GENEWIZ NGS Data Report

1. Project Summary

Customer	GENEWIZ NGS
Email	ngs@genewiz.com
Quote Number	GW0101001
Configuration	Illumina HiSeq, PE 2x150

2. Description of Workflow

2.1 WES library preparation workflow

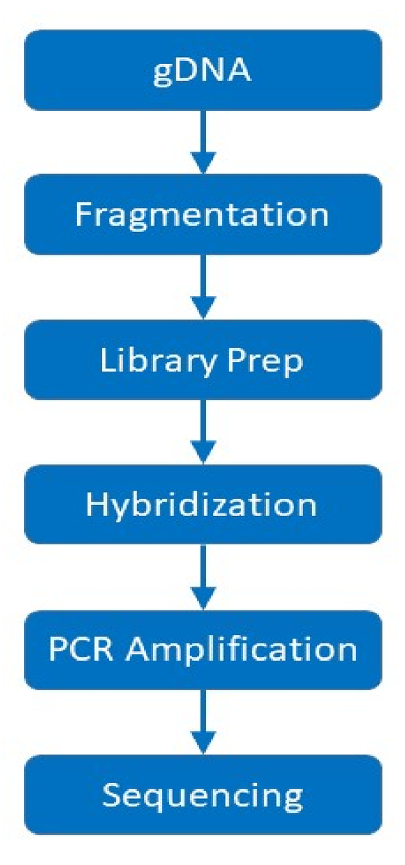

Figure 2.1 WES library preparation and sequencing workflow

2.2 Bioinformatics Workflow

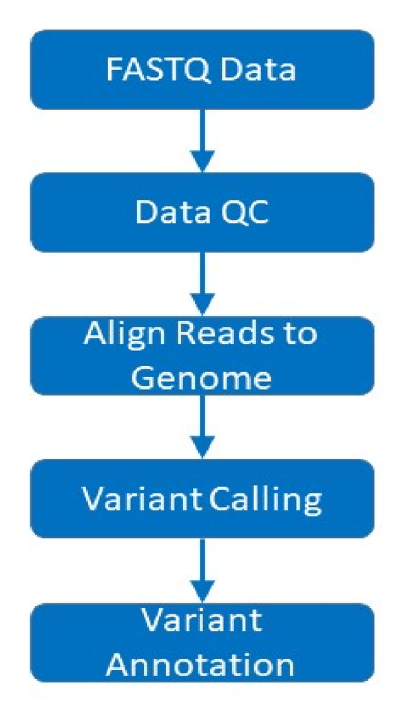

Figure 2.2 WES bioinformatics analysis workflow

3. Analysis

3.1 Sample sequencing statistics

Raw BCL files generated by the sequencer were converted to fastq files for each sample using bcl2fastq v.2.19. The summary statistics for the raw data are shown in Table 3.1.

Show 10 entries | Search:

Table 3.1 Sample sequencing statistics

Project	Sample ID	Barcode Sequence	# Reads	Yield (Mbases)	Mean Quality Score	% Bases >= 30
GW0101001	T18	TCCGCGAA+TAAGATTA	52,825,656	15,847	39.15	95.01
GW0101001	B17	CGGCTATG+AGGATAGG	37,698,603	11,309	38.60	93.13
GW0101001	T17	TCCGCGAA+CTTCGCCT	48,948,991	14,685	38.97	94.40
GW0101001	T1	TCCGCGAA+GCCTCTAT	45,604,517	13,681	38.97	94.41
GW0101001	B1	TAATGCGC+GTCAGTAC	41,215,620	12,365	38.49	92.79

Select Columns | Download

Showing 1 to 5 of 5 entries — Previous 1 Next

3.2 Overall sequencing statistics

Overall sequencing statistics are shown in table 3.2.

Table 3.2 Overall sequencing statistics

Project	# Reads	Yield (Mbases)	Mean Quality Score	% Bases >= 30
GW0101001	984,630,696	295,388	38.75	93.65

3.3 Mapping sequence reads to the referene genome

Sequence reads were trimmed to remove possible adapter sequences and nucleotides with poor quality using Trimmomatic v.0.38. The trimmed reads were mapped to the reference genome using the Illumina Dragen Bio-IT Platform . BAM files were generated as a result of this step. Table 3.3 shows the alignment statistics generated by Picard Tools.

