WES analysis of PDO and tissue samples in PDAC

Kane Toh*

2024-01-16

Contents

1 Methods

1.1 Analysis of whole Exome sequencing data

1.1.1 SNP/Indels

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`.

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

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

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 `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.

1.1.2 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.

1.1.3 Microsatellite instability

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

1.2 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.

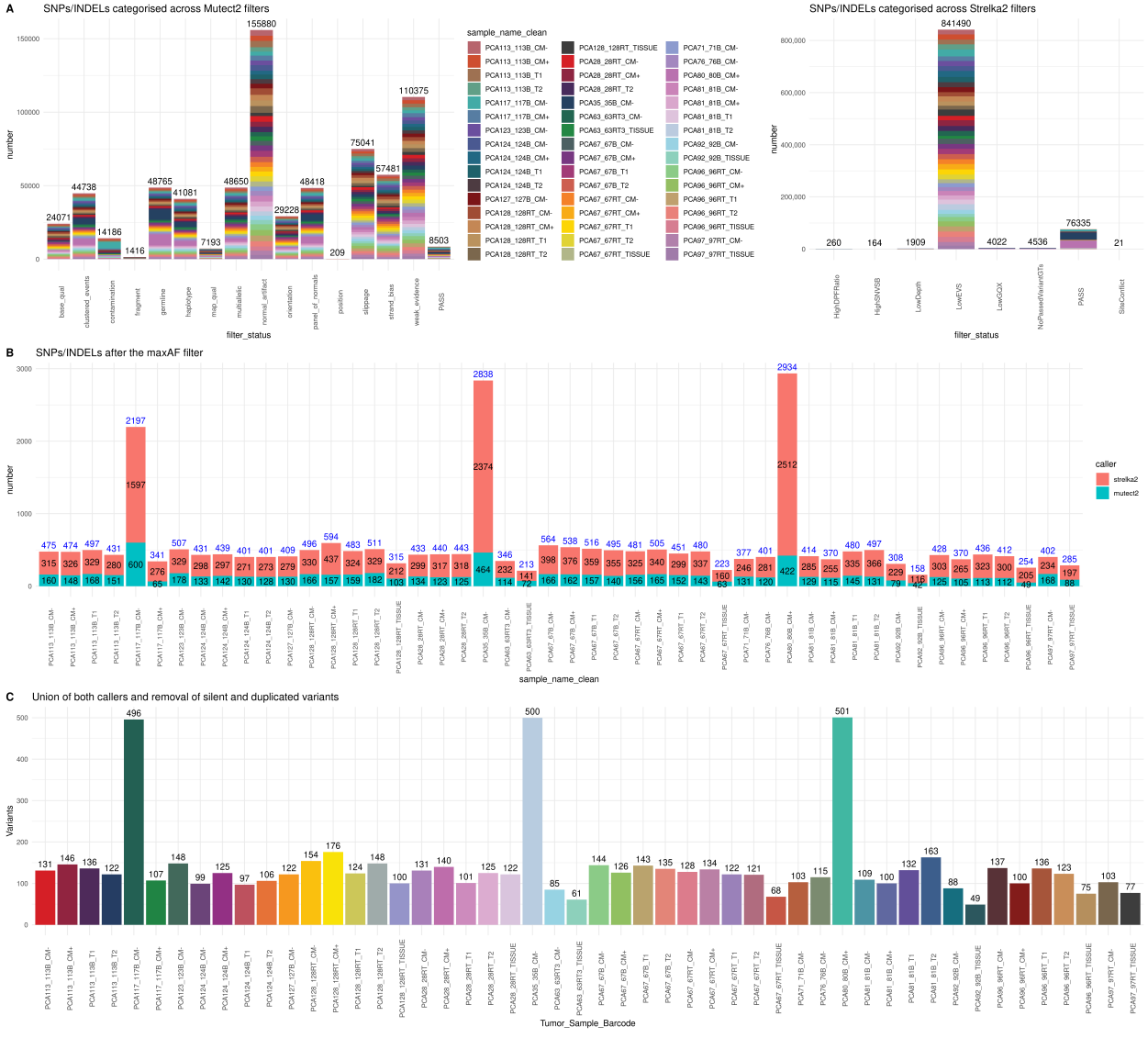


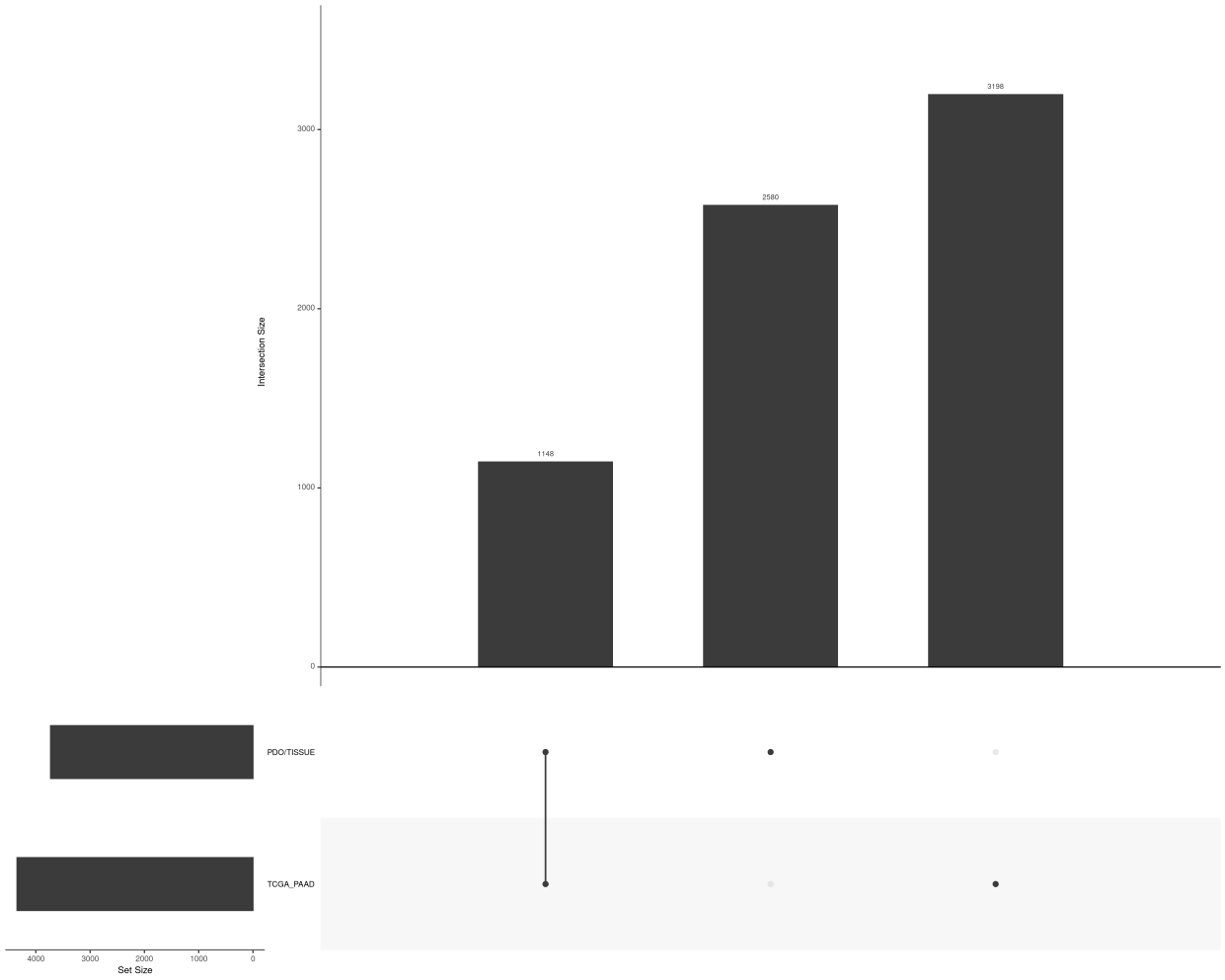
Figure 1: Summary of filtering steps used in short

