eurofins | Genomics

Data Analysis Report: ONCOPANEL ALL-IN-ONE v2.6

Project / Study: EF-DEMO

Date: August 30, 2019

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 ($\geq$ 50x, or $\geq$ 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 sample_1, sample_2.

	sample_1	sample_2
Total SNV	87038	87252
Known SNV	77053	77008
Unknown SNV	9985	10244
Total InDel	38742	39725
Known InDel	26400	27208
Unknown InDel	12342	12517

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 28. Jan 2019) [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 sample_1 Results

Table 2: Variants (SNV and InDels) in sample - **sample_1.** 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	64.3 %	1175x	rs166622	drug response
chr17:17216394	AC055811.2	.	c.*119insC	6.5 %	92x	rs3363	pathogenic

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr5:132595759	AC116366.3	. . . .	c.*2341delA c.*1321delA c.*2025delA c.*2155delA	13.5 %	953x	rs408407	pathogenic
chr14:104780214	AKT1	. p.E17K	n.80G>A c.49G>A	3.8 %	499x	rs13983	pathogenic
chr10:94780653	AL583836.1	.	c.*394G>A	3.9 %	978x	rs16899	drug response
chr10:94781859	AL583836.1	.	c.*439G>A	21.9 %	603x	rs16897	drug response
chr20:32434638	ASXL1	p.G641fs p.G646fs	c.1919insG c.1934insG	4.3 %	376x	rs426927	pathogenic
chr11:108335105	ATM	p.V2716A	c.8147T>C	3.6 %	1399x	rs142700	pathogenic
chr17:65536466	AXIN2	p.G600fs p.G665fs	c.1799delG c.1994delG	6.8 %	147x	rs5880	pathogenic
chr7:140753336	BRAF	. p.V207E p.V600E p.V28E	c.*1249T>A c.620T>A c.1799T>A c.83T>A	17.6 %	1687x	rs13961	pathogenic
chr17:43082434	BRCA1	.	c.*4110C>T	4.3 %	1074x	rs17675	pathogenic
chr13:32339421	BRCA2	p.K1691fs	c.5073delA	5.0 %	923x	rs51762	pathogenic
chr13:32339699	BRCA2	p.N1784fs	c.5351delA	13.8 %	979x	rs37961	pathogenic
chr13:32363217	BRCA2	p.I2675fs	c.8021insA	6.1 %	723x	rs267050	pathogenic
chr9:21971187	CDKN2A	. p.P72L	c.*95C>T c.215C>T	7.9 %	89x	rs376310	pathogenic
chr15:93002203	CHD2	p.Q1392fs . .	c.4173insA c.*406insA c.*344insA	16.3 %	423x	rs218395	pathogenic
chr3:41224610	CTNNB1	p.S33Y p.S26Y	c.98C>A c.77C>A	6.1 %	968x	rs17577	pathogenic
chr15:51210647	CYP19A1	.	c.*161T>G	57.2 %	222x	rs316467	drug response
chr19:41006936	CYP2B6	p.Q172H	c.516G>T	21.2 %	1338x	rs29671	drug response
chr19:41009358	CYP2B6	p.K262R	c.785A>G	25.7 %	350x	rs120171	drug response
chr10:94942290	CYP2C9	p.R144C	c.430C>T	11.5 %	801x	rs8409	drug response
chr10:94981296	CYP2C9	p.I359L	c.1075A>C	6.5 %	1025x	rs8408	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	35.2 %	565x	rs16889	drug response
chr7:55174771	EGFR	p.E746_ A750del p.E701_ A705del .	c.2235_ 2249delGGAATTAAGAGAAGC c.2100_ 2114delGGAATTAAGAGAAGC c.*225_ *239delGGAATTAAGAGAAGC	2.3 %	1212x	rs163343	drug response
chr7:55181305	EGFR	p.A767_ V769ins . p.A722_ V724ins	c.2300_ 2308insCCAGCGTGG c.*290_ *298insCCAGCGTGG c.2165_ 2173insCCAGCGTGG	3.4 %	948x	rs177678	drug response