Show 10 entries | Search:

Table 3.3 Sample alignment summary

Sample	Total Reads	Total Cleaned Reads	Unique Reads	% Unique Reads	% Aligned Unique Reads	Mean Bait Coverage	% Target Bases above 20X	Target Size
B1	82,431,240	82,372,120	68,310,081	82.93	99.72	131	88.35	60456963
B17	75,397,206	75,355,262	63,852,187	84.73	99.72	124	87.69	60456963
T1	91,209,034	91,189,696	63,637,347	69.79	99.67	158	58.10	60456963
T17	97,897,982	97,876,858	63,512,712	64.89	99.64	172	48.09	60456963
T18	105,651,312	105,633,450	55,357,018	52.40	99.53	185	75.88	60456963

Select Columns | Download

Showing 1 to 5 of 5 entries — Previous 1 Next

3.4 Variant calling

Somatic variants were called using the Illumina Dragen Bio-IT Platform in somatic mode. Paired normal samples were used in the process if provided. A panel of normal (PON) if contains over 50 non-related sasmples was also used to reduce false positives. Variants were further filtered and any variants in the follow categories were considered as false positives and removed: (1) marked as common variants in dbSNP build 151 and (2) non_cancer_AC > 5 in gnomad exome database r2.1.1.

The filtered VCF was then annotated with Ensembl Variant Effect Predictor (VEP) v95. Table 3.4 summarize the variant calling results of all the samples.

Show 10 entries | Search:

Table 3.4 Variant calling summary

Sample	Total variants	SNV	insertion	deletion	Known variants	Novel variants
T1	260	232	3	25	53	207
T17	446	418	7	21	114	332
T18	902	808	30	64	625	277

Select Columns | Download

Showing 1 to 3 of 3 entries — Previous 1 Next

For each variant that is mapped to the reference genome, all overlapping Ensembl transcripts were identified, and the effects that each allele of the variant may have on each transcript were predicted by VEP. The set of consequence terms was defined by the Sequence Ontology (SO). Table 3.5 summarize the effects of variants for samples in the cohort. Note that each allele of each variant may have a different effect in different transcripts. Effects are color coded for severity level (High, Moderate, Low, Modifier).

Show 10 entries | Search:

Sample	splice_acceptor_variant	splice_donor_variant	stop_gained	frameshift_variant	stop_lost	inframe_ins
T1	10	13	22	25	2	
T17	6	8	25	36	2	
T18	6	33	34	33		

Select Columns | Download

Showing 1 to 3 of 3 entries — Previous 1 Next

3.5 Cohort Analysis

The most severe impact was selected for each variant and they are used for downstream cohort analysis.

3.5.1 Summary statistics of variants at cohort level.

Impact of the variants were classified based on MAF document spcifications. Figure 3.1 shows the variant classification of samples in the cohort.