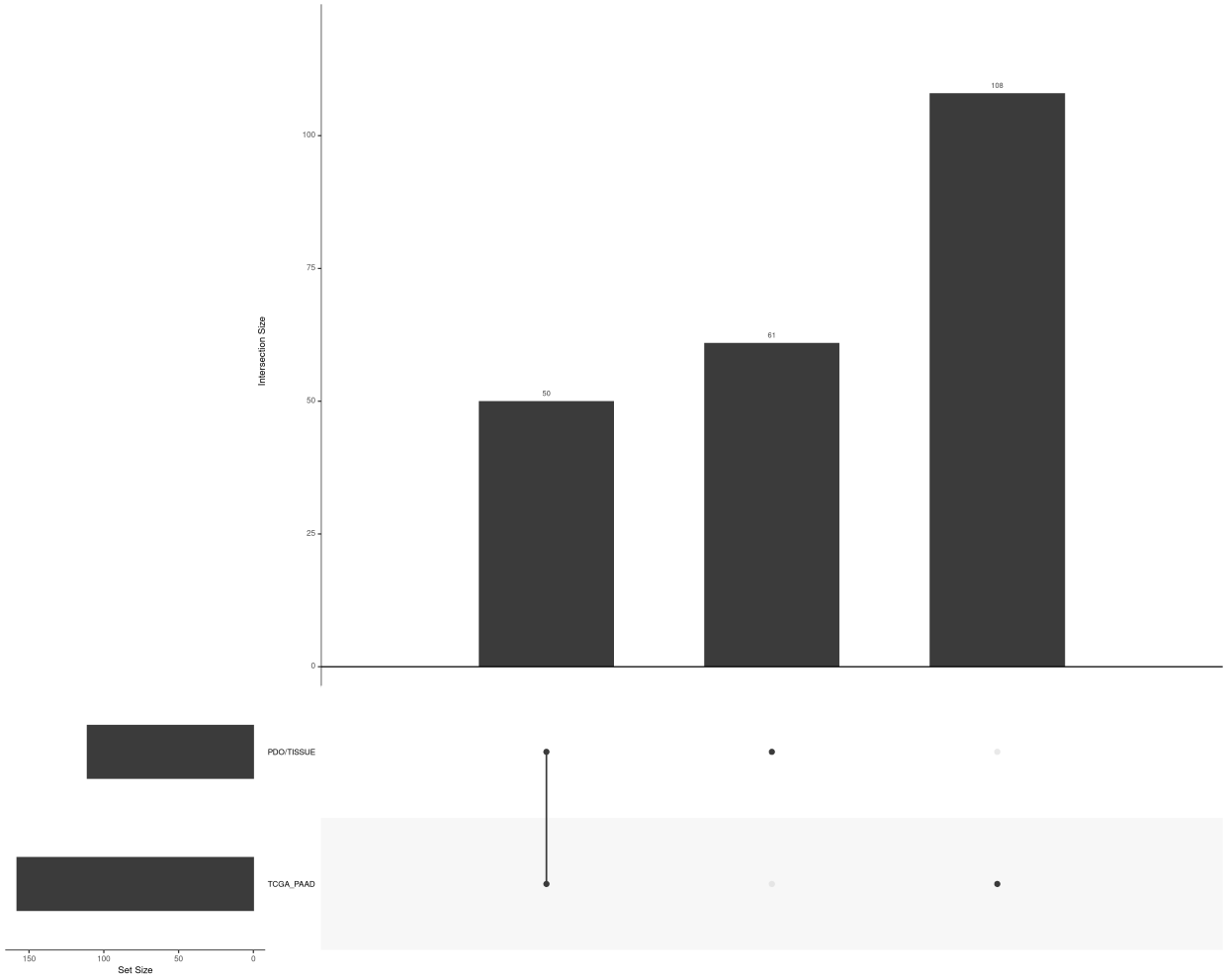
2.2 Comparing against the TCGA PAAD dataset

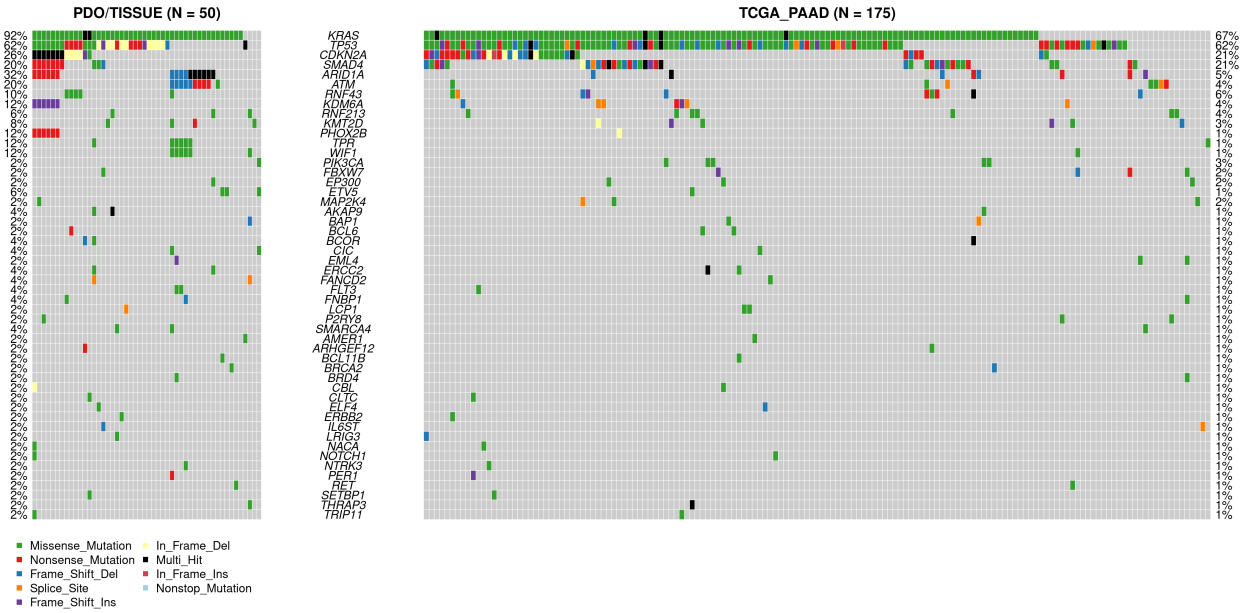
To provide external validation of the variants called in our PDAC_TISSUE dataset, we compared our somatic variants and gene mutations against the corresponding public TCGA_PAAD dataset. As shown in the upSet plot (Figure **??**), we found that around 26% of all mutated genes were shared between both datasets. Of the 111 mutated genes in our dataset that were also listed in the Catalogue of Somatic Mutations in Cancer (COSMIC) database (Figure **??**), 50 of them (31%) were also shared with the COSMIC genes in the TCGA