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr7:55174014	EGFR	p.G719S . p.G674S	c.2155G>A c.*145G>A c.2020G>A	4.1 %	533x	rs16612	drug response
chr22:41169525	EP300	p.D1399N	c.4195G>A	9.6 %	741x	rs376401	likely pathogenic
chr8:117837145	EXT1	p.R129H p.R340H	c.386G>A c.1019G>A	12.0 %	624x	rs265129	pathogenic
chr16:89792518	FANCA	. p.E345fs	c.*21_ *22delAG c.1034_ 1035delAG	10.5 %	257x	rs558653	likely pathogenic
chr5:177093242	FGFR4	p.G23R p.G388R	c.67G>A c.1162G>A	28.0 %	239x	rs16326	pathogenic
chr17:17216394	FLCN	p.H429fs	c.1285insC	6.5 %	92x	rs3363	pathogenic
chr19:3118944	GNA11	p.Q57L p.Q209L	c.170A>T c.626A>T	5.6 %	503x	rs376002	pathogenic
chr11:67585218	GSTP1	. p.I105V	c.*137A>G c.313A>G	59.2 %	586x	rs37340	drug response
chr12:120994311	HNF1A	p.P291fs p.G226fs .	c.864delG c.677delG c.*304delG	4.2 %	330x	rs435424	pathogenic
chr11:118473470	KMT2A	p.P806fs p.P773fs	c.2417delC c.2318delC	10.0 %	1820x	rs522154	pathogenic
chr12:25245347	KRAS	p.G13D	c.38G>A	6.1 %	1570x	rs12580	pathogenic
chr15:66435113	MAP2K1	p.Q56P	c.167A>C	4.6 %	412x	rs375978	pathogenic
chr15:66436809	MAP2K1	p.H119Y	c.355C>T	4.9 %	937x	rs40741	pathogenic
chr5:80675095	MSH3	p.K383fs	c.1148delA	29.2 %	942x	rs8738	pathogenic
chr2:47803500	MSH6	p.F958fs p.F1088fs . p.F786fs p.F56fs	c.2871insC c.3261insC c.*2608insC c.2355insC c.165insC	5.6 %	1494x	rs89364	pathogenic
chr17:31226459	NF1	p.P678fs p.P344fs . p.P712fs	c.2033delC c.1031delC c.*1434delC c.2135delC	5.0 %	958x	rs428991	pathogenic
chr17:31226459	NF1	p.I345fs p.I679fs . p.I713fs	c.1031insC c.2033insC c.*1434insC c.2135insC	5.1 %	958x	rs141513	pathogenic
chr3:179218303	PIK3CA	p.E545K	c.1633G>A	4.2 %	830x	rs13655	pathogenic
chr3:179230077	PIK3CA	p.G914R	c.2740G>A	4.9 %	1711x	rs39703	pathogenic
chr3:179234297	PIK3CA	p.H1047R	c.3140A>G	16.7 %	1486x	rs13652	pathogenic
chr5:132595759	RAD50	p.K722fs . . p.?661fs	c.2165delA c.*1791delA c.*351delA c.1982delA	13.5 %	953x	rs408407	pathogenic
chr12:21178615	SLCO1B1	p.V174A	c.521T>C	17.4 %	1074x	rs37346	drug response
chr7:141972804	TAS2R38	p.I296V	c.886A>G	60.2 %	1378x	rs2906	drug response
chr7:141973545	TAS2R38	p.A49P	c.145G>C	52.9 %	1201x	rs2904	drug response

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr17:7674241	TP53	p.S82F p.S230F p.S202F p.S109F p.S148F p.S241F	c.245C>T c.689C>T c.605C>T c.326C>T c.443C>T c.722C>T	5.5 %	470x	rs12359	likely pathogenic
chr17:7674241	TP53	p.S148C p.S82C p.S241C p.S230C p.S109C p.S202C	c.443C>G c.245C>G c.722C>G c.689C>G c.326C>G c.605C>G	5.3 %	470x	rs177791	likely pathogenic
chr17:7676154	TP53	p.P33R p.P72R	c.98C>G c.215C>G	80.0 %	210x	rs12351	drug response
chr3:14145949	XPC	p.Q939K .	c.2815C>A c.*2268C>A	36.5 %	384x	rs190215	drug response

