WES analysis of PDO and tissue samples in PDAC

Kane Toh*

2024-01-18

Contents

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,

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

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`.

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. The list of COSMIC genes were taken from the `COSMIC.67` R package.

Mutational data from the TCGA_PAAD cohort was retrieved from maftools with the `tcgaLoad` 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.

To perform the unsupervised clustering of our samples, we first constructed a binary matrix that indicates the absence or presence of each mutation. The heatmap was plotted using the `ComplexHeatmap` R package, with a custom-defined Jaccard distance metric to measure the dissimilarity between samples based on the binary matrix.

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) <= 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

We compared our somatic variants and gene mutations against the corresponding public TCGA_PAAD dataset. Here, we examine the PDOs cultured in the CM- condition for all 17 patients.

As shown in the upSet plot (Figure 2), we found that around 26% of all mutated genes were shared between both datasets. 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-oncoplots in Figure 4 show that the known driver genes such as KRAS (67% TCGA vs 94% CM-), TP53 (62% TCGA and 76% CM-), CDKN2A (21% TCGA and 35% CM-) and SMAD4 (21% TCGA and 29% CM-) driver gene mutations were recapitulated in our CM- PDO samples. On the other hand, our CM- samples have a greater proportion of ARID1A (5% TCGA and 24% CM-) and ATM (4% TCGA and 18% CM-) mutations.

Next, we assessed the tumor mutational burden (TMB) estimates between all our 50 samples and the 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`).

2.3 Somatic SNP/INDEL mutational landscape

Here, we discuss the SNP/INDEL mutational landscape for our samples.

