

WES analysis of PDO and tissue samples in PDAC

Kane Toh*

2024-01-17

Contents

1 Methods	1
1.1 Analysis of SNPs/INDELS in whole Exome sequencing data	1
1.2 Copy number alterations	2
1.3 Microsatellite instability	2
1.4 HLA class I typing	2
2 Results	3
2.1 Documentation of somatic variant calling steps	3
2.2 Comparing against the TCGA PAAD dataset	3
2.3 Somatic SNP/INDEL mutational landscape	3
2.4 Validation of KRAS mutations with Sanger Sequencing	7
2.5 Unsupervised clustering of PDO and patient samples	7
2.6 Copy number alterations	13
2.7 Tumor mutational burden (TMB) score	13
2.8 Microsatellite instability	13
2.9 HLA typing result	13

1 Methods

1.1 Analysis of SNPs/INDELS in whole Exome sequencing data

The nf-core/sarek (v3.1.2) pipeline from the nf-core collection of workflows, which follows the GATK best practices, was used to call short somatic variants (SNVs/INDELS) and copy number alterations (CNAs). The complete parameter configuration file is stored in the `nf-params-pad.json` file in JSON format, and the pipeline was run with `nextflow run nf-core/sarek -r 3.1.2 --input wes_input_full.csv -params-file nf-params-pad.json -profile docker -resume`.

*Genomics and Data Analytics Core (GeDaC), Cancer Science Institute of Singapore, National University of Singapore; knetoh@nus.edu.sg

In brief, reads were aligned to the human reference genome (GRCh38) using the Burrows Wheeler Aligner (BWA) with default parameters. Next, the bam files were processed by marking duplicates and carrying out quality recalibration at the base level. Mutect2 and Strelka2 were used to identify short mutations which include single nucleotide variants (SNVs) and insertion/deletions (INDELs), and subsequently annotated with VEP. Variants identified in either caller were retained for downstream analysis if they met the following criteria:

1. Passed the caller's internal filters
2. Have a MAX_AF score less than 0.001, if the value is present in the CSV INFO column ("maxAF < 0.01 or not MAX_AF"). The MAX_AF score represents the highest allele frequency observed in any population from 1000 genomes, ESP or gnomAD.

The filtered VCF files across both callers were then combined and converted to MAF format with the `vcf2maf` tool. Downstream analysis was carried out in R using the maftools packages, and several heatmaps were drawn with the ComplexHeatmap package. The tumor mutational burden (TMB) score was computed with the `tmb` function in maftools, with the twist exome panel capture size of 36.5 MB supplied to the `captureSize` parameter.

Mutational data from the TCGA_PAAD cohort was retrieved from maftools with the `tgcLoad` function. The sample with the tumor sample barcode of 'TCGA-IB-7651-01' was excluded from the analysis as it has around two orders of magnitude more mutations identified than the rest of the samples, making it a clear outlier.

1.2 Copy number alterations

ASCAT (v3.0.0) was used to detect copy number alterations in the 48 WES samples with tumor-normal pairs, by accounting for the admixture of non-neoplastic cells and tumor ploidy levels.

Our copy number profile summarisation follows the classification defined by Steele et al., (2022). We classified segments based on the sum of the major and minor alleles (TCN), and defined the CNAs as follows with slight modification from Steele for clarity:

- Deletion: TCN=1
- Gain: TCN=3-4 (equivalent to minor gain as defined by Steele and colleagues)
- Amp: TCN=5-8 (equivalent to major gain as defined by Steele and colleagues)
- Wild type: TCN=2.

1.3 Microsatellite instability

MSIsensorpro (v1.2.0) was used to detect instances of microsatellite instability in the 48 WES samples with tumor-normal pairs.

1.4 HLA class I typing

HLA class I typing was performed with Optitype v1.3.1, implemented using nf-core/hlatyping v2.0.0 of the nf-core collection of workflows (Ewels et al., 2020). The coverage plots (`*coverage_plot.pdf`) that were output from Optitype show clear exome enrichment. Coverage is high on both exons 2 and 3 of the HLA Class I loci, as shown by the large green areas (paired-end reads where both ends are aligned to the allele sequence with without any mismatches) that concentrate on the grey bands representing the locations of exons 2 and 3. This indicates that high quality data was used for predicting the HLA genotype.

2 Results

2.1 Documentation of somatic variant calling steps

Here, we document the short somatic variant calling filtering steps in greater detail. We ran the nf-core/sarek pipeline with the Mutect2 and Strelka2 callers in matched tumor-normal mode for 48 out of 50 samples. 2 of the samples from patients PCA35 and PCA80 did not have a matched PBMC normal sample. Here, we assess the number of called variants that pass each of the caller's internal filters in Figure 1A. A total of 8503 and 76335 variants passed Mutect2 and Strelka2's internal filters respectively. Next, amongst known variants annotated in population databases, we applied a max population allele frequency (maxAF) filter to discard common genetic polymorphisms (Figure 1B). This filter retains rare variants at a variant allele frequency (VAF) $\leq 0.1\%$, and thus guards against germline artefacts and emphasizes clinically relevant mutations (Refer to the [Methods](#) section). The number in black within each colored vertical bar shows the number of mutations retained for each caller per sample, whilst the blue number at the top of each bar sums the number of mutations from both callers. Finally, Figure 1 shows the number of variants that are retained after removing duplicates from both callers and removing non-synonymous/silent variants. Overall, we notice that the 2 samples without the paired normal have a greater number of mutations called, with 500 and 501 SNPs/INDELS called for PCA35 and PCA80 respectively. In addition, we note that the PCA117_117B_CM-sample had a large number of SNPs/INDELS called.

2.2 Comparing against the TCGA PAAD dataset

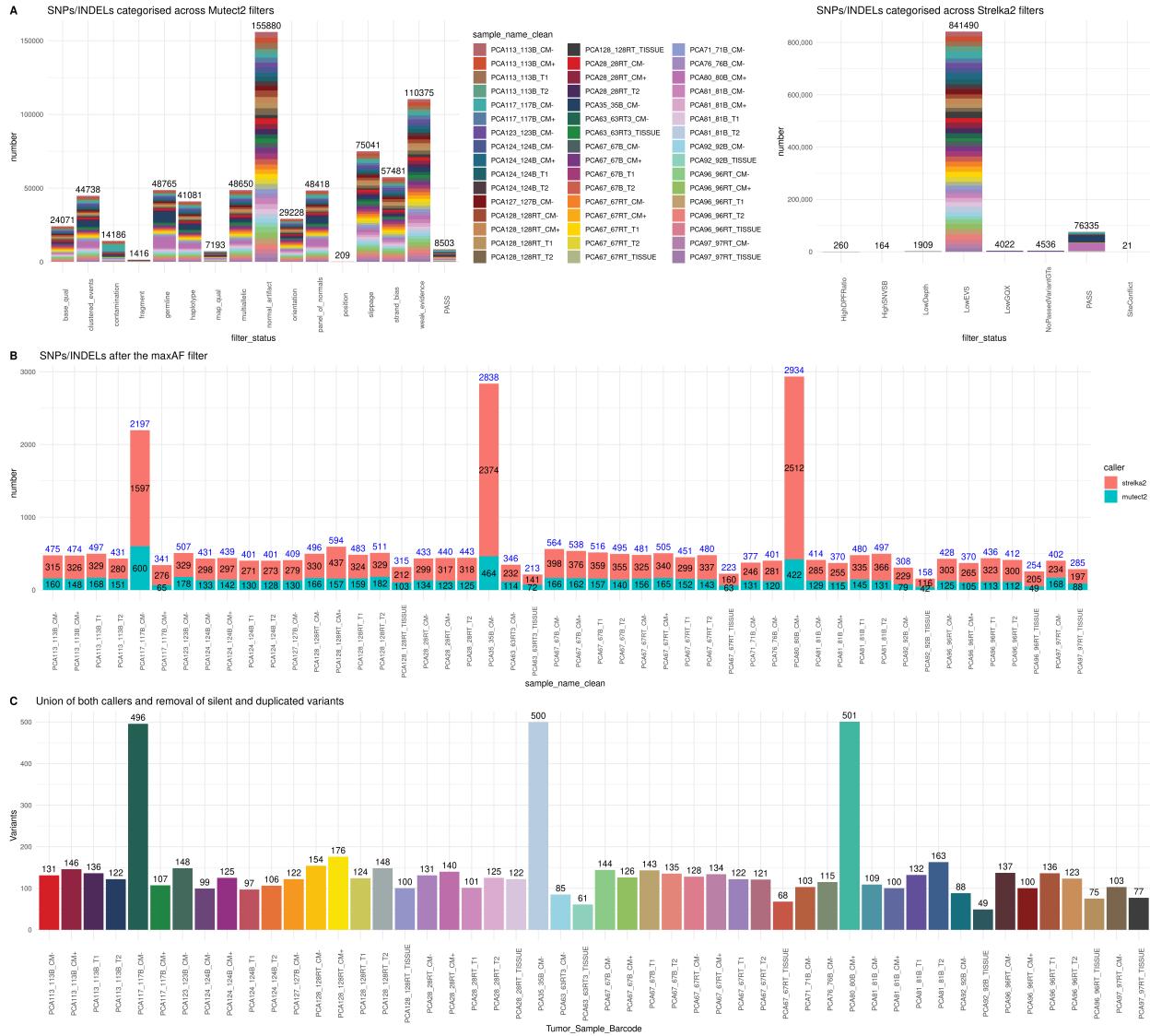
To provide external validation of the variants called in our PDAC_TISSUE dataset, we compared our somatic variants and gene mutations against the corresponding public [TCGA_PAAD dataset](#). As shown in the upSet plot (Figure 2), we found that around 26% of all mutated genes were shared between both datasets. A gene is considered to be shared across both datasets if it is present in at least one sample in each study. Of the 111 mutated genes in our dataset that were also listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (Figure 3), 50 of them (31%) were also shared with the COSMIC genes in the TCGA dataset.