1.2.2 sample_2 Results

Table 3: Variants (SNV and InDels) in sample - **sample_2.** 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	63.3 %	1225x	rs166622	drug response
chr17:17216394	AC055811.2	.	c.*119insC	16.0 %	81x	rs3363	pathogenic
chr5:132595759	AC116366.3	. . . .	c.*2341delA c.*1321delA c.*2025delA c.*2155delA	14.3 %	883x	rs408407	pathogenic
chr14:104780214	AKT1	. p.E17K	n.80G>A c.49G>A	4.5 %	404x	rs13983	pathogenic
chr10:94780653	AL583836.1	.	c.*394G>A	4.0 %	1004x	rs16899	drug response
chr10:94781859	AL583836.1	.	c.*439G>A	20.1 %	656x	rs16897	drug response
chr20:32434638	ASXL1	p.G641fs p.G646fs	c.1919insG c.1934insG	5.1 %	351x	rs426927	pathogenic
chr11:108335105	ATM	p.V2716A	c.8147T>C	2.9 %	1368x	rs142700	pathogenic
chr7:140753336	BRAF	. p.V207E p.V600E p.V28E	c.*1249T>A c.620T>A c.1799T>A c.83T>A	16.9 %	1635x	rs13961	pathogenic
chr17:43082434	BRCA1	.	c.*4110C>T	4.0 %	1001x	rs17675	pathogenic
chr13:32339421	BRCA2	p.K1691fs	c.5073delA	6.6 %	1040x	rs51762	pathogenic
chr13:32339699	BRCA2	p.N1784fs	c.5351delA	17.7 %	979x	rs37961	pathogenic
chr13:32363217	BRCA2	p.I2675fs	c.8021insA	5.5 %	747x	rs267050	pathogenic
chr9:21971187	CDKN2A	. p.P72L	c.*95C>T c.215C>T	8.5 %	94x	rs376310	pathogenic
chr15:93002203	CHD2	p.Q1392fs . .	c.4173insA c.*406insA c.*344insA	15.9 %	428x	rs218395	pathogenic
chr3:41224610	CTNNB1	p.S33Y p.S26Y	c.98C>A c.77C>A	4.9 %	871x	rs17577	pathogenic
chr15:51210647	CYP19A1	.	c.*161T>G	49.5 %	210x	rs316467	drug response
chr19:41006936	CYP2B6	p.Q172H	c.516G>T	21.9 %	1272x	rs29671	drug response
chr19:41009358	CYP2B6	p.K262R	c.785A>G	23.2 %	349x	rs120171	drug response
chr10:94942290	CYP2C9	p.R144C	c.430C>T	10.8 %	719x	rs8409	drug response
chr10:94981296	CYP2C9	p.I359L	c.1075A>C	5.5 %	969x	rs8408	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	35.1 %	524x	rs16889	drug response

