Data Analysis Report: Oncoprofiling (version 2.8)

Project / Study: EF-DEMO

Date: November 26, 2021

This report is not a diagnostic / clinical report and is intended for Research Use Only!

Table of Contents

1 Results

1.1 Variant discovery

Single nucleotide variants (SNVs), Insertions and deletions (InDel) are detected in each sample using LoFreq[1], and are filtered based on mutation allele frequency (>1%) and coverage (≥ 50x, or ≥ 10% of average coverage excluding duplicated fragments; coverage metrics can be found in chapter 2.3). Variants that pass these thresholds are summarised in the following table(s).

Table 1: Variant metrics for sample1, sample2, sample3.

	sample1	sample2	sample3
Total SNV	4078	4007	4850
Known SNV	4041	3953	4796
Unknown SNV	37	54	54
Total InDel	1734	1638	2090
Known InDel	1685	1593	2036
Unknown InDel	49	45	54

Known SNV / InDel: in reference variant databases (dbSNP, COSMIC[2] and / or ClinVar[3]).
Unknown SNV / InDel: currently not listed in reference variant database (as aforementioned).

1.2 Sample-wise known clinical significant variants

Variants detected are screened for known clinical significance in ClinVar (released 20. Nov 2021) [3] database. The ClinVar database aggregates information about genomic variation and its relationship to human health. It is hosted by the National Center for Biotechnology Information (NCBI). Detailed explanation of clinical significance in ClinVar database can be found at https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/.

Variants which have clinical significance state as ”Likely pathogenic”, ”Pathogenic” and ”Drug response” are filtered from the complete list of variants and are reported in following table(s). For more detailed information navigate to the Clinvar database and type in the dbIDs of your variant of interest. Variant effects for multiple transcripts for the same variant are listed as separate entries. In case of multiple transcripts, transcripts which have missense, splice junction, UTR, frameshift, disruptive frameshift insertion / deletion variant types are listed.

1.2.1 sample1 Results

Table 2: Variants (SNV and InDels) in sample - **sample1.** Entries are sorted by gene.

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr7:87600185	ABCB1	.	c.-1A>G	100.0 %	582x	829326	drug response
chr10:94942290	CYP2C9	p.R144C	c.430C>T	45.7 %	1328x	8409	drug response
chr11:67585218	GSTP1	. p.I105V	c.*137A>G c.313A>G	47.7 %	707x	37340	drug response
chr12:21178615	SLCO1B1	p.V174A	c.521T>C	47.9 %	1500x	37346	drug response

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr7:141972804	TAS2R38	p.I296V	c.886A>G	49.3 %	1680x	2906	drug response
chr17:7676154	TP53	p.P33R p.P72R	c.98C>G c.215C>G	99.3 %	437x	12351	drug response
chr3:14145949	XPC	p.Q939K .	c.2815C>A c.*2268C>A	43.6 %	456x	190215	drug response

Table 3: miRNA variations in sample - **sample1.** .

miRNA	Variant ID	Detected (Mut. Freq., Depth)
hsa-mir-149	rs71428439	No
hsa-mir-196a-2	rs11614913	No
hsa-mir-423	rs6505162	No
hsa-mir-603	rs11014002	No
hsa-mir-605	rs2043556	No
hsa-mir-618	rs2682818	No
hsa-mir-646	rs6513497	No
let-7b	rs10877887	No
let-7c	rs10877887	No
miR-137	rs1625579	No
miR-143	rs4705342	No
miR-155	rs1893650	No
miR-17-92 cluster	—	No
miR-21	—	No
miR-30a	rs2222722	No
miR141	rs34385807	No
miR200a	rs7521584	No
miR200b	rs7521584	No
miR200c	rs7521584	No
miR210	rs1062099, rs10902173	No
miR31	—	No
miR34a	rs72631823	No
miR34b	rs4938723	No
miR34c	rs4938723	No
miR429	rs7521584	No

1.2.2 sample2 Results

Table 4: Variants (SNV and InDels) in sample - **sample2.** Entries are sorted by gene.

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr7:87504154	ABCB1	.	c.*89A>T	61.5 %	135x	828786	drug response
chr7:87531302	ABCB1	p.S893A p.S829A	c.2677T>G c.2485T>G	99.9 %	1164x	166622	drug response
chr7:87600185	ABCB1	.	c.-1A>G	99.7 %	648x	829326	drug response
chr19:41006936	CYP2B6	p.Q172H	c.516G>T	59.2 %	1489x	29671	drug response
chr19:41009358	CYP2B6	p.K262R	c.785A>G	48.0 %	254x	120171	drug response
chr22:42127503	CYP2D6	. p.G322S p.G370S p.G373S	c.*192G>A c.964G>A c.1108G>A c.1117G>A	30.9 %	658x	828893	drug response
chr17:17219082	FLCN	p.S333fs	c.997_ 998delTC	1.5 %	274x	529991	pathogenic
chr18:51065549	SMAD4	p.R94H p.R361H p.R265H	c.281G>A c.1082G>A c.794G>A	22.2 %	1406x	24832	pathogenic
chr7:141972804	TAS2R38	p.I296V	c.886A>G	51.6 %	1299x	2906	drug response
chr7:141973545	TAS2R38	p.A49P	c.145G>C	51.7 %	1332x	2904	drug response
chr17:7673803	TP53	p.R114C p.R273C p.R234C p.R262C p.R141C	c.340C>T c.817C>T c.700C>T c.784C>T c.421C>T	26.5 %	896x	43594	pathogenic
chr17:7676154	TP53	p.P33R p.P72R	c.98C>G c.215C>G	99.4 %	314x	12351	drug response
chr3:14145949	XPC	p.Q939K .	c.2815C>A c.*2268C>A	36.3 %	410x	190215	drug response

Table 5: miRNA variations in sample - **sample2.** .

miRNA	Variant ID	Detected (Mut. Freq., Depth)
hsa-mir-149	rs71428439	No
hsa-mir-196a-2	rs11614913	No
hsa-mir-423	rs6505162	No
hsa-mir-603	rs11014002	No
hsa-mir-605	rs2043556	No
hsa-mir-618	rs2682818	No
hsa-mir-646	rs6513497	No
let-7b	rs10877887	No
let-7c	rs10877887	No
miR-137	rs1625579	No
miR-143	rs4705342	No
miR-155	rs1893650	No
miR-17-92 cluster	—	No
miR-21	—	No

miRNA	Variant ID	Detected (Mut. Freq., Depth)
miR-30a	rs2222722	No
miR141	rs34385807	No
miR200a	rs7521584	No
miR200b	rs7521584	No
miR200c	rs7521584	No
miR210	rs1062099, rs10902173	No
miR31	—	No
miR34a	rs72631823	No
miR34b	rs4938723	No
miR34c	rs4938723	No
miR429	rs7521584	No

1.2.3 sample3 Results

Table 6: Variants (SNV and InDels) in sample - **sample3.** Entries are sorted by gene.