Importantly, focusing only on the mutations in the COSMIC database, the co-oncoplot in Figure 4 show that the known driver genes such as KRAS (67% TCGA vs 92% PDO/TISSUE), TP53 (62% TCGA and PDO/TISSUE), CDKN2A (21% TCGA and 26% PDO/TISSUE) and SMAD4 (21% TCGA and 20% PDO/TISSUE) driver gene mutations were recapitulated in our PDAC/TISSUE samples. On the other hand, our PDO/TISSUE samples have a greater proportion of ARID1A (5% TCGA and 32% PDO/TISSUE) and ATM (4% TCGA and 20% PDO/TISSUE) mutations.

When we assessed the tumor mutational burden (TMB) estimates between the PDAC/TISSUE and TCGA datasets (Figure 5), we find that our dataset has a higher median TMB score than the TCGA dataset (3.38MB: see `tmb_pdac.png`). Our 7 tissue samples are in the upper range of the TMB scores of the TCGA tissue samples (see `tcgaCompare_tmb_tissueOnly.png`) In conclusion, our somatic variant calling results are consistent with the TCGA_PAAD dataset in terms of the mutations, and have a higher tumor mutational burden than the TCGA samples.

2.3 Somatic SNP/INDEL mutational landscape

The oncplot shown in Figure 6 shows the mutated genes that were also present in the COSMIC database, emphasising known mutations with causal implications in carcinogenesis. Across the 50 PDAC samples, we observed that 46 samples (92%) possess the KRAS mutation . Of the 46 samples with the KRAS mutation, 24 samples (52%) exhibit G12V, 16 (35%) G12V and 5 G12R (11%). 62% of the samples have mutations in the TP53 gene, with each patient harboring a different mutation (See `TP53_lollipop.png` for the range of mutations displayed). Other genes that were mutated in relatively high proportion include the ARID1A (32%), ATM (20%), CDKN2A (26%) and SMAD4 (20%) mutations.