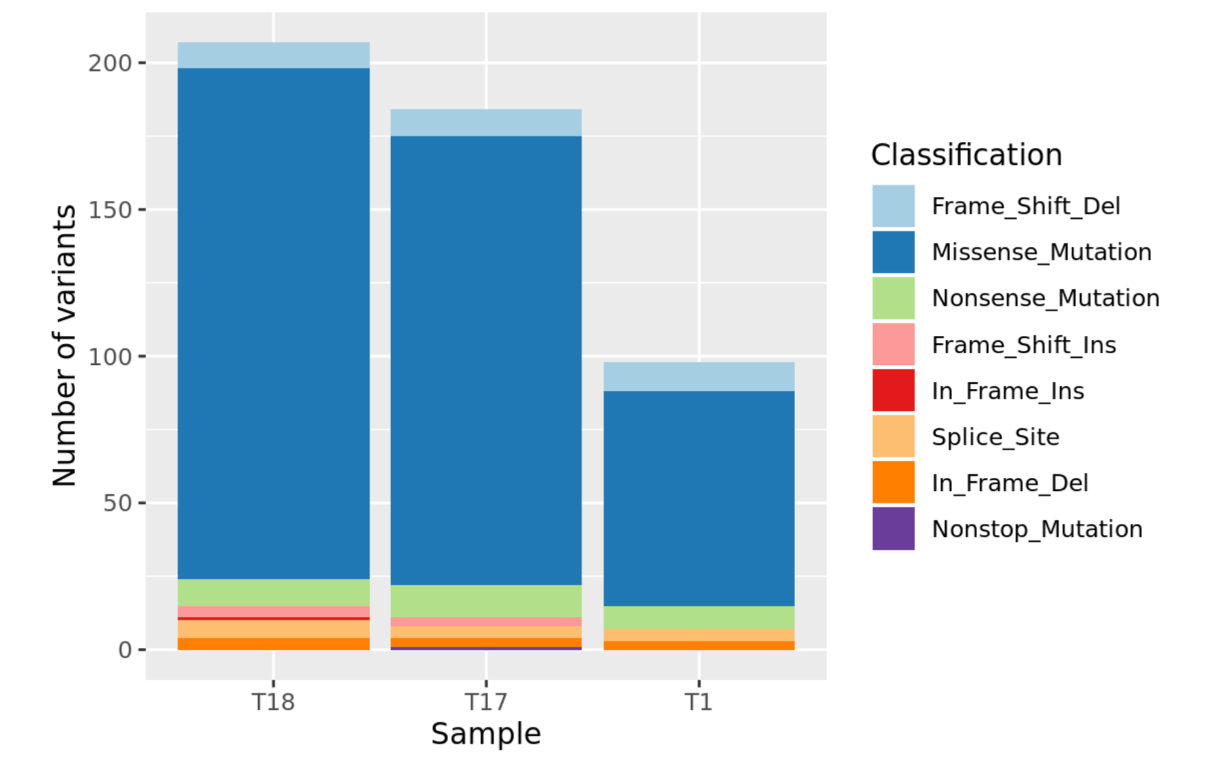

Figure 3.1 Distributaion of variant classification

DNA substitution mutations are of two types. Transitions are interchanges of purines or pyrimidines. Transversions are interchanges of purine for pyrimidine bases. Figure 3.2 shows the classification of the base substituions on the cohort level.