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr7:55174771	EGFR	p.E746_A750del p.E701_A705del .	c.2235_2249delGGAATTAAGAGAAGC c.2100_2114delGGAATTAAGAGAAGC c.*225_*239delGGAATTAAGAGAAGC	2.3 %	1105x	rs163343	drug response
chr7:55181305	EGFR	p.A767_V769ins . p.A722_V724ins	c.2300_2308insCCAGCGTGG c.*290_*298insCCAGCGTGG c.2165_2173insCCAGCGTGG	2.5 %	947x	rs177678	drug response
chr7:55174014	EGFR	p.G719S . p.G674S	c.2155G>A c.*145G>A c.2020G>A	5.1 %	505x	rs16612	drug response
chr22:41169525	EP300	p.D1399N	c.4195G>A	7.1 %	743x	rs376401	likely pathogenic
chr8:117837145	EXT1	p.R129H p.R340H	c.386G>A c.1019G>A	12.0 %	607x	rs265129	pathogenic
chr16:89792518	FANCA	. p.E345fs	c.*21_*22delAG c.1034_1035delAG	7.9 %	240x	rs558653	likely pathogenic
chr5:177093242	FGFR4	p.G23R p.G388R	c.67G>A c.1162G>A	30.3 %	201x	rs16326	pathogenic
chr17:17216394	FLCN	p.H429fs	c.1285insC	16.0 %	81x	rs3363	pathogenic
chr19:3118944	GNA11	p.Q57L p.Q209L	c.170A>T c.626A>T	5.5 %	457x	rs376002	pathogenic
chr11:67585218	GSTP1	. p.I105V	c.*137A>G c.313A>G	58.7 %	513x	rs37340	drug response
chr11:118473470	KMT2A	p.P806fs p.P773fs	c.2417delC c.2318delC	8.5 %	1812x	rs522154	pathogenic
chr12:25245347	KRAS	p.G13D	c.38G>A	6.6 %	1635x	rs12580	pathogenic
chr15:66435113	MAP2K1	p.Q56P	c.167A>C	7.2 %	403x	rs375978	pathogenic
chr15:66436809	MAP2K1	p.H119Y	c.355C>T	5.3 %	889x	rs40741	pathogenic
chr5:80675095	MSH3	p.K383fs	c.1148delA	27.7 %	982x	rs8738	pathogenic
chr2:47803500	MSH6	p.F958fs p.F1088fs . p.F786fs p.F56fs	c.2871insC c.3261insC c.*2608insC c.2355insC c.165insC	5.5 %	1388x	rs89364	pathogenic
chr17:31226459	NF1	p.I345fs p.I679fs . p.I713fs	c.1031insC c.2033insC c.*1434insC c.2135insC	5.5 %	873x	rs141513	pathogenic
chr17:31226459	NF1	p.P678fs p.P344fs . p.P712fs	c.2033delC c.1031delC c.*1434delC c.2135delC	4.7 %	873x	rs428991	pathogenic
chr3:179218303	PIK3CA	p.E545K	c.1633G>A	4.6 %	831x	rs13655	pathogenic
chr3:179230077	PIK3CA	p.G914R	c.2740G>A	5.0 %	1670x	rs39703	pathogenic
chr3:179234297	PIK3CA	p.H1047R	c.3140A>G	16.5 %	1471x	rs13652	pathogenic

Location	Gene	AA Change	Codon Change	Mutation Freq.	Depth	ClinVar ID	ClinVar Significance
chr5:132595759	RAD50	p.K722fs . . p.?661fs	c.2165delA c.*1791delA c.*351delA c.1982delA	14.3 %	883x	rs408407	pathogenic
chr12:21178615	SLCO1B1	p.V174A	c.521T>C	18.2 %	1089x	rs37346	drug response
chr7:141972804	TAS2R38	p.I296V	c.886A>G	59.5 %	1386x	rs2906	drug response
chr7:141973545	TAS2R38	p.A49P	c.145G>C	54.7 %	1172x	rs2904	drug response
chr17:7674241	TP53	p.S148C p.S82C p.S241C p.S230C p.S109C p.S202C	c.443C>G c.245C>G c.722C>G c.689C>G c.326C>G c.605C>G	4.5 %	404x	rs177791	likely pathogenic
chr17:7674241	TP53	p.S82F p.S230F p.S202F p.S109F p.S148F p.S241F	c.245C>T c.689C>T c.605C>T c.326C>T c.443C>T c.722C>T	6.4 %	404x	rs12359	likely pathogenic
chr17:7676154	TP53	p.P33R p.P72R	c.98C>G c.215C>G	82.4 %	205x	rs12351	drug response
chr9:93289548	WNK2	p.D1204fs p.D1638fs p.D1596fs p.D397fs . p.D123fs p.D1601fs	c.3610insG c.4912insG c.4786insG c.1189insG c.-105insG c.367insG c.4801insG	7.0 %	115x	rs520975	likely pathogenic
chr3:14145949	XPC	p.Q939K .	c.2815C>A c.*2268C>A	40.3 %	330x	rs190215	drug response

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 used according to the published works[4, 5] with some exclusions. 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 $<$ 0.05
- 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 4: TMB values for each sample

Sample	TMB1	TMB2	TMB3
sample_1	120.29	90.13	66.75
sample_2	118.26	90.81	69.46

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 5: Case vs Control setup.