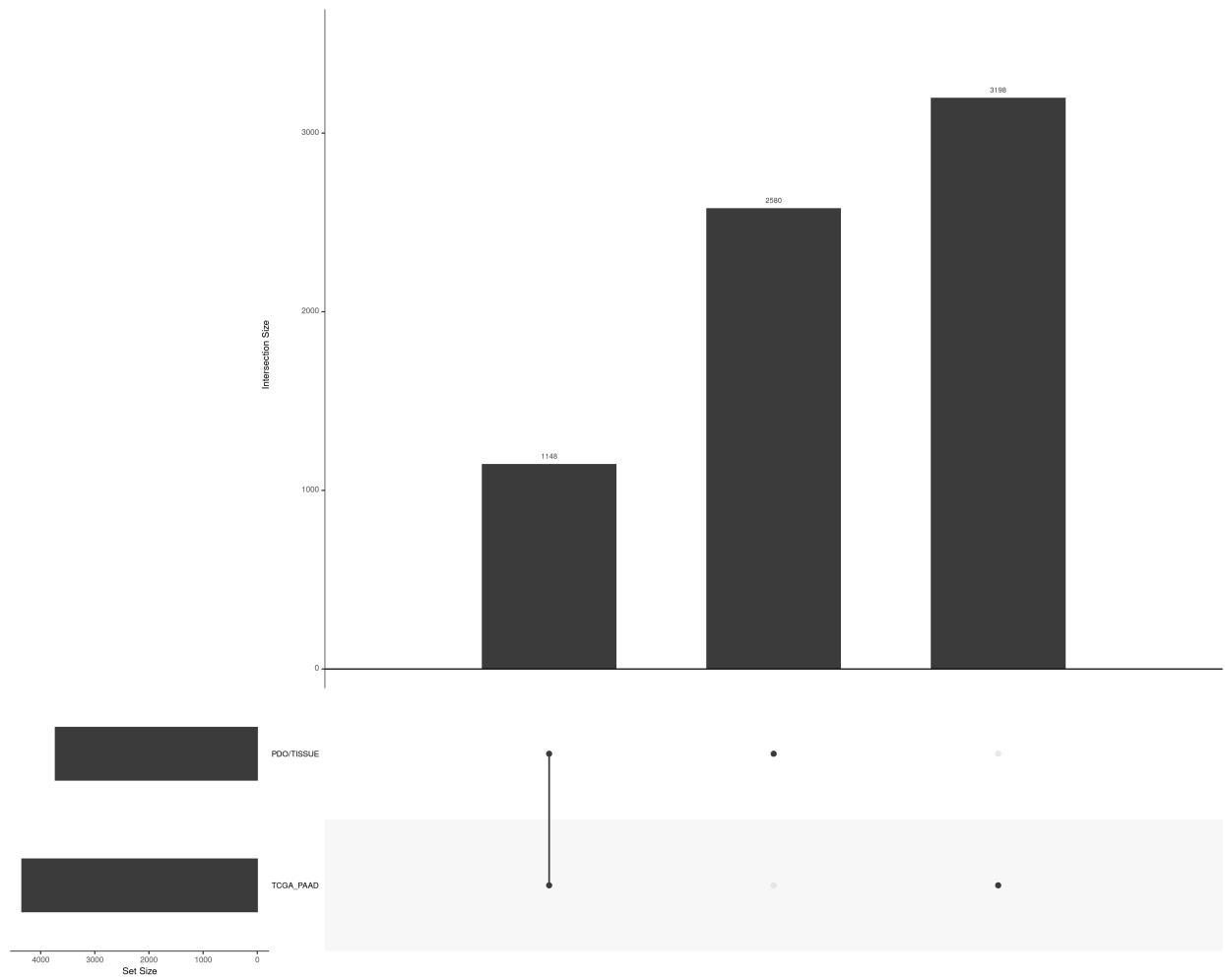


Figure 2: Number of mutations found in TCGA vs PDAC/TISSUE data

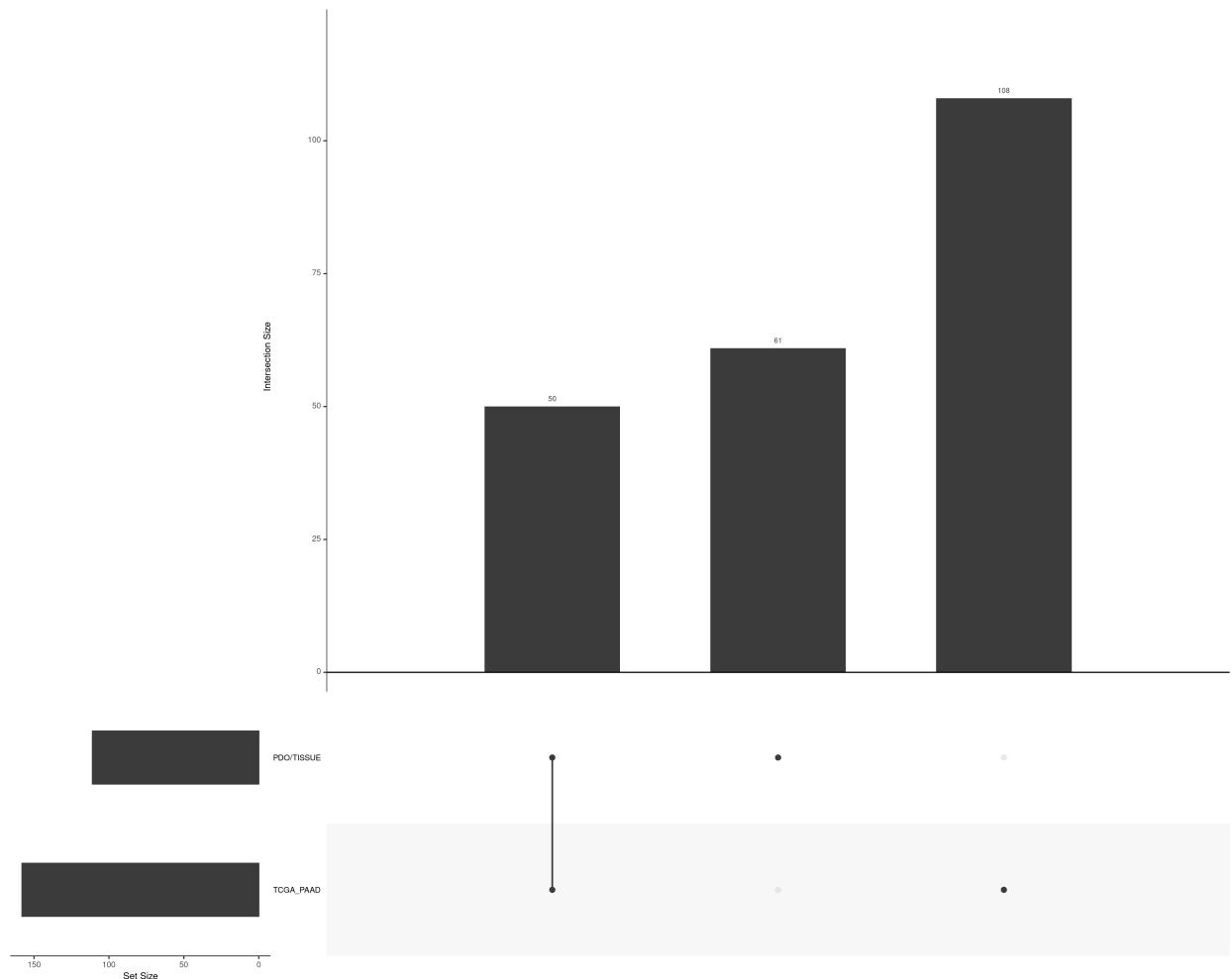


Figure 3: Number of COSMIC mutations found in TCGA vs PDAC/TISSUE data

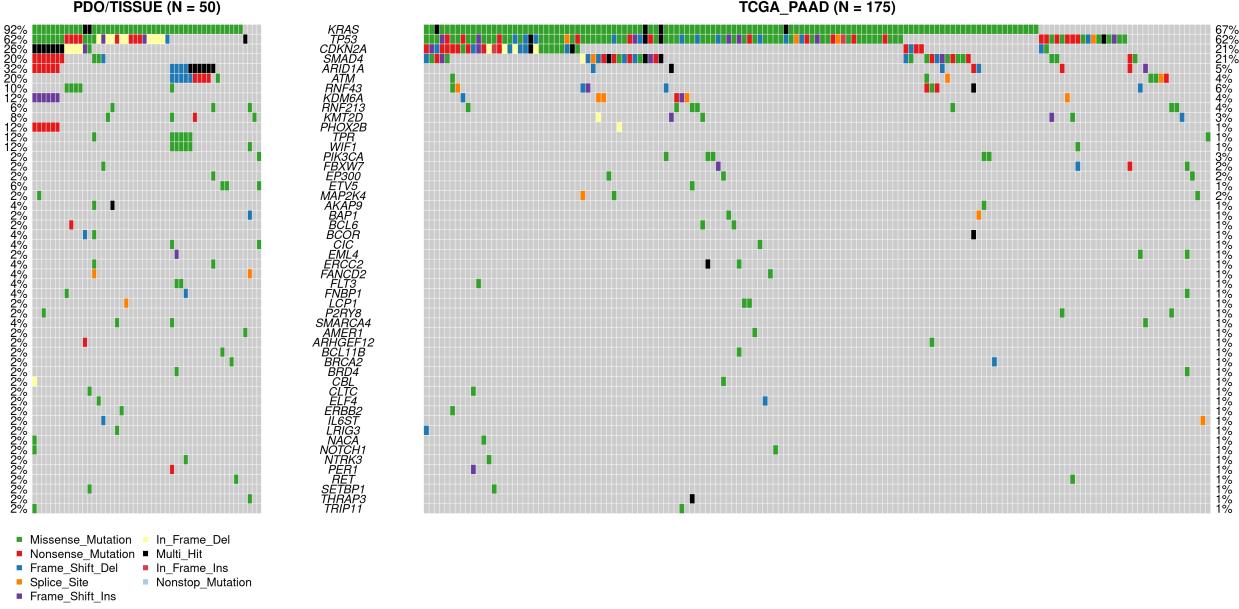


Figure 4: Co-oncoplot examining the proportion of TCGA or PDAC/TISSUE samples displaying the COSMIC mutations

We find that the majority of mutations identified (2964) are unique to each patient, with only 1 gene shared across 11 (OR8U1), 12 (TP53) and 15 (KRAS) patients (Figure 7). This suggests that the mutational landscape is highly heterogeneous, with a few conserved driver genes shared across distinct individuals.

2.4 Validation of KRAS mutations with Sanger Sequencing

We validated our WES results with Sanger Sequencing at the KRAS loci. As shown in Figure 8), most of the WES-identified KRAS mutations are identical to the Sanger Sequencing KRAS result, with the following exceptions:

- An additional P34R mutation was identified in WES in the PCA76B sample.
- Multiple hits on top of a G12I mutation, including G12V and G12S, were identified in WES for PCA71B.
- For 28R, WES identified a G12V mutation whereas Sanger-sequencing reported a G12D mutation.

In total, 14/17 (82%) samples have identical KRAS mutations identified by both WES and Sanger sequencing. The KRAS wild-type alleles in 2 samples were concordant in both sequencing methods.

2.5 Unsupervised clustering of PDO and patient samples

We carried out an unsupervised clustering of the 7 patient tissue samples along with the remaining PDO samples, based on their somatic mutational profiles (Figure 9). Specifically, a binary matrix was constructed based on whether a sample possesses a specific mutation within the full set of mutated genes, and the spearman correlation was chosen a measure of dissimilarity between the samples. We found that 3 of the tissue samples clustered together with the corresponding PDO samples, whereas the remaining 4 of them had relatively fewer mutations and tended to cluster together with other samples with similarly lower TMB.

