1. **Introduction**

Immune checkpoint inhibitors (ICI) have revolutionized the paradigm for cancer therapy. However, only some patients are responsive to ICI and suitable for treatment. Determining what subset of patients would benefit from ICI is an important problem subject to intensive research. ICI response prediction involves the use of multiple biomarkers, including neoantigen, T-cell fraction, and tumor mutational burden. With advances in next generation sequencing, prediction of these biomarkers became possible through whole exome sequencing (WES)1-3. However, a method for predicting T-cell clonality from WES data is still to be developed, despite its prognostic value in efficacy of immunotherapy4. Although clonality can be predicted from TCR sequencing methods, it requires additional testing and comparatively large DNA input (250–500 ng for a typical clinical assessment)5, adding to the cost of immunotherapy.

This project focuses on predicting T-cell clonality from WES data, which would provide cost-effective methods for ICI response prediction inclusive of clonality. The first aim of this project is to align WES reads and analyze their read counts. This is essential because the read count trends will determine T-cell clonality. The second aim is to detect for homology in TCR sequences. This is i because homology in TCR would suggest that there is low statistical power in the association between clonality and response. The third aim of this project is to validate results using TCR repertoire analysis tools. Validation with other tools is an integral part of confirming the viability of a new approach.

This study concerns a cancer cell line derived from a lymphoblast sample which contains very few clones. Therefore, the scope of T-cell clonality assessment is limited to homogenous or weakly heterogeneous T-cell samples. Results cannot be applied to highly heterogeneous samples such as tumors. Additionally, validation of T-cell clonality is limited to indirect methods via existing software. Thorough validation of WES based approaches would require matched WES and TCR sequence samples, but they were unavailable in this study.

V(D)J recombination is an important process in T-cell development. Genes undergo random deletion or insertion prior to being translated and assembled into receptors. T-cell clonality can thus be identified from this clone-specific DNA deletion pattern in TCR.6 Assuming extreme ends of the TCR gene to have normal background rates of deletion unaffected by V(D)J recombination, deletion patterns could be identified by aligning reads to reference TCR genes and measuring read depth in terms of genome position.

TCR genes are classified into four subtypes: alpha (TRA), beta (TRB), delta (TRD) and gamma (TRG). A heterodimeric αβ T-cell receptor or γδ T-cell receptor is formed by the combination of pairs. In all four chains, CDR3 regions exhibit highest variability in sequence and are the principal determinants of receptor binding specificity. This is because CDR3 is determined by both V and J genes, as well as insertions and deletions at their junctions7. Naturally, CDR3 regions are key in determining T-cell clonality.

1. **Materials and Methods**

**Sample data**

The study was conducted with PEER cell line data of hematopoietic and lymphoid tissue. PEER was established from peripheral blood of T-cell acute lymphoblastic leukemia (T-ALL) patients and cultured. Next generation sequencing was performed using the Twist Bioscience for Illumina Exome 2.0 Plus Panel, with read lengths of 150bp. WES data was obtained from the Cancer Cell Line Encyclopedia (CCLE) dataset8 and downloaded via the Sequence Read Archive9 (SRA) under accession number SRR8619000. The cell line was confirmed to have abundant T-cells (95.7% T-cell fraction) and gone through V(D)J recombination.

Paired-end WES data was deinterleaved using in-house code. All reads were aligned to the indexed human reference genome (GRCh38) using the Burrows-Wheeler Aligner10 (BWA) algorithm. Reads were extracted using SAMtools11 and downsampled according to alignment rate (>0.2). As a general control step, exons with less than 15 were excluded from analysis due to low coverage. BED files were downloaded from Twist Bioscience and used for recalibration. False positive alignments to regions that were not targeted in the exome sequencing panel were excluded from analysis.

**Bioinformatics Pipeline**

TRAV and TRAJ genes were retrieved from Ensembl Genes 108 using the BioMart12 R package. Annotated data of these genes were obtained using the UCSC table browser13. Read counts were plotted and sorted according to genomic positions. Results were visualized with the Matplotlib python library. In order to test for homology between genes, TRAV and TRAJ genes were mutually aligned according to BWA scoring system with the pairwise2 module of Biopython. Alignment scores were normalized and results of the alignment were visualized using the Seaborn python Library.

