specificity restriction, reducing the number of HLA-restricted meta-clones. 16 *tcrdist*-units optimised this trade-off between sensitivity and specificity (Figure 3).

All analyses were performed by me under the guidance of Prof. Benny Chain, unless stated otherwise.

7. Acknowledgements

First and foremost, I would like to thank my supervisor Professor Benny Chain for his continued kindness, guidance, and support through this whole project. His mentorship not only propelled my understanding of immunology to new heights but also served as a constant source of inspiration. I would also like to give a special thanks to Professor Richard Milne for connecting me to my supervisor and making this entire project possible. Finally, I would like to extend my gratitude to all the patients from the TST study and COVIDSortium who have kindly and bravely donated their TCR sequence repertoires to science. I hope that the collective efforts of the scientific community can continue to make good use of their data to save more lives in the future.

8. References

1. Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. J. (2013). *T Cell-Mediated Immunity*. Nih.gov; Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK10762/

2. Mora, T., & Walczak, A. M. (2016). Quantifying lymphocyte receptor diversity. *BioRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/046870

3. DeWitt, W. S., Smith, A., Schoch, G., Hansen, J. A., Matsen, F. A., & Bradley, P. (2018). Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *ELife*, *7*. https://doi.org/10.7554/elife.38358

4. Hogan, S. A., Courtier, A., Cheng, P. F., Jaberg-Bentele, N. F., Goldinger, S. M., Manuel, M., Perez, S., Plantier, N., Mouret, J.-F., Dan, T., Marieke I.G. Raaijmakers, Kvistborg, P., Pasqual, N., John B.A.G. Haanen, Dummer, R., & Levesque, M. P. (2019). Peripheral Blood TCR Repertoire Profiling May Facilitate Patient Stratification for Immunotherapy against Melanoma. *Cancer Immunology Research*, *7*(1), 77–85. https://doi.org/10.1158/2326-6066.cir-18-0136

5. Zornikova, K. V., Sheetikov, S. A., Alexander Yu Rusinov, Iskhakov, R. N., & Bogolyubova, A. V. (2023). Architecture of the SARS-CoV-2-specific T cell repertoire. *Frontiers in Immunology*, *14*. https://doi.org/10.3389/fimmu.2023.1070077

6. Yuval Elhanati, Sethna, Z., Curtis Gove Callan, Mora, T., & Walczak, A. M. (2018). Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological Reviews*, *284*(1), 167–179. https://doi.org/10.1111/imr.12665

7. Soto, C., Bombardi, R. G., Branchizio, A., Kose, N., Matta, P., Sevy, A. M., Sinkovits, R. S., Pavlo Gilchuk, Finn, J. A., & Crowe, J. E. (2019). *High frequency of shared clonotypes in human B cell receptor repertoires*. *566*(7744), 398–402. https://doi.org/10.1038/s41586-019-0934-8

8. Mikhail Shugay, Bagaev, D. V., Turchaninova, M. A., Bolotin, D. A., Britanova, O. V., Putintseva, E. V., Pogorelyy, M. V., Nazarov, V. I., Zvyagin, I. V., Kirgizova, V. I., K.I. Kirgizov, Skorobogatova, E. V., & Chudakov, D. M. (2015). VDJtools: Unifying Post-analysis of T Cell Receptor Repertoires. *PLOS Computational Biology*, *11*(11), e1004503–e1004503. https://doi.org/10.1371/journal.pcbi.1004503

9. Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., Ji, X., Han, A., Krams, S. M., Pettus, C., Haas, N., Arlehamn, C. S. L., Sette, A., Boyd, S. D., Scriba, T. J., Martinez, O. M., & Davis, M. M. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature*, *547*(7661), 94–98. https://doi.org/10.1038/nature22976

10. Pogorelyy, M. V., Minervina, A. A., Mikhail Shugay, Chudakov, D. M., Lebedev, Y. B., Mora, T., & Walczak, A. M. (2019). Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biology*, *17*(6), e3000314–e3000314. https://doi.org/10.1371/journal.pbio.3000314

11. Mayer-Blackwell, K., Schattgen, S., Cohen-Lavi, L., Crawford, J. C., Souquette, A., Gaevert, J. A., Hertz, T., Thomas, P. G., Bradley, P., & Fiore-Gartland, A. (2021). TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *ELife*, *10*, e68605. https://doi.org/10.7554/eLife.68605

12. Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., Crawford, J. C., Clemens, E. B., Nguyen, T. H. O., Kedzierska, K., La Gruta, N. L., Bradley, P., & Thomas, P. G. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, *547*(7661), 89–93. https://doi.org/10.1038/nature22383

13. World Health Organization. (2023). *Global tuberculosis report 2023*. Www.who.int. https://www.who.int/publications/i/item/9789240083851

14. Kuan, R., Muskat, K., Peters, B., & Lindestam Arlehamn, C. S. (2020). Is mapping the BCG vaccine-induced immune responses the key to improving the efficacy against tuberculosis? *Journal of Internal Medicine*, *288*(6), 651–660. https://doi.org/10.1111/joim.13191

15. Tanner, R., Villarreal-Ramos, B., Vordermeier, H. M., & McShane, H. (2019). The Humoral Immune Response to BCG Vaccination. *Frontiers in Immunology*, *10*. https://doi.org/10.3389/fimmu.2019.01317

16. Bruchfeld, J., Correia-Neves, M., & Källenius, G. (2015). Tuberculosis and HIV Coinfection. *Cold Spring Harbor Perspectives in Medicine*, *5*(7), a017871. https://doi.org/10.1101/cshperspect.a017871

17. Huang, H., Wang, C., Rubelt, F., Scriba, T. J., & Davis, M. M. (2020). Analyzing the M. tuberculosis immune response by T cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nature Biotechnology*, *38*(10), 1194. https://doi.org/10.1038/s41587-020-0505-4

18. Musvosvi, M., Huang, H., Wang, C., Xia, Q., Rozot, V., Krishnan, A., Acs, P., Cheruku, A., Obermoser, G., Leslie, A., Behar, S. M., Hanekom, W. A., Bilek, N., Fisher, M., Kaufmann, S. H. E., Walzl, G., Hatherill, M., Davis, M. M., & Scriba, T. J. (2023). T cell receptor repertoires associated with control and

disease progression following Mycobacterium tuberculosis infection. *Nature Medicine*, *29*(1), 258–269. https://doi.org/10.1038/s41591-022-02110-9