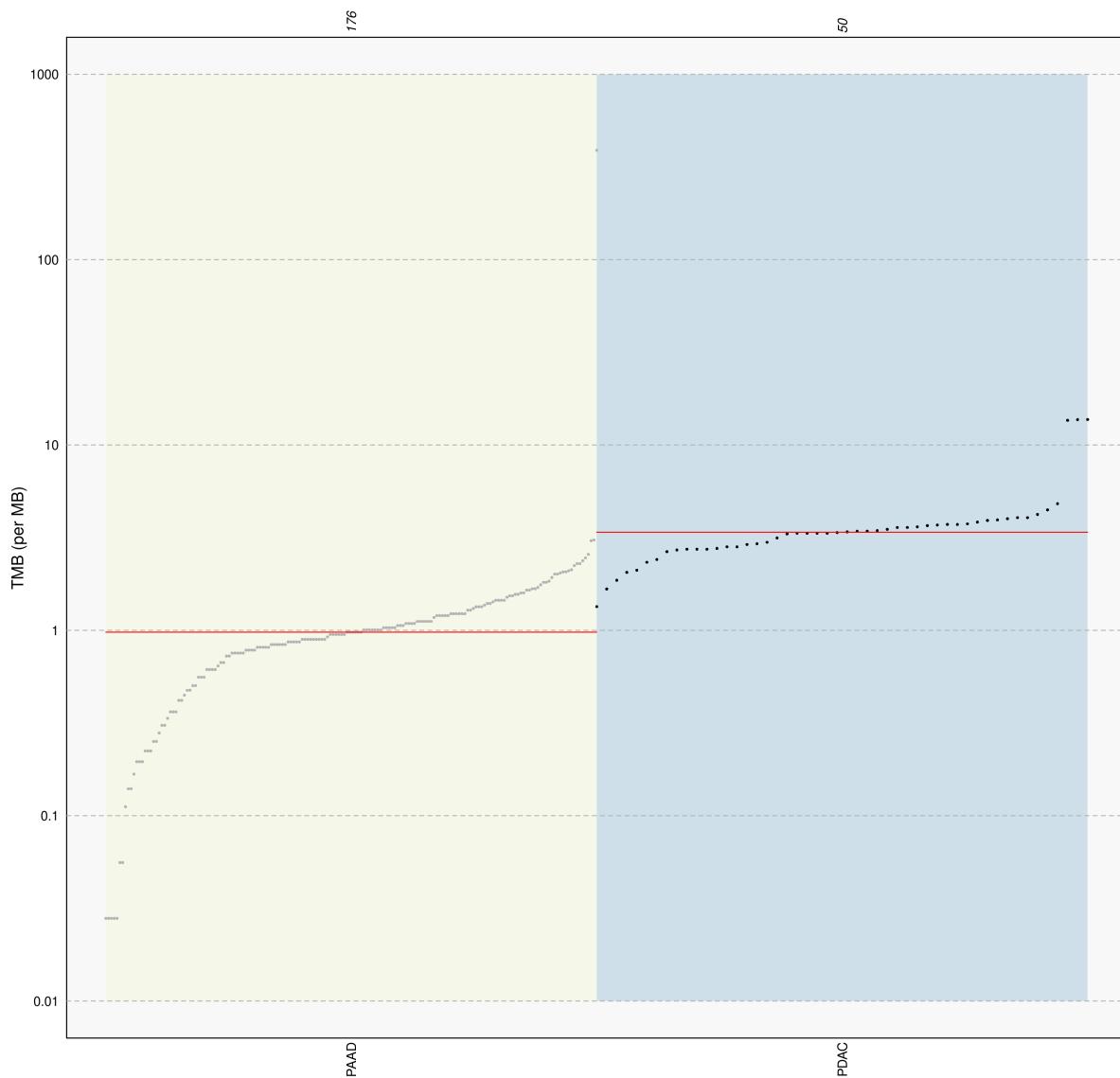


Figure 5: Comparing the TMB score between TCGA and PDAC/TISSUE data

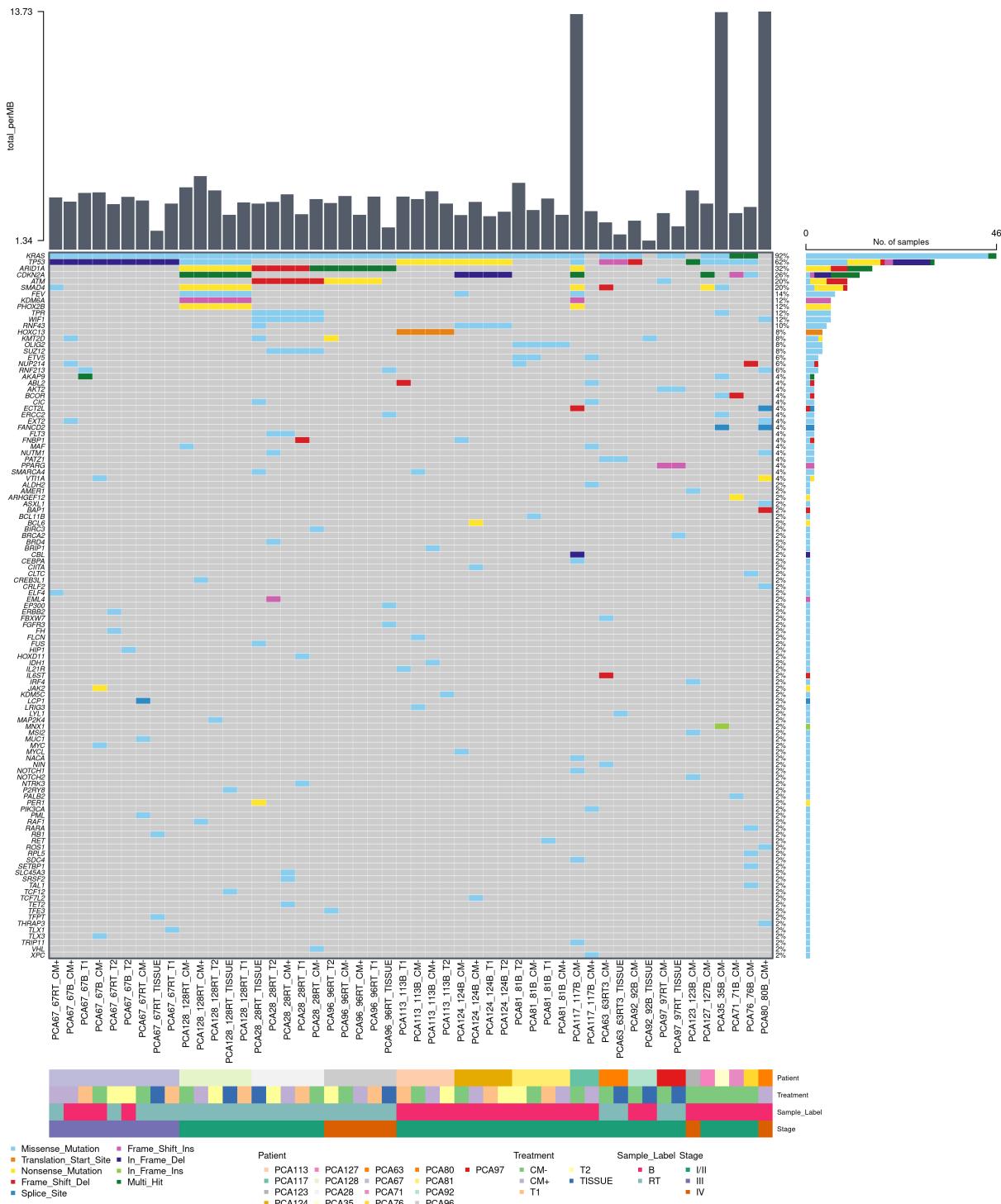


Figure 6: Oncoplot showing the SNPs/INDELs in the COSMIC genes

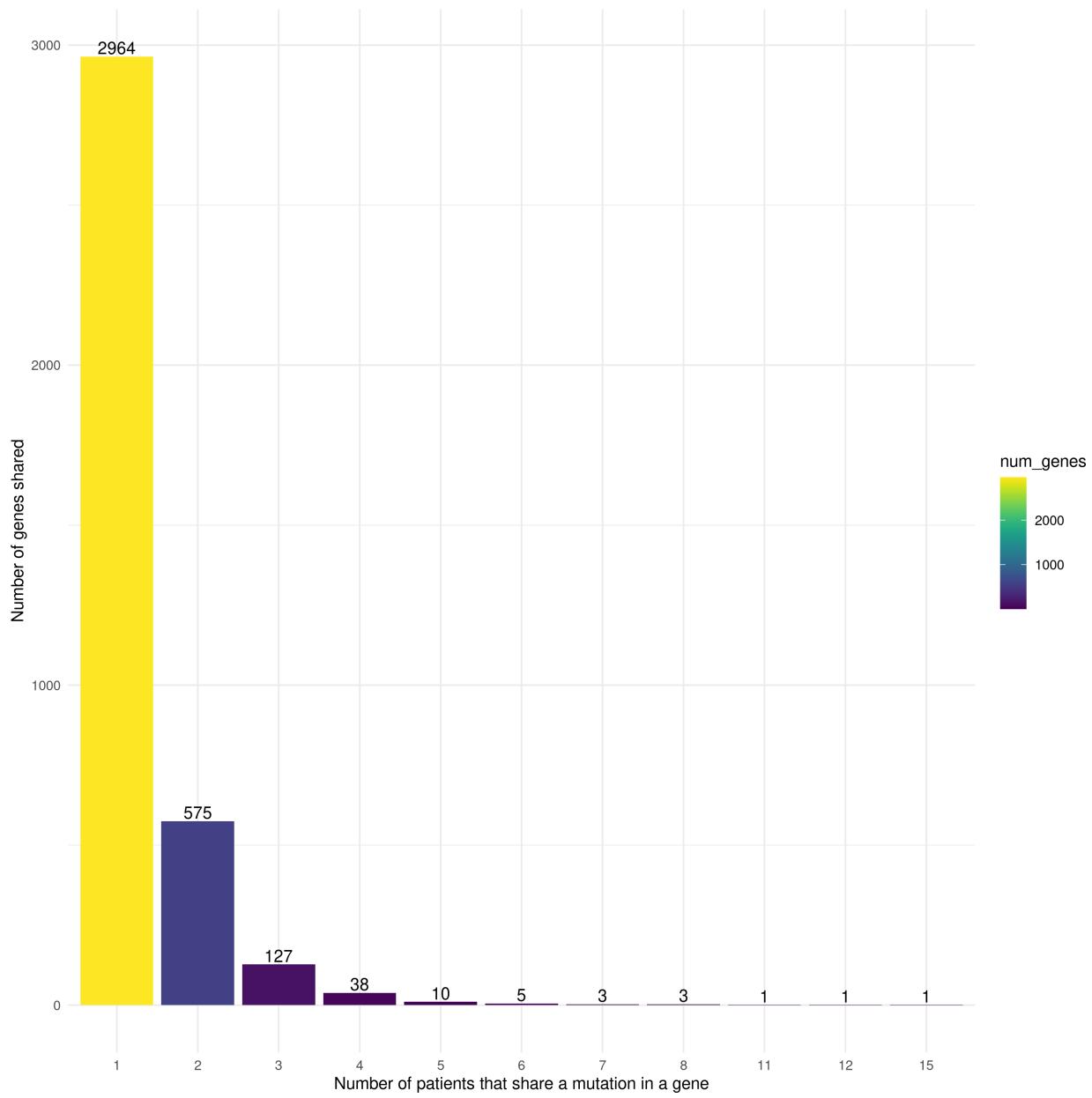