Clonality scores were calculated using read counts aligned to TRAV and TRAJ genes. Thus clonal candidates were determined by the peaks in the read count bar graph. Results were validated using MiXCR14 software though procedure outlined in Bolotin et al. WES reads were aligned to V(D)J and constant region reference segments. Then T-cell clonality was determined by assembling the CDR3 region. Non-targeted partial assembly was performed as sequences were not enriched nor targeted. The resulting CDR3 sequence with highest CDR3 scores were used for validation. MiXCR infers insertion and deletion sequences with best matches in existing CDR3 sequence databases. Therefore, insertion in the N region and deletions at the end of V genes and beginning of J genes were not penalized.

**Code availability**

All raw data and code can be accessed from github:

<https://github.com/zunuan/Junior-Independent-Study>

1. **Figures and Figure Legends**
2. **B.**

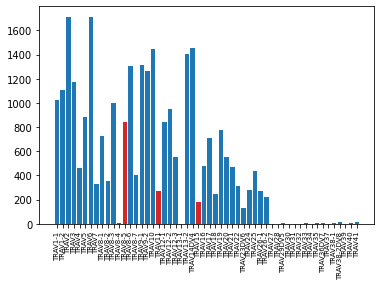
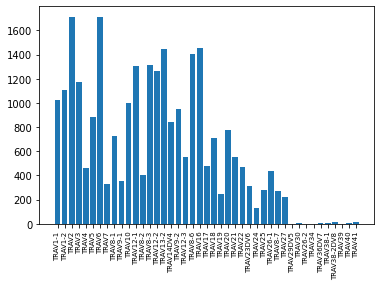
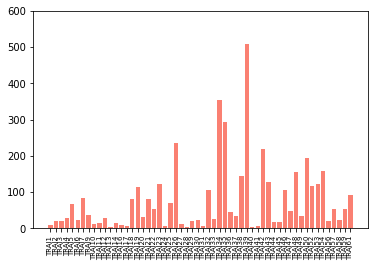
****

Fig 1. Raw alignment and recalibration of WES reads.

A. Reads before recalibration. Untargeted genes by the exome panel are shown in red.

B. Read alignment after recalibration with BED file. Reads have been excluded and realigned.

1.  **B.**

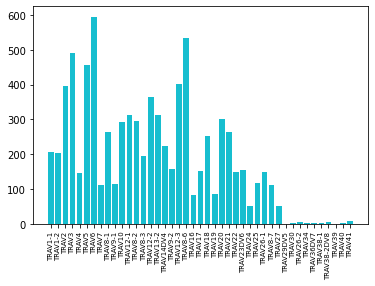


Fig 2. Read count of TRAV and TRAJ genes

A. Read counts of TRAV genes sorted according to genome position. (r = 0.55)

B. Read counts of TRAJ genes sorted according to genome position. (r = 0.32)

1. **B.**

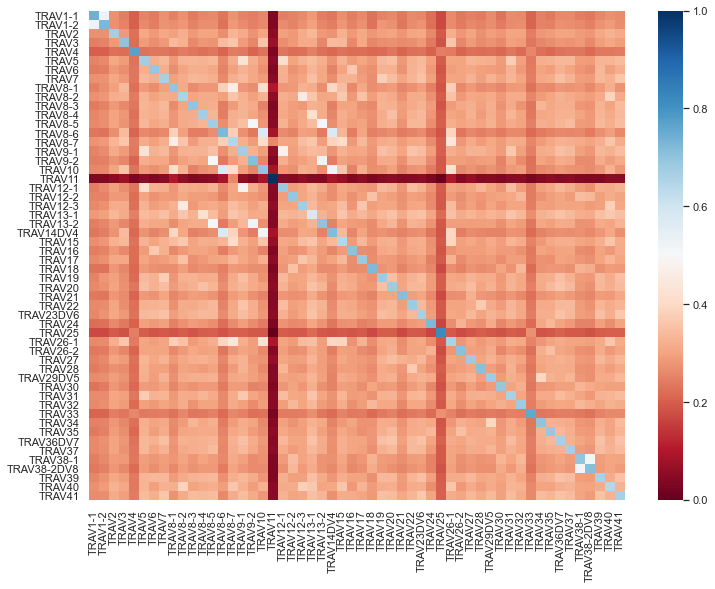
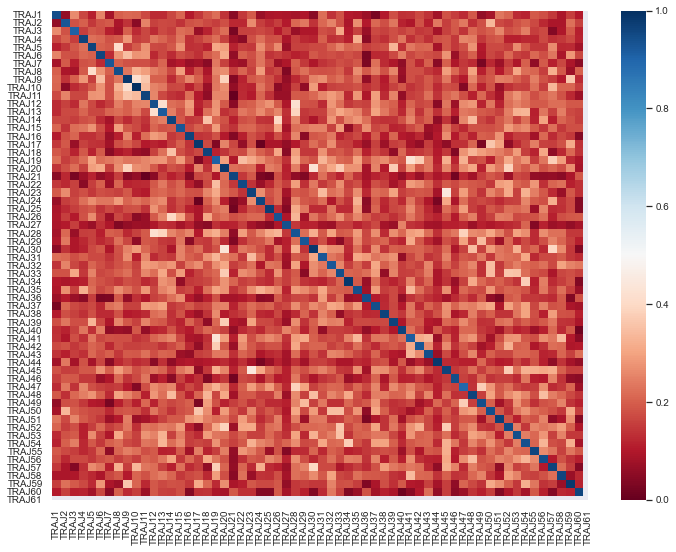