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr7:87531302	ABCB1	p.S893A p.S829A	c.2677T>G c.2485T>G	99.9 %	1221x	166622	drug response
chr7:87600185	ABCB1	.	c.-1A>G	100.0 %	637x	829326	drug response
chr10:94781859	AL583836.1	.	c.*439G>A	44.7 %	799x	16897	drug response
chr15:51210647	CYP19A1	.	c.*161T>G	44.4 %	99x	316467	drug response
chr22:42127526	CYP2D6	p.R365H p.R314H p.R362H .	c.1094G>A c.941G>A c.1085G>A c.*169G>A	32.2 %	1105x	828892	drug response
chr22:42128945	CYP2D6	. . . . .	c.440-1G>A c.506-1G>A n.1230-1G>A c.353-1G>A c.173-1G>A	63.8 %	886x	16889	drug response
chr22:42129809	CYP2D6	p.H72R p.H94R p.H34R	c.215A>G c.281A>G c.101A>G	98.6 %	144x	829654	drug response
chr22:42129819	CYP2D6	p.L69M p.L31M p.L91M	c.205C>A c.91C>A c.271C>A	99.3 %	143x	829652	drug response
chr11:67585218	GSTP1	. p.I105V	c.*137A>G c.313A>G	43.4 %	862x	37340	drug response
chr12:21178615	SLCO1B1	p.V174A	c.521T>C	98.7 %	1350x	37346	drug response
chr7:141972804	TAS2R38	p.I296V	c.886A>G	50.1 %	1460x	2906	drug response
chr7:141973545	TAS2R38	p.A49P	c.145G>C	49.8 %	1427x	2904	drug response
chr17:7676154	TP53	p.P33R p.P72R	c.98C>G c.215C>G	98.1 %	413x	12351	drug response
chr3:14145949	XPC	p.Q939K .	c.2815C>A c.*2268C>A	51.7 %	582x	190215	drug response

Table 7: miRNA variations in sample - **sample3.** .

miRNA	Variant ID	Detected (Mut. Freq., Depth)
hsa-mir-149	rs71428439	No
hsa-mir-196a-2	rs11614913	No
hsa-mir-423	rs6505162	No
hsa-mir-603	rs11014002	No
hsa-mir-605	rs2043556	No
hsa-mir-618	rs2682818	No
hsa-mir-646	rs6513497	No
let-7b	rs10877887	No
let-7c	rs10877887	No
miR-137	rs1625579	No

miRNA	**Variant ID**	**Detected (Mut. Freq., Depth)**
miR-143	rs4705342	No
miR-155	rs1893650	No
miR-17-92 cluster	—	No
miR-21	—	No
miR-30a	rs2222722	No
miR141	rs34385807	No
miR200a	rs7521584	No
miR200b	rs7521584	No
miR200c	rs7521584	No
miR210	rs1062099, rs10902173	No
miR31	—	No
miR34a	rs72631823	No
miR34b	rs4938723	No
miR34c	rs4938723	No
miR429	rs7521584	No

1.3 Tumor mutational burden

Tumor mutational burden (TMB) is defined as the number of somatic, coding, base substitution, and indel mutations per megabase of genome examined. All base substitutions and indels in the coding region of targeted genes, including synonymous mutations, are initially counted before filtering as described below.

The filter settings were performed according to the published works[4, 5] with some exlusions. The following mutations are excluded from the TMB calculation:

- Non-coding mutations
- Mutations listed as known somatic mutations in COSMIC v71[2] and ClinVar[3]
- Known germline mutations in dbSNP[6]
- Mutations with depth $<$ 50X and allele frequency $<$ 0.03
- Germline mutations occurring with 2 or more counts in the ExAC (gnomAD) database[7]
- Mutations predicted to be germline by the somatic-germline-zygosity algorithm[8]
- Mutations in tumor suppressor genes (TSG, list in appendix D) were not counted, since the Oncopanel assay genes are biased toward genes with functional mutations in cancer.

To calculate the TMB per megabase, the total number of mutations counted is divided by the size of the coding region of the targeted region in megabase. Due to the lack of standardization of TMB computing, various TMB values are computed and reported[5].

Mutations included	Mutation Type	TMB1	TMB2	TMB3
missense, non-synonymous	SNP	YES	YES	YES
silent, synonymous	SNP	YES	NO	NO
stop-gain, stop-loss, frameshift, inframe	INDEL	YES	YES	NO

Table 8: TMB values for each sample

Sample	TMB1	TMB2	TMB3
sample1	0.68	0.68	0.34
sample2	1.02	1.02	0.68
sample3	3.05	2.37	2.03

1.4 Copy number analysis

Copy number variations (CNV) are detected using the software package CNVkit[9] which uses normalized read depths to infer copy number evenly across the exome/genome. CNVkit uses both the on-target reads and the nonspecifically captured off-target reads to calculate log2 copy ratios across the genome for each sample. Briefly, off-target bins are assigned from the genomic positions between targeted regions, with the average off-target bin size being much larger than the average on-target bin to match their read counts. Both the on and off target locations are then separately used to calculate the mean read depth within each interval. The on and off target read depths are then combined, normalized to a reference derived from control samples, corrected for several systematic biases (GC content, sequence complexity and targets) to result in a final table of log2 copy ratios. Then, the segmentation algorithm uses log2 ratio values to infer discrete copy number events.Copy number events with minimum 100 x coverage are reported.

Note: For the detection of CNVs a reference sample set is required. The CNV is calculated based on the average coverage distribution of the reference samples. The reference sample set should consist of at least 7 samples. Nonetheless, a bias in the reference due to over- or underrepresentation of sequencing data is possible. Thus, the sample set has to be chosen carefully and providing more than 8 samples leads to higher robustness of the data and higher confidence of the CNVs. As the detection of CNVs always strongly depends on the selected sample set / control group, validation of the results is strongly recommended.

Table 9: Case vs Control setup.

Case	Control(s)
sample1	control1, control2, control3, control4, control5, control6, control7
sample2	control1, control2, control3, control4, control5, control6, control7
sample3	control1, control2, control3, control4, control5, control6, control7

Table 10: Summary of CNV events detected in each sample.

Sample	Duplication Events	Deletion Events
sample1	1	10
sample2	26	0
sample3	1	7

1.4.1 sample1 Results

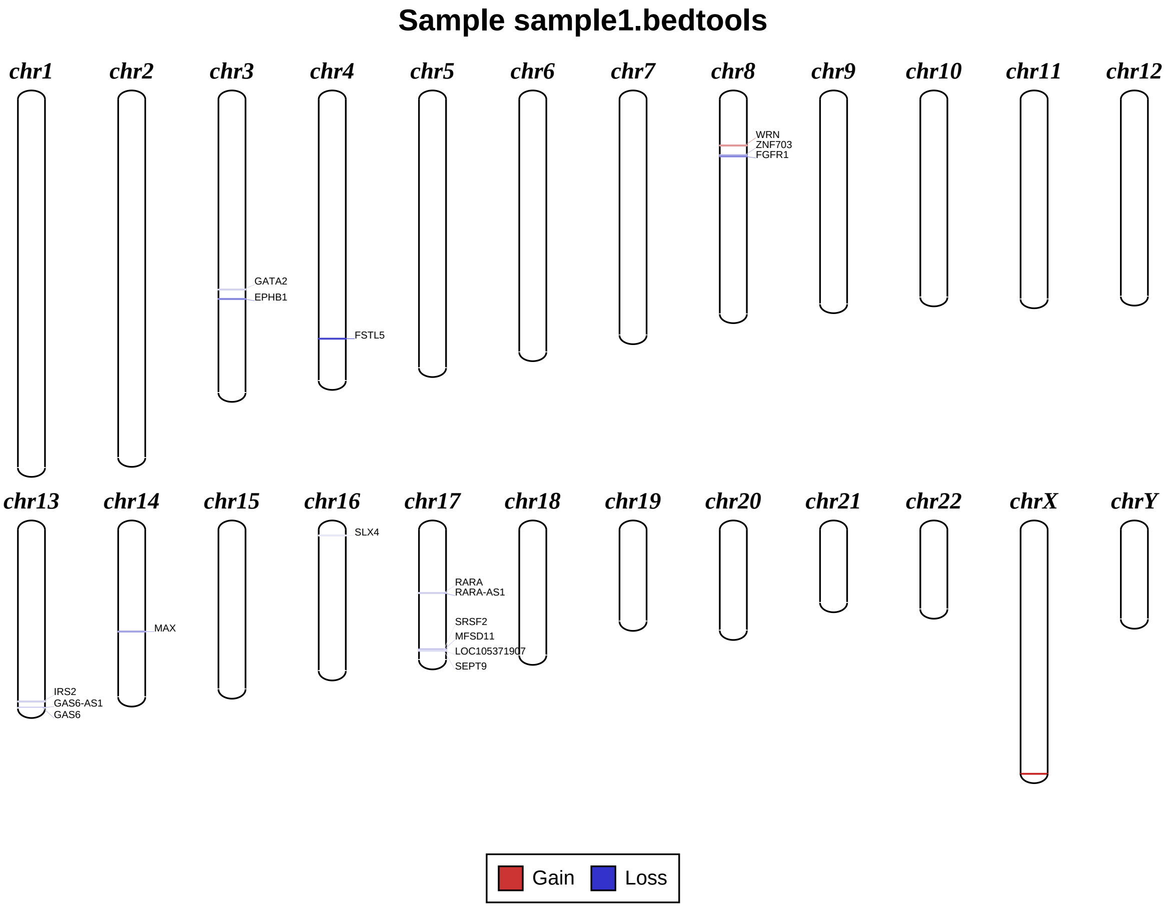

Figure 1: Ideogram representing chromosome wise copy number events observed in sample sample1. Copy gain events are drawn in red and copy loss events are drawn in blue.