Figure 7: Barplots showing the number of mutations that are shared by patients

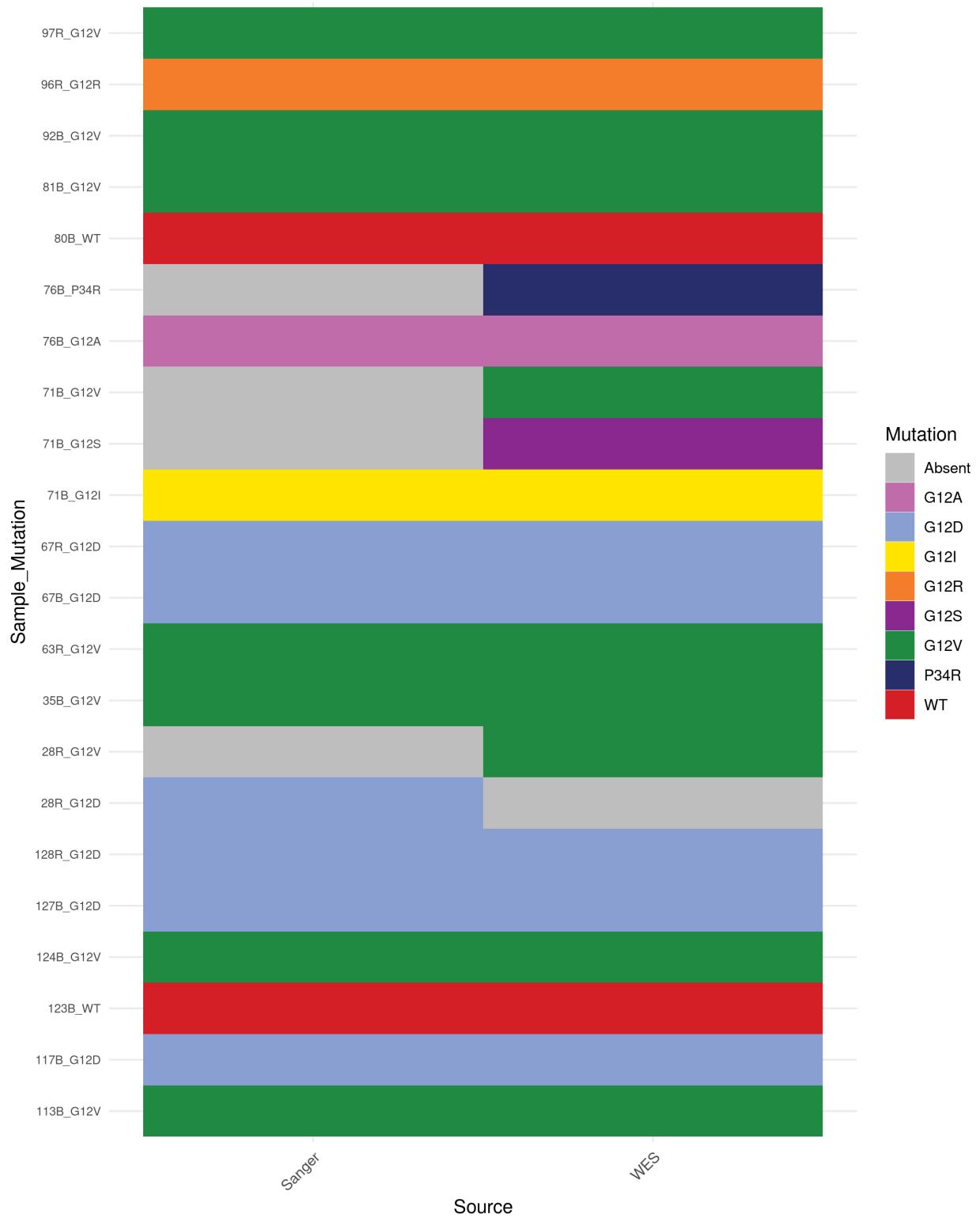


Figure 8: Assessing the concordance between the KRAS mutations identified by WES and Sanger Sequencing

Unsupervised clustering of PDO and patient tissue samples based on the SNPs and INDELs mutational landscape