19. Pan, Y.-G., Benjamas Aiamkitsumrit, Bartolo, L., Wang, Y., Lavery, C., Marc, A., Holec, P. V., Rappazzo, C., Eilola, T., Gimotty, P. A., Hensley, S., Antia, R., Zarnitsyna, V. I., Birnbaum, M. H., & Su, L. F. (2021). *Vaccination reshapes the virus-specific T cell repertoire in unexposed adults*. *54*(6), 1245-1256.e5. https://doi.org/10.1016/j.immuni.2021.04.023
20. Xia, M., Blazevic, A., Fiore-Gartland, A., & Hoft, D. F. (2023). Impact of BCG vaccination on the repertoire of human γδ T cell receptors. *Frontiers in Immunology*, *14*. https://doi.org/10.3389/fimmu.2023.1100490
21. Lindestam, C. S., McKinney, D. M., Carpenter, C., Paul, S., Virginie Rozot, Makgotlho, E., Gregg, Y., Michele van Rooyen, Ernst, J. D., Hatherill, M., Hanekom, W. A., Peters, B., Scriba, T. J., & Sette, A. (2016). A Quantitative Analysis of Complexity of Human Pathogen-Specific CD4 T Cell Responses in Healthy M. tuberculosis Infected South Africans. *PLOS Pathogens*, *12*(7), e1005760–e1005760. https://doi.org/10.1371/journal.ppat.1005760
22. Lynch, J. N., Donermeyer, D. L., Weber, K. S., Kranz, D. M., & Allen, P. M. (2013). Subtle changes in TCRα CDR1 profoundly increase the sensitivity of CD4 T cells. *Molecular Immunology*, *53*(3), 283–294. https://doi.org/10.1016/j.molimm.2012.08.020
23. Springer, I., Tickotsky, N., & Louzoun, Y. (2021). Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.664514

24. Chung Eun Ha, & N.V. Bhagavan. (2023). Immunology. *Elsevier EBooks*, *3*, 695–726. https://doi.org/10.1016/b978-0-323-88541-6.00019-3

25. Van Rossum, G. (2020). Python programming language, version 3.8.5. Retrieved from https://www.python.org

26. Hagberg, A., Gov) -Los, Schult, D., & Swart, P. (2008). *Exploring Network Structure, Dynamics, and Function using NetworkX*. https://aric.hagberg.org/papers/hagberg-2008-exploring.pdf

27. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

28. Tareen, A., & Kinney, J. B. (2019). Logomaker: beautiful sequence logos in Python. *Bioinformatics*, *36*(7), 2272–2274. https://doi.org/10.1093/bioinformatics/btz921

29. Walt, Stéfan Van Der, S. Chris Colbert, and Gaël Varoquaux. 2011. “The NumPy Array: A Structure for Efficient Numerical Computation.” Computing in Science and Engineering 13 (2): 22–30. https://doi.org/10.1109/MCSE.2011.37.

30. McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In Proceedings of the 9th Python in Science Conference, 1697900:51–56. http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

31. Milighetti, M., Peng, Y., Tan, C., Mark, M., Nageswaran, G., Byrne, S., Ronel, T., Peacock, T., Mayer, A., Chandran, A., Rosenheim, J., Whelan, M., Yao, X., Liu, G., Felce, S. L., Dong, T., Mentzer, A. J., Knight, J. C., Balloux, F., & Greenstein, E. (2023). Large clones of pre-existing T cells drive early immunity against SARS-COV-2 and LCMV infection. *IScience*, *26*(6), 106937. https://doi.org/10.1016/j.isci.2023.106937

32. Thomas, N., Heather, J., Ndifon, W., Shawe-Taylor, J., & Chain, B. (2013). Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics*, *29*(5), 542–550. https://doi.org/10.1093/bioinformatics/btt004

33. Holm, L. L., Vukmanovic-Stejic, M., Blauenfeldt, T., Benfield, T., Andersen, P., Akbar, A. N., & Ruhwald, M. (2018). A Suction Blister Protocol to Study Human T-cell Recall Responses In Vivo. *Journal of Visualized Experiments*, *138*. https://doi.org/10.3791/57554

34. Mayer-Blackwell, K. (2024, March 14). *kmayerb/tcrdist3*. GitHub. https://github.com/kmayerb/tcrdist3

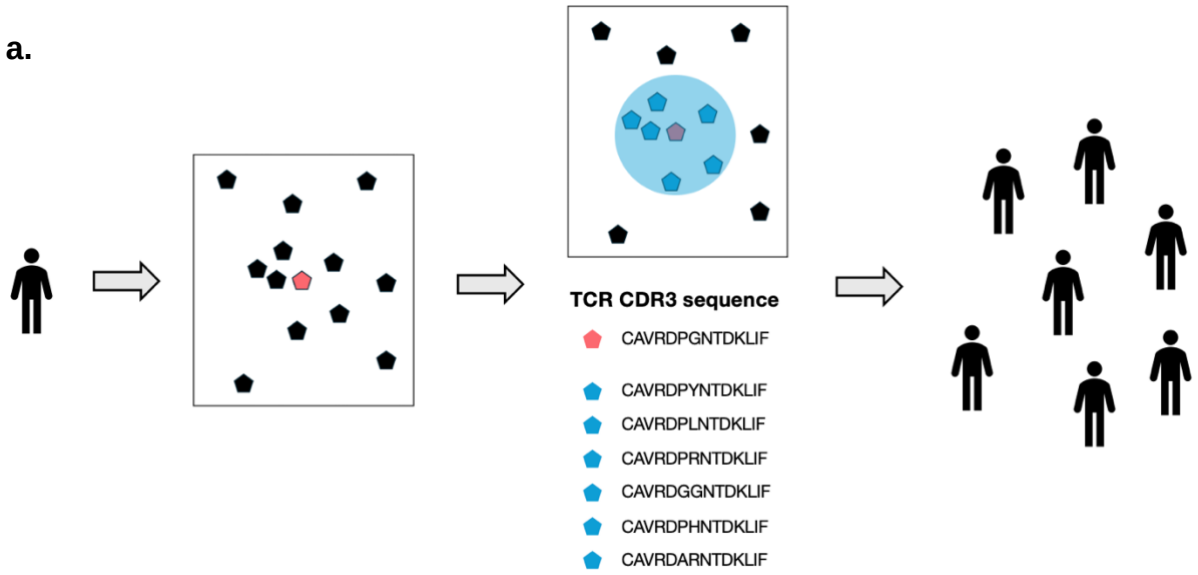
35. Aoki, T., Shinichiro Motohashi, & Haruhiko Koseki. (2023). Regeneration of invariant natural killer T (iNKT) cells: application of iPSC technology for iNKT cell-targeted tumor immunotherapy. *Inflammation and Regeneration*, *43*(1). https://doi.org/10.1186/s41232-023-00275-5

36. Johnston, A., & Gudjonsson, J. E. (2014). Psoriasis and the MAITing Game: A Role for IL-17A+ Invariant TCR CD8+ T Cells in Psoriasis? *Journal of Investigative Dermatology*, *134*(12), 2864–2866. https://doi.org/10.1038/jid.2014.361

9. Figures and Tables

Figure 1

a.



b.