Case	Control(s)
sample_1	sample_2

Table 6: Summary of CNV events detected in each sample.

Sample	Duplication Events	Deletion Events
sample_1	6	0

1.4.1 sample_1 Results

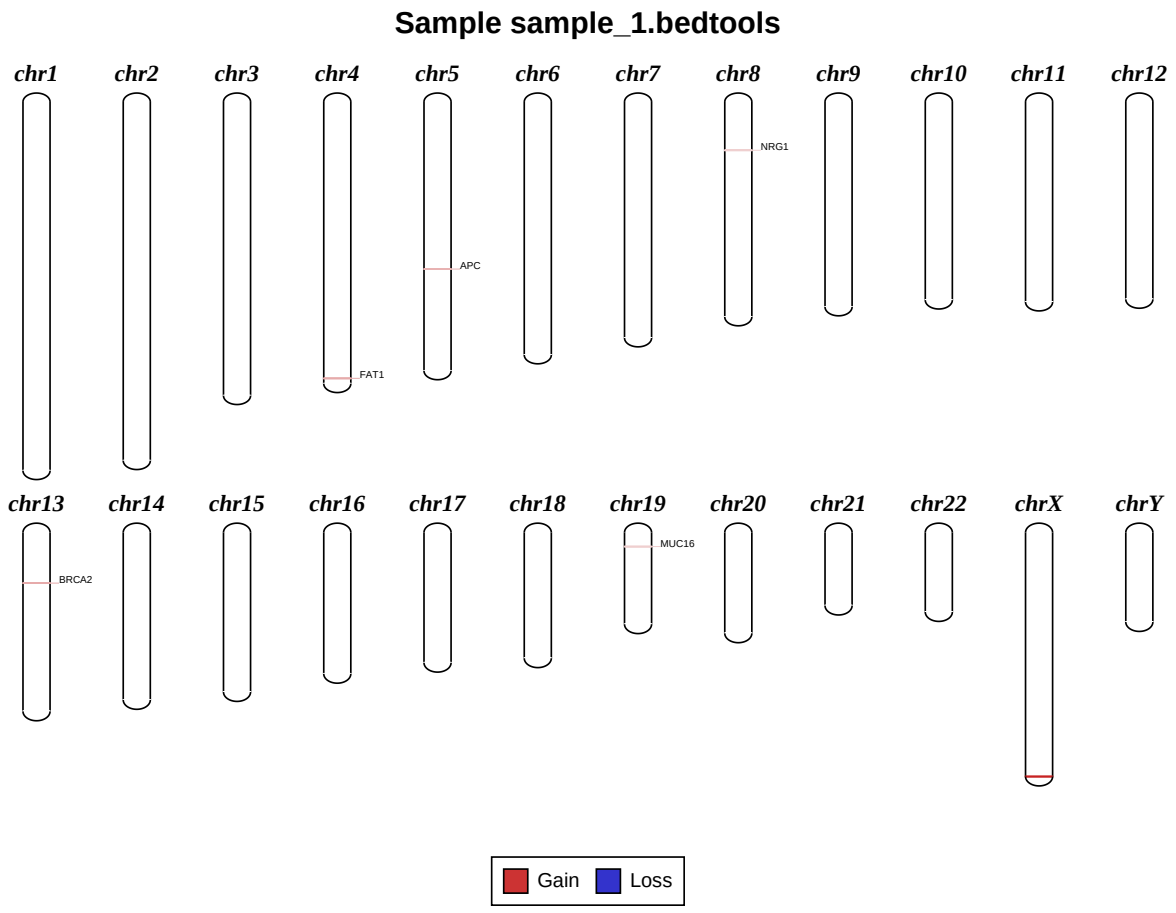

Figure 1: Ideogram representing chromosome wise copy number events observed in sample sample_1. Copy gain events are drawn in red and copy loss events are drawn in blue.