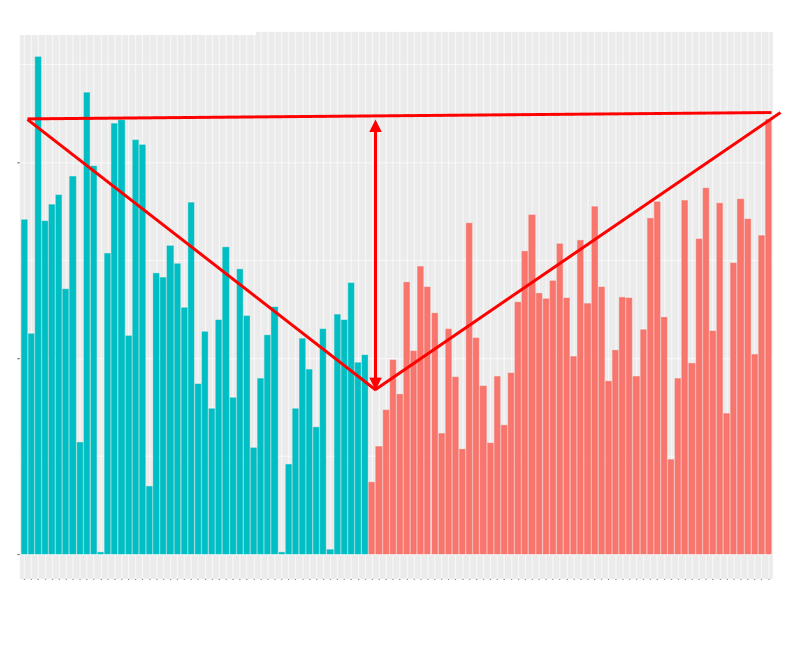
Fig 3. Homology heatmap of TRAV and TRAJ genes

A. Mutual alignment of TRAV genes to detect homology.

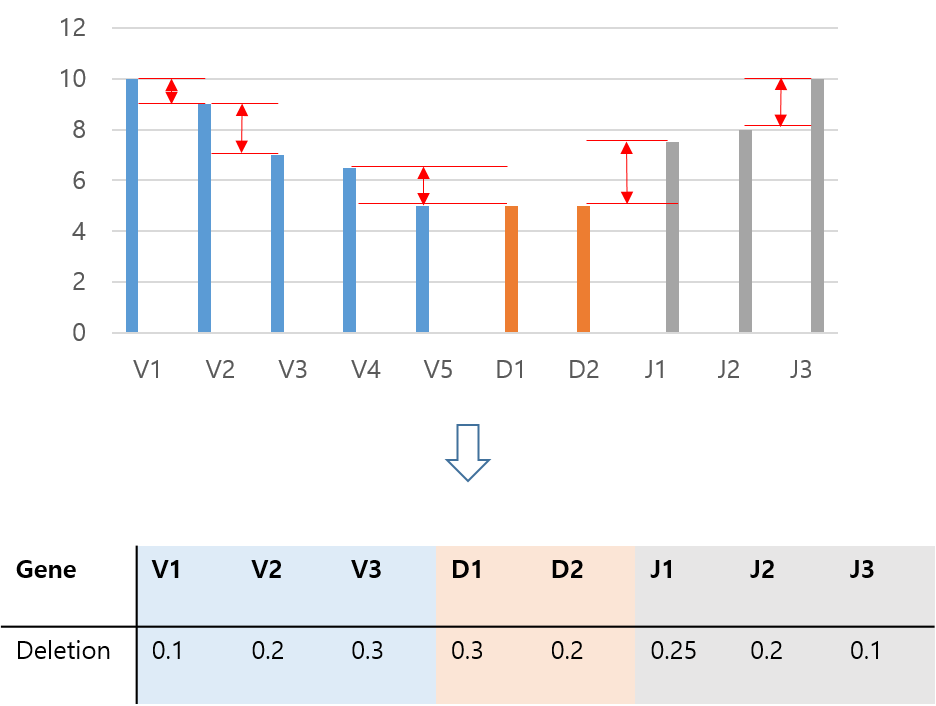
B. Mutual alignment of TRAj genes to detect homology.