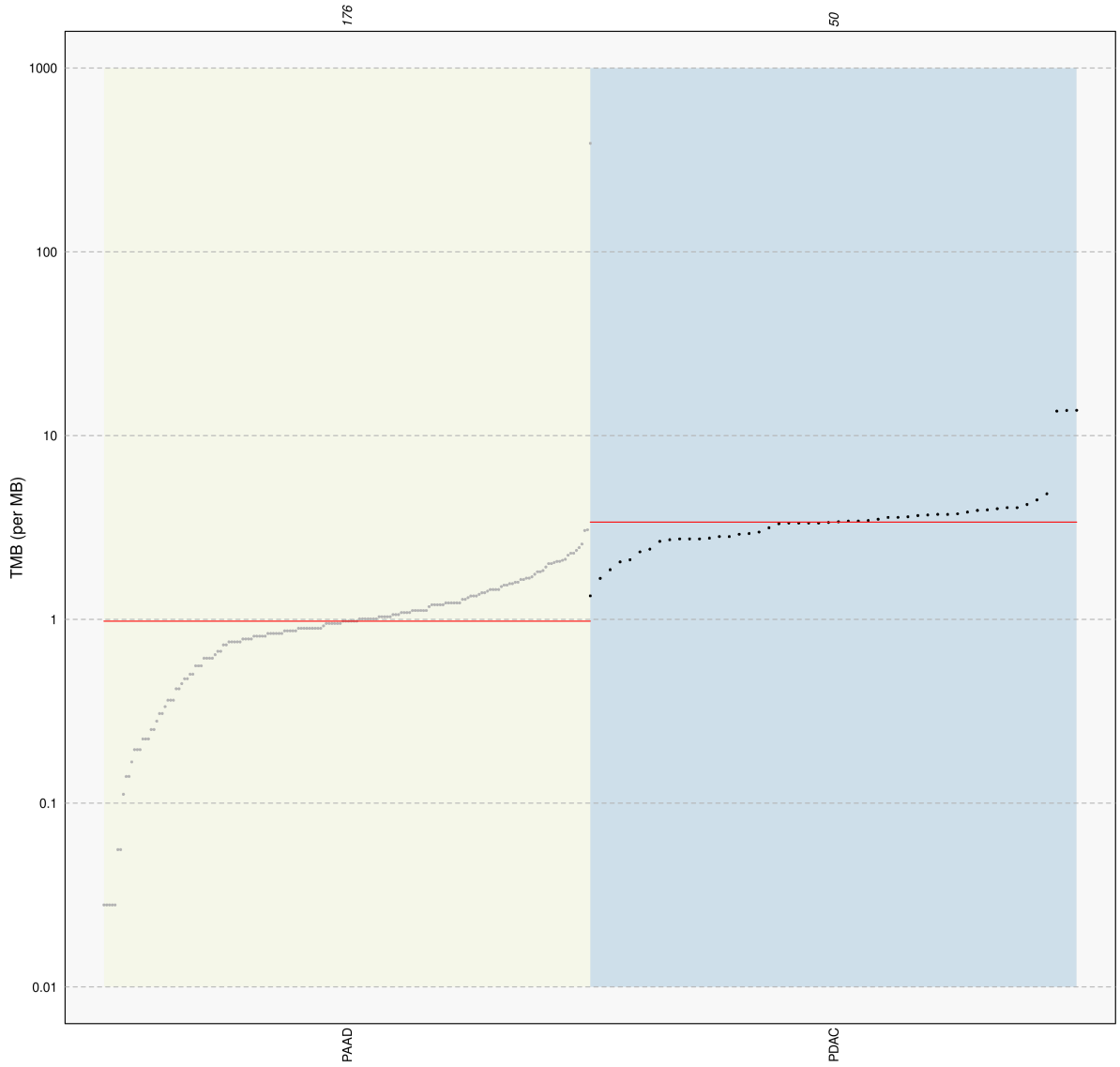
dataset. Importantly, the co-oncoplots in Figure **??** show that the known driver genes such as KRAS (67% TCGA vs 92% PDO/TISSUE), TP53 (62% TCGA and PDO/TISSUE), CDKN2A (21% TCGA and 26% PDO/TISSUE) and SMAD4 (21% TCGA and 20% PDO/TISSUE) driver gene mutations were recapitulated in our PDAC/TISSUE 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 **??**), 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.




Intersection Size
100
75
50
25
0
50
61
108
PDO/TISSUE
TCGA_PAAD
150
100
50
0
Set Size


PDO/TISSUE (N = 50)
TCGA_PAAD (N = 175)
KRAS
TP53
CDKN2A
SMAD4
ARID1A
ATM
RNF43
KDM6A
RNF213
KMT2D
PHOX2B
TPR
WIF1
PIK3CA
FBXW7
EP300
ETV5
MAP2K4
AKAP9
BAP1
BCL6
BCOR
CIC
EML4
ERCC2
FANCD2
FLT3
FNBP1
LCP1
P2RY8
SMARCA4
AMER1
ARHGEF12
BCL11B
BRCA2
BRD4
CBL
CLTC
ELF4
ERBB2
IL6ST
LRIG3
NACA
NOTCH1
NTRK3
PER1
RET
SETBP1
THRAP3
TRIP11
Missense_Mutation
Nonsense_Mutation
Frame_Shift_Del
Splice_Site
Frame_Shift_Ins
In_Frame_Del
Multi_Hit
In_Frame_Ins
Nonstop_Mutation