v_a_gene	j_a_gene	Centroid: CAGLSGTYKYIF	Edit Distance	TCRdist	Substitution 1	Penalty 1	Substitution 2	Penalty 2	CDR3 Weight
TRAV27*01	TRAJ40	CAGQSGTYKYIF	1	6			(L,Q)	2	3
TRAV27*01	TRAJ40	CAVISGTYKYIF	2	15	(G,V)	3	(L,I)	2	3
TRAV27*01	TRAJ40	CAVLSGTYKYIF	1	9	(G,V)	3			3
TRAV27*01	TRAJ40	CAGTSGTYKYIF	1	3			(L,T)	1	3
TRAV27*01	TRAJ40	CAVFSGTYKYIF	2	9	(G,V)	3	(L,F)	0	3
TRAV27*01	TRAJ40	CAATSGTYKYIF	2	3	(G,A)	0	(L,T)	1	3
TRAV27*01	TRAJ40	CAGVSGTYKYIF	1	3			(L,V)	1	3
TRAV27*01	TRAJ40	CAALSGTYKYIF	1	0	(G,A)	0			3
TRAV27*01	TRAJ40	CAGPSGTYKYIF	1	9			(L,P)	3	3
TRAV27*01	TRAJ40	CAASSGTYKYIF	2	6	(G,A)	0	(L,S)	2	3
TRAV27*01	TRAJ40	CAVFSGTYKYIF	2	9	(G,V)	3	(L,F)	0	3

Figure 1: Schematic Diagram of *tcrdist3*'s contribution towards generating meta-clones
(a) *Transforming Private TCR Clonotypes into Public Meta-clones*. *Tcrdist3* takes an input data set of TCR sequences where pairwise quantification of their biochemical distances allows for the grouping of similar TCR sequences. These clusters are based on a centroid TCR sequence (red pentagon) and a pre-set *tcrdist* distance threshold (blue radius) which determines another TCR sequence's similarity (blue pentagons) or dis-similarity (black pentagons) to the centroid. In grouping through similarity, this allows for a cluster to encompass multiple TCR sequences and individuals. **(b)** *Example of how tcrdist3 calculates TCR sequence biochemical distance.* Here, TCR sequences are compared to the centroid TRAV27*01, TRAJ40, and CAGLSGTYKYIF, the V-gene, J-gene and CDR3 region respectively. The TCR sequences have matching V and J-genes, so the CDR1, CDR2, and CDR2.5 contribution to *tcrdist3* is 0. TCR sequences' distances from the centroid will rely specifically on the CDR3 region. Penalties were calculated as BLOSUM62 substitutions. TCRdist = Penalty1*CDR3 Weight + Penalty2*CDR3 Weight.

Figure 2

a.

TCR sequences	TCR Sequences' Respective Individuals	Non-redundant List of Individual's HLAs	Collated Redundant List of the HLAs	Most Common HLA
8192, 3969, 5633	HTB_0021, HTB_0093, HTB_0014	HTB_0021: B*12, B*35, A*34, DRB1*15	B*12, B*35, A*34, DRB1*15, C*05, A*30, C*02, DRB1*15, B*07, B*15, DRB1*15, A*23	DRB1*15
		HTB_0093: C*05, A*30, C*02, DRB1*15		
		HTB_0014: B*07, B*15, 'DRB1*15, A*23		
9986, 7812, 3096	HTB_0016, HTB_0083, HTB_0011	HTB_0016: DPB1*01, B*05, B*05, DRB1*15	DPB1*01, B*05, DRB1*15, DRB1*15, B*15, DPB1*02, DRB1*15, A*23, DPB1*01	DRB1*15
		HTB_0083: DRB1*15, B*15, DPB1*02, DPB1*02		
		HTB_0011: DRB1*15, A*23, A*23, DPB1*01		
1414, 1416, 8464	HTB_0022, HTB_0022, HTB_0777	HTB_0022: B*44, B*19, A*04, DRB1*15	B*44, B*19, A*04, DRB1*15, B*44, B*19, A*04, DRB1*15, DRB1*15, B*20, DPB1*11	DRB1*15
		HTB_0022: B*44, B*19, A*04, DRB1*15		
		HTB_0777: DRB1*15, B*20, B*20, DPB1*11		

b.