**Table 1.** Underlined sequences are preserved, red sequences have been deleted after V(D)J recombination, and blue sequences have been inserted after recombination. Clonality scores are read counts followed by MiXCR hit scores in parenthesis.

|  |  |  |
| --- | --- | --- |
| **Gene** | **Clonality score** | **Sequence** |
| **TRAV6** | 1714  (495.7) | **GGGTGAAGAGCCAAAAGATAGAACAGAATTCCGAGGCTCTGA**  **ACATTCAGGAGGGTAAAACGGCCACCCTGACCTGCAACTATAC**  **AAACTATTCTCCAGCATACTTACAGTGGTACCGACAAGATCCAG**  **GAAGAGGCCCTGTTTTCTTGCTACTCATACGTGAAAATGAGAAA**  **GAAAAAAGGAAAGAAAGACTGAAGGTCACCTTTGATACCACCC**  **TTAAACAGAGTTTGTTTCATATCACAGCCTCCCAGCCTGCAGAC**  **TCAGCTACCTACCTCTGTGCTCTAGACA** |
| **TRAJ39** | 508  (220.8) | **GGAATAACAATGCCAGACTCATGTTTGGAGATGGAACTCAGCT**  **GGTGGTGAAGCCCA** |
| **CDR3** | - | **TGTGCTCGCTGGTAACAATGCCAGACTCATGTTT** |

****

Supplementary Fig 1. Schematic of ideal read count trends

****

Supplementary Fig 2. Schematic of clonality assessment with non-negative matrix factorization

1. **Results and Discussion**

**Data cleaning and recalibration**

In somatic cells, chromosomes are assumed to be diploid and have two copies of most genes. Variations in these copy numbers may often indicate development of tumor or certain diseases2. However, this assumption does not hold in T-cells, where V(D)J recombination occurs during development. Therefore, it is likely that there is a region that is almost always lost in recombination. Raw WES alignment results of TRAV genes clearly exhibit this trough region of deletion (Fig. 1a). TRAV genes located downstream only exhibit background levels of alignment (<12 reads). This is consistent with previous findings that endonucleases recognize recombination signal sequences (RSSs) and cleave genes for recombination15. RSSs cleave downstream TRAV genes and upstream TRAJ genes. Therefore, determining the regions in which RSSs had bound could be determined with read count analysis.

Yet, there are three upstream TRAV genes that had background alignment rates (TRAV8-4, TRAV13-1). Comparing aligned gene regions with targeted regions of the Illumina Exome 2.0 Plus Panel, some regions appeared to be untargeted in the sequencing process. These regions included the upstream TRAV genes with background alignment and false positive alignments in 9 additional TRAV genes (marked in red; Fig. 1a). Recalibration of reads were performed by excluding these reads from analysis. Additionally, as a general control step, exons with low read depth of less than 15X were also excluded from analysis due to low coverage (Fig. 1b). Recalibration results suggest that RSSs have bound to a region between TRAV27 or TRAV28 during recombination, but there is uncertainty to which as TRAV28 was not targeted. Resolution of determining clonality through WES is thus dependent on the target coverage of exome sequencing panels over T-cell receptor genes.

**Read count analysis**

In T-cell fraction determination tools such as ExTRECT2, the read depth ratio is calculated based on copy numbers of genes at the extreme beginning and end of the TCRA gene. This is based on the assumption that genomic regions at the extreme ends of the TCRA gene would have background copy number variations values unaffected by any V(D)J recombination. This assumption was adapted for purposes of this study. In analyzing read counts, the extreme ends of upstream TRAV and downstream TRAJ genes were assumed to have read counts unaffected by V(D)J recombination. Since RSS binding at the extreme beginning (TRAV1-1) would cause all V genes to experience deletion and fail to develop a functioning T-cell receptor, only background rates of variation should occur. Similarly, the extreme downstream TRAJ gene (TRAJ60) would also be unaffected by V(D)J recombination.