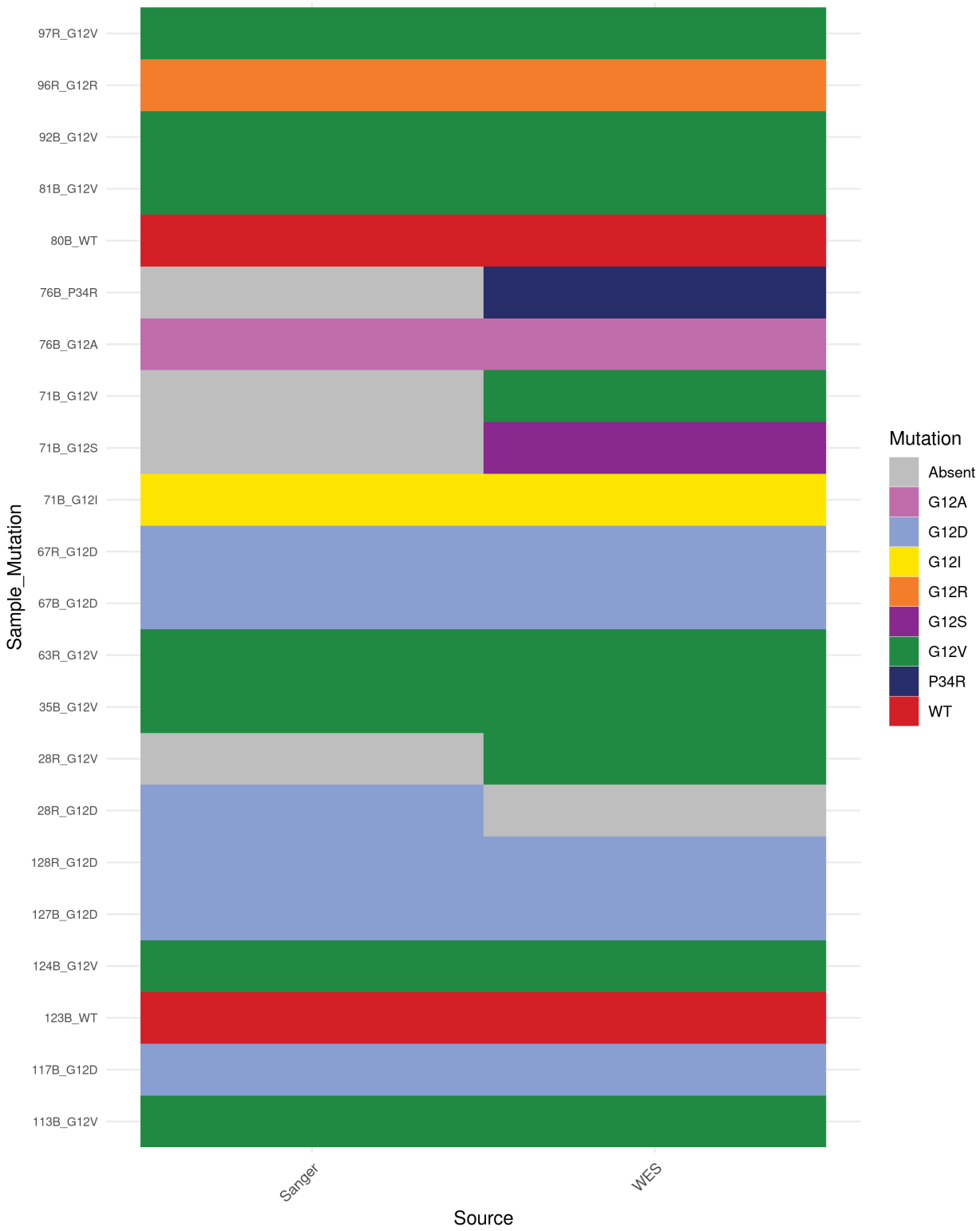
2.3 Somatic SNP/INDEL mutational landscape

Across the PDAC samples, we observed that all samples possess the KRAS mutation. Of the 46 samples with the KRAS mutation, 24 samples (52%) exhibit G12V, 16 (35%) G12V and 5 G12R (11%). We validated our WES results with Sanger Sequencing at the KRAS loci. As shown in Figure **??**), 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.

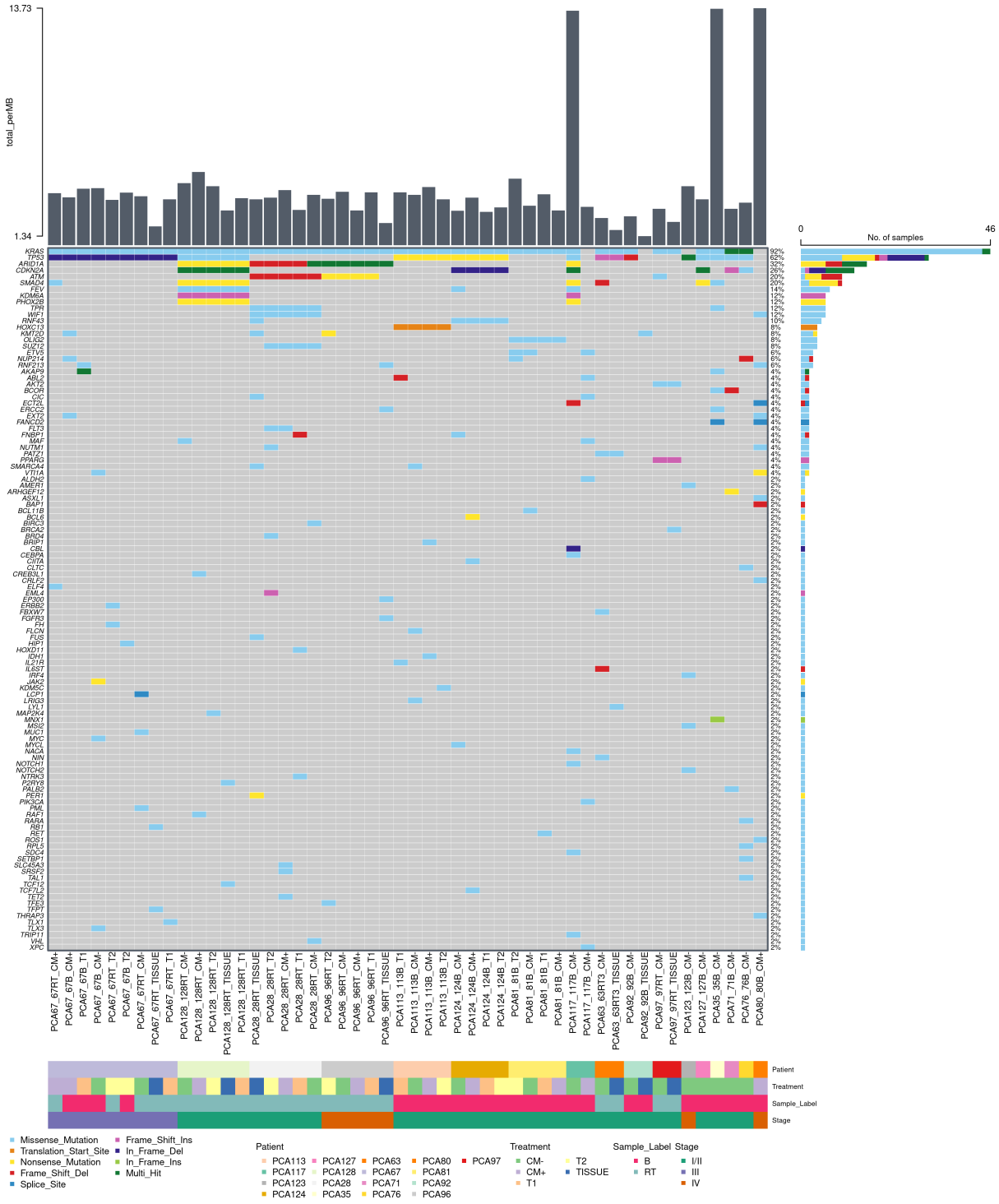
In total, 14/17 (82%) samples have identical KRAS mutations identified by both WES and Sanger sequencing.

The KRAS wild-type alleles in 2 samples were concordant in both sequencing methods.