Table 11: Duplication events detected in sample sample1. Gene column lists the name of genes (HGNC convention), CN column contains copy number observed and Depth column displays the coverage depth at the location (Loci column).

Gene	CN	Depth	Loci
WRN	3	1095.49	chr8:31143396-31143776

Table 12: Deletion events detected in sample sample1. Gene column lists the name of genes (HGNC convention), CN column contains copy number observed and Depth column displays the coverage depth at the location (Loci column).

Gene	CN	Depth	Loci
ZNF703	1	133.89	chr8:37695867-37698447
SLX4	1	1037.71	chr16:3584709-3594745
RHBDF2, SRSF2, SRSF2, MIR636, SRSF2, MIR636, MFSD11, SRSF2, MFSD11, LOC105371907, LOC105371907, SEPT9, SEPT9	1	429.41	chr17:76481559-77373714
RARA, RARA-AS1, RARA	1	1973.73	chr17:40342650-40350074
RARA, RARA-AS1	1	734.83	chr17:40341243-40342650

Gene	CN	Depth	Loci
MAX	1	173.26	chr14:65101368-65102492
IRS2, GAS6-AS1, GAS6	1	128.04	chr13:109781905-113821050
GATA2, EPHB1	1	219.34	chr3:128485587-134795850
FSTL5	1	765.9	chr4:161656300-161656630
FGFR1	1	435.4	chr8:38467635-38468738

1.4.2 sample2 Results

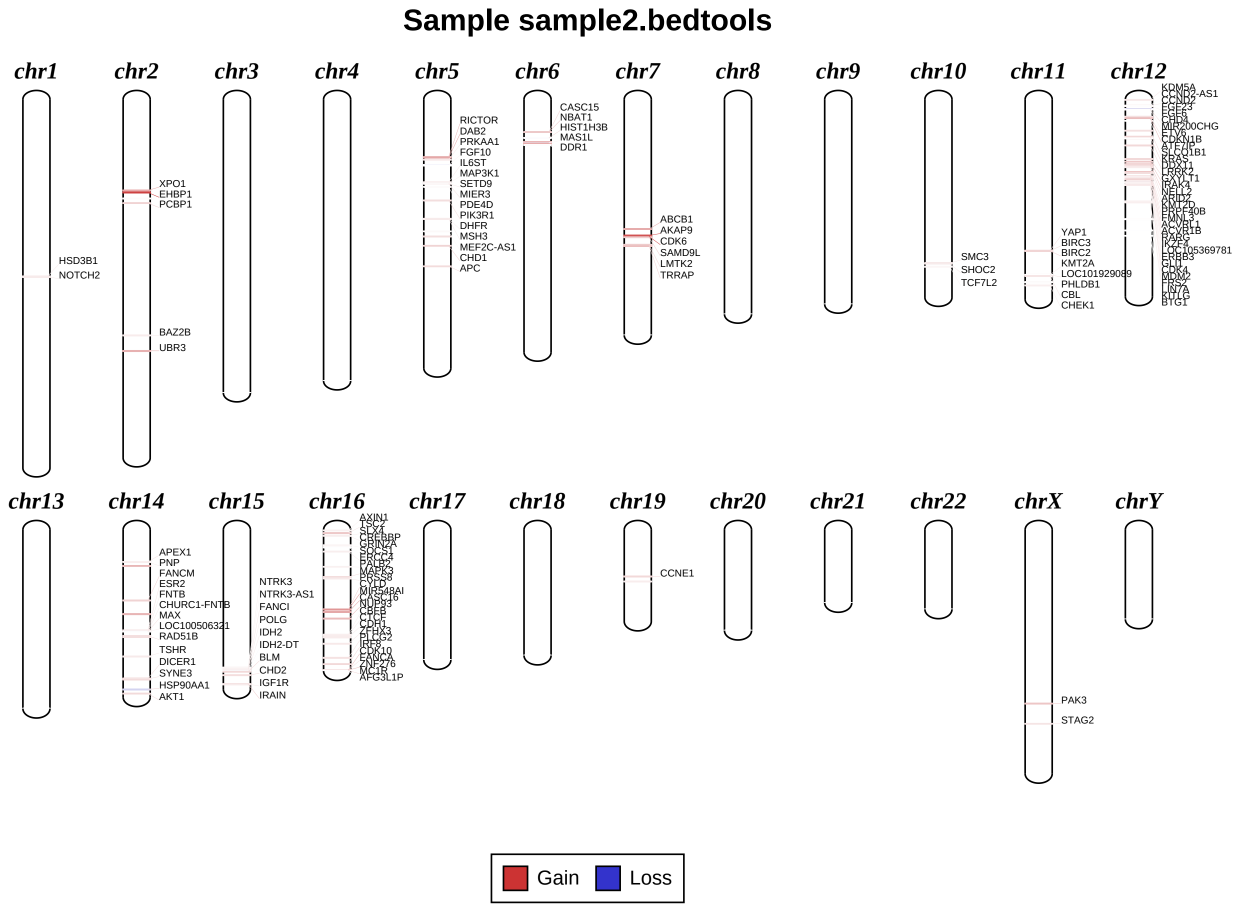

Figure 2: Ideogram representing chromosome wise copy number events observed in sample sample2. Copy gain events are drawn in red and copy loss events are drawn in blue.

Table 13: Duplication events detected in sample sample2. Gene column lists the name of genes (HGNC convention), CN column contains copy number observed and Depth column displays the coverage depth at the location (Loci column).

Gene	CN	Depth	Loci
ABCB1, AKAP9, CDK6	4	1787.69	chr7:87595599-92774983
YAP1, BIRC3, BIRC2	3	1626.69	chr11:102186074-102378098
XPO1, EHBP1, PCBP1	3	1418.7	chr2:61478672-70088123
UBR3	3	1389.84	chr2:169872085-170080724
SMC3, SHOC2, TCF7L2	3	1160.23	chr10:110567728-113166005
RICTOR, DAB2	3	1410.1	chr5:38942150-39395076
PRKAA1, FGF10, FGF10, FGF10-AS1, FGF10-AS1, IL6ST, MAP3K1, SETD9, MIER3, MIER3, PDE4D, PIK3R1, DHFR, MSH3, MSH3, MEF2C-AS1, CHD1	3	1284.96	chr5:40791532-98858504
PAK3, STAG2	3	1148.35	chrX:111196541-124083704