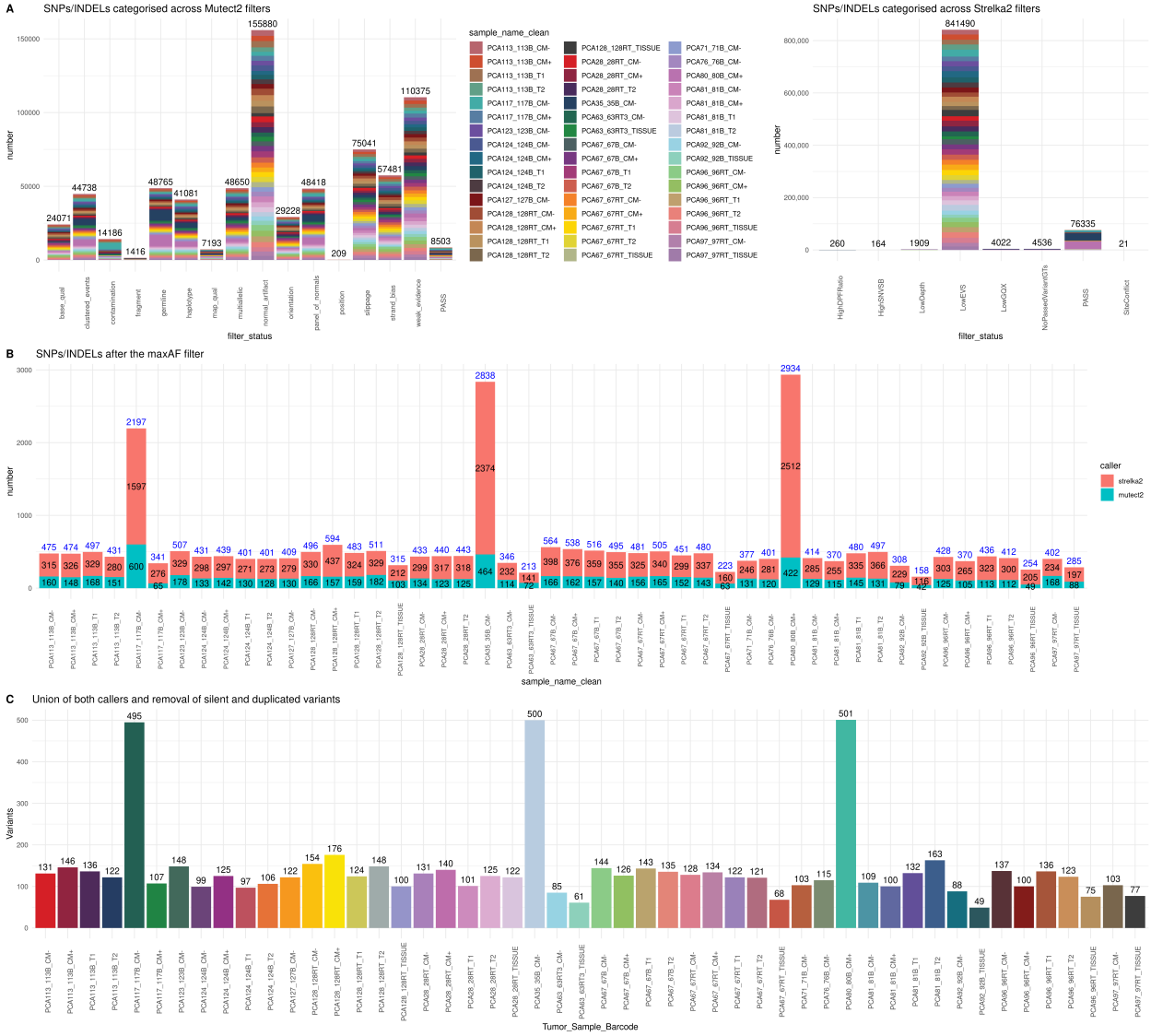

Figure 1: Summary of filtering steps used in variant callers

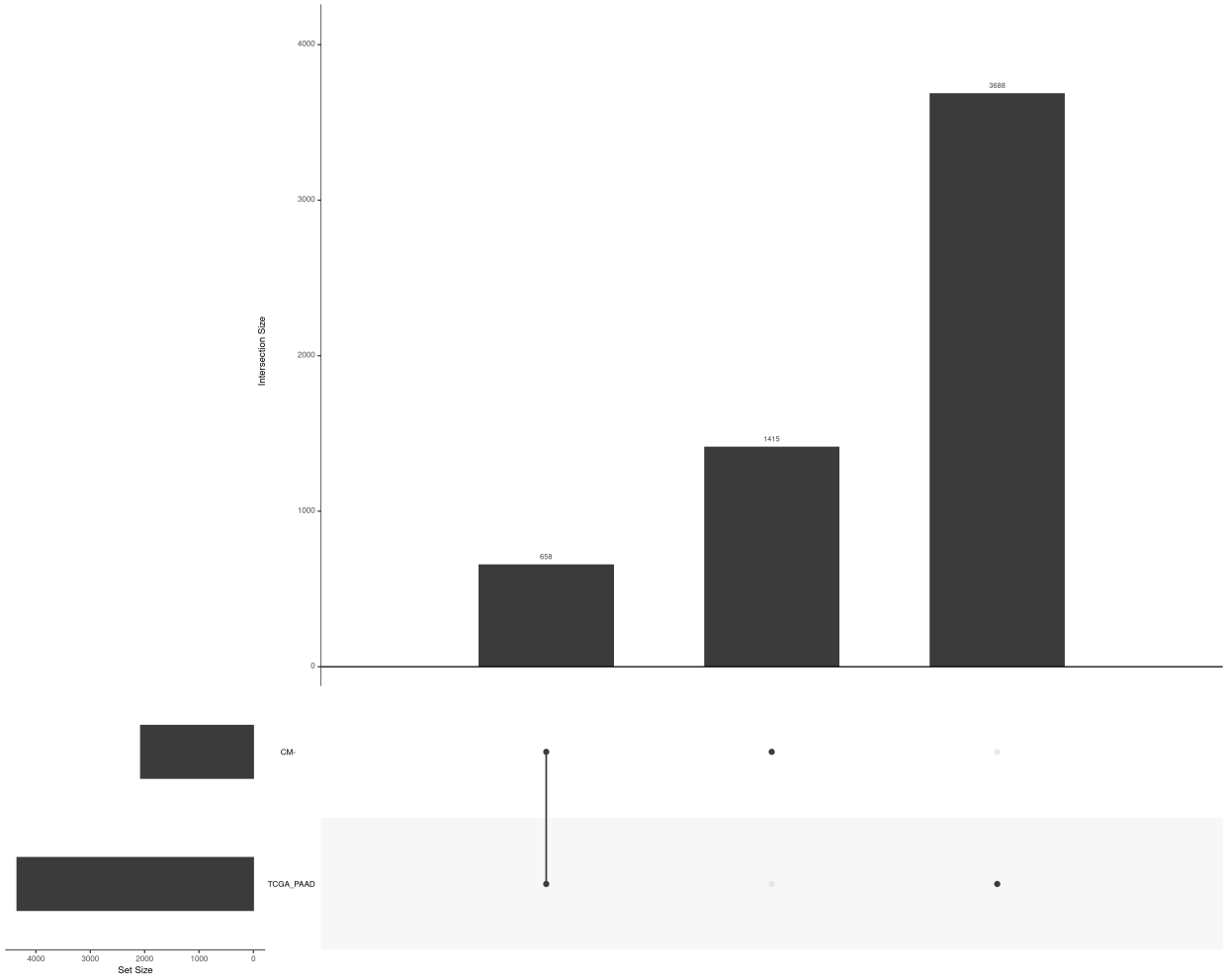

Figure 2: Number of mutations found in TCGA vs PDAC/TISSUE data

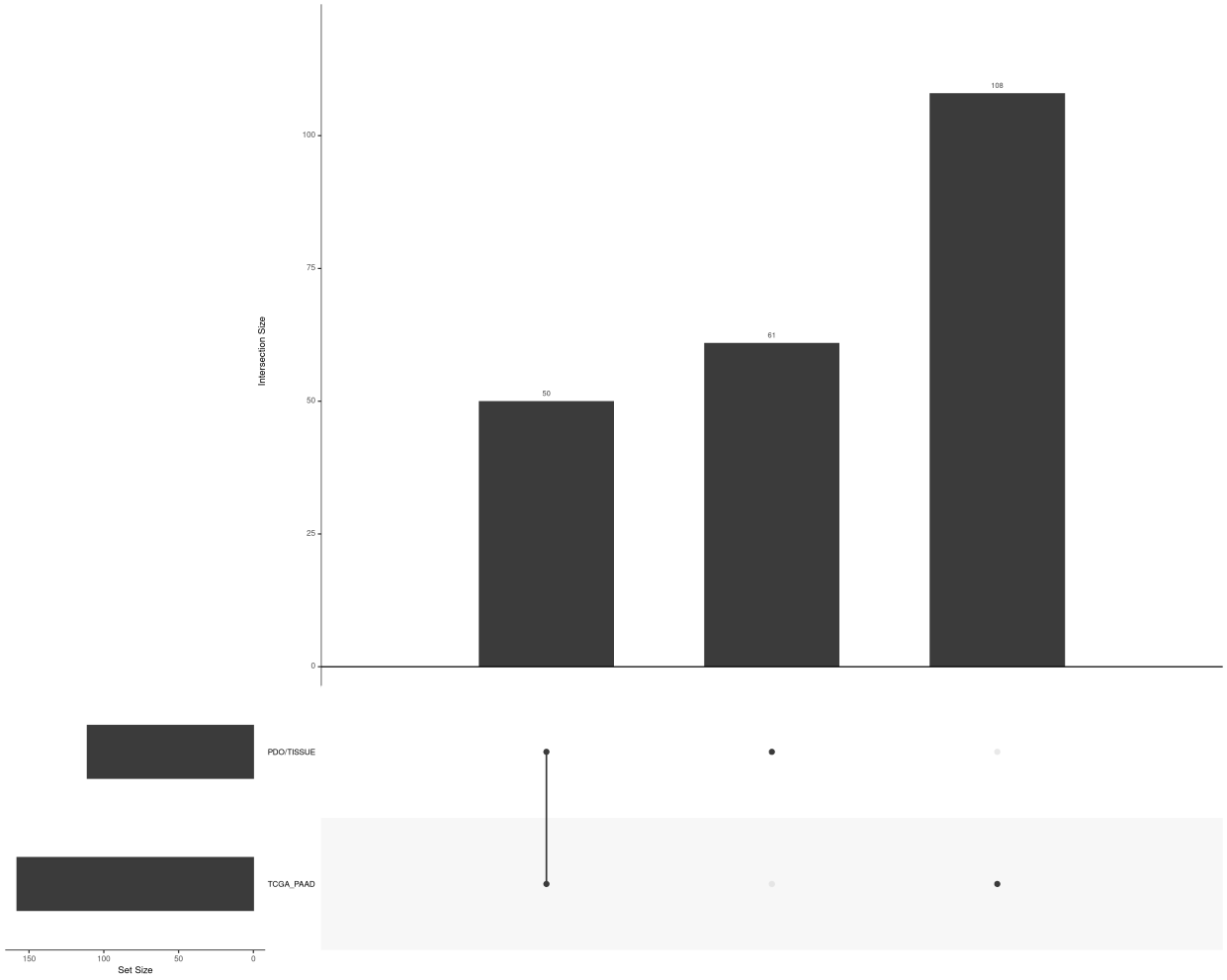

Figure 3: Number of COSMIC mutations found in TCGA vs PDAC/TISSUE data

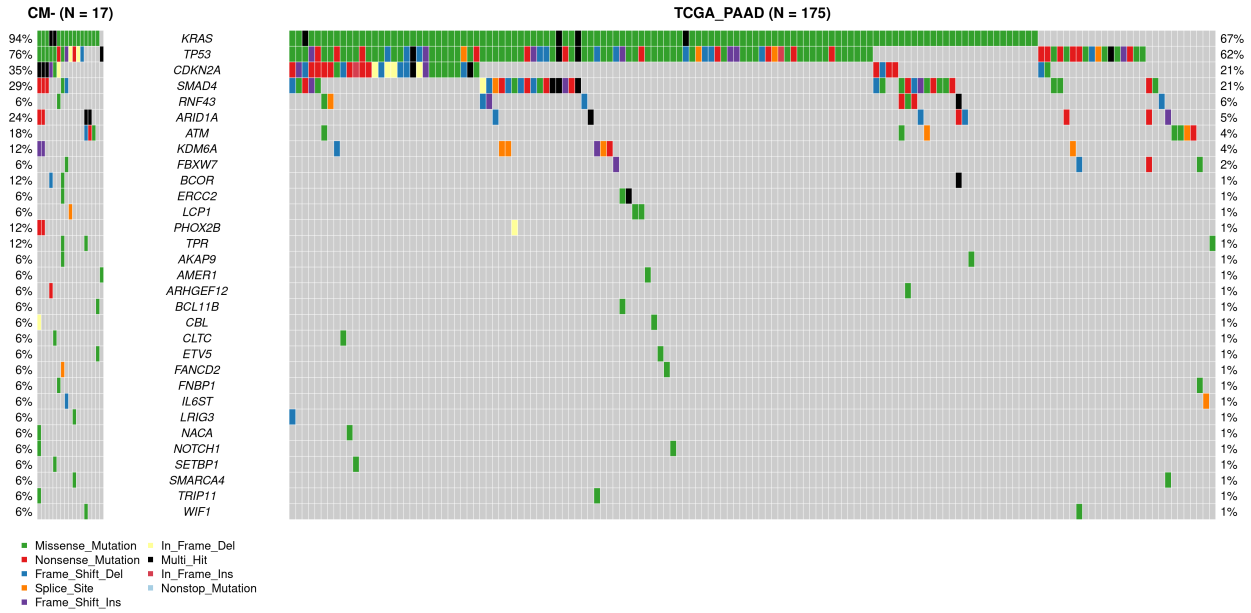

Figure 4: Co-oncoplot examining the proportion of TCGA or PDAC/TISSUE samples displaying the COSMIC mutations

2.3.1 Directory organisation for oncoplots

All oncoplots are kept in two folders: `oncoplot_snps_indels/` and `oncoplot_cn/`. The file names are similar for both directories. Taking `oncoplot_snps_indels` to illustrate, there are 8 oncoplots, grouped into 4 pairs:

1. CM- samples only: all mutated genes or COSMIC genes only
2. CM- samples with matched tissue samples: all mutated genes or COSMIC genes only
3. PDOs with all 4 (CM-, CM+, T1, T2) conditions present: all mutated genes or COSMIC genes only
4. All samples: all mutated genes or COSMIC genes only

2.3.2 Discussion of CM- SNP/INDEL landscape

In this discussion, we focus only on the CM- samples. The oncoplot shown in Figure 6 shows the mutated genes that were present in the COSMIC database,
emphasising known mutations with causal implications in carcinogenesis. COSMIC genes were included in the oncoplot as long as at least a single sample amongst the 50 samples analysed show a mutation. Hence, the COSMIC genes without mutations in the CM- samples were included in the plot as they were mutated in other non CM- samples that are not displayed.