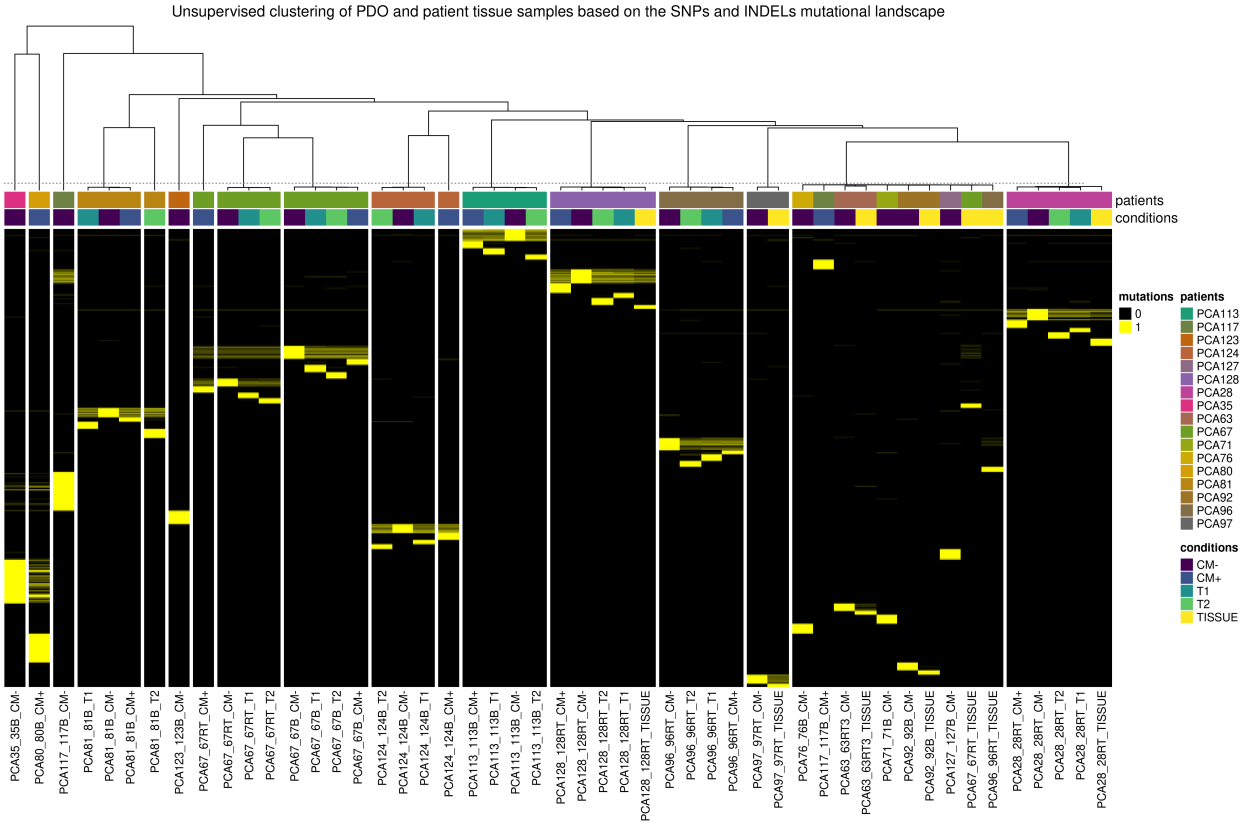
62% of the samples have mutations in the TP53 gene, with each patient harboring a different mutation (See `TP53_lollipop.png` for the range of mutations displayed).

Other genes that were mutated in relatively high proportion include the ARID1A (32%), ATM (20%), CDKN2A (26%) and SMAD4 (20%) mutations. The oncoplot shown in Figure **??** shows the mutated genes that were also present in the COSMIC database, thus emphasising known mutations with causal implications in carcinogenesis.


13.73
1.34
total_perMB
0
No. of samples
46
Patient
Treatment
Sample_Label
Stage
Missense_Mutation
Translation_Start_Site
Nonsense_Mutation
Frame_Shift_Del
Splice_Site
Frame_Shift_Ins
In_Frame_Del
In_Frame_Ins
Multi_Hit
Patient
PCA113
PCA117
PCA123
PCA124
PCA127
PCA128
PCA28
PCA35
PCA63
PCA67
PCA71
PCA76
PCA80
PCA81
PCA92
PCA96
PCA97
Treatment
CM-
CM+
T1
T2
TISSUE
Sample_Label
B
RT
Stage
I/II
III
IV

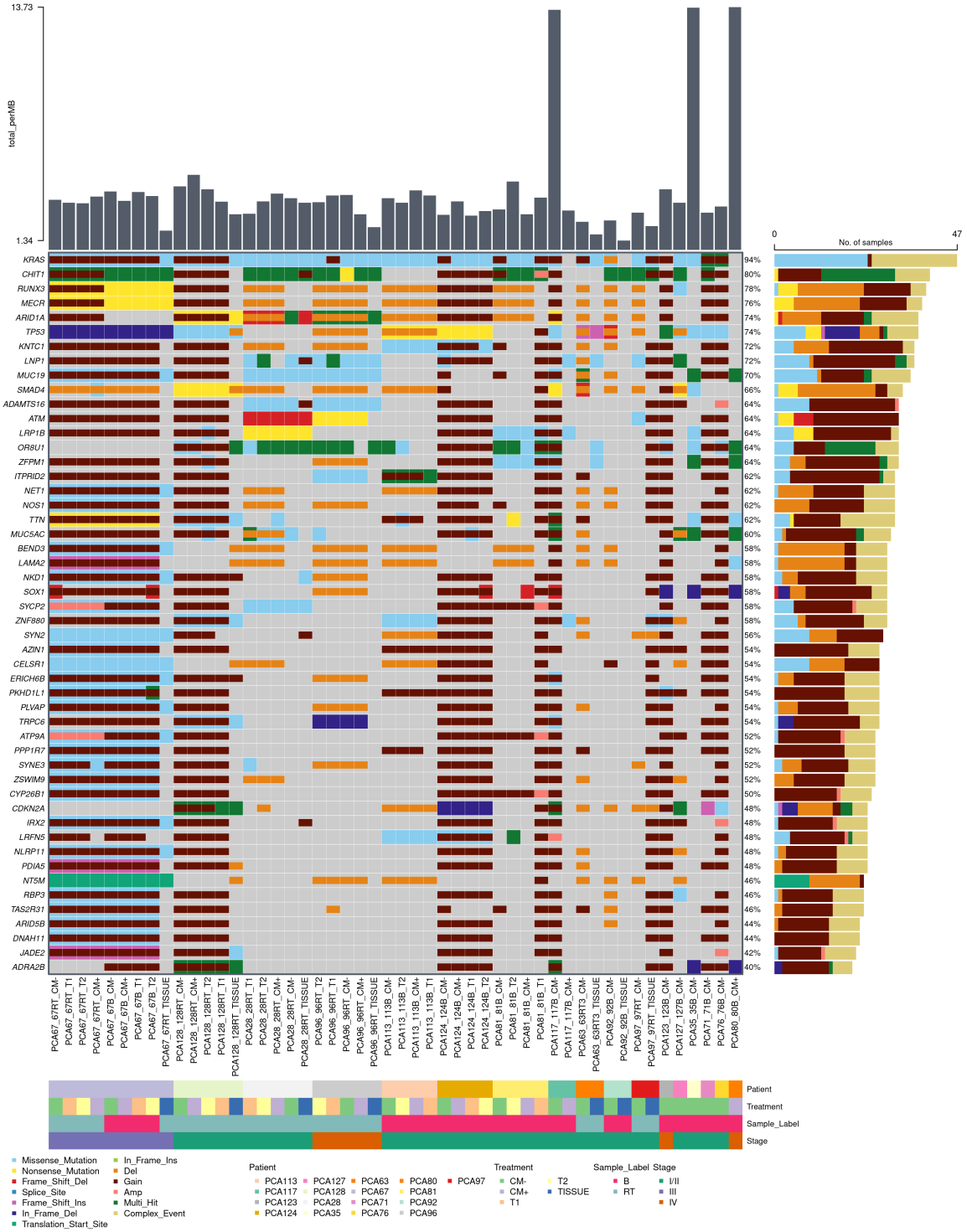
2.4 Unsupervised clustering of PDO and patient samples