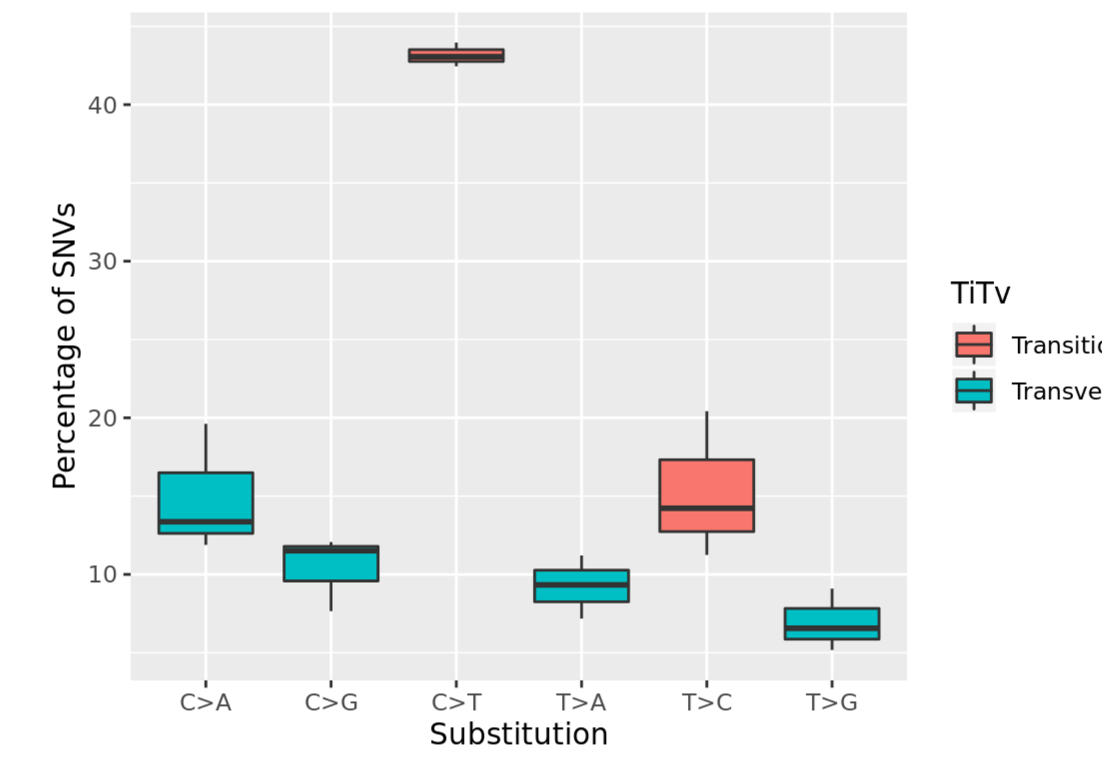

Figure 3.2 Distribution of base substitution

3.5.2 Analysis of top mutated genes

Figure 3.3 shows the mutation classification of the most mustated genes in the cohort across all samples.

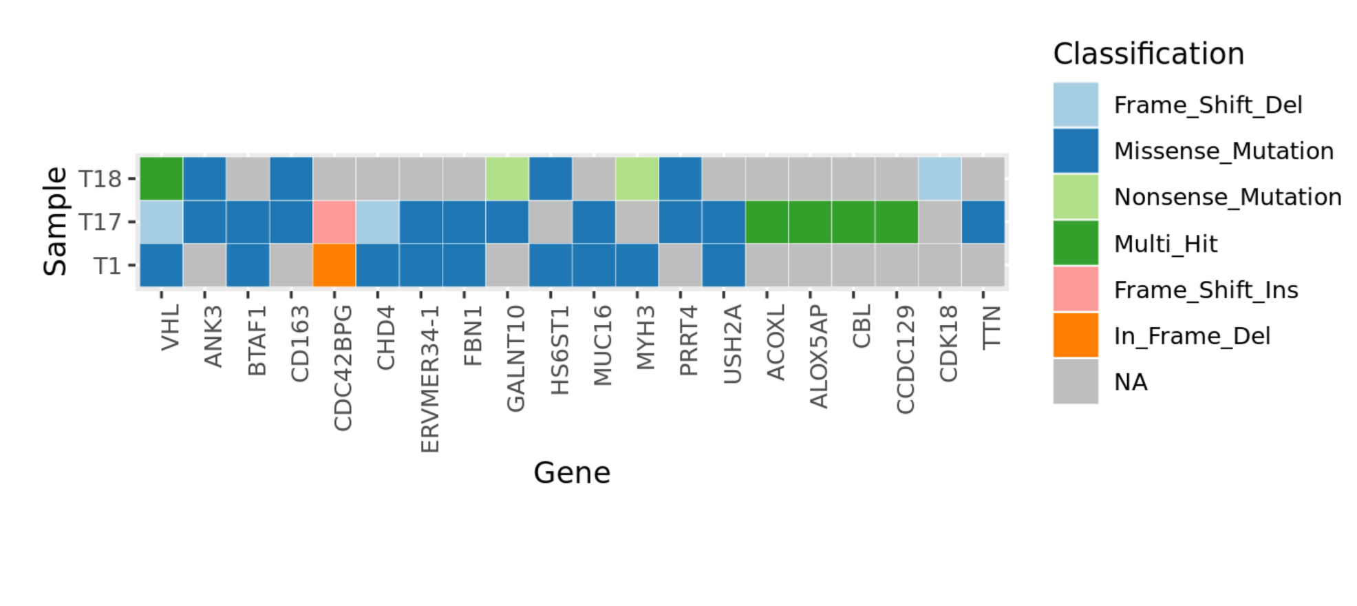

Figure 3.3 The most mutated genes in the cohort

Figure 3.4 shows the mutation profiles of these top mutated genes.