Across the 17 PDAC CM- samples corresponding to all 17 patients in the study, we observed that 16/17 samples (94%) possess the KRAS mutation. The exception is PCA123B, which in had the wild type KRAS allele in the patient as shown in our WES analysis and Sanger Sequencing result (Figure 8). Of the 17 samples with the KRAS mutation, 8/17 samples exhibit G12V (47%) and 5/17 G12D (29%). The remaining mutations present only in a single sample are P34R, G12A, G12I and G12R. 76% of the CM- samples have mutations in the TP53 gene. With the exception of the G199E mutation which was shared between patient samples

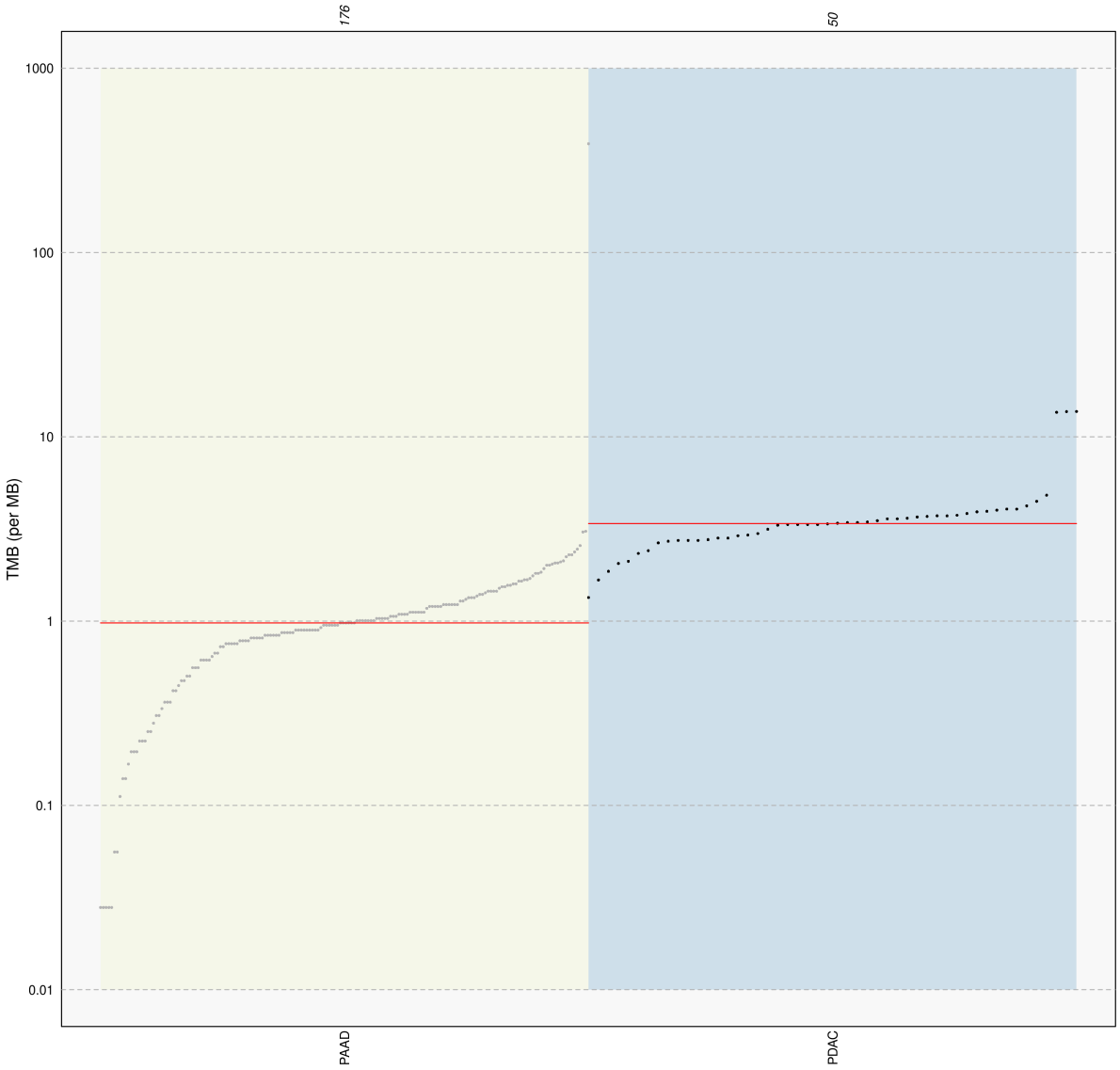

Figure 5: Comparing the TMB score between TCGA and PDAC/TISSUE data

PCA117_B and PCA128_RT, each patient harbors a different mutation (See `TP53_CM-_mutations.csv` for the list of mutations in the CM- samples).

Other genes that were mutated in relatively high proportion include the CDKN2A (35%), SMAD4 (29%) , ARID1A (24%) and ATM (18%) mutations.

Again considering only the CM- samples in our dataset, we find that the majority of mutations identified (1779) are unique to each sample, with only 1 gene shared across 15 patients (KRAS), 1 gene across 12 patients (TP53), 2 genes across 8 patients (CHIT1, OR8U1), 2 genes across 6 patients (CDKN2A, MUC19), 2 genes across 5 patients (SMAD4, LNP1) and 8 genes across 4 patients (ADGRV1, ARID1A, ATG2A, MUC5AC, PXDNL, RAI1, SOX1, TTN). This suggests that the mutational landscape is highly heterogeneous, with a few conserved driver genes shared across distinct individuals. The full list of mutations shared across multiple patients in the CM- samples can be found in `num_patients_shared_cm-.csv`. For more detailed exploration to answer questions such as "Which patients have a mutation in gene X", the full list of mutations can be further explored in `full_mutation_list.csv`. For the list of mutations shared across multiple patients in all samples, refer to `num_patients_shared.csv` and `num_gene_category.png`.

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.

To double check that this is not a false positive call due to issues with the variant caller, we examined the reads at amino acid glycine codon 12 (specifically, position 25,245,350 on chromosome 12 in grch38) and amino acid proline at codon 34 (specifically, position 25,245,284 in chromosome 12 on grch38) in the KRAS gene in the integrated genome viewer (IGV) as these correspond to the discrepancies between both methods. For 76B_CM-, we note that of the 64 reads mapped to the site, around 59% show a transversion of G->C (see `igv_sanger_wes/76B_CM-.png`). For the other samples shown, almost all the reads at this site display the wild type allele. The reads are also mapped with high confidence. Thus, this supports the proline to arginine substitution for this sample. At codon 12, for 28RT_CM+, we find around 40% of the 108 reads indicate a C->A transition at position 25,245,350 and thus the codon GTT, supporting the G12V substitution (instead of the G12D substitution indicated by Sanger sequencing). See `igv_sanger_wes/4samples_wes.png`. Interestingly, for 71B_CM-, we found that there is a dinucleotide variation spanning from positions 25,245,350 and 25,245,351, such that all reads that carry a mutation also carry both substitutions at the same time (see `igv_sanger_wes/dinucleotide_isoleucine.png`). Therefore, in fact, the mutation for codon 12 in 71B should be G12I (glycine to isoleucine) as the codon now reads ATT. WES mistakenly called an additional two separate codon substitutions, on top of the G12I mutation, as the variant caller did not consider whether nearby variants are part of the same haplotype. In general, the WES mutation calls are strongly supported by the mapped reads.