Gene	CN	Depth	Loci
PAK3	3	1304.83	chrX:111123080-111196541
NUP93, CBFB, CTCF, CDH1, ZFHX3, PLCG2, IRF8, CDK10, ZNF276, FANCA, FANCA, MC1R, AFG3L1P	3	1002.15	chr16:56844639-90000717
NTRK3, NTRK3, NTRK3-AS1, FANCI, FANCI, POLG, IDH2, IDH2, IDH2-DT, BLM, CHD2, IRAIN, IGF1R, IGF1R	3	1170.88	chr15:88043022-98957561
KMT2A, KMT2A, LOC101929089, PHLDB1, CBL, CHEK1	3	1436.35	chr11:118438882-125655462
KDM5A, CCND2-AS1, CCND2, CCND2, FGF23, FGF6, CHD4, CHD4, SCARNA11, MIR200CHG, MIR200C, MIR200CHG, MIR200CHG, MIR141, ETV6	3	1202.51	chr12:285338-11752514
HSD3B1, NOTCH2	3	1154.29	chr1:119507438-120069444
FANCM, ESR2, CHURC1-FNTB, FNTB, MAX, MAX, MAX, LOC100506321, RAD51B, TSHR	3	1296.02	chr14:45136020-81144466
ETV6, CDKN1B, ATF7IP, SLCO1B1, KRAS, DDX11, LRRK2, GXYLT1, IRAK4, NELL2, ARID2	3	1084.97	chr12:11752514-45729979
DICER1, SYNE3, HSP90AA1, HSP90AA1, WDR20, AKT1	3	1137.97	chr14:95090360-104792810
CYLD, MIR548AI, CASC16, NUP93	3	1293.6	chr16:50749580-56844639
CHD1, APC	3	1142.45	chr5:98858799-112775902
CDK6, SAMD9L, LMTK2, TRRAP, TRRAP, SCARNA28	3	2402.07	chr7:92832956-98895981
CCNE1, RHPN2	3	1320.85	chr19:29805755-33041393
CASC15, CASC15, NBAT1, HIST1H3B, MAS1L, DDR1	3	1091.18	chr6:22125544-30885415
BAZ2B, UBR3	3	1359.64	chr2:159320121-169828169
AXIN1, TSC2, TSC2, PKD1, SLX4, CREBBP, GRIN2A, SOCS1, ERCC4, PALB2, MAPK3, PRSS8	3	868.05	chr16:287993-31135645
ARID2, KMT2D, PRPF40B, PRPF40B, FMNL3, ACVRL1, ACVR1B, RARG, MIR196A2, IKZF4, LOC105369781, IKZF4, ERBB3, GLI1, CDK4, CDK4, MIR6759, MDM2, FRS2, LIN7A, LIN7A, MIR618, KITLG, BTG1	3	960.46	chr12:45729979-92145654
APEX1, PNP	3	1049.04	chr14:20455489-23016822

No deletion events found!

1.4.3 sample3 Results

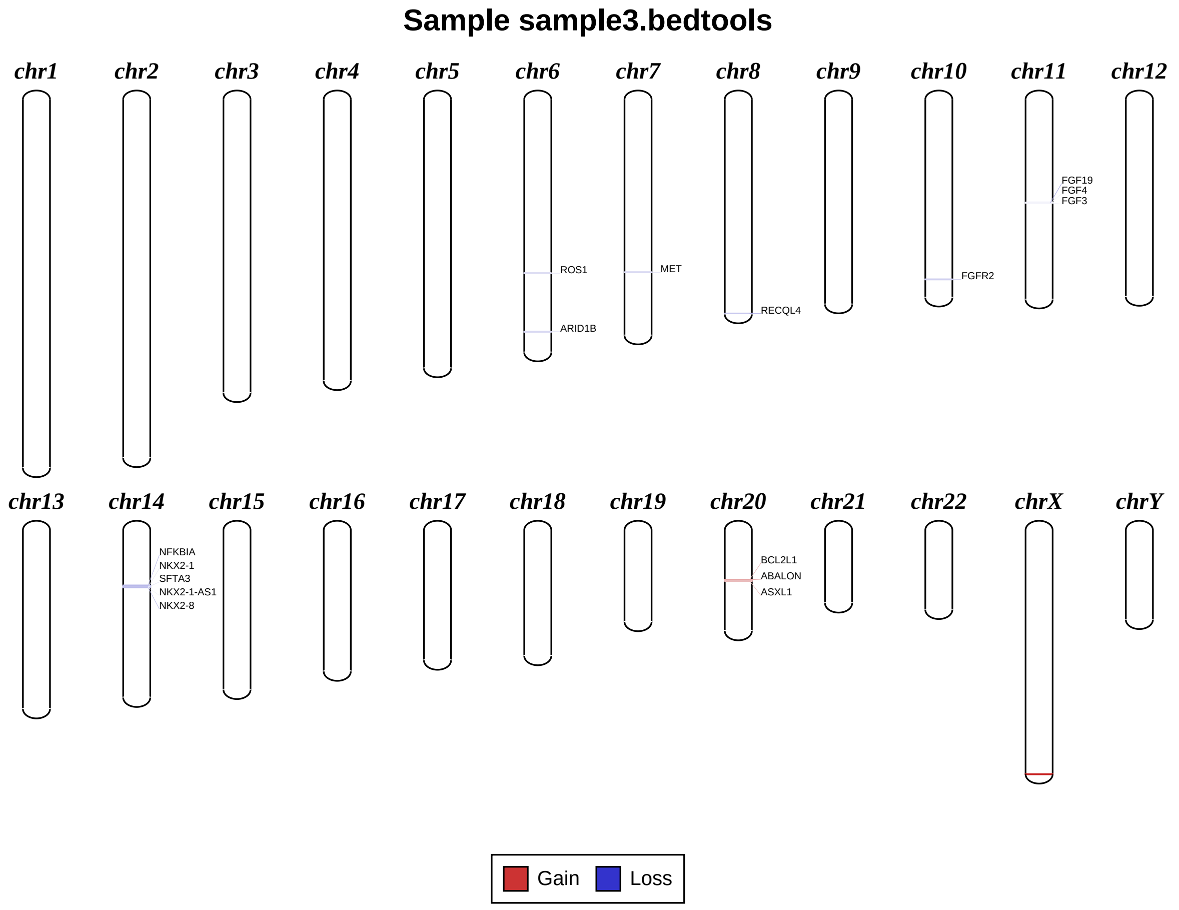

Figure 3: Ideogram representing chromosome wise copy number events observed in sample sample3. Copy gain events are drawn in red and copy loss events are drawn in blue.

Table 14: Duplication events detected in sample sample3. Gene column lists the name of genes (HGNC convention), CN column contains copy number observed and Depth column displays the coverage depth at the location (Loci column).

Gene	CN	Depth	Loci
BCL2L1, BCL2L1, ABALON, ASXL1	3	1897.76	chr20:31665797-32437352

Table 15: Deletion events detected in sample sample3. Gene column lists the name of genes (HGNC convention), CN column contains copy number observed and Depth column displays the coverage depth at the location (Loci column).

Gene	CN	Depth	Loci
ROS1	1	3174.51	chr6:117325088-117329304
RECQL4	1	121.64	chr8:144516899-144517828
NFKBIA, SFTA3, NKX2-1, SFTA3, NKX2-1, NKX2-1-AS1, NKX2-1, NKX2-1-AS1, NKX2-8	1	326	chr14:35404371-36582512
MET	1	3102.34	chr7:116672675-116700308

Gene	CN	Depth	Loci
FGFR2	1	2175.43	chr10:121479435-121483672
FGF19, FGF4, FGF3	1	311.03	chr11:69703602-69819043
ARID1B	1	123.12	chr6:156777787-156829213

1.5 Fusion gene discovery

Fusion events are detected using the software DELLY2[10]. From the genome alignments, DELLY2 discovers fusion events (translocations and inversions) by integrating insert distances determined by the paired-end reads and split-read alignments to accurately detect genomic rearrangements at single nucleotide resolution. Fusion events are tagged as "Known fusions" if they match the entry in ChimerDB[11] (collection of known fusion events). Known fusion events with minimum 7 x coverage are reported. Complete lists of fusion events can be found in supplementary deliverables.

Table 16: Summary of fusion events detected in each sample.

Sample	Known events	Unknown events
sample1	0	2
sample2	0	1
sample3	0	1

1.5.1 sample1 Results

No known events found!

1.5.2 sample2 Results

No known events found!

1.5.3 sample3 Results

No known events found!

2 Quality Metrics

2.1 Sequence Quality Metrics

The base quality of each sequence read is inspected. Low quality calls are removed before proceeding with further processing. Using a sliding window approach, bases with low quality are removed from the 3' and 5' ends. Bases are removed if the average phred quality is below 15. Finally only mate pairs (forward and reverse read) were used for the next analysis step. The total amount of raw sequence data and the results of the quality filtering is collected and reported in the following table.