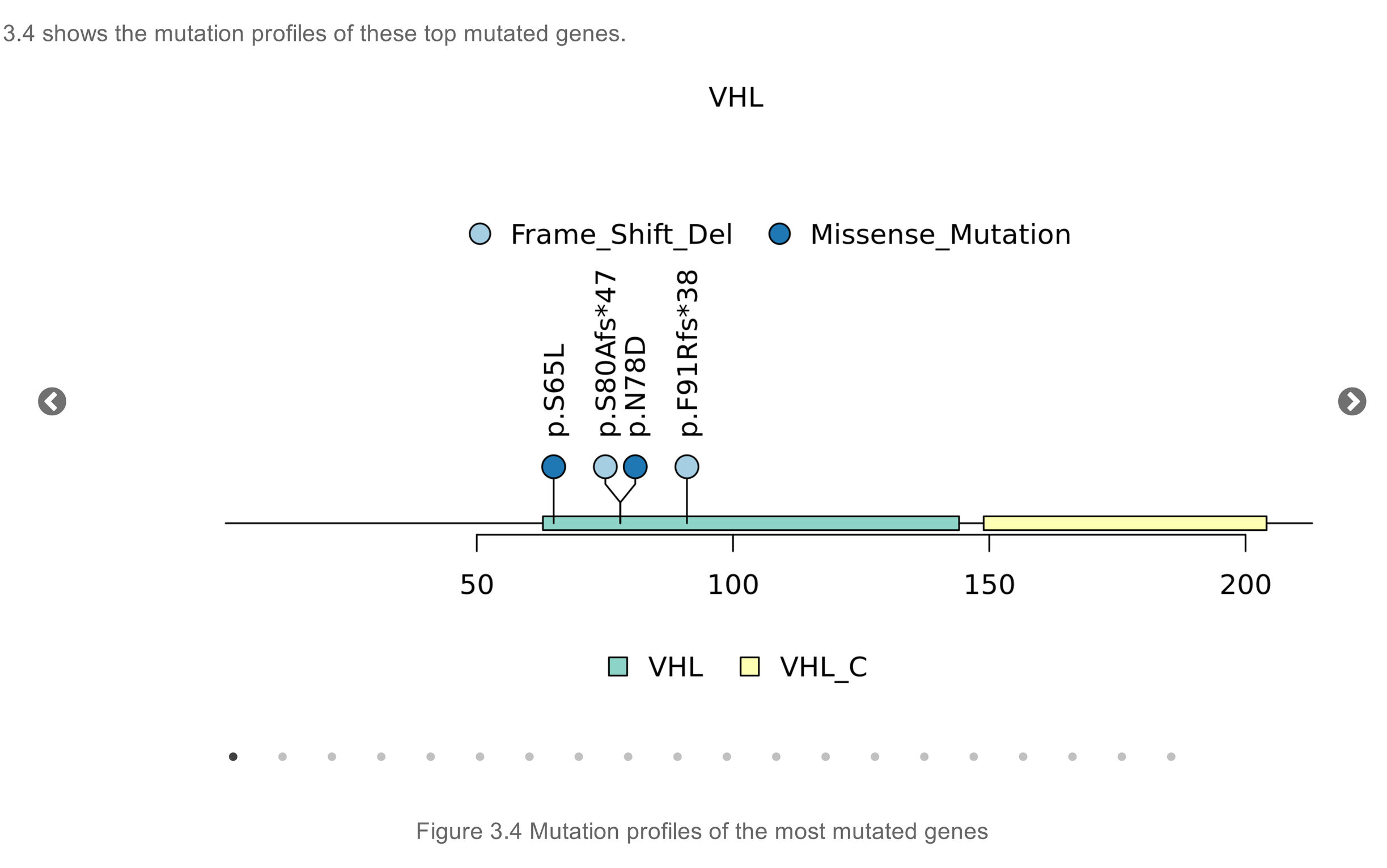

Figure 3.4 Mutation profiles of the most mutated genes

3.5.3 Tumor mutation burden

Tumor mutation load is calculated based on number of mutations in the genome region that targeted. The result is compared to the TCGA dataset in Figure 3.5.

Note: The TMB calculation need further validation.