2.5 Note on TP53 mutations

Whilst the KRAS gain of function mutations are prodominantly at codon 12 as expected, the TP53 inactivating mutations do not overlap between different patients. This is expected for loss of function inactivating

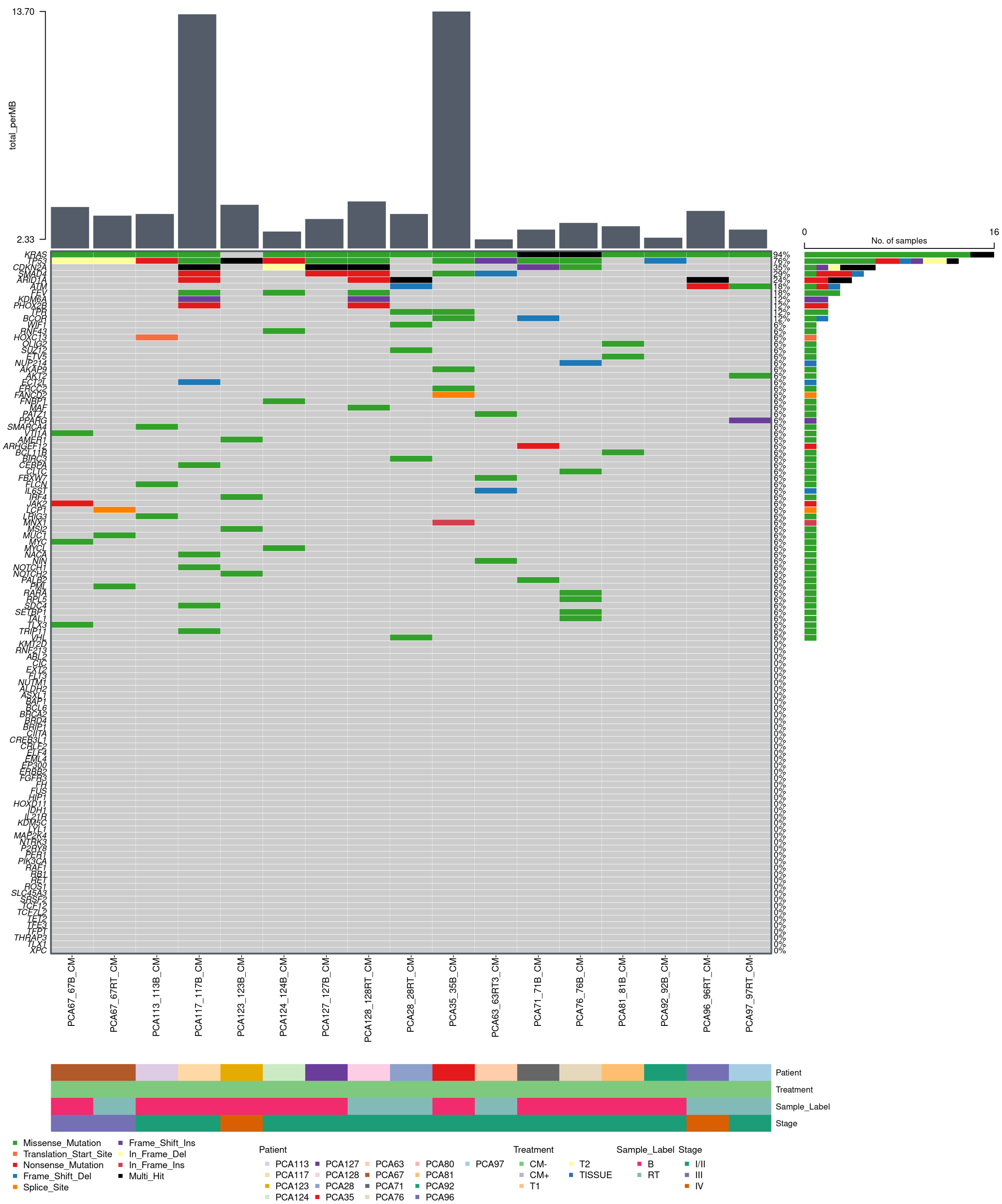

Figure 6: Oncoplot showing the SNPs/INDELs in the COSMIC genes for the CM- samples only.