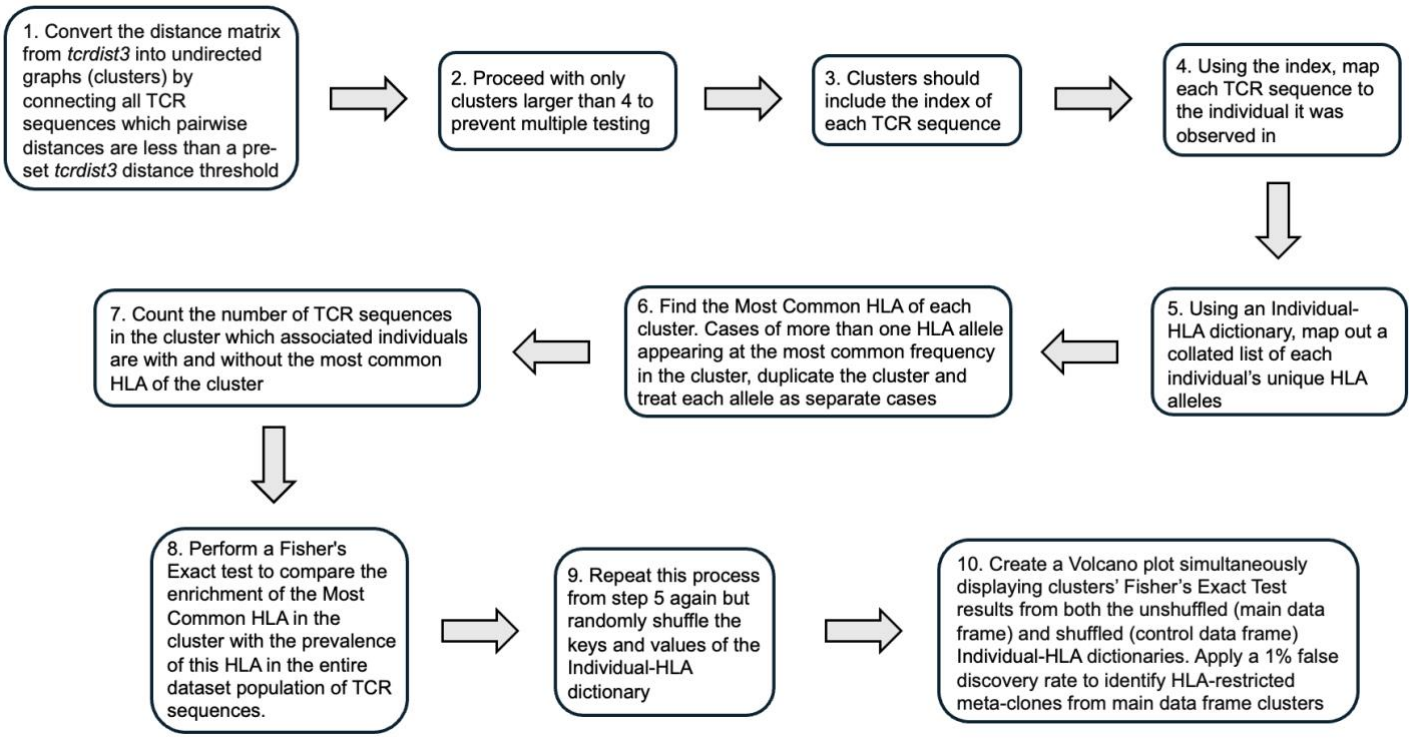


Figure 2: HLA-restriction pipeline
(a) *Inexact Representative Example of Individual-HLA dictionary Mapping to find the cluster's Most Common HLA*. Each row is one cluster, where each cluster has 3 TCR sequences, each with an associated Individual. Listing each Individual's HLA alleles for each TCR sequence is done non-redundantly, but the collation of the cluster's HLA alleles is redundant. The most common HLA of the cluster is marked in red. **(b)** *Flowchart of the HLA-restriction Pipeline.* Following the hypothesis that if biochemically similar TCR sequences in a meta-clone recognize the same p-MHC complex, meta-clones could then be identified through their TCR sequences' restriction to a specific HLA. This pipeline maps how to identify whether a cluster of TCR sequences is significantly restricted to a specific HLA. It uses an Individual-HLA dictionary and the Benjamini-Hochberg Procedure. In this study, HLA-restriction was used to evaluate both the TST and PCR Negative clusters to later highlight characteristics of true antigen challenge-derived meta-clones from a false positive meta-clone.

Figure 3

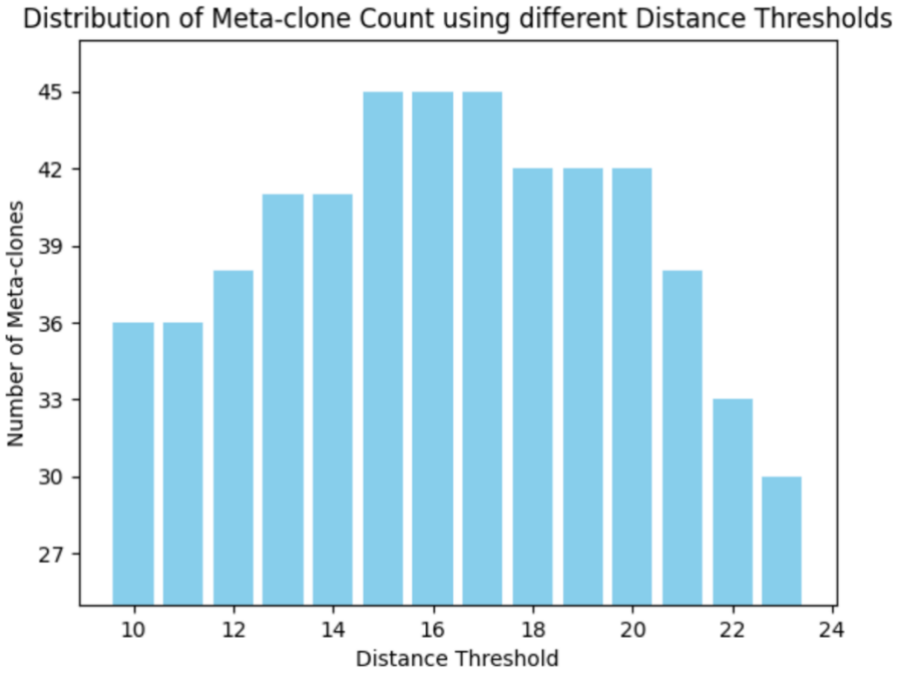
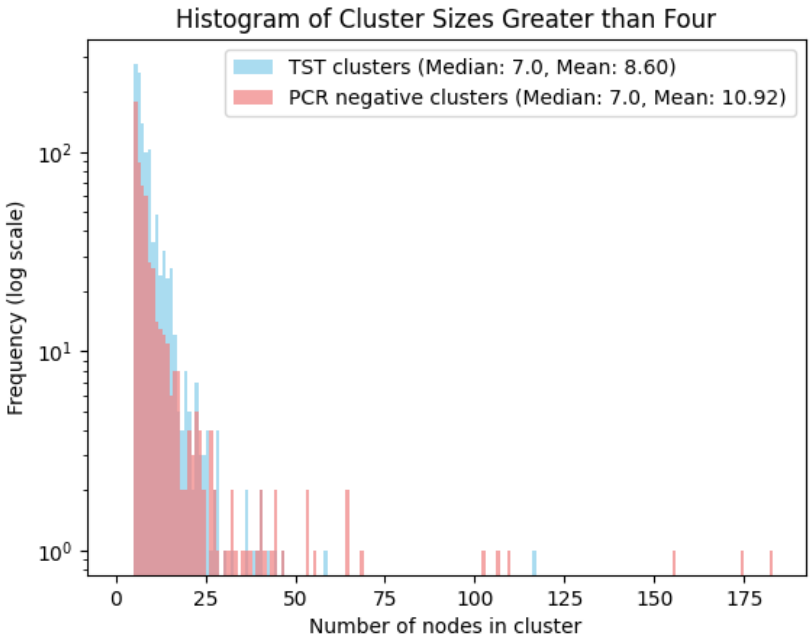


Figure 3: Varying Distance Thresholds results in a Normal Distribution of the Resultant TST Meta-clone Count

To decide on this study's *tcrdist3* distance threshold, the objective was to find the threshold which maximised the number of significant meta-clones generated in the analysis. The x-axis shows the *tcrdist*-unit we chose as the distance threshold, ranging from 10 to 23 units. The y-axis shows the number of meta-clones identified after HLA-restriction at the distance threshold chosen. A threshold of 15-17 *tcrdist*-units was found to optimise discovery, providing the greatest number (45) of TST meta-clones.

Figure 4

a.



b.