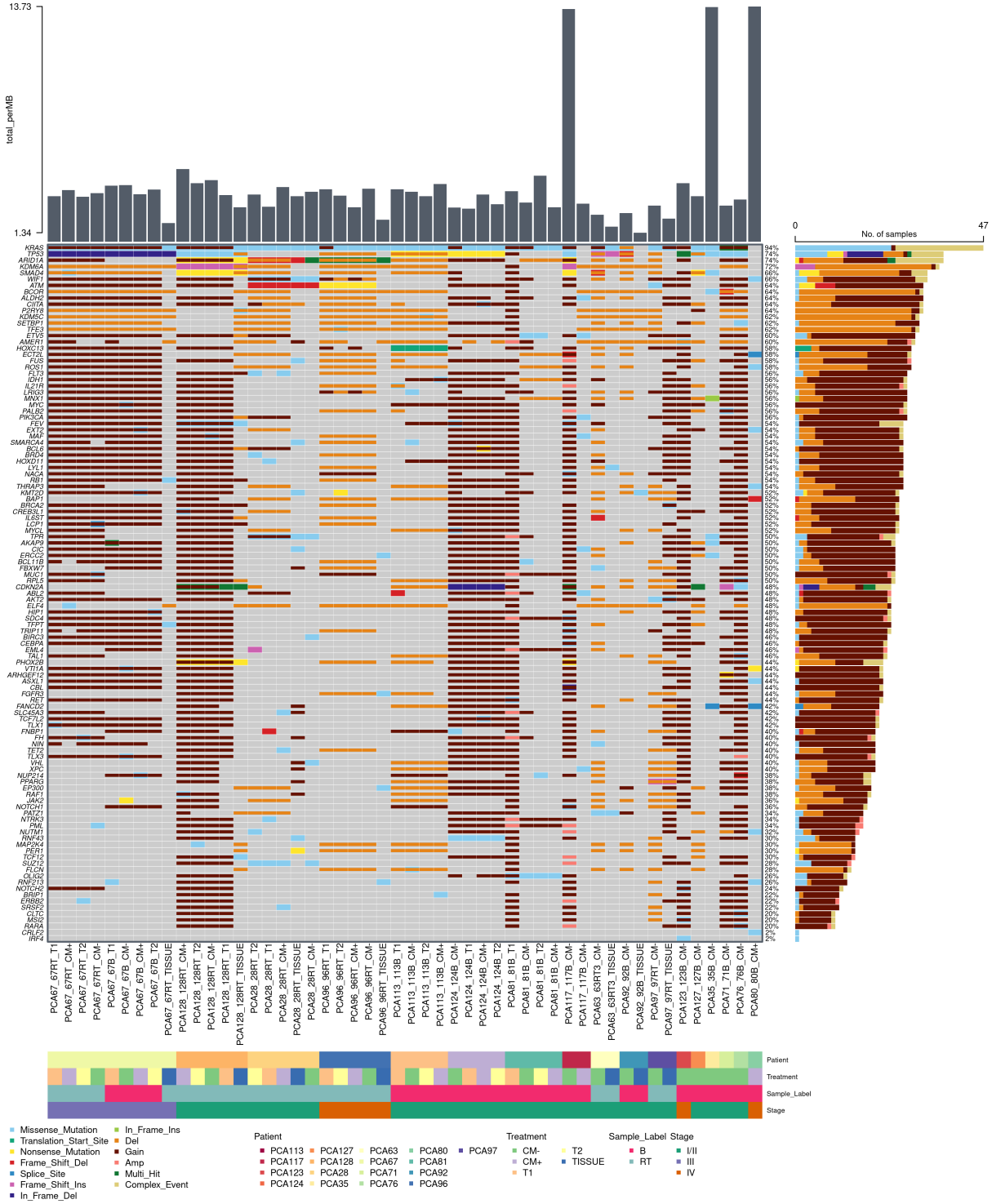
We carried out an unsupervised clustering of the 7 patient tissue samples along with the remaining PDO samples, based on their somatic mutational profiles (Figure **??**) . Specifically, a binary matrix was constructed based on whether a sample possesses a specific mutation within the full set of mutated 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.



2.5 Copy number alterations

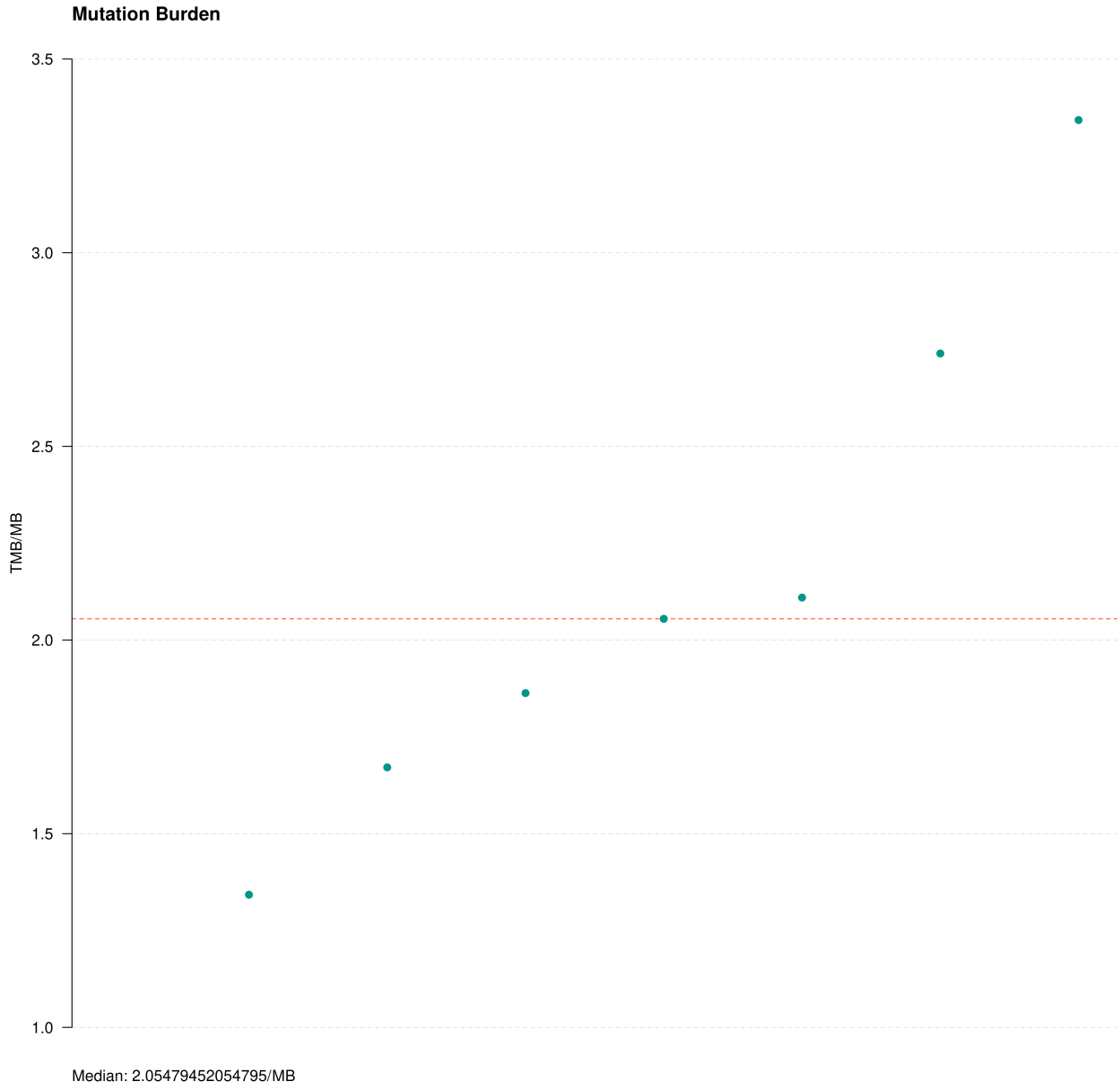
We incorporated the CNA results into the oncoplot and display them in this report. Two oncoplots are shown, one for the top 50 genes that are mutated in the PDO/TISSUE samples (Figure **??**). Another oncoplot highlights all the COSMIC genes that are mutated in the samples (Figure **??**). The TMB score is indicated in the top panel (identical result to `tmb_pdac.png`).