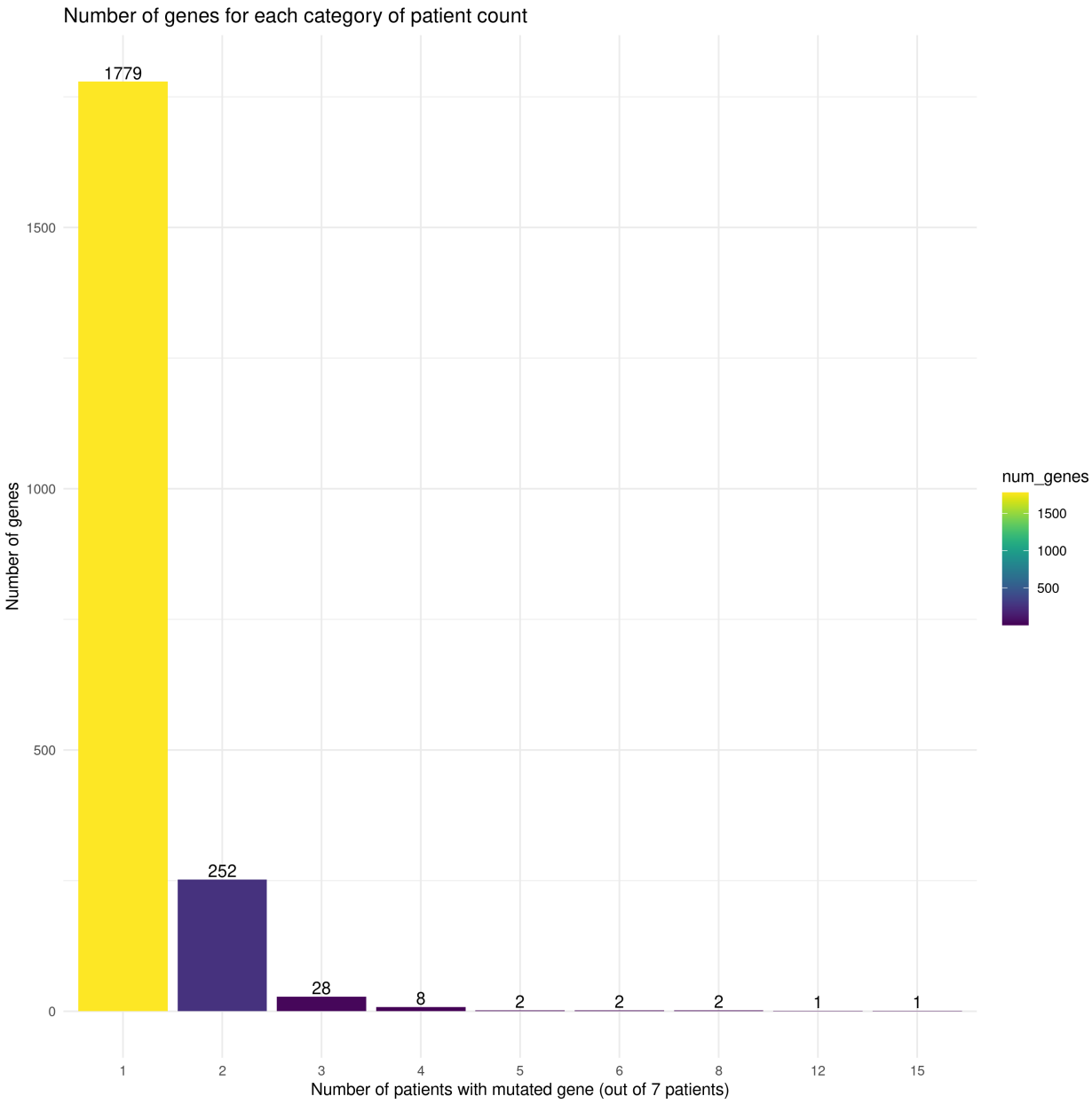

Figure 7: Barplots showing the number of mutations that are shared by patients

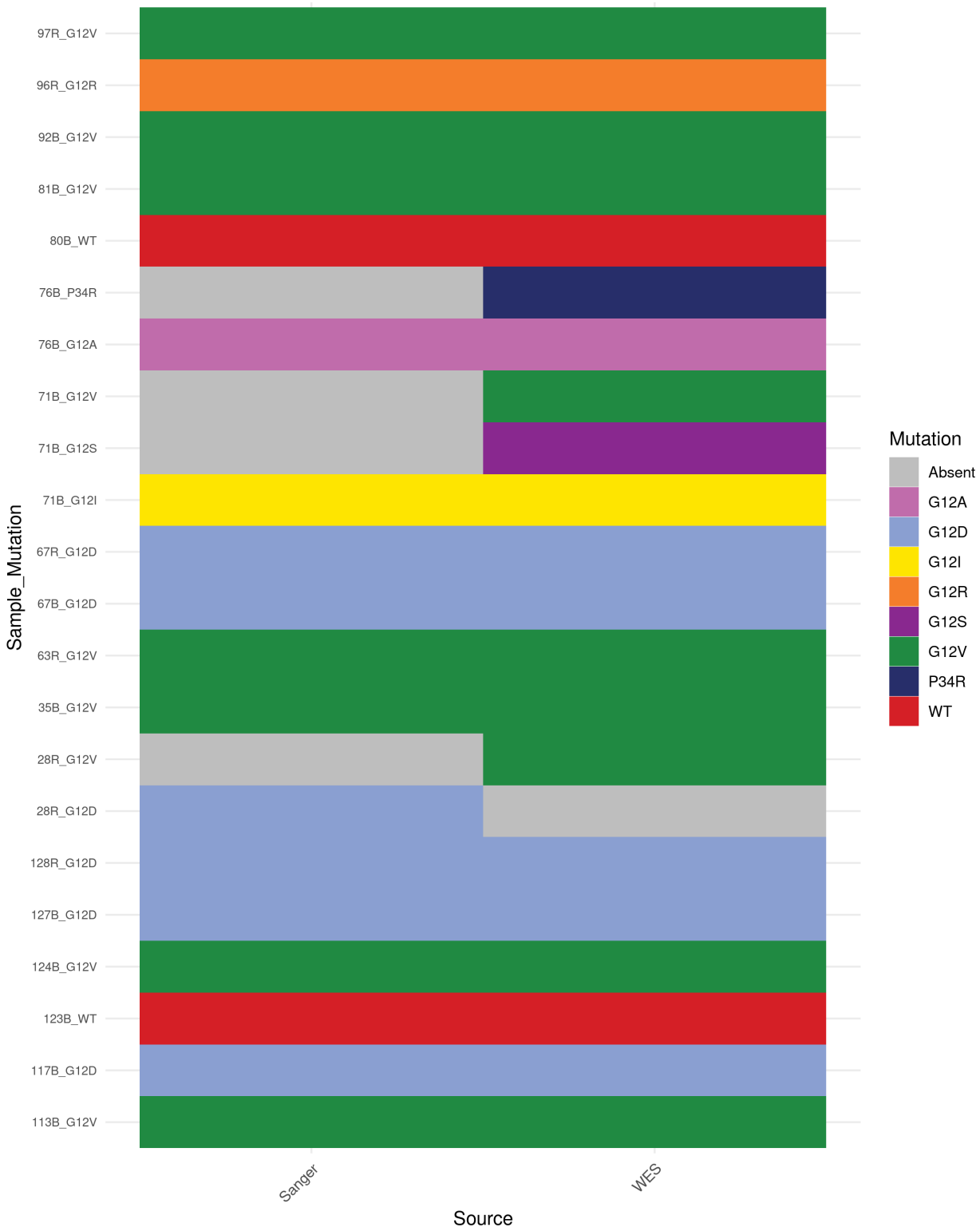

Figure 8: Assessing the concordance between the KRAS mutations identified by WES and Sanger Sequencing

mutations, as there are many more ways to inactivate the TP53 protein than to lead to a gain of function for the KRAS protein. Importantly, all TP53 mutations are shared within the PDOs and if present, the tissue sample, for each patient. See `TP53_mutations_list.csv` for the full list of TP53 mutations called.

2.6 Unsupervised clustering of PDO and patient samples

We carried out an unsupervised clustering of the PDO samples that have a matched patient tissue sample, using mutations in the COSMIC genes. (Figure 9) . In general, we find that the PDO samples cluster together with the patient tissue samples. For the full clustering heatmap including all samples with or without a matched tissue sample, see `unsupervised_clustering_snps_indels_shortName_COSMIC.png`. For the heatmap clustered using all genes, see `unsupervised_clustering_snps_indels_shortName.png`.

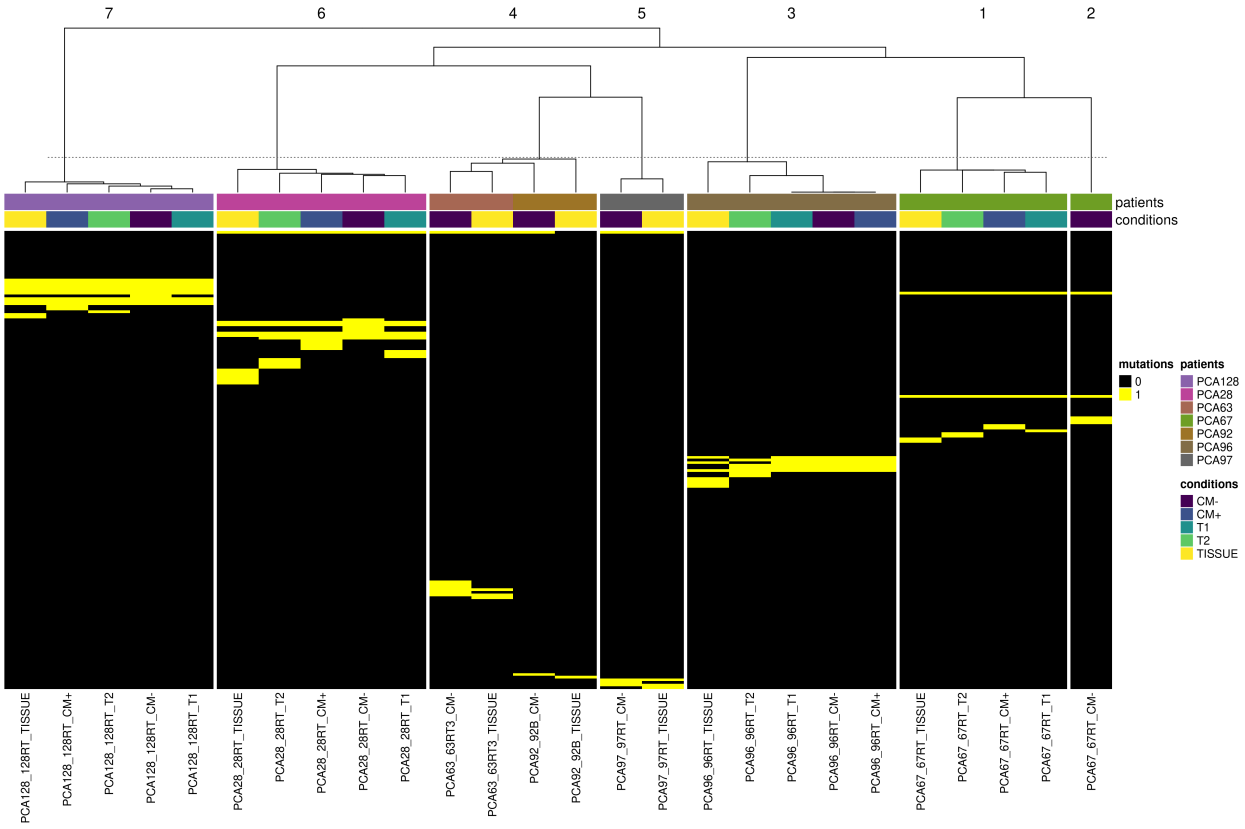

Figure 9: Unsupervised clustering of samples with a matched patient tissue sample. Only mutations in genes present in the COSMIC database are used.

2.7 Copy number alterations

We incorporated the CNA results into the oncoplot and display them in this report. Similar to the oncoplots containing only the SNPs/INDELs, Note that CNA results are absent from 2 patients, PCA80 and PCA35 as they did not have matched blood samples for accurate copy number profile identification with ASCAT. The oncoplot for all CM- samples are shown in ((Figure 10). Further figures can be found in the `oncoplots_cn` folder. The TMB score is indicated in the top panel (identical to `tmb_pdac.png`). Further results from this analysis to explore can be found in the `ascat_cn_signatures` folder. Ploidy and purity estimates for samples are shown in the plots `purity.png` and `ploidy.png`.