Table 7: Duplication events detected in sample sample_1. 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
NRG1	3	1109.62	chr8:32622747-32719140
MUC16	3	1236.01	chr19:8935059-8981169
FAT1	3	1658.77	chr4:186617671-186621810
FAT1	3	1498.12	chr4:186706644-186709944
BRCA2	3	1179.68	chr13:32332365-32341220
APC	3	1524.13	chr5:112837538-112844038

No deletion events found!

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 1 x coverage are reported. Complete lists of fusion events can be found in supplementary deliverables.

Table 8: Summary of fusion events detected in each sample.

Sample	Known events	Unknown events
sample_1	2	3
sample_2	2	2

1.5.1 sample_1 Results

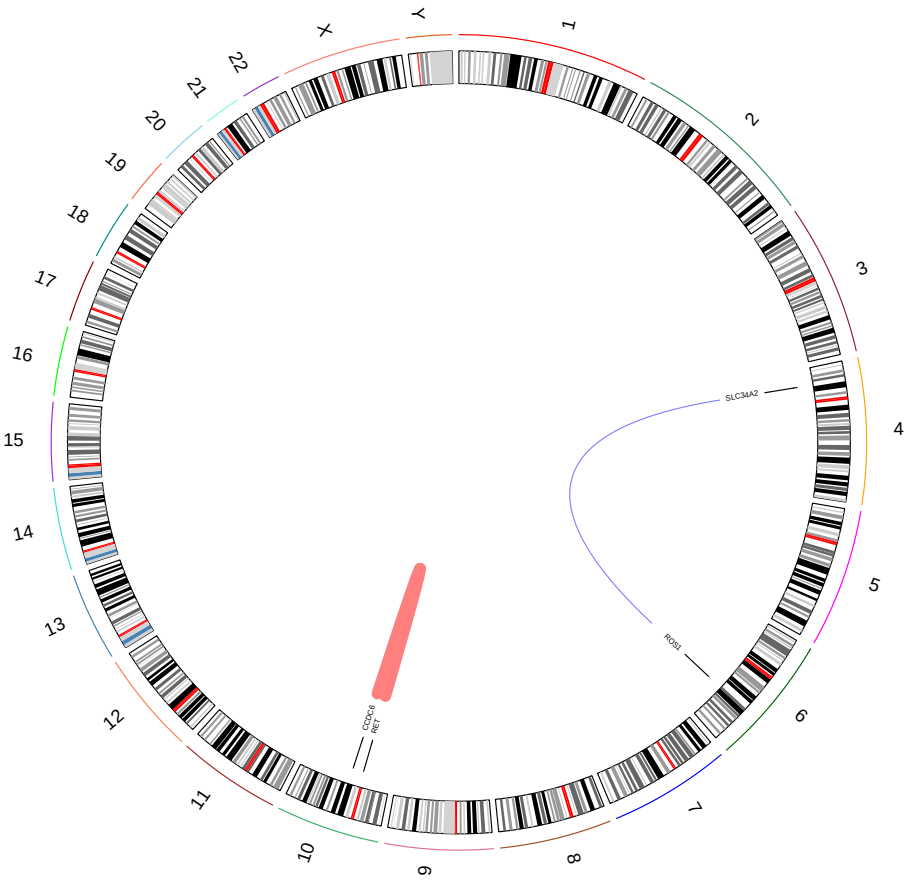

Figure 2: Circos plot displaying fusion events in relation to chromosome location for sample sample_1. Fusion events observed on the same chromosome are drawn in red whereas fusion events that are on different chromosomes are drawn in blue. Gene annotations are drawn at the tip of the arcs.

Table 9: Fusion events detected in sample sample_1. Associated disease and source of annotation are mentioned in Disease and Source column, respectively.

Fusion genes	Fusion location	Supporting fusion reads	Supporting paired reads	Disease	Source
RET–CCDC6	chr10:43114504-chr10:59878853	27	32	adenocarcinoma	Mitelman,OMIM,GenBank

Fusion genes	Fusion location	Supporting fusion reads	Supporting paired reads	Disease	Source
ROS1–SLC34A2	chr6:117337163-chr4:25665007	7	12	non small cell lung cancer, lung cancer, gastric adenocarcinoma, lung adenocarcinoma	Cosmic

1.5.2 sample_2 Results