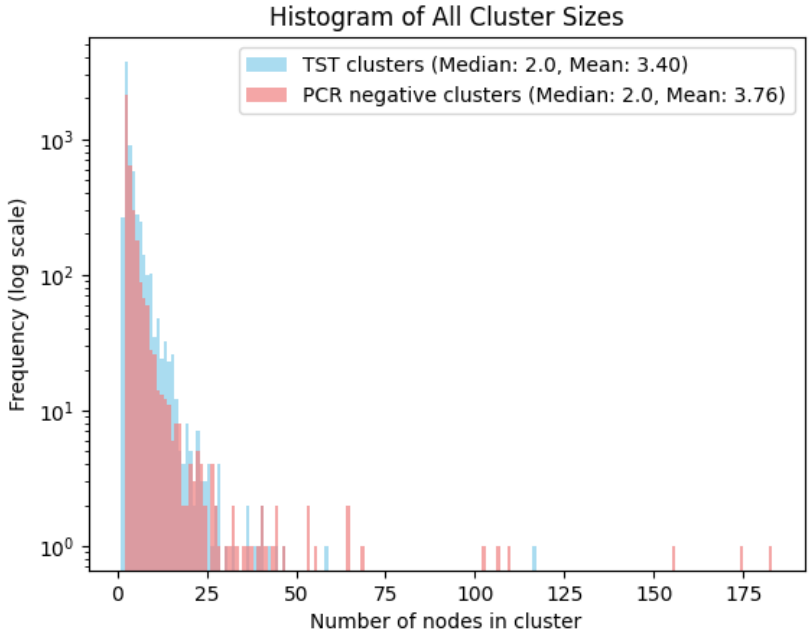
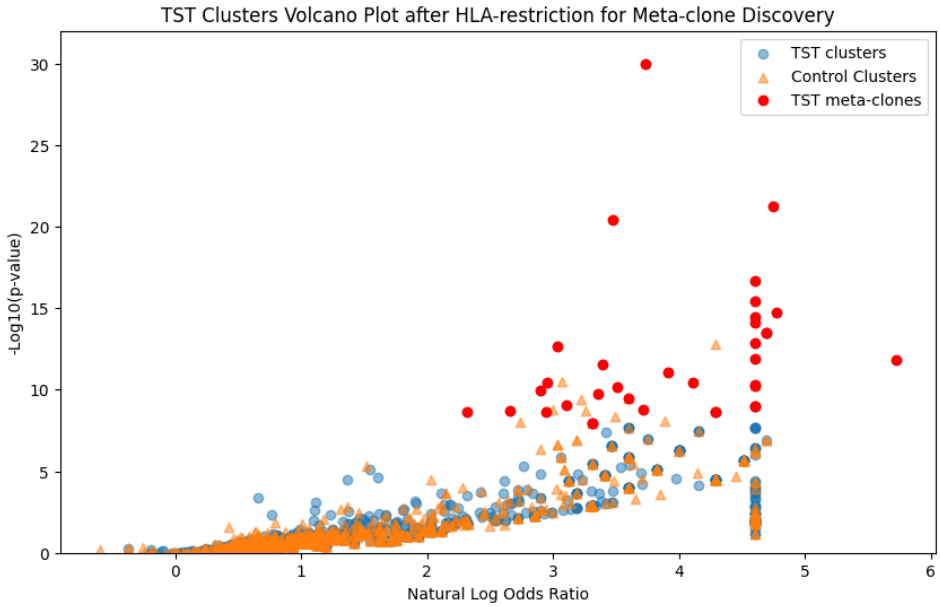


Figure 4: Quantitative Similarity seen between TST and PCR Negative Cluster Sizes after *tcrdist3* analysis

(a) and **(b)** The x-axis shows the number of TCR sequences within a cluster. The y-axis shows the logarithmic scale of the number of clusters containing the corresponding number of TCR sequences within. **(a)** *Histogram comparing the TST and PCR Negative Cluster Sizes of those clusters Greater than 4*. The histogram shows a general overlap in the TST (blue) and PCR Negative (red) cluster Sizes. They share a median of 7 TCR sequences in a cluster and similar cluster size means of 8.60 and 10.92. **(b)** *Histogram comparing TST and PCR Negative Cluster Sizes for all clusters.* All clusters are presented here to demonstrate the statistical pattern in cluster size similarity are real quantitative observations and not a result of data manipulation through a deliberate focus of clusters with more than 4 TCR sequences. They share a median of 2 TCR sequences in a cluster and similar cluster size means of 3.40 and 3.76.

Figure 5

a.



b.

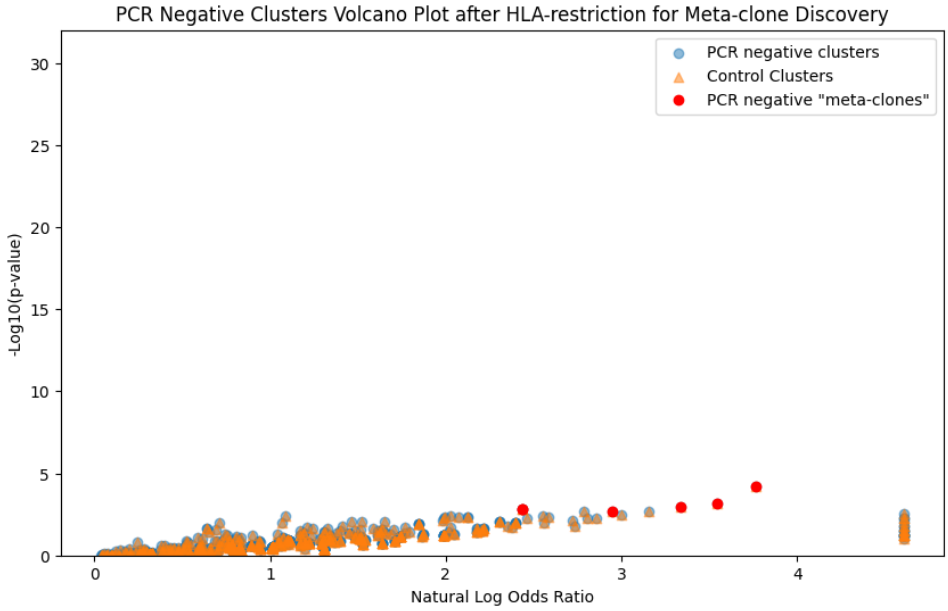
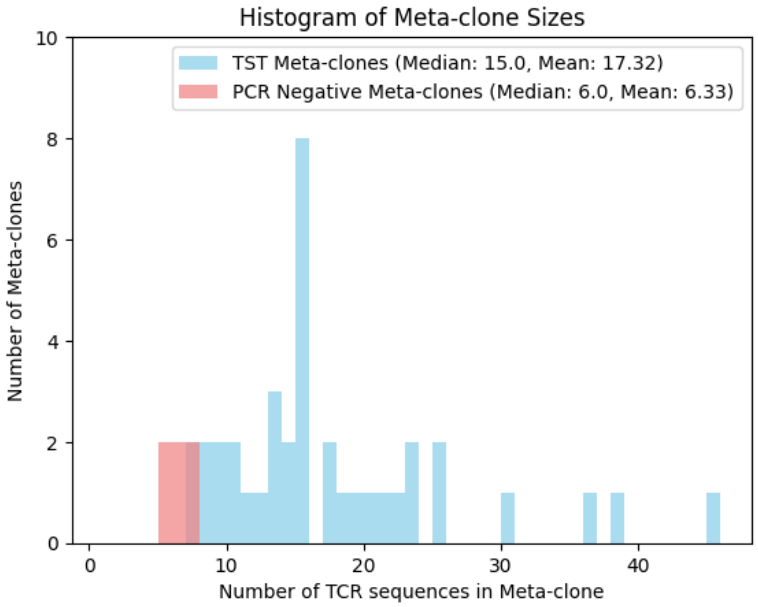


Figure 5: HLA-restriction on TST and PCR Negative Clusters results in a differing Meta-clone Counts