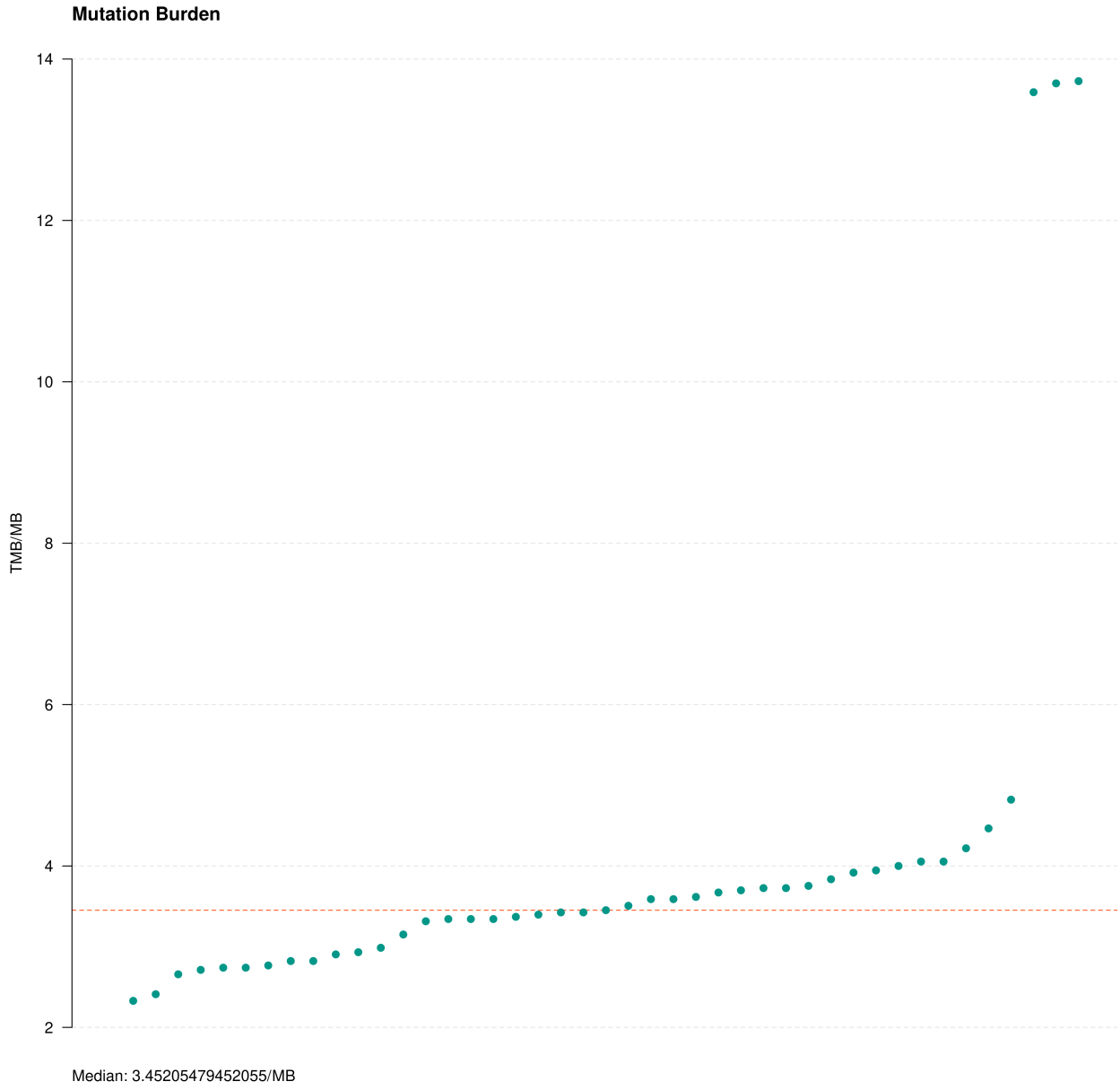

13.73
1.34
total_perMB
0
No. of samples
47
KRAS 94%
CHIT1 80%
RUNX3 78%
MECR 76%
ARID1A 74%
TP53 74%
KNTC1 72%
LNP1 72%
MUC19 70%
SMAD4 66%
ADAMTS16 64%
ATM 64%
LRP1B 64%
OR8U1 64%
ZFPM1 64%
ITPRID2 62%
NET1 62%
NOS1 62%
TTN 62%
MUC5AC 60%
BEND3 58%
LAMA2 58%
NKD1 58%
SOX1 58%
SYCP2 58%
ZNF880 58%
SYN2 56%
AZIN1 54%
CELSR1 54%
ERICH6B 54%
PKHD1L1 54%
PLVAP 54%
TRPC6 54%
ATP9A 52%
PPP1R7 52%
SYNE3 52%
ZSWIM9 52%
CYP26B1 50%
CDKN2A 48%
IRX2 48%
LRFN5 48%
NLRP11 48%
PDIA5 48%
NT5M 46%
RBP3 46%
TAS2R31 46%
ARID5B 44%
DNAH11 44%
JADE2 42%
ADRA2B 40%
PCA67_67RT_CM-
PCA67_67RT_T1
PCA67_67RT_T2
PCA67_67RT_CM+
PCA67_67B_CM-
PCA67_67B_CM+
PCA67_67B_T1
PCA67_67B_T2
PCA67_67RT_TISSUE
PCA128_128RT_CM-
PCA128_128RT_CM+
PCA128_128RT_T2
PCA128_128RT_T1
PCA128_128RT_TISSUE
PCA28_28RT_T1
PCA28_28RT_T2
PCA28_28RT_CM+
PCA28_28RT_CM-
PCA28_28RT_TISSUE
PCA96_96RT_T2
PCA96_96RT_T1
PCA96_96RT_CM-
PCA96_96RT_CM+
PCA96_96RT_TISSUE
PCA113_113B_CM-
PCA113_113B_T2
PCA113_113B_CM+
PCA113_113B_T1
PCA124_124B_CM-
PCA124_124B_CM+
PCA124_124B_T1
PCA124_124B_T2
PCA81_81B_CM-
PCA81_81B_T2
PCA81_81B_CM+
PCA81_81B_T1
PCA117_117B_CM-
PCA117_117B_CM+
PCA63_63RT3_CM-
PCA63_63RT3_TISSUE
PCA92_92B_CM+
PCA92_92B_TISSUE
PCA97_97RT_CM-
PCA97_97RT_TISSUE
PCA123_123B_CM-
PCA127_127B_CM+
PCA35_35B_CM-
PCA71_71B_CM-
PCA76_76B_CM-
PCA80_80B_CM+
Patient
Treatment
Sample_Label
Stage
Missense_Mutation
Nonsense_Mutation
Frame_Shift_Del
Splice_Site
Frame_Shift_Ins
In_Frame_Del
Translation_Start_Site
In_Frame_Ins
Del
Gain
Amp
Multi_Hit
Complex_Event
Patient
PCA113 PCA127 PCA63 PCA80 PCA97
PCA117 PCA128 PCA67 PCA81
PCA123 PCA28 PCA71 PCA92
PCA124 PCA35 PCA76 PCA96
Treatment
CM- T2
CM+ TISSUE
T1
Sample_Label
B
RT
Stage
I/II
III
IV


13.73
1.34
total_perMB
0
47
No. of samples
KRAS 94%
TP53 74%
ARID1A 74%
KDM6A 72%
SMAD4 66%
WIF1 66%
ATM 64%
BCOR 64%
ALDH2 64%
CIITA 64%
P2RY8 64%
KDM5C 62%
SETBP1 62%
TFE3 62%
ETV5 60%
AMER1 60%
HOXC13 58%
ECT2L 58%
FUS 58%
ROS1 58%
FLT3 56%
IDH1 56%
IL21R 56%
LRIG3 56%
MNX1 56%
MYC 56%
PALB2 56%
PIK3CA 56%
FEV 54%
EXT2 54%
MAF 54%
SMARCA4 54%
BCL6 54%
BRD4 54%
HOXD11 54%
LYL1 54%
NACA 54%
RB1 54%
THRAP3 54%
KMT2D 52%
BAP1 52%
BRCA2 52%
CREBBP 52%
IL6ST 52%
LCP1 52%
MYCL 52%
TPR 50%