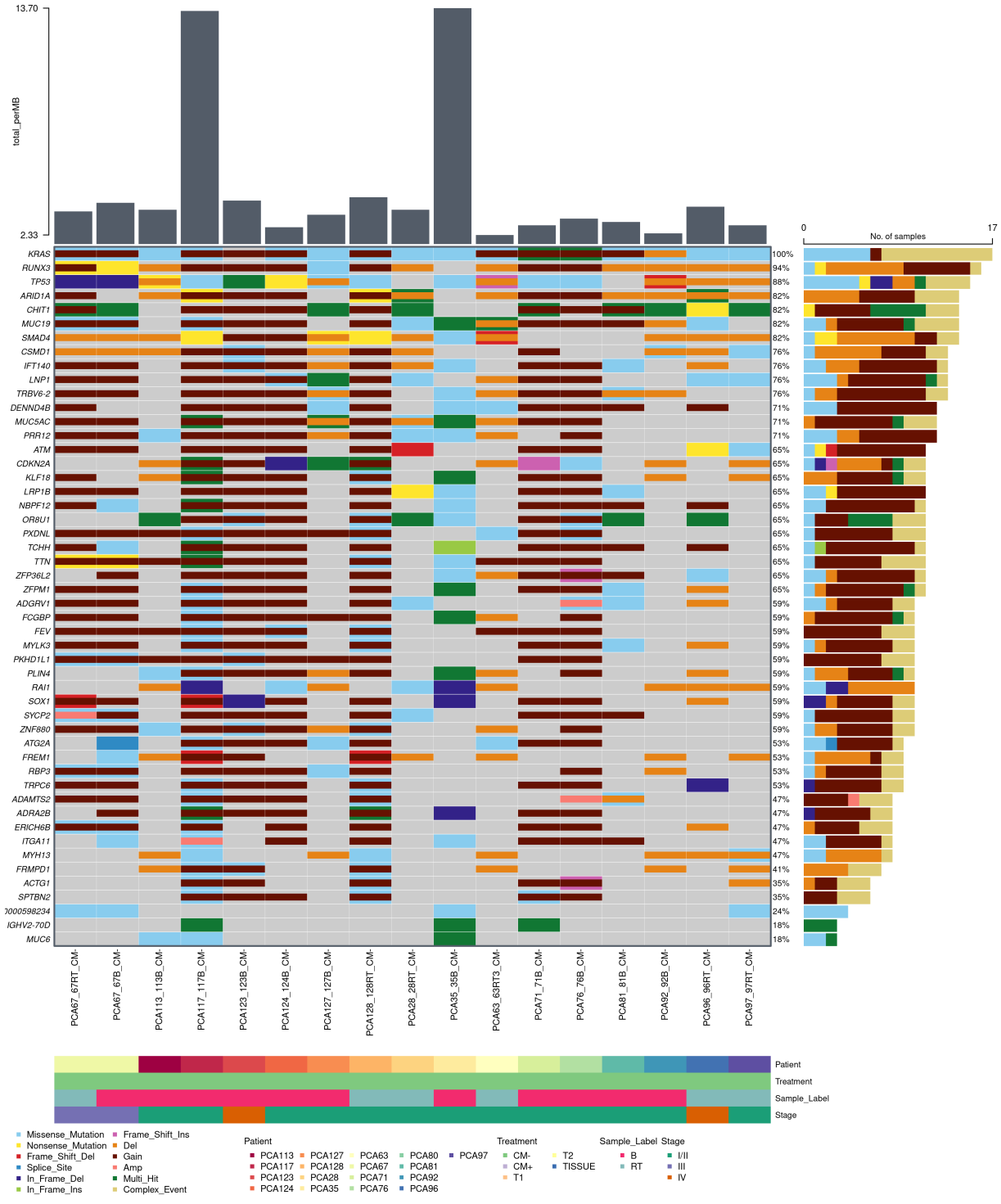

Figure 10: Oncoplot showing the SNPs/INDELs and copy number alterations in the COSMIC genes for the CM- samples

2.8 KRAS copy number alterations, grouped by condition

We also assessed whether the CM- and T1 conditions lead to a greater proportion of samples possessing CNAs in the KRAS gene ((Figure 11). Of the 15 PDOs cultured in CM- (not including the PCA35_CM- and PCA80_CM- samples as they do not have a matched normal), we find that 10/15 have a gain in copy number for KRAS. For T1 treatment, 5/8 samples have a gain in copy number. For CM+ and T2, 3/9 samples have a gain in copy number.

Comparing across the samples for each patient in Figure 12, we make the following observations. First, samples that possess CNAs in the CM+ and T2 conditions also possess CNAs in the CM- and T1 conditions, however the converse is not the case. In addition, for PCA124B and PCA117B where samples were grown in both CM+ and CM- media, only the CM- condition led to CNA for KRAS. Both these observations suggest that the CM- and T1 conditions likely encourage tumor outgrowth relative to the CM+ and T2 media conditions. In the PCA67_RT sample, we note that only the T1 treatment condition led to CNA. Also, several samples such as PCA28_RT and PCA113B did not have a CNA at the KRAS locus across all cultured media conditions. We also include a similar heatmap for the KRAS, TP53, SMAD4 and CDKN2A genes in `complexheatmap_CN_group_condition.png` plot.

2.9 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 13 vs 3.45 mutations/MB 14) in the organoid samples.

2.10 Microsatellite instability

With the exception of 117B_CM-, all the other samples have very low levels of microsatellite instability (Figure 15).

2.11 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 16). This suggests that the HLA class I alleles are distinctively characteristic of individual patients.

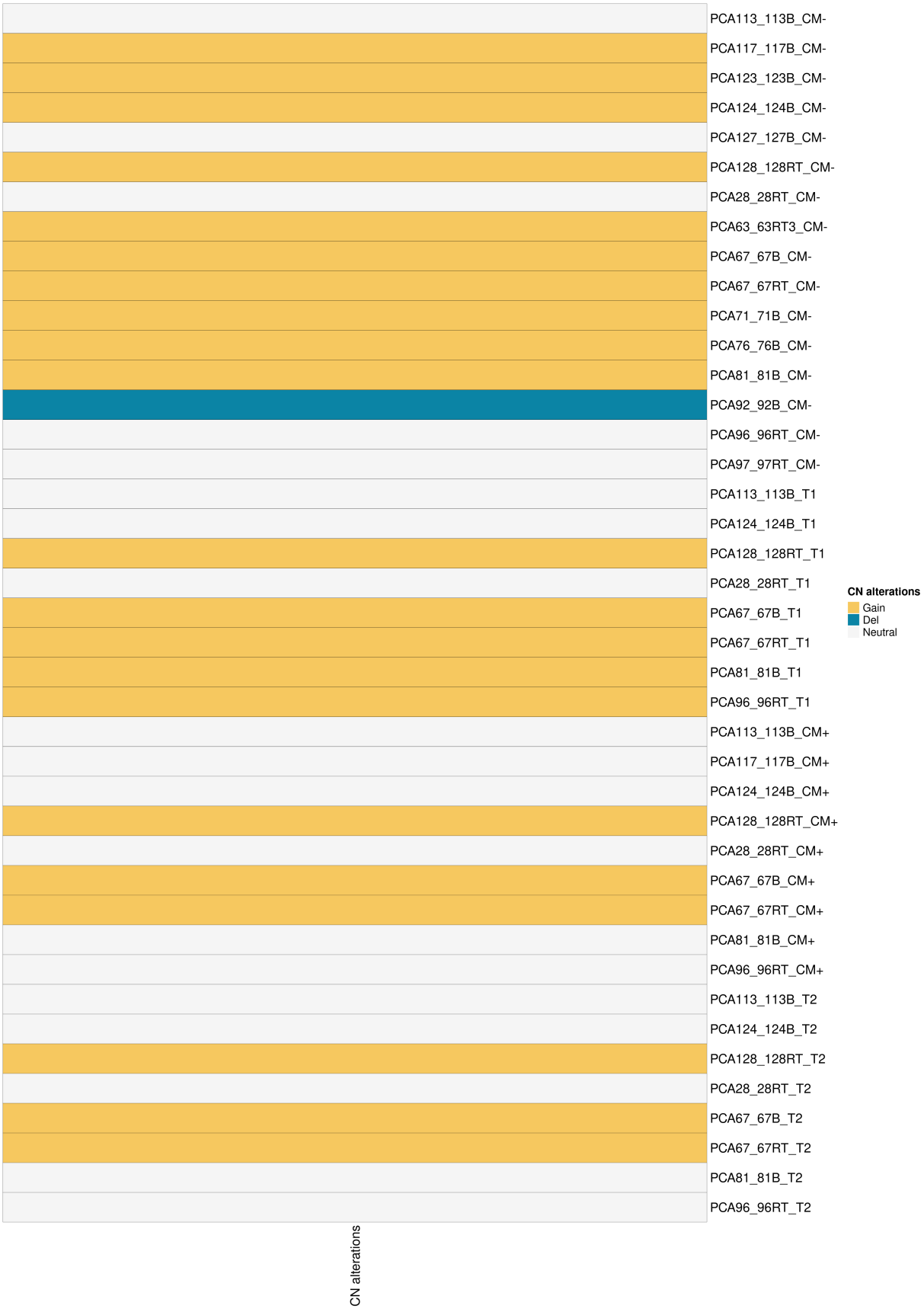

Figure 11: CNA for KRAS, where samples are grouped according to culture media condition