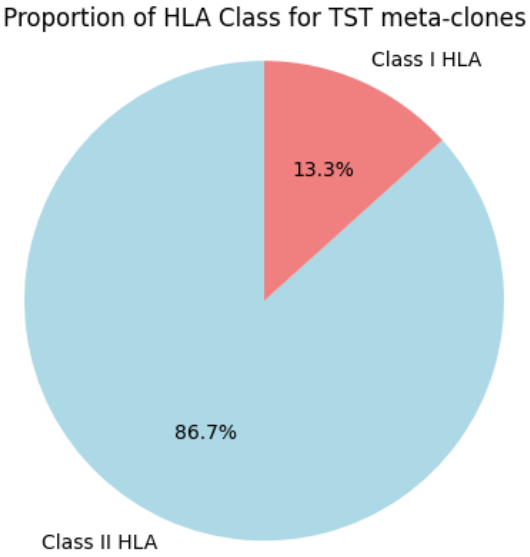
(a) and **(b)** Fisher's Exact Test results were calculated for each cluster, by counting the proportion of TCRs in the cluster associated with the most common HLA allele, compared to that proportion of the whole repertoire. The x-axis shows the natural logarithm of the odds ratio. Infinite odds ratios were replaced with a value of 100 so the points could be plotted. The y-axis shows the p-value at negative $\log_{10}$. Red circles are identified HLA-enriched meta-clones based on a 1% false discovery rate for shuffled TCR-HLA linkage. Blue Circles are clusters of biochemically similar TCR sequences not showing HLA-enrichment. Orange Triangles are the control TST or PCR Negative clusters, where their HLA-restriction was mapped by an Individual-HLA dictionary with its keys and values randomly shuffled. **(a)** *TST meta-clone discovery.* 45 TST meta-clones were identified where clusters' restriction for a specific HLA was significantly greater than a random case of HLA-enrichment. The p-value threshold for meta-clone discovery was found at 1.94×10^{-8}. **(b)** *PCR Negative meta-clone discovery.* Only 6 HLA-enriched meta-clones were identified in the control repertoires. The p-value threshold for meta-clone discovery was found at 4.33×10^{-4}.

Figure 6

a.



b.



c.

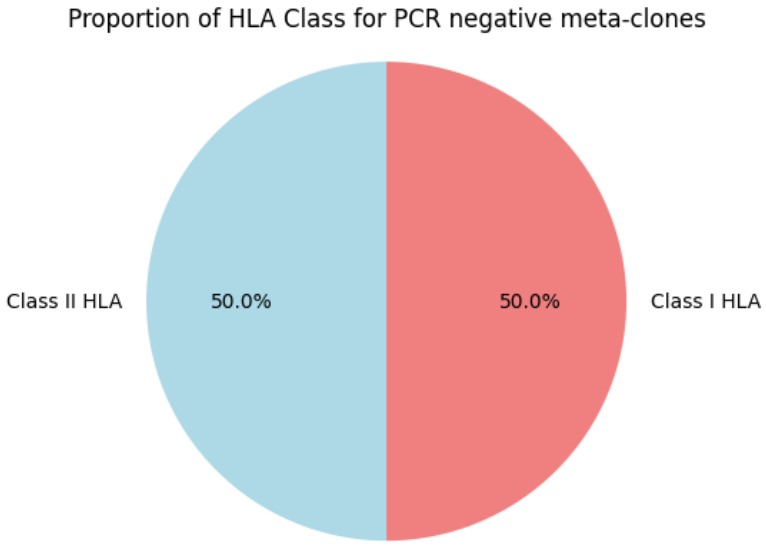
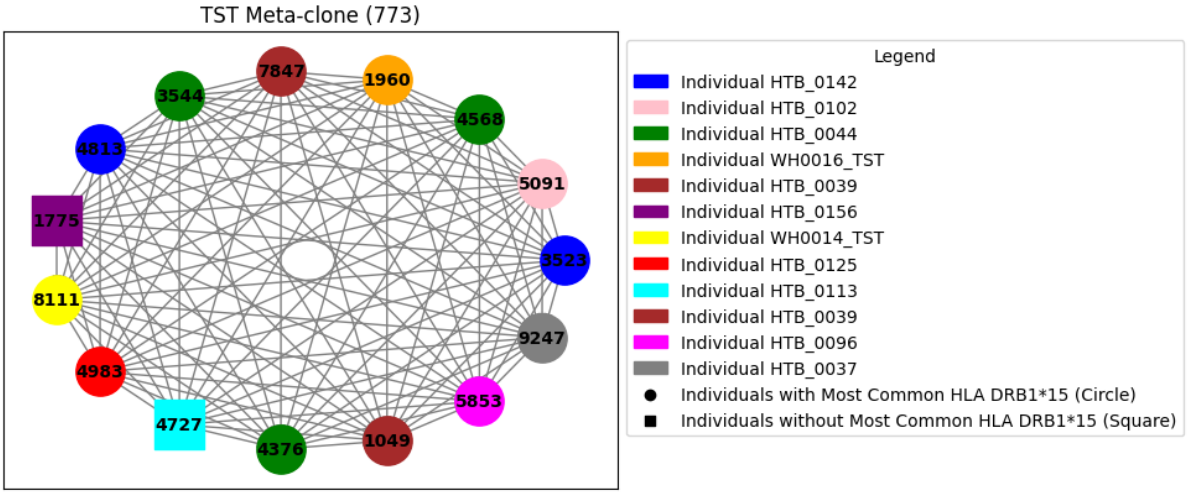


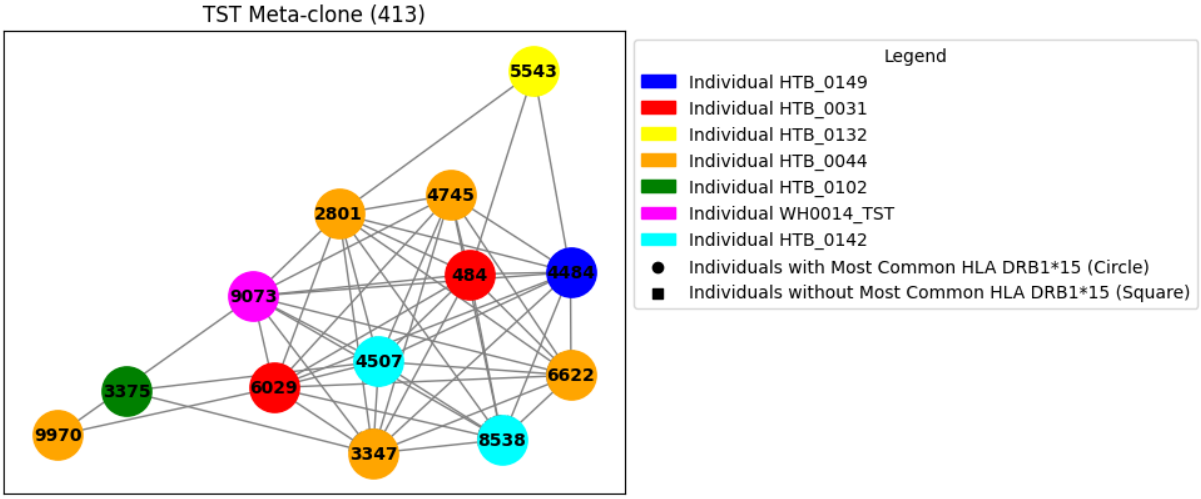
Figure 6: Quantitative and Qualitative Comparison of the TST and PCR Negative Meta-clones furthers biological confidence
(a) *Histogram comparing the TST and PCR Negative Meta-clone Sizes.* The x-axis shows the number of TCR sequences within a cluster. The y-axis shows the number of clusters containing the corresponding number of TCR sequences within. The number of TCR sequences making up TST Meta-clones are consistently equal to or greater than the number making up PCR Negative Meta-clones. Respectively, TST and PCR meta-clones have a median size of 15 and 6, and a mean size of 17.32 and 6.33. **(b)** *Pie chart showing the Class II HLA preference in TST meta-clones.* The TST Meta-clones display the expected preference towards Class II HLA restriction. 86.7% (39) of TST meta-clones were restricted to a Class II HLA and 13.3% (6) of TST meta-clones restricted to a Class I HLA **(c)** *Pie chart showing lacking Class HLA preference in PCR Negative meta-clones.* Of a total 6 PCR negative meta-clones, 3 were Class II HLA restricted and the other 3 were Class I HLA restricted.