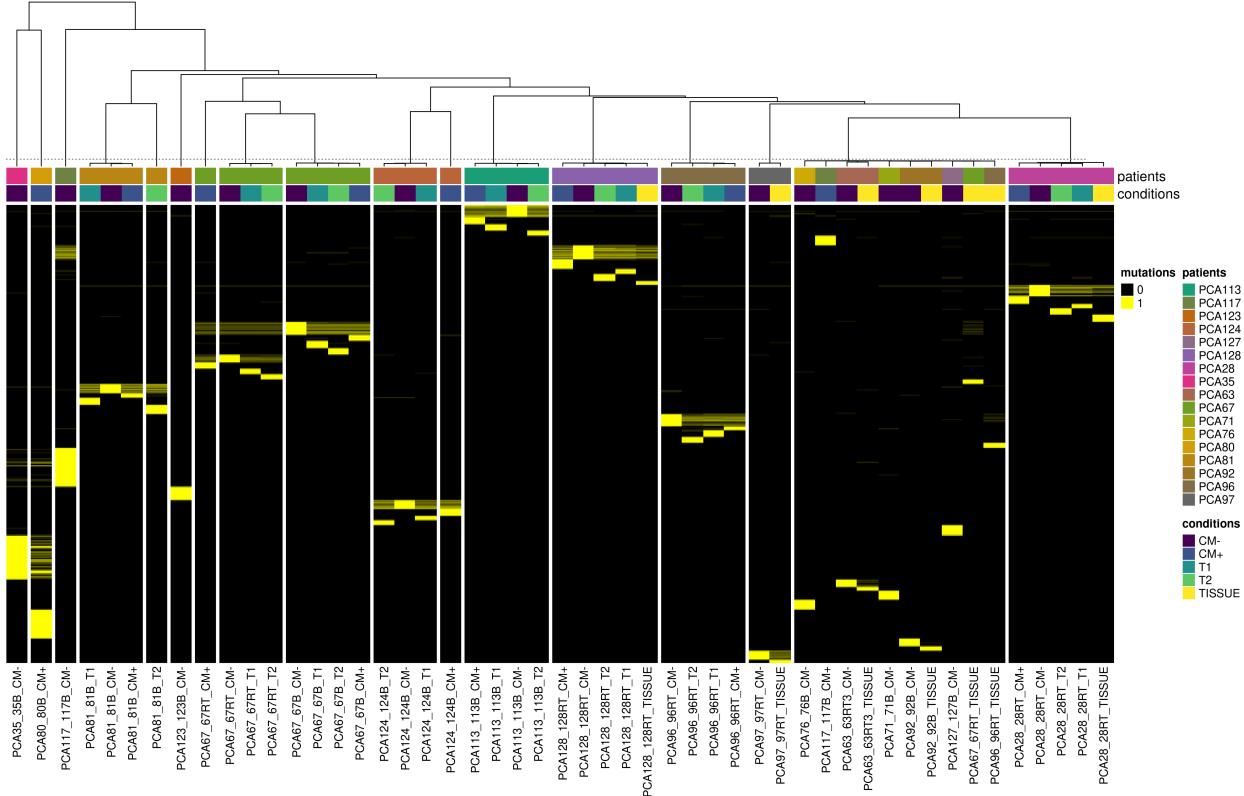


Figure 9: Unsupervised clustering of PDO and tissue samples

2.6 Copy number alterations

We incorporated the CNA results into the oncoplot and display them in this report. Two oncoplots are shown, one for the top 50 genes that are mutated in the PDO/TISSUE samples (Figure 10). Another oncoplot highlights all the COSMIC genes that are mutated in the samples (Figure 11). The TMB score is indicated in the top panel (identical result to `tmb_pdac.png`). Further results from this analysis to explore can be found in the `ascat_cn_signatures` folder.

2.7 Tumor mutational burden (TMB) score

We find that the patient tissue samples have a lower TMB score relative to the PDOs, with a median of 2.05 mutations/MB (Figure 12 vs 3.45 mutations/MB 13) in the organoid samples.

2.8 Microsatellite instability

With the exception of 117B_CM-, all the other samples have very low levels of microsatellite instability (Figure 14).

2.9 HLA typing result

Based on the analysis of HLA class I alleles across various samples using the Jaccard similarity index, we note a notable correlation between the HLA class I allele profiles and their respective patient origins (Figure 15). This suggests that the HLA class I alleles are distinctively characteristic of individual patients.

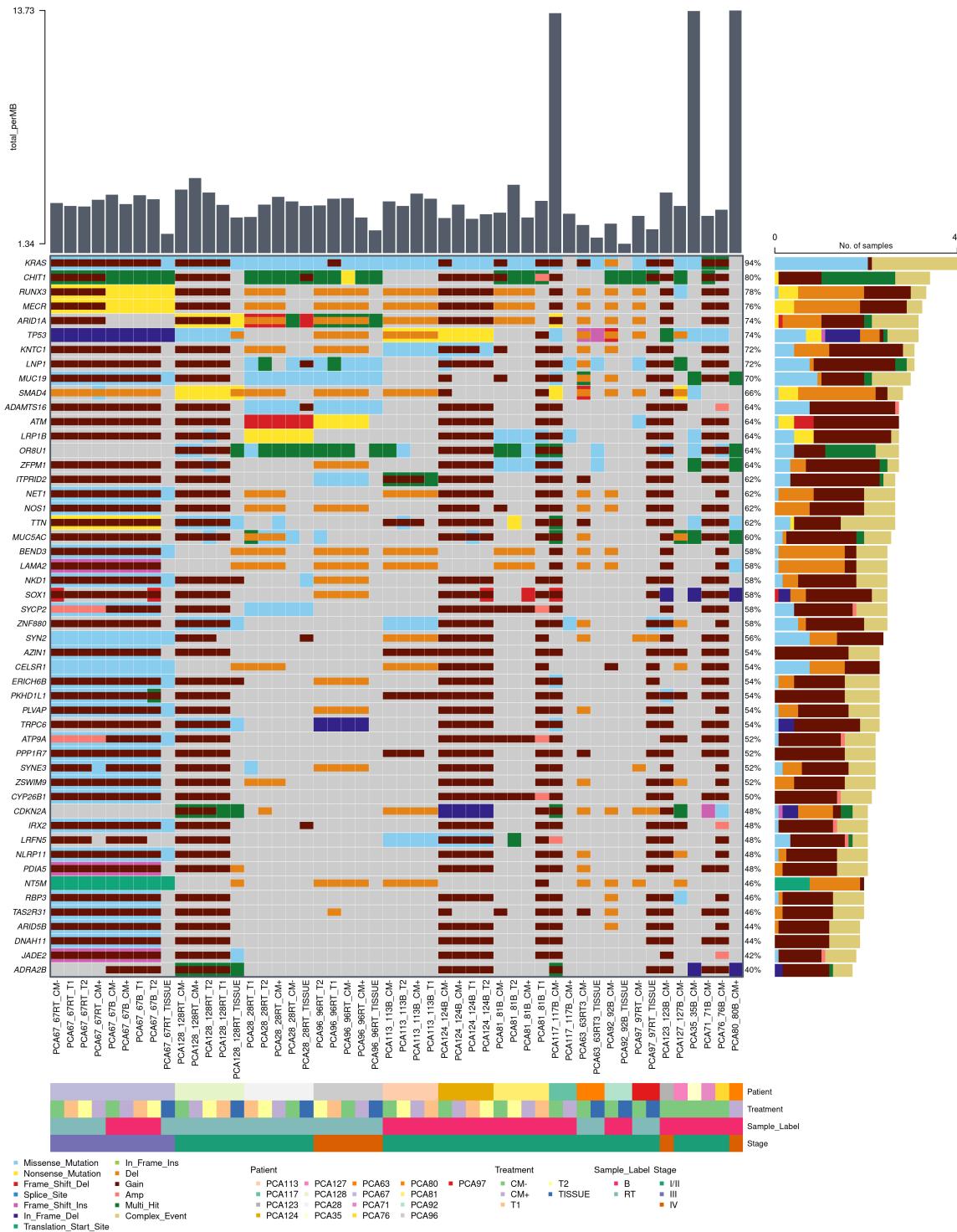


Figure 10: Oncoplot showing the SNPs/INDELs and copy number alterations in the top 50 mutated genes