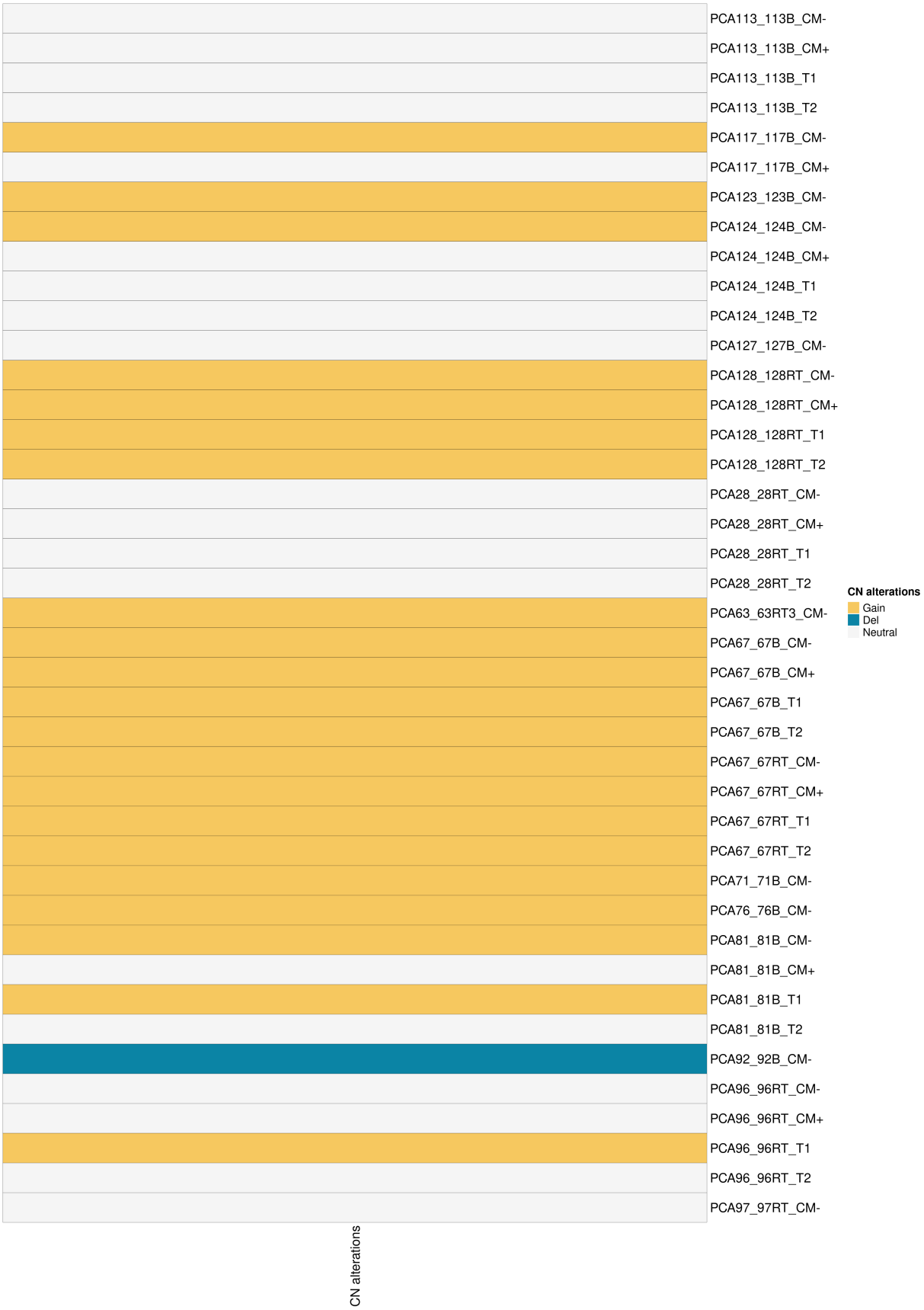

Figure 12: CNA for KRAS, where samples are grouped by patient sample

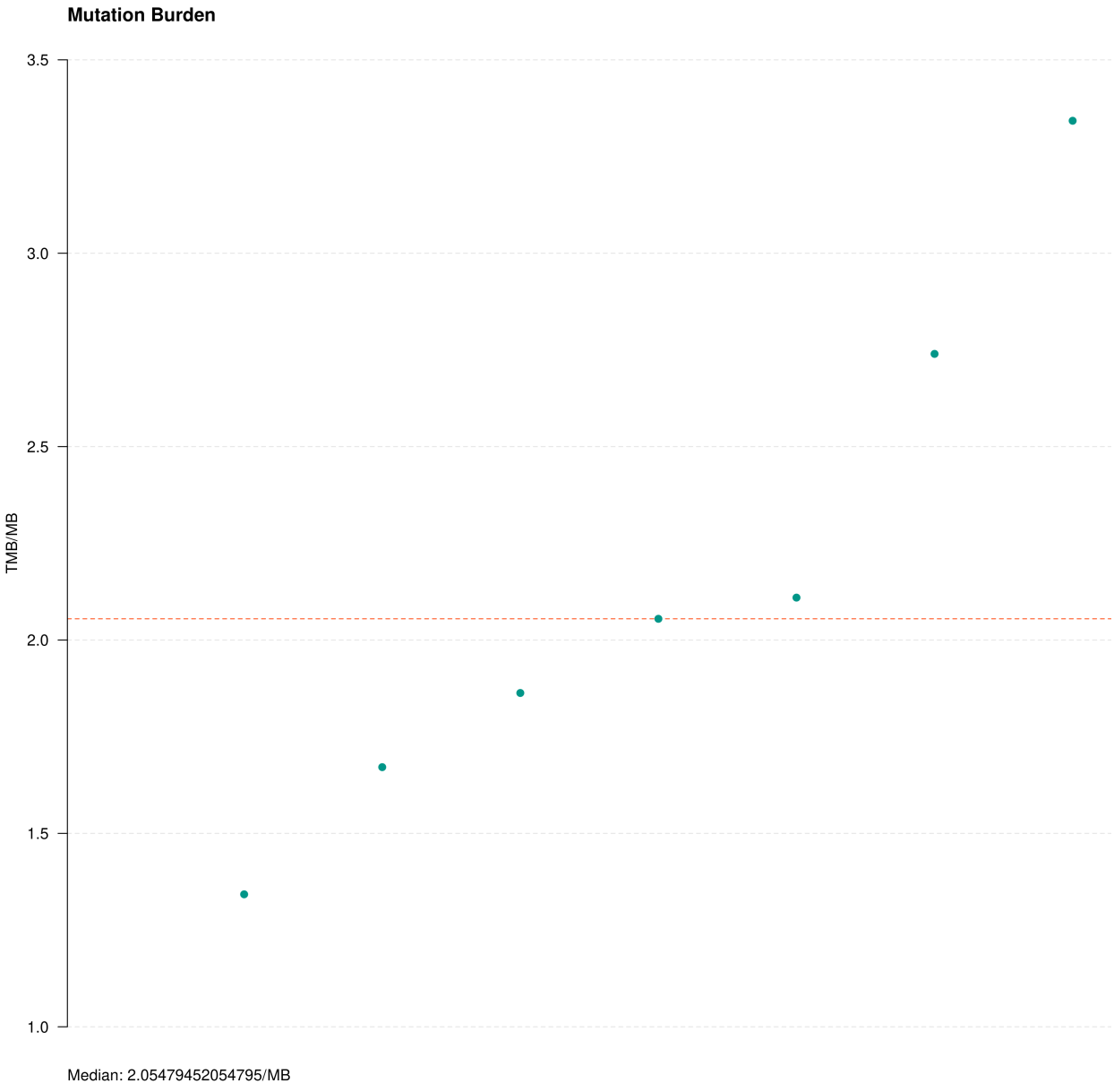

Figure 13: TMB scores in the tissue samples

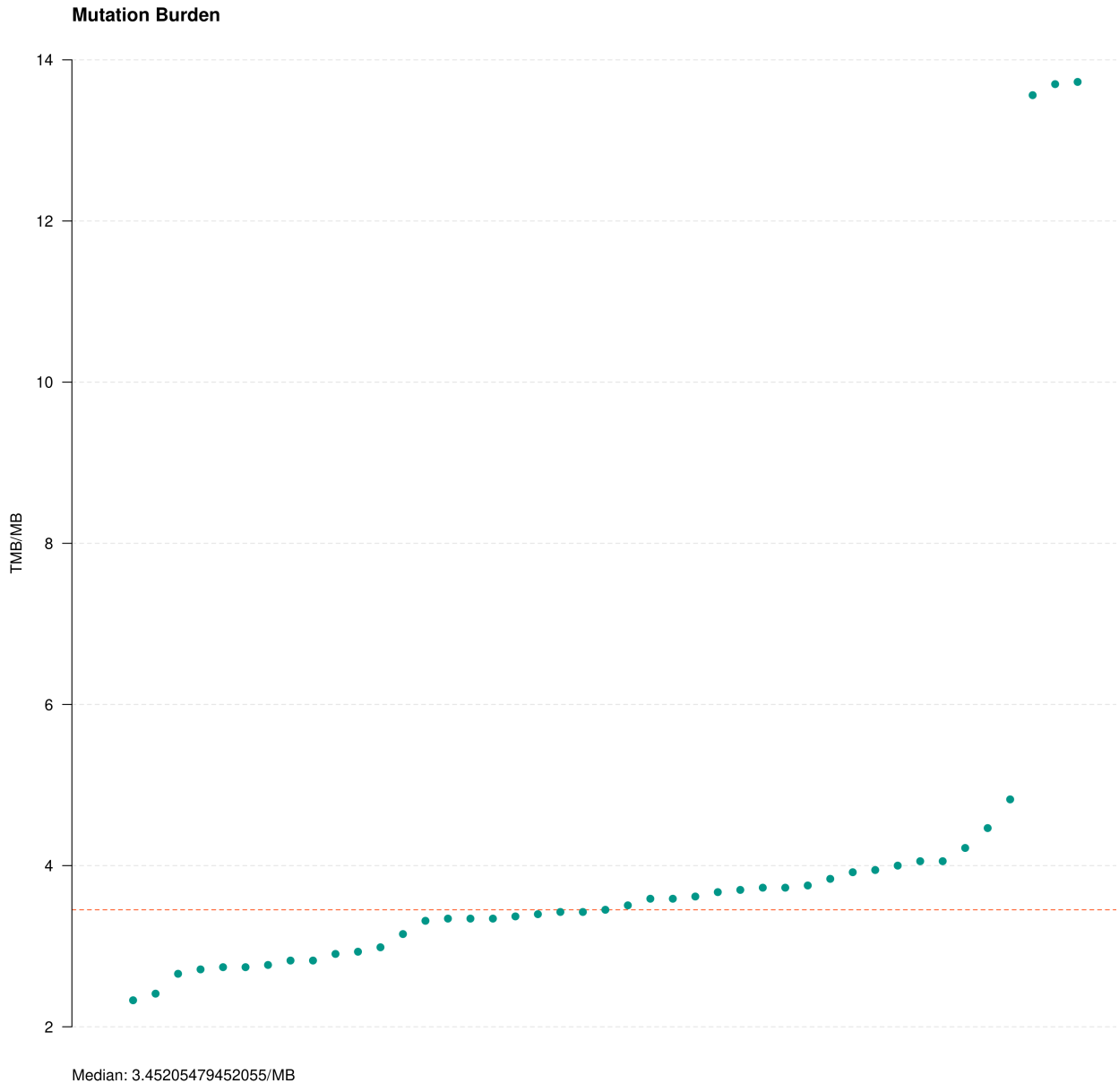

Figure 14: TMB scores in the PDOs