AKAP9 50%
CIC 50%
ERCC2 50%
BCL11B 50%
FBXW7 50%
MUC1 50%
RPL5 50%
CDKN2A 48%
ABL2 48%
AKT2 48%
ELF4 48%
HIP1 48%
SDC4 48%
TPPT 48%
TRIP11 48%
BIRC3 48%
CEBPA 46%
EML4 46%
TAL1 46%
PHOX2B 44%
VTI1A 44%
ARHGEF12 44%
ASXL1 44%
CBL 44%
FGFR3 44%
PLCG1 44%
FANCD2 42%
SLC45A3 42%
TCF7L2 42%
TLX1 42%
PNBP1 40%
FH 40%
NIN 40%
TET2 40%
TLX3 40%
VHL 40%
XPC 40%
NUP214 38%
PPARG 38%
EP300 38%
RAF1 38%
JAK2 36%
NOTCH1 36%
PATZ1 34%
NTRK3 34%
PML 34%
NUTM1 32%
RNF43 30%
MAP2K4 30%
PER1 30%
TCF12 30%
SUZ12 28%
FLCN 28%
OLIG2 26%
RNF213 26%
NOTCH2 24%
BRIP1 22%
ERBB2 22%
SRSF2 22%
CLTC 20%
MSI2 20%
NARA 20%
CRLF2 2%
IRF4 2%
PCA67_67RT_T1
PCA67_67RT_CM+
PCA67_67RT_T2
PCA67_67RT_CM-
PCA67_67B_T1
PCA67_67B_CM-
PCA67_67B_CM+
PCA67_67B_T2
PCA67_67RT_TISSUE
PCA128_128RT_CM+
PCA128_128RT_T2
PCA128_128RT_CM-
PCA128_128RT_T1
PCA128_128RT_TISSUE
PCA28_28RT_T2
PCA28_28RT_T1
PCA28_28RT_CM+
PCA28_28RT_TISSUE
PCA28_28RT_CM-
PCA96_96RT_T1
PCA96_96RT_T2
PCA96_96RT_CM+
PCA96_96RT_CM-
PCA96_96RT_TISSUE
PCA113_113B_T1
PCA113_113B_CM-
PCA113_113B_T2
PCA113_113B_CM+
PCA124_124B_CM-
PCA124_124B_T1
PCA124_124B_CM+
PCA124_124B_T2
PCA81_81B_T1
PCA81_81B_CM-
PCA81_81B_T2
PCA81_81B_CM+
PCA117_117B_CM-
PCA117_117B_CM+
PCA63_63RT3_CM-
PCA63_63RT3_TISSUE
PCA92_92B_CM-
PCA92_92B_TISSUE
PCA97_97RT_CM-
PCA97_97RT_TISSUE
PCA123_123B_CM-
PCA127_127B_CM-
PCA35_35B_CM-
PCA71_71B_CM-
PCA76_76B_CM-
PCA80_80B_CM+
Patient
Treatment
Sample_Label
Stage
Missense_Mutation
Translation_Start_Site
Nonsense_Mutation
Frame_Shift_Del
Splice_Site
Frame_Shift_Ins
In_Frame_Del
In_Frame_Ins
Del
Gain
Amp
Multi_Hit
Complex_Event
Patient
PCA113
PCA117
PCA123
PCA124
PCA127
PCA128
PCA28
PCA35
PCA63
PCA67
PCA71
PCA76
PCA80
PCA81
PCA92
PCA96
PCA97
Treatment
CM-
CM+
T1
T2
TISSUE
Sample_Label
B
RT
Stage
I/II
III
IV

2.6 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 **??** vs 3.45 mutations/MB **??**) in the organoid samples.





3 Microsatellite instability

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


PCA113B
PCA117B
PCA123B
PCA124B
PCA127B
PCA128R
PCA28R
PCA63R
PCA67B
PCA67R
PCA71B
PCA76B
PCA81B
PCA92B
PCA96R
PCA97R
pert
sample_name
Patient
PCA113B
PCA117B
PCA123B
PCA124B
PCA127B
PCA128R
PCA28R
PCA63R
PCA67B
PCA67R
PCA71B
PCA76B
PCA81B
PCA92B
PCA96R
PCA97R