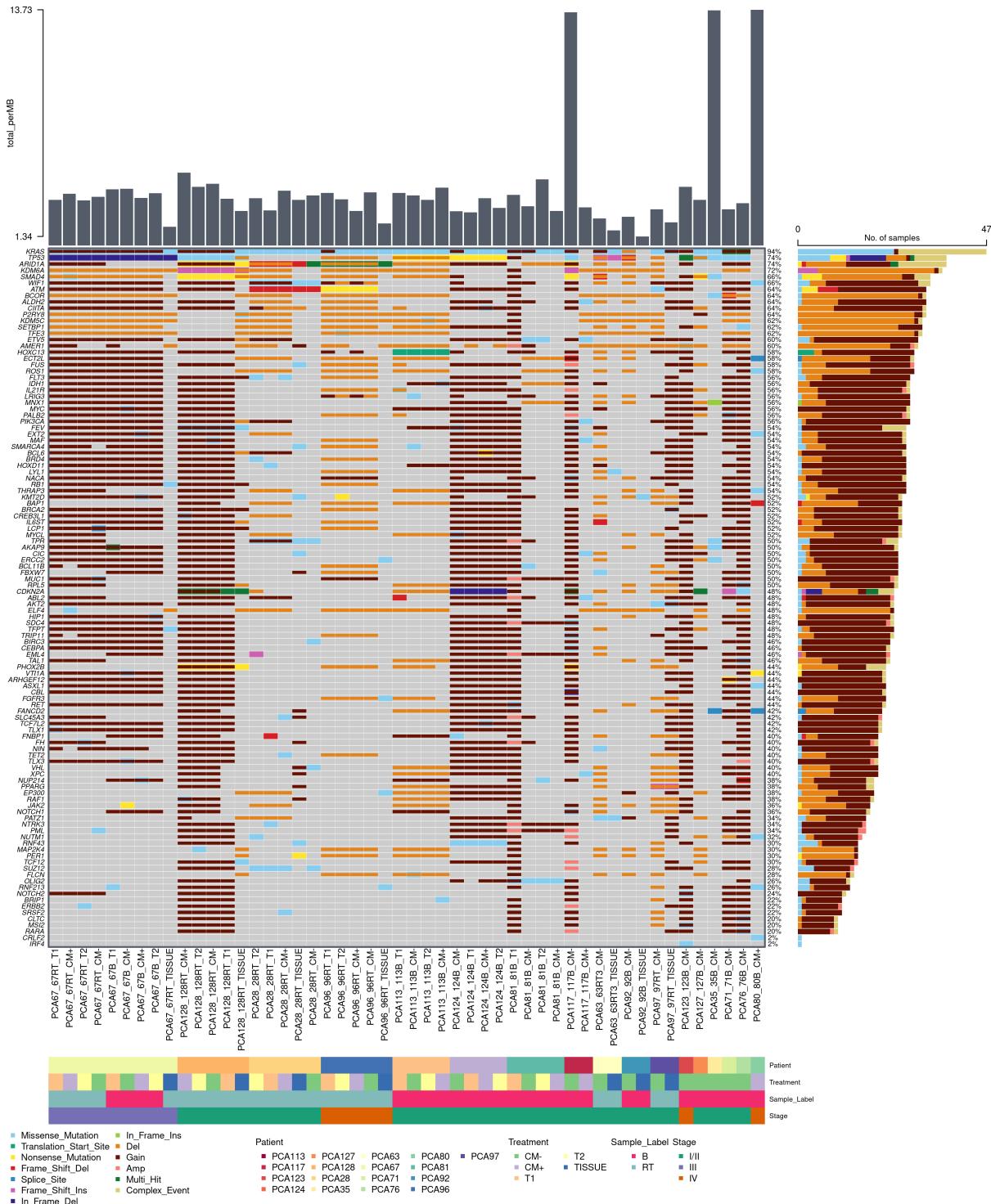


Figure 11: Oncoplot showing the SNPs/INDELS and copy number alterations in the COSMIC genes