Figure 15: Msisensorpro results show low levels of microsatellite instabillity, with most samples have a score less than 1%

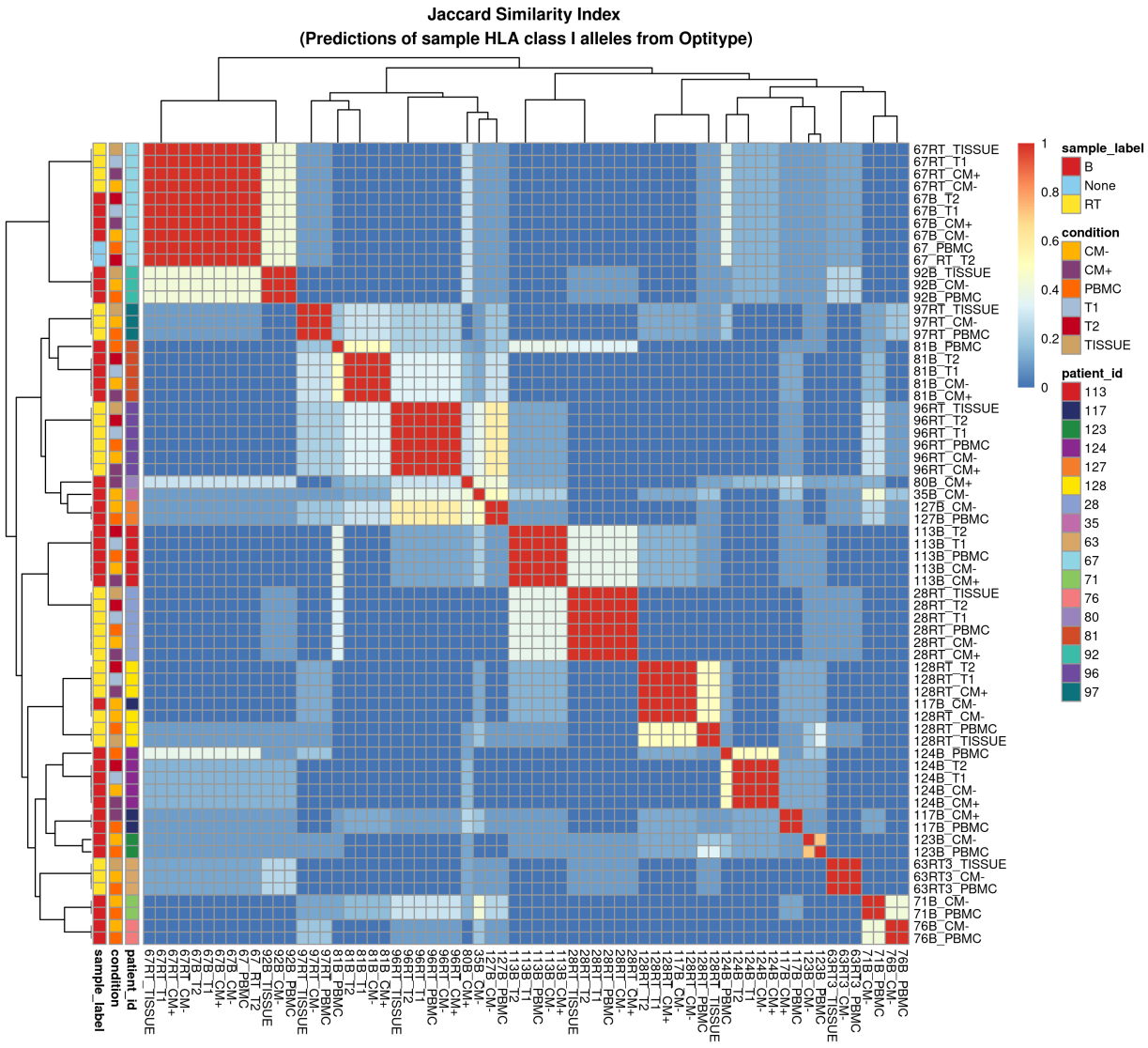

Figure 16: Clustering samples by the Jaccard similarity index by the predicted HLA class I alleles.