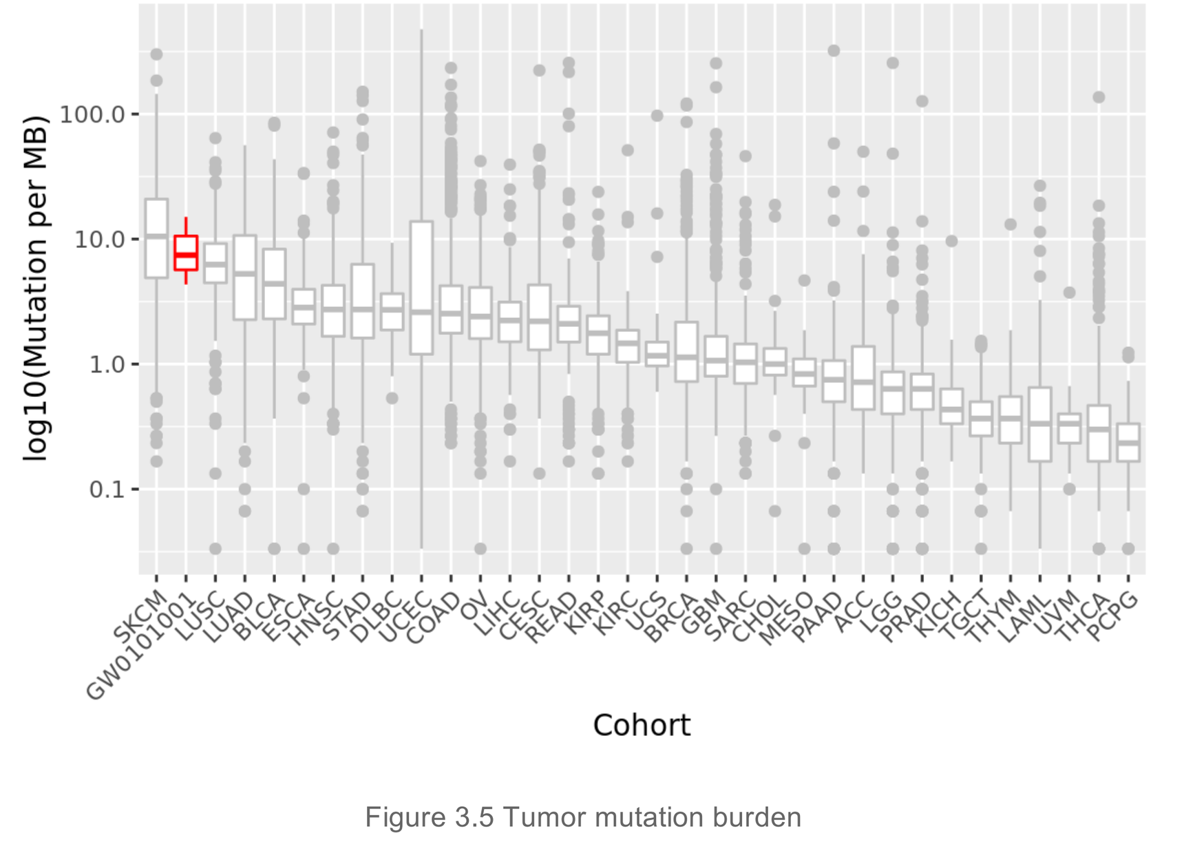

Figure 3.5 Tumor mutation burden

4. Deliverables

- Fastq
  - **Sample_R1/2.fastq.gz**: One pair of raw fastq files for each sample
- BAM
  - **Sample.aln.bam**: One mapped BAM file for each sample
- SNP Indel analysis
  - **Sample.hard-filtered.vcf.gz**: One VCF file with Dragen hard filter information for each tumor sample
  - **Sample.somatic.vcf.gz**: One VCF file filtered to remove known germline mutations for each tumor sample
  - **Sample.somatic.filtered.vep.vcf.gz**: One post-filter VCF file annotated using VEP for each tumor sample
  - **Sample.maf.gz**: Variants in MAF (Mutation Annotation Format) format for each sample
- CNV analysis (If applicable)
  - **Sample.target.counts**: Target regions reads count file for each sample
  - **Sample.target.counts.bw**: Target regions reads count file in bigwig format for each sample
  - **Sample.target.counts.gc-corrected**: Target regions reads count file after GC content correction for each sample
  - **Sample.cnv.vcf.**: CNV VCF file for each tumor sample
  - **Sample.cnv.gff3.**: CNV calling in gff3 format for each tumor sample
  - **Sample.seg.called.merged**: CNV segments for each tumor sample
- SV analysis (If applicable)
  - **Sample.sv.vcf.**: SV VCF file for each tumor sample
- Cohort analysis (If applicable)
  - **all_sample.maf.gz**: Combined SNP calls in MAF format
- Reports
  - **Sample_sample_report.html**: Sample sequencing and alingment report for each sample
  - **Sample_variant_report.html**: Sample variant calling report for each tumor sample
  - **project_report.html**: Project summary report