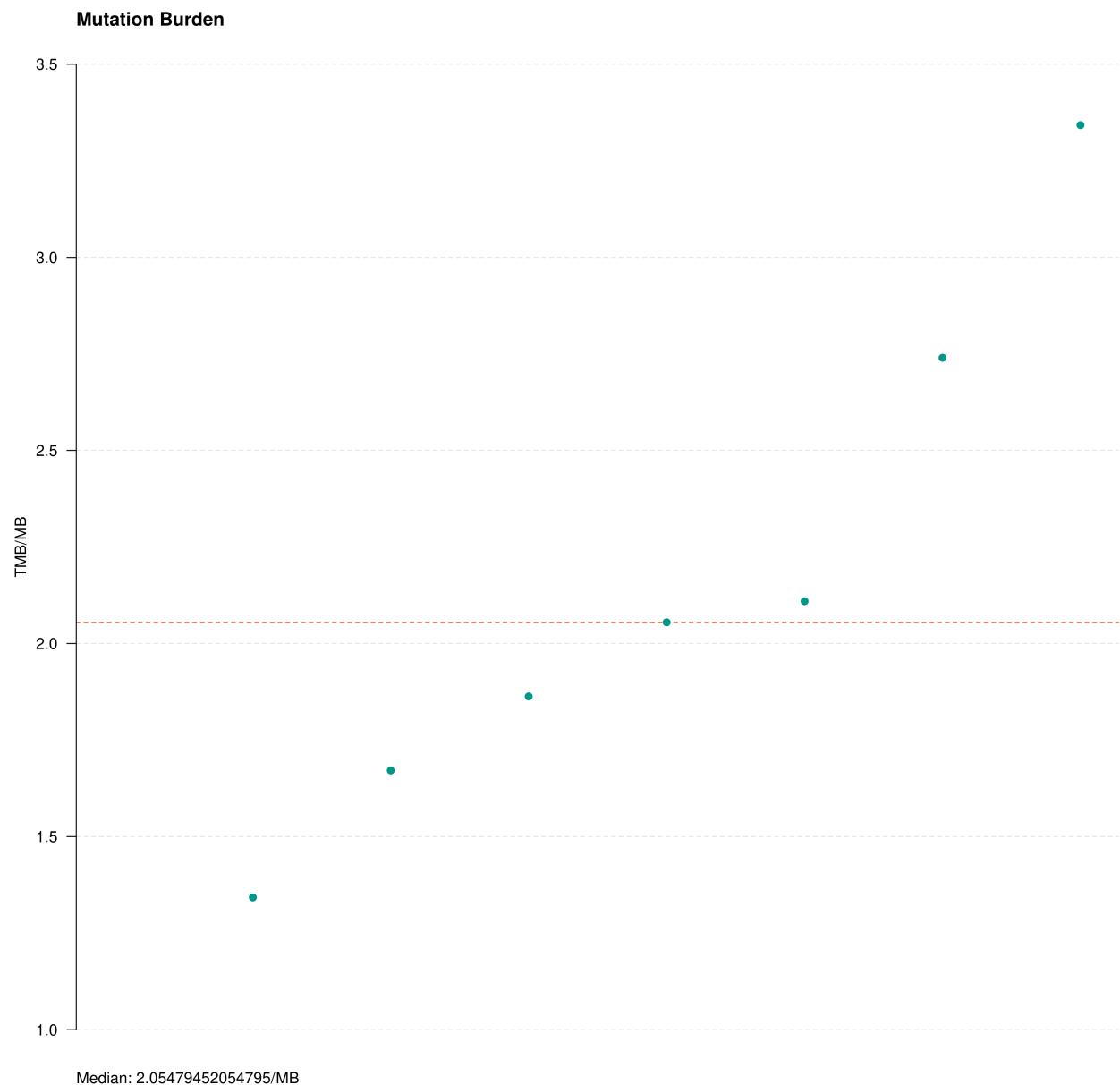


Figure 12: TMB scores in the tissue samples

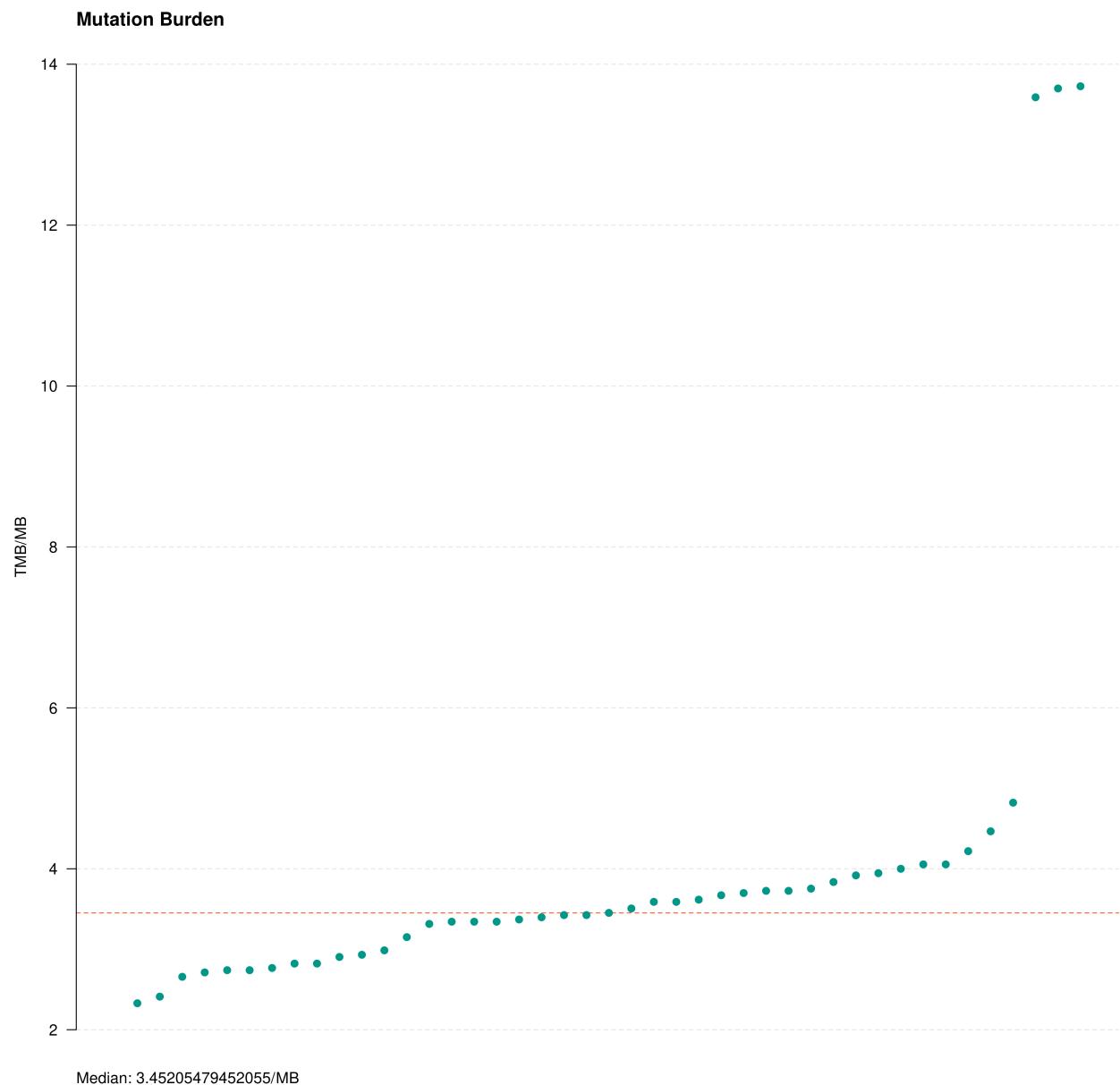


Figure 13: TMB scores in the PDOs



Figure 14: Msisensorpro results show low levels of microsatellite instability, with most samples have a score less than 1% 18

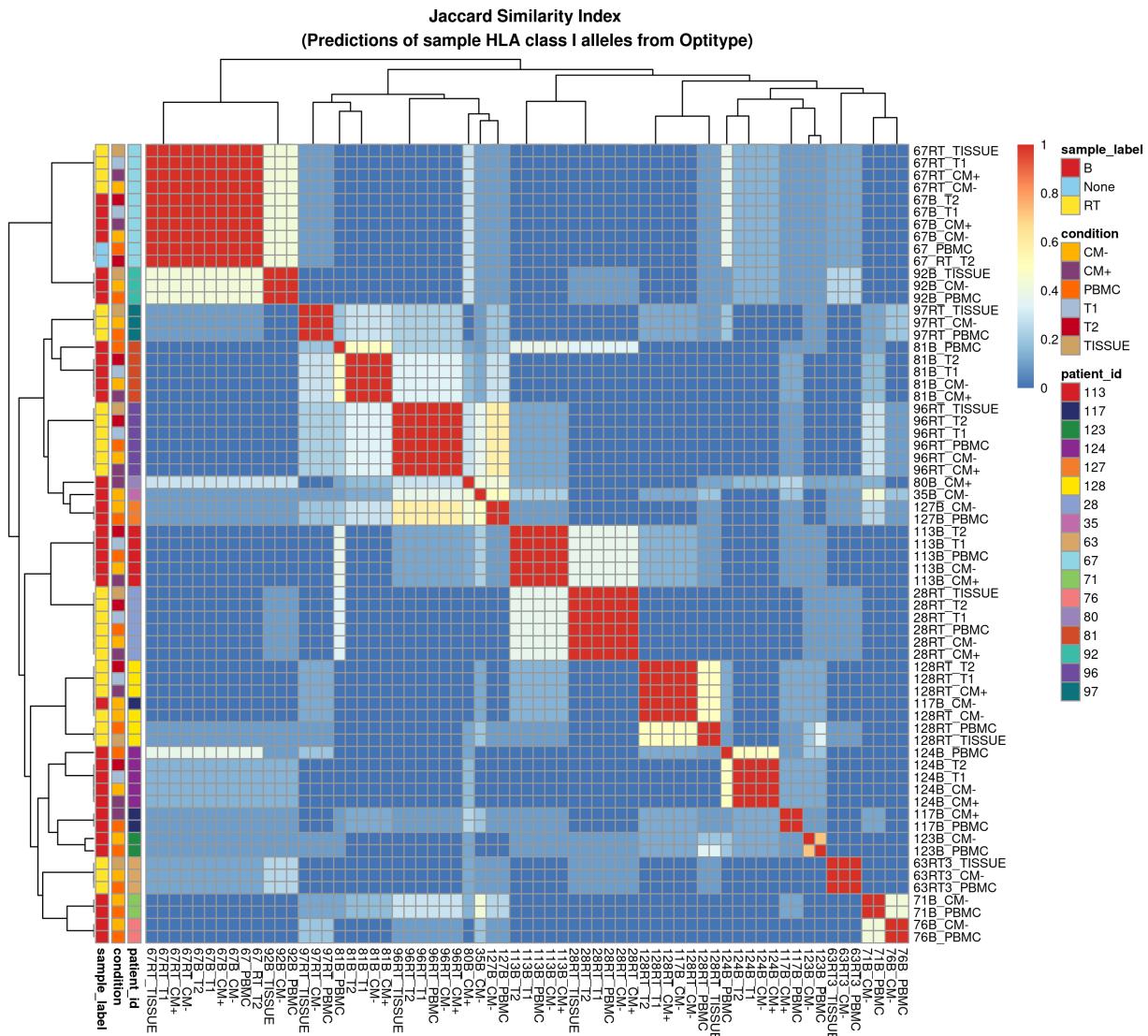


Figure 15: Clustering samples by the Jaccard similarity index by the predicted HLA class I alleles.