Figure 7

a.


TST Meta-clone (773)
7847
1960
3544
4568
4813
5091
1775
3523
8111
9247
4983
5853
4727
1049
4376
Legend
Individual HTB_0142
Individual HTB_0102
Individual HTB_0044
Individual WH0016_TST
Individual HTB_0039
Individual HTB_0156
Individual WH0014_TST
Individual HTB_0125
Individual HTB_0113
Individual HTB_0039
Individual HTB_0096
Individual HTB_0037
Individuals with Most Common HLA DRB1*15 (Circle)
Individuals without Most Common HLA DRB1*15 (Square)

b.


TST Meta-clone (413)
5543
2801
4745
484
4484
9073
4507
6622
3375
6029
9970
3347
8538
Legend
Individual HTB_0149
Individual HTB_0031
Individual HTB_0132
Individual HTB_0044
Individual HTB_0102
Individual WH0014_TST
Individual HTB_0142
Individuals with Most Common HLA DRB1*15 (Circle)
Individuals without Most Common HLA DRB1*15 (Square)

c.

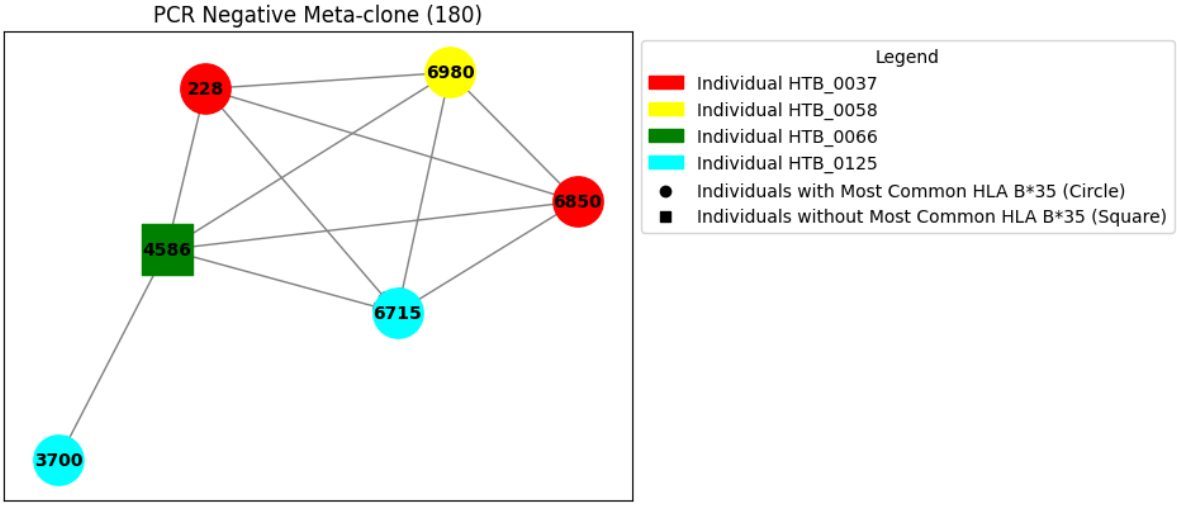

PCR Negative Meta-clone (180)
6980
228
6850
4586
6715
3700
Legend
Individual HTB_0037
Individual HTB_0058
Individual HTB_0066
Individual HTB_0125
Individuals with Most Common HLA B*35 (Circle)
Individuals without Most Common HLA B*35 (Square)

Figure 7 (count.)

d.

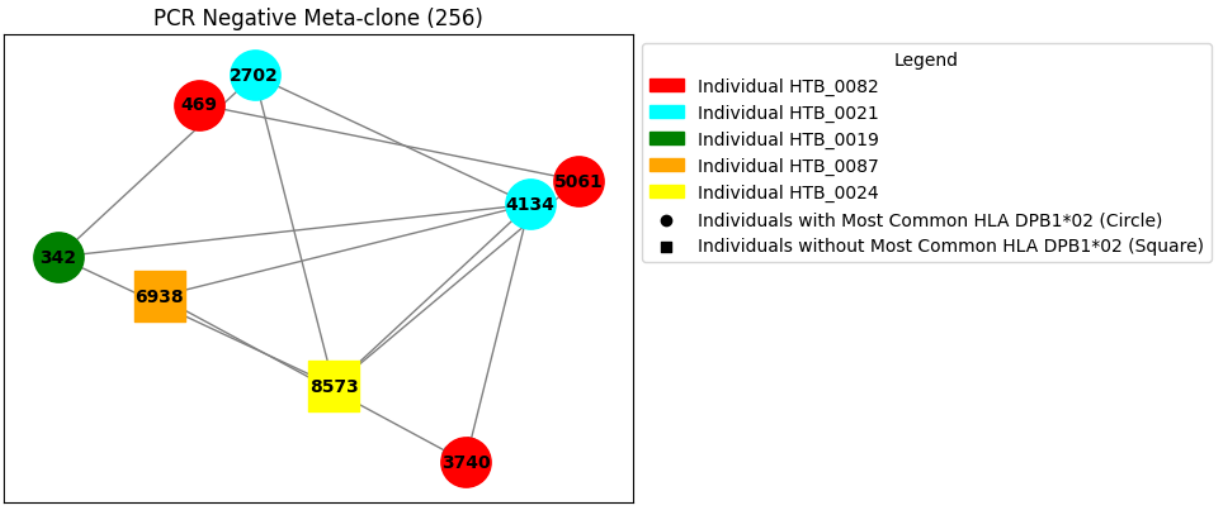
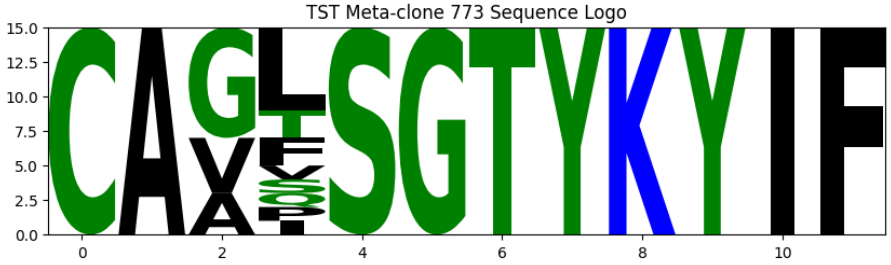