It was hypothesized that TRAV genes should exhibit identical read count patterns due to frequent deletion closer to the CDR3 region. Together with similar patterns in TRAJ genes, a V-shape read count trend was expected to be observed (Supplementary Fig. 1). Fitting a linear model to assess read count trends, TRAV genes were observed to have moderate negative correlation (r=-0.55; Fig. 2a), indicating variance in read counts for each gene. However, no significance was observed in TRAJ genes as it showed weak positive correlation (r=0.32; Fig. 2b). These results suggest errors in read alignment.

There appears to have been softclips in mapping reads in the TCR region. Extracting unmapped reads with SAMtools show that bases in the 5’ and 3’ ends of the reads are not part of the alignment. This causes a read sequence to not be aligned from the beginning base pair to the to the end, clipping off either end. Softclipping is most likely the result of V(D)J recombination near the TCR region.

**Homology detection**

Underlying features for this error in alignment were explored. One candidate source of error was homology between genes. Homology between genes may cause reads to be mapped to incorrect genes as their sequences are similar. DNA sequences, rather than protein standard sequences, were compared to detect sequence similarity. As BWA attempts semi-global alignment, local aligners such as BLAST were unavailable. Therefore, the BWA penalty score system was implemented through the pairwise2 module of Biopython and global alignment was performed to calculate scores.

The heatmap shows that alignment scores for all mutual comparisons were negative, excluding self-alignments (Fig. 3). Alignment scores were normalized to account for the different lengths of each TRAV and TRAJ gene. It should be noted that the significantly low alignment score of TRAV 11 is due to the length of its sequence (1520bp; Fig 3a) in comparison to the mean (600bp). Other relatively long sequences such as TRAV4, TRAV25, and TRAV33 also show significantly lower alignment scores. Unlike well-defined identity measurement in protein sequences through scoring matrices such as PAM250 or BLOSUM62, nucleotide sequences lack such measurement systems. Yet, largely negative alignment scores (mean, -637) suggest that there is no homology between any TRAV or TRAJ genes, indicating dissimilarity between both TRAV and TRAJ sequences. This suggests that errors in alignment were not caused by sequence similarity, but by limitations in the sequencing method.

**Limitations in WES**

Conventional methods of T-cell clonality assessment is done by determining composite rearrangements of the specific V and J genes. This can be determined by TCR sequencing, which enriches the target sequences with amplicons or hybrid capture and sequences these 250bp-length reads. However, WES reads are only of lengths 50 ~ 150bp. Despite the PEER cell line having read lengths of 150bp, there seems to be limitations in alignment quality due to relatively short read lengths.

In targeted methods such as TCR sequencing, TRD sequences are often sequenced instead of TRA. Colonotypes can be determined by sequences of any chain: TRA, TRB, TRD, and TRG. TRD is used because it has the least amount of V genes (8 in TRD; 54 in TRA) and J genes (4 in TRD; 61 in TRA), requiring the least amount of primers5. In untargeted approaches such as WES, TRA genes are suitable for clonality assessment due to the lack in need for specific primers and the polymorphic nature of TRA. However, the tradeoff is that TRAJ genes are very short in length (50 ~ 70bp). Alignment to such short references is innately more difficult with untargeted methods which are lower in accuracy. This is consistent with the study results that exhibited good alignment for TRAV genes with longer reference (250 ~ 1500bp) but bad alignment for TRAJ (Fig. 2).

Finally, the whole exome panel may be prone to capture kit bias. This bias has been previously described by Wang et.al in TCGA samples, which is inclusive of TRA genes16. This bias leads to coverage in certain exons to be lost, hindering read counts as can be seen in Fig. 2a. Possible solutions to these problems will be discussed in later sections.

**Validation with MiXCR**

Results of read count alignment were validated with MiXCR, a T-cell receptor repertoire analysis tool. Read count peaks of the PEER cell line sample are TRAV6 (1714 reads) and TRAJ39 (508 reads). Results were validated using MiXCR software as described previously. Multiple CDR3 candidates were assembled, and the CDR3 sequence with highest integrated V hit scores and J hit scores was extracted for validation. This validation sequence was locally aligned against all TRAV and TRAJ genes.

Results showed that the validation sequence had the highest alignment scores with TRAV6 and TRAJ39, identical to that of the read count peaks of the PEER cell line sample. Table 1 describes the entire sequence of these genes and the assembled CDR3 region. Regions downstream of TRAV6 and upstream of TRAJ39 have been deleted (marked in red; Table 1). This is consistent with previous discussions in read count analysis. Furthermore, a string of sequences has been added in between V and J genes in CDR3, a result of V(D)J recombination. Most importantly, sequences from V and J genes have been completely conserved in the CDR3 region (underlined; Table 1). This evidence strongly suggests that peaks determined by read count analysis may be used to determine what genes compose the CDR3 region, the key to clonality assessment.