Table 17: Sequence quality metrics per sample

Sample	Total Reads	LQ Reads	Single Reads	HQ Reads
sample1	133,309,002	2,238,975 (1.7%)	1,906,691 (1.4%)	129,163,336 (96.9%)
sample2	134,164,916	2,348,219 (1.8%)	2,006,583 (1.5%)	129,810,114 (96.8%)
sample3	121,002,458	2,166,177 (1.8%)	1,816,709 (1.5%)	117,019,572 (96.7%)

Total Reads: Total number of sequence reads analysed for each sample.

LQ Reads: Number of low quality reads.

Single Reads: Number of high quality reads without mates (2nd read).

HQ Reads: Number of high quality reads used for further analysis.

2.2 Mapping and Alignment Processing

Mapping to the reference sequence / database is done using BWA[12] with default parameters. Please note that the mapping efficiency depends on the accuracy of the reference and the quality of sequence reads. Reads are then classified according to the following categories:

- Mapped: Reads mapped to reference.
- Unique: Reads mapped to exactly one site on the reference.
- Non-unique: Reads mapped to more than one site on the reference.
- Singletons: Mapped reads with unmapped mates.
- Cross-Contig: Mapped reads with mates mapped to a different contig / chromosome.
- On-target: Uniquely mapped reads that mapped to a target region with +/- 100 bp tolerance.

For targeted sequencing (e. g. exome sequencing, amplicon panels), the targeted regions are subregions of the reference sequence. For whole genome sequencing, the target region is the full reference sequence. Unmapped reads, non-unique reads, singletons, cross-contig reads, and off-target reads are discarded. Only uniquely mapped on-target reads are processed further.

Remaining reads are deduplicated using sambamba[13] in order to remove the artificial coverage caused by the PCR amplification step during the library preparation and / or sequencing. If a read maps to the same genomic location and has the same orientation as another already mapped read, the reads are considered as duplicates. For paired-end data, all mates of compared pairs have to fulfill the criteria in order to be designated as PCR duplicates. One copy of the duplicated reads is kept for futher analyses, the others are discarded.

As a next step, a base quality recalibration is performed to improve the base quality scores of reads. A base quality score represents the probability of a particular base mismatching the reference genome. After recalibration, quality scores are more accurate in that they are closer to the true probability of a mismatch. This process is achieved by analysing the covariation among several different features of a base. The reported quality score, sequencing cycle, and sequencing context are considered for this step. Base quality recalibration is done using GATK[14, 15] modules.

Detailed alignment metrics for each sample can be found in file `*.alignment_metrics.tsv`. (see Deliverables, chapter 3).

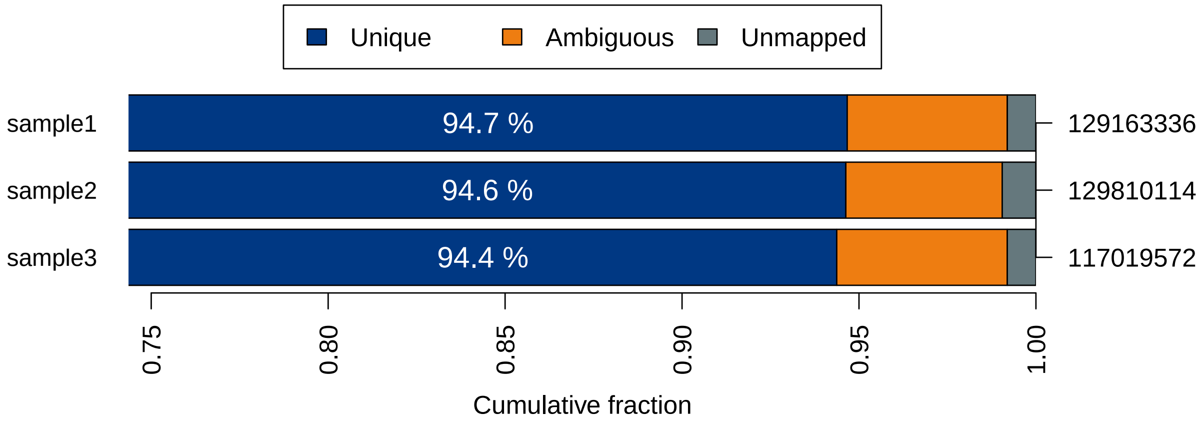

Figure 4: Summary of alignment results. For each sample, the fraction of uniquely mapped, non-uniquely mapped (ambiguous) and unmapped reads relative to the total number of reads per sample (right y-axis) is shown.

Table 18: Mapped read metrics observed per sample. Percentage of reads in category **Unique** is calculated based on the number of reads mapping to entire reference. Percentage of reads in category **On-target** is calculated based on the number of reads mapped uniquely. Percentage of reads in category **Deduplicated** is calculated based on the number of on-target reads.

No.	Sample	Mapped HQ Reads	Unique	On-Target	Deduplicated
1	sample1	128,118,084 (99.19%)	122,274,813 (95.44%)	95,193,024 (77.85%)	42,751,738 (44.91%)
2	sample2	128,572,919 (99.05%)	122,836,537 (95.54%)	98,587,762 (80.26%)	35,932,292 (36.45%)
3	sample3	116,071,924 (99.19%)	110,432,425 (95.14%)	82,206,426 (74.44%)	38,394,190 (46.70%)

2.3 Coverage Report

The coverage plot showing the base coverage distribution from the HQ aligned data. Depth of coverage is plotted on X-axis and the percentage of the respective reference covered is plotted on Y-axis. The coverage plot is restricted to the target region without extension. The shape of the curve defines the uniformity of the reference coverage in the samples analysed. Samples with high uniformity usually have >90% covered at 0.2x average coverage (e.g. 100x for 500x average coverage)

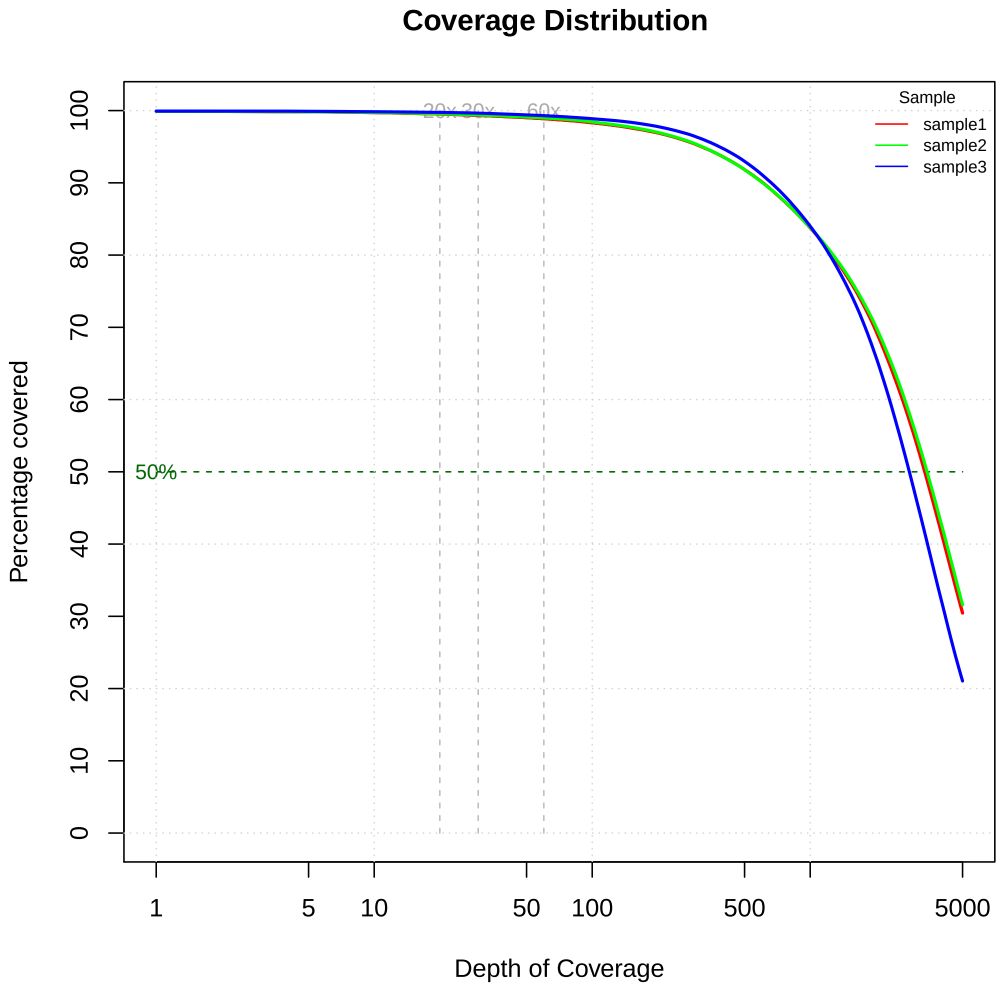

Figure 5: Coverage plot (including duplicated fragments).