Figure 7: Visualizing TST and PCR Negative Meta-clones gives insight into the Meta-clones' Structures and Compositions

(a), **(b)**, **(c)**, and **(d)** TST meta-clones 773 and 413 and PCR negative meta-clones 180 and 256 were chosen because they showed the most significant HLA restriction, based on their p-value, within their respective TST or control meta-clones. Each node represents a TCR sequence, and the edges connect two similar TCR sequences. Each node is colored based on the individual the TCR sequence was observed in, and the shape of the nodes represent whether that TCR sequence's originating individual carries the restricted HLA of the meta-clone (circles yes and squares no). **(a)** and **(b)** *Visualised TST Meta-clones 773 and 413.* Both meta-clones show a high degree of interconnectedness between the TCR sequences. Even if there are some repeated individuals, the TCR sequences are found to come from a range of different individuals. (**(a)** each TCR sequence is connected to every other TCR sequence of the meta-clone). **(c)** and **(d)** *Visualised PCR Negative Meta-clones 180 and 256*. The TCR sequences are less connected to each other (**(c)** PCR Negative Meta-clone 180 TCR sequence 3700 is only similar to one other TCR sequence, 4586, in the whole meta-clone).

Figure 8

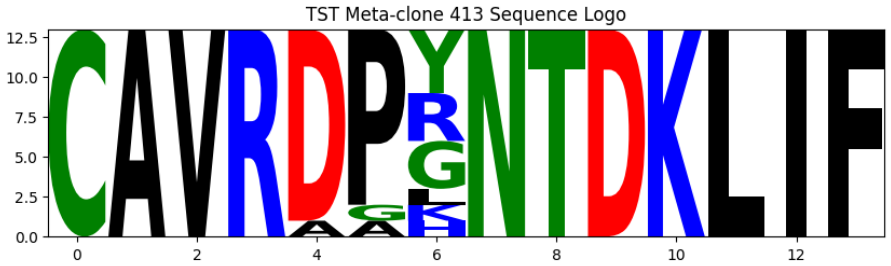
a.



TST Meta-clone 773 Regex Pattern:

[C][A][AGV][FILOPSTV][S][G][T][Y][K][Y][I][F]

b.



TST Meta-clone 413 Regex Pattern:

[C][A][V][R][AD][AGP][GHKLRY][N][T][D][K][L][I][F]

Figure 8: CDR3 Amino Acid Sequence Logos and Regex Patterns Envision How Meta-clones could be Future Biological Tools

(a) and **(b)** polar amino acids are green, basic amino acids are blue, acidic amino acids are red, and hydrophobic amino acids are black. Given that TST meta-clones are consistently identical between respective V and J gene regions, the Sequence logo and Regular Expression are based on the CDR3 amino acid sequences of the TCRs which make up the TST Meta-clone. **(a)** *The CDR3 Sequence Logo and Regex Pattern for TST Meta-clone 773.* Between TCR sequences, amino acids vary in position 2 and 3 of the CDR3 region. **(b)** *The CDR3 Sequence Logo and Regex Pattern for TST Meta-clone 413*. Between TCR sequences, amino acids vary in position 4, 5 and 6 of the CDR3 region.

Table 1

Subgraph	Most Common HLA	p-value	Odds Ratio
773	DRB1*15	9.99630069495029E-31	41.6147843369713
413	DRB1*15	5.46779512062207E-22	115.30763160035800
881	DRB1*15	3.61630956328955E-21	32.211443645444100
32	DRB1*15	2.07060216319365E-17	Inf
470	DRB1*15	3.97569964015759E-16	Inf
260	DRB1*15	1.96895080051988E-15	118.34204295826200
781	DQA1*06_DQB1*03	3.29171842889159E-15	Inf
241	DRB1*15	7.6337953787174E-15	Inf
653	DRB1*15	3.27843722145842E-14	109.23880888455000
901	DRB1*15	3.27843722145842E-14	109.23880888455000
395	DRB1*15	1.46580941992845E-13	Inf
30	DRB1*15	2.10667304415054E-13	20.80739216848570
781	DRB1*12	1.37722013236589E-12	Inf
467	DRB1*09	1.61154722248193E-12	304.83168513641100
193	DRB1*15	2.83514912989478E-12	29.58551073956550
1491	DRB1*15	8.56921553556813E-12	50.0677874054186
273	DQA1*05_DQB1*02	3.54506955527967E-11	19.12068948539690
141	DRB1*15	3.62474962177438E-11	60.68822715808320
691	DRB1*15	5.4048249526735E-11	Inf
1212	DRB1*08	5.7351893017979E-11	Inf
169	DRB1*15	7.04643275791172E-11	33.37852493694570
103	DRB1*15	1.13173774711215E-10	18.206468147424900
726	DRB1*15	1.72562296862597E-10	28.61016423166780
23	DRB1*15	3.45463594789929E-10	36.41293629484990
0	DRB1*15	8.26085911464736E-10	22.252349957963800
830	DRB1*15	1.03788387458407E-09	Inf
101	DRB1*15	1.03788387458407E-09	Inf
917	DRB1*15	1.66112818131701E-09	40.96455333170610
273	DRB1*03	1.99791059801162E-09	14.2401123670077
284	DQA1*05_DQB1*03	2.1266526842331E-09	10.12511955790150
366	DQA1*05_DQB1*02	2.18017262259963E-09	18.89574019733340
243	DRB1*15	2.21227606526998E-09	72.82587258969980
1256	DRB1*15	2.21227606526998E-09	72.82587258969980
9	DRB1*15	1.05657988378911E-08	27.309702221137400
231	DRB1*15	1.05657988378911E-08	27.309702221137400
781	A*24	9.73435078615522E-07	Inf
1491	DQA1*01_DQB1*06	0.0142734288811296	4.174490249787200
1491	DPB1*04	0.0343216228873431	3.479025437755650
467	C*04	0.0468629174978552	6.603287131009690
169	C*07	0.0516709360012928	2.6441581107979800
23	C*07	0.0629450774378154	2.884536120870530
231	DQA1*05_DQB1*03	0.1176094173083630	2.301163535886690
9	DQA1*01_DQB1*03	0.1220876822014790	2.2507698027589000
103	C*07	0.4256549782024960	1.442268060435260
103	C*04	0.6953237786298740	0.8254108913762110

Table 1: Class II HLA restricted TST Meta-clones dominate as more Significantly HLA-Restricted Meta-clones

Ranking the TST Meta-clones by their p-values (smallest to largest) shows those restricted to a Class II HLA majorly rank as meta-clones which HLA-restriction is highly significant, high confidence meta-clones. TST Meta-clones restricted to Class I HLAs only appear as meta-clones with relatively low significance in the HLA restriction, low confidence meta-clones.