**Conclusions**

In summary, this study aligned WES reads of the PEER cell line sample to determine T-cell clonality. TRAV genes had moderate correlation in reads, while there was poor alignment in TRAJ genes. The cause for such poor alignment was hypothesized to be homology between sequences, but no sequence similarity was observed. Therefore, is seems that the challenge lies in innate limitations of WES methods. Yet the WES-based approach still exhibits potential in determining clonality, as genes with peak read counts matched with CDR3 assemblies with MiXCR..

1. **Future work**

Two approaches can be taken to improve on the results of this study—sample diversification approach and a bioinformatics approach. First, the sample could be diversified. As mentioned previously, the PEER cell line sample used in this study contains very few clones as it was derived from a lymphoblast. Further, there is no public TCR sequencing data for validation. Ideally, a monoclonal T-cell sample sequenced by WES with matched TCR sequencing data would aid in determining the relationship between read counts and T-cell clonality. If such a quality sample could be obtained, simply going though similar protocols in this study would be very insightful.

Second, improvements in bioinformatics methodology may be used. In a general sense, WES-specific alignment methods can be developed. Developed TCR detection tools such as MiXCR and IMSEQ are based on transcriptome data. Development of TCR detection methodology for genomic data, perhaps utilizing De Brujin graphs or k-mer methods, would be a possible solution.

More specifically, attempts to improve on the results of read counts can be made. Methodology to redistribute unaligned reads and reads with low alignment scores to the appropriate gene can be developed. Ideally, read counts would follow a linear scheme (Supplementary Fig. 1). Another method to redistribute the reads is to perform N-sequence insertion before alignment. One of the major challenges in alignment is variable insertions and deletions in the sequences due to V(D)J recombination. A greedy approach to this problem would be adding a string of N-sequences that map to all nucleotides without penalty. This way, we can mitigate the effect of large penalties being assigned to correct alignments, improving the read count scheme.

Once read counts have been well-distributed, clonality of cancer samples can also be determined. While cancer samples would indeed be heterogeneous, the difference between read counts of V and J genes could be transformed into a matrix though non-negative matrix factorization (Supplementary Fig. 2). Reads could be clustered to determine clonalities of a heterogeneous sample.

Finally, once clonality of a cancer sample can be determined through WES, a TCR sequencing-free ICI prediction workflow could be constructed. Further, this technology could be applied to previous WES samples for determination of clonality. A collective validation of research done prior to TCR-seq development would also be made possible.

1. **References**

1. McGranahan, N. *et al.* (2016) Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, 351, 1463-1469.

2. Bentham, R. *et al.* (2021) Using DNA sequencing data to quantify T cell fraction and therapy response. *Nature*, 597, 555-560.

3. Vilimas, T. (2020) Measuring Tumor Mutational Burden Using Whole-Exome Sequencing. *Methods mol biol*, 2055, 63-91.

4. Valpione, S. *et al.* (2021) The T cell receptor repertoire of tumor infiltrating T cells is predictive and prognostic for cancer survival. *Nat Commun*, 12, 4098.

5. Mahe, E. *et al.* (2018) T cell clonality assessment: past, present, and future. *J Clin Pathol*, 71, 195-200.

6. Schatz, D. *et al.* (2011) Recombination centres and the orchestration of V(D)J recombination. *Nat Rev Immunol* 11, 251–263.

7. Springer, I. *et al.* (2021) Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Front Immunol* 12:664514.

8. Barretina, J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607.

9. Leinonen R. *et al.* (2010) International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res* D19-21.

10. Li, H. *et al.* (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5), 589–595.

11. Danecek P. *et al.* (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, 10:2.

12. Smedley, D. *et al.* (2009) BioMart – biological queries made easy. *BMC Genomics*, 10:22.

13. Karolchik, D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, D493-6.

14. Bolotin, D. *et al.* (2015) MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods, 12, 380-381.

15. Blanco B. *et al.* (2011) Control of V(D)J Recombination through Transcriptional Elongation and Changes in Locus Chromatin Structure and Nuclear Organization. Genet Res Int.

16. Wang G. *et al.* (2018) Whole-exome sequencing capture kit biases yield false negative mutation calls in TCGA cohorts. *PLoS One* 3;13(10).