Table 19: Depth of coverage summary (including duplicated fragments).

	target coverage		% of target covered with at least				
sample	total bases	average (x)	2x	50x	100x	300x	500x
sample1	12.38 GB	4199.00	99.9	99.0	98.3	95.3	91.8
sample2	12.57 GB	4261.26	99.9	99.1	98.4	95.4	91.8
sample3	10.34 GB	3506.19	99.9	99.4	98.9	96.4	93.0

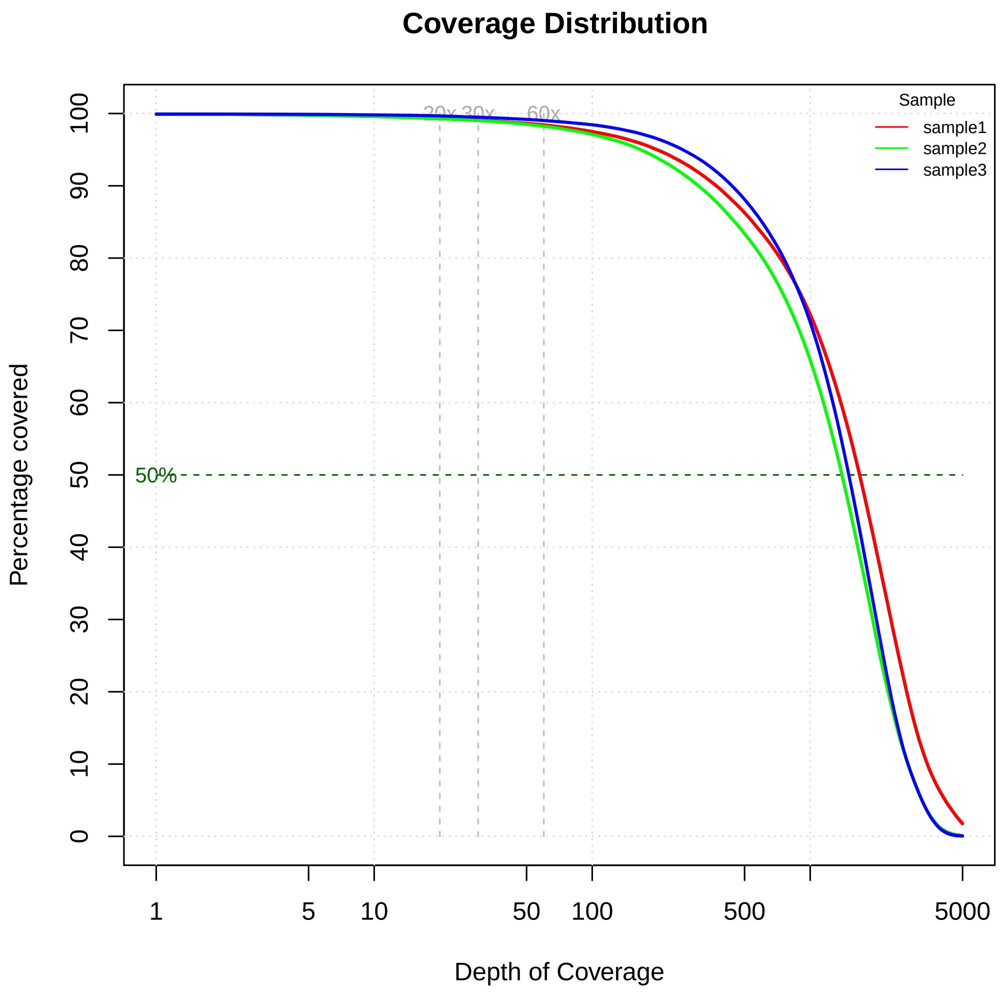

Figure 6: Coverage plot (excluding duplicated fragments).

Table 20: Depth of coverage summary (excluding duplicated fragments).

	target coverage		**% of target covered with at least**				
sample	total bases	average (x)	2x	50x	100x	300x	500x
sample1	5.41 GB	1834.82	99.9	98.6	97.5	92.1	86.3
sample2	4.41 GB	1495.37	99.9	98.5	97.1	90.2	83.4
sample3	4.67 GB	1583.93	99.9	99.2	98.4	94.0	88.1

2.4 Library Report

Fragment insert size histogram of the paired-end library observed from all the samples analysed. The insert size is determined by mapping individual read pairs on the reference sequence. The distance between 5'prime ends of both sequenced reads in a pair that are mapped to the reference is the observed length of the sequenced fragment. By performing this operation for all mapped reads the distribution can be generated. X-axis shows the insert size in bp and Y-axis shows the number of fragments with the observed fragment insert sizes.

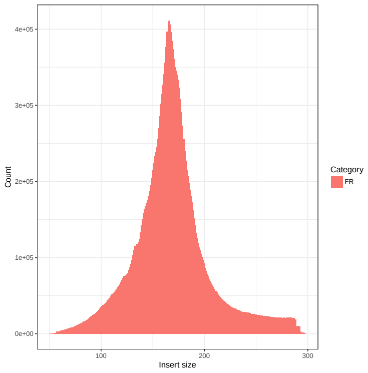

Figure 7: sample1 .

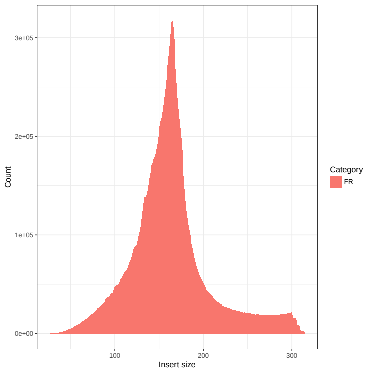

Figure 8: sample2 .

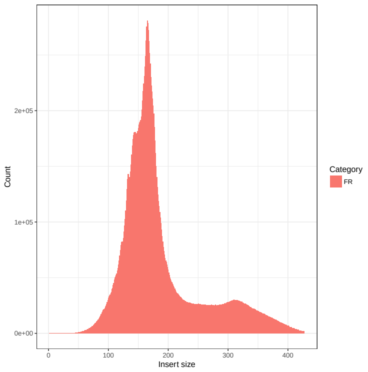

Figure 9: sample3 .

Table 21: Sample wise insert size metrics for HQ aligned reads. The mean insert size (Mean) and its standard deviation (Stddev) is given in base pairs.

Sample	Pair orientation	Mean	Stddev	# Read pairs
sample1	FR	168	36	21,376,620
sample2	FR	163	44	17,964,962
sample3	FR	188	68	19,193,716

3 Deliverables

Table 22: List of delivered files, format and recommended programs to access the data.

File	Format	Program To Open File
PROJECT.Variant_Analysis_Report.pdf	PDF	PDF reader
PROJECT.alignment_metrics.tsv	TSV	Spreadsheet Editor
PROJECT.cleaning_metrics.tsv	TSV	Spreadsheet Editor
PROJECT_supplementary_tables.tar.gz	GZ	Unzip tool
SAMPLE.CNV_deletion.tsv	TSV	Spreadsheet Editor
SAMPLE.CNV_duplication.tsv	TSV	Spreadsheet Editor
SAMPLE.fusion_events.tsv	TSV	Spreadsheet Editor
SAMPLE.hg38.HQ.alignment.bam	BAM	IGV, Tablet
SAMPLE.hg38.HQ.alignment.bam.bai	BAI	None
SAMPLE.hg38.alignment.bam	BAM	IGV, Tablet
SAMPLE.hg38.alignment.bam.bai	BAI	None
SAMPLE.indels.tsv	TSV	Spreadsheet Editor
SAMPLE.indels.vcf	VCF	Text Editor
SAMPLE.snps.tsv	TSV	Spreadsheet Editor
SAMPLE.snps.vcf	VCF	Text Editor

SAMPLE.hg38.alignment.bam was used for Fusion Gene discovery (see chapter 1.5)

SAMPLE.hg38.HQ.alignment.bam was used for Variant discovery (see chapter 1.1) and for Copy number analysis (see chapter 1.4)

PROJECT_supplementary_tables.tar.gz contains the variant calls (SNVs and InDels) that were observed in the sample(s) but filtered out due to QC checks.

4 Formats

Table 23: References and descriptions of file format.

Format	Description
BAM[16]	Compressed binary version of the Sequence Alignment / Mapping (SAM) format, a compact and index-able representation of nucleotide sequence alignments.
TSV	Tab separated table style text file. This can be imported into spreadsheet processing software like MS OFFICE Excel.
VCF[17]	Variant Call Format (VCF) is a format to describe and report the variants.

5 FAQ

Q: How can I open a TSV file in Excel?
A: Start Excel and click File -> Open and select the TSV file you want to open. Next an assistant dialog should show up. Make sure that you select tab as separator. Set the format of all rows without numbers to text. The TSV files use the dot as decimal separator and comma as thousands separator. Make sure that you set both correctly.

6 Bibliography

[1] Andreas Wilm, Pauline Poh Kim P. Aw, Denis Bertrand, Grace Hui Ting H. Yeo, Swee Hoe H. Ong, Chang Hua H. Wong, Chiea Chuen C. Khor, Rosemary Petric, Martin Lloyd L. Hibberd, and Niranjan Nagarajan. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*, 40(22):11189–11201, December 2012.

[2] Simon A. Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Y. Kok, Mingming Jia, Tisham De, Jon W. Teague, Michael R. Stratton, Ultan McDermott, and Peter J. Campbell. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):gku1075–D811, October 2014.

[3] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, Wonhee Jang, Kenneth Katz, Michael Ovetsky, George Riley, Amanjeev Sethi, Ray Tully, Ricardo Villamarin-Salomon, Wendy Rubinstein, and Donna R. Maglott. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, January 2016.

[4] Michael Allgäuer, Jan Budczies, Petros Christopoulos, Volker Endris, Amelie Lier, Eugen Rempel, Anna-Lena Volckmar, Martina Kirchner, Moritz von Winterfeld, Jonas Leichsenring, Olaf Neumann, Stefan Fröhling, Roland Penzel, Michael Thomas, Peter Schirmacher, and Albrecht Stenzinger. Implementing tumor mutational burden (tmb) analysis in routine diagnostics-a primer for molecular pathologists and clinicians. *Translational lung cancer research*, 7(6):703–715, Dec 2018. 30505715[pmid].

[5] Bárbara Meléndez, Claude Van Campenhout, Sandrine Rorive, Myriam Remmelink, Isabelle Salmon, and Nicky D'Haene. Methods of measurement for tumor mutational burden in tumor tissue. *Translational lung cancer research*, 7(6):661–667, Dec 2018. 30505710[pmid].

[6] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311, 01 2001.

[7] M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, (...), and Exome Aggregation Consortium. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, Aug 2016.

[8] James X. Sun, Yuting He, Eric Sanford, Meagan Montesion, Garrett M. Frampton, Stéphane Vignot, Jean-Charles Soria, Jeffrey S. Ross, Vincent A. Miller, Phil J. Stephens, Doron Lipson, and Roman Yelensky. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLOS Computational Biology*, 14(2):1–13, 02 2018.

[9] Eric Talevich, A. Hunter Shain, Thomas Botton, and Boris C. Bastian. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*, 12(4):e1004873+, April 2016.

[10] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M. Stütz, Vladimir Benes, and Jan O. Korbel. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, September 2012.

[11] Pora Kim, Suhyeon Yoon, Namshin Kim, Sanghyun Lee, Minjeong Ko, Haeseung Lee, Hyunjung Kang, Jaesang Kim, and Sanghyuk Lee. ChimerDB 2.0 - a knowledgebase for fusion genes updated. *Nucleic acids research*, 38(suppl 1):D81–D85, 2010.

[12] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760, July 2009.

[13] Artem Tarasov, Albert J. Vilella, Edwin Cuppen, Isaac J. Nijman, and Pjotr Prins. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, February 2015.

[14] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, 2010.

[15] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43:491–498, 2011.

[16] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[17] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[18] Derek Barnett, Erik Garrison, Aaron Quinlan, Michael Strömberg, and Gabor Marth. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, 27(12):btr174–1692, April 2011.

[19] Mary Kate Wing. ”bamUtil is a repository that contains several programs that perform operations on SAM/BAM files.”. http://genome.sph.umich.edu/wiki/BamUtil, 2015.

[20] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010.

[21] Picard. http://picard.sourceforge.net.

[22] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[23] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[24] Pablo Cingolani. ”snpEff: Variant effect prediction”. http://snpeff.sourceforge.net, 2012.

[25] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, 30(15):2114–2120, August 2014.

A Analysis Workflow

The schematic diagram of the data analysis steps that have been performed is shown in figure 1.

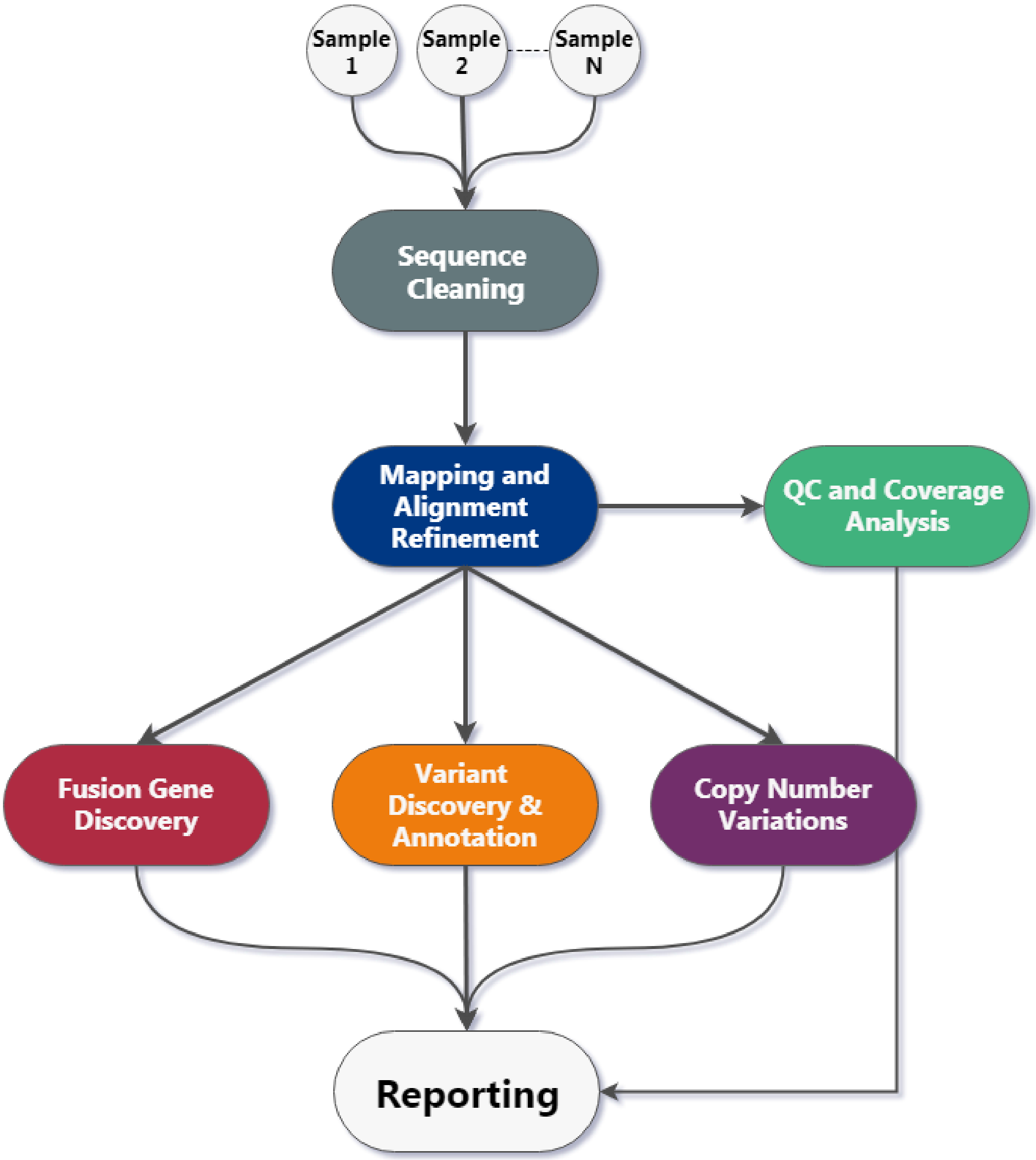

Figure 10: ONCOPANEL ALL-IN-ONE v2.8 Workflow

B Sequence Data Used

Table 24: Analysed samples (SE = single end, PE = paired end).

Sample	Read Type	File Name
sample1	PE	EF-12345_sample1_lib123453_1234_1_1.fastq.gz.gz EF-12345_sample1_lib123453_1234_1_2.fastq.gz.gz
sample2	PE	EF-12345_sample2_lib123454_1234_1_1.fastq.gz.gz EF-12345_sample2_lib123454_1234_1_2.fastq.gz.gz
sample3	PE	EF-12345_sample3_lib123455_1234_1_1.fastq.gz.gz EF-12345_sample3_lib123455_1234_1_2.fastq.gz.gz

C Reference Database

Table 25: Information about the Homo sapiens Reference Database.

Tag	Description
Name	Homo sapiens
Version	hg38.chronly
Source	UCSC
Size (bp)	3.088 GB
Sequences	23

Table 26: Information about additional reference data used.

Type	Version	Source
Annotation	22	GENCODE
dbSNP[6]	151	NCBI
ClinVar[3]	20.11.21	NCBI
COSMIC[2]	71	Sanger Institute
gnomAD[7]	2.1.1	Broad Institute
ChimerDB[11]	2.0	ERCSB

Table 27: Information about the target region used.

Tag	Description
Name	Eurofins Genomics Europe All in One
Size (bp)	2,951,184
Source	Eurofins Genomics Europe Sequencing GmbH

D Tumor Supressor Genes

APC, ARHGEF12, ATM, BCL11B, BLM, BMPR1A, BRCA1, BRCA2, CARS, CBFA2T3, CDH1, CDH11, CDK6, CDKN2C, CEBPA, CHEK2, CREB1, CREBBP, CYLD, DDX5, EXT1, EXT2, FBXW7, FH, FLT3, FOXP1, GPC3, IDH1, IL2, JAK2, MAP2K4, MDM4, MEN1, MLH1, MSH2, NF1, NF2, NOTCH1, NPM1, NR4A3, NUP98, PALB2, PML, PTEN, RB1, RUNX1, SDHB, SDHD, SMARCA4, SMARCB1, SOCS1, STK11, SUFU, SUZ12, SYK, TCF3, TNFAIP3, TP53, TSC1, TSC2, VHL, WRN, WT1.

E Relevant Programs

Table 28: Name, version and description of relevant programs.

Program	Version	Description
bamtools[18]	2.3.0	BamTools provides a small, but powerful suite of command-line utility programs for manipulating and querying BAM files for data.
BamUtil[19]	1.0.10	BamUtil is a repository that contains several programs that perform operations on SAM/BAM files
bedtools[20]	2.26.0	Bedtools allows one to intersect, merge, count, complement, and shuffle genomic intervals from multiple files in widely-usedgenomic file formats such as BAM, BED, GFF/GTF, VCF
BWA[12]	0.7.15	BWA is a software package for mapping low-divergent sequences against a large reference genome
CNVkit[9]	0.9.1.dev0	CNVkit is a Python library and command-line software toolkit to infer and visualize copy number from targeted DNA sequencing data
Delly2[10]	0.7.6	DELLY2: Structural variant discovery by integrated paired-end and split-read analysis
GATK[14, 15]	3.7	GATK is a java-based command-line toolkit that process SAM / BAM / VCF files.
LoFreq[1]	2.1.3.1	Lofreq is a fast and sensitive variant caller for inferring SNVs and indels from next-generation sequencing data.
Picard[21]	1.131	Picard is a java-based command-line utilities for processing SAM / BAM files.
R[22]	3.2.4	R is a programming language and environment for statistical computing.
sambamba[13]	0.6.6	Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files.
SAMTools[23]	0.1.18	SAMtools provide various utilities for manipulating alignments in the SAM format.
snpEff[24]	4.3	SnpEff is a genetic variant annotation and effect prediction toolbox.
SnpSift[24]	4.3	SnpSift helps filtering and manipulating genomic annotated files .
Trimmomatic[25]	0.33	Trimmomatic performs a variety of useful trimming tasks for Illumina paired-end and single-end data.

F Tables

Table 29: Definition of fields of the tab delimited variant report (Sample.indels.tsv and Sample.snps.tsv).

Name	Meaning
Ref ID	Name of chromosome or reference contig where the variant occurs.
Position	Position of reference contig or chromosome where the variant occurs.
Reference Base (s)	The reference base at the variant site.
Modified Base (s)	Alternative (observed) base in the samples in general [VARIANT].
Mutation Frequency (%)	The mutation frequency with which a particular mutation occurs in a population.
Coverage Depth (x)	The total depth of the reads that passed the internal quality control metrics from all reads present at this site.
dbID	Known variant indentifier.
FILTER	Variants passing the filters will be tagged as ”PASS” and the variants failing the filters will be tagged by the respective filter names.
AF	Allele (Mutation) frequency.
DP	Counts for ref-forward bases, ref-reverse, alt-forward and alt-reverse bases.
CLNDSDBID	Variant disease database ID.
CLNSIG	Variant Clinical Significance, 0 - unknown, 1 - untested, 2 - non-pathogenic, 3 - probable-non-pathogenic, 4 - probable-pathogenic, 5 - pathogenic, 6 - drug-response, 7 - histocompatibility, 255 - other.
ExAC_AF	Allele frequency in Exome Aggregation Consortium (gnomAD) database.
ExAC_AC	Allele counts in Exome Aggregation Consortium (gnomAD) database.

Table 30: Definition of genomic annotations as produced by snpEff (Sample.indels.tsv and Sample.snps.tsv).

Name	Meaning
EFFECT	Variant’s effect on protein.
IMPACT	Predicted impact from variant’s protein effect.
HGVS_C	Variant’s codon change (DNA level).
HGVS_P	Variant’s codon change (Protein level).
GENE	The gene entry associated with the location of the variant call.
BIOTYPE	Variant’s coding status.
TRID	Associated transcript IDs.
CDS_POS	Variant’s codon change position.
AA_POS	Variant’s amino acid position.

© Eurofins Genomics Europe Sequencing GmbH, 11-2021, K-1482, Version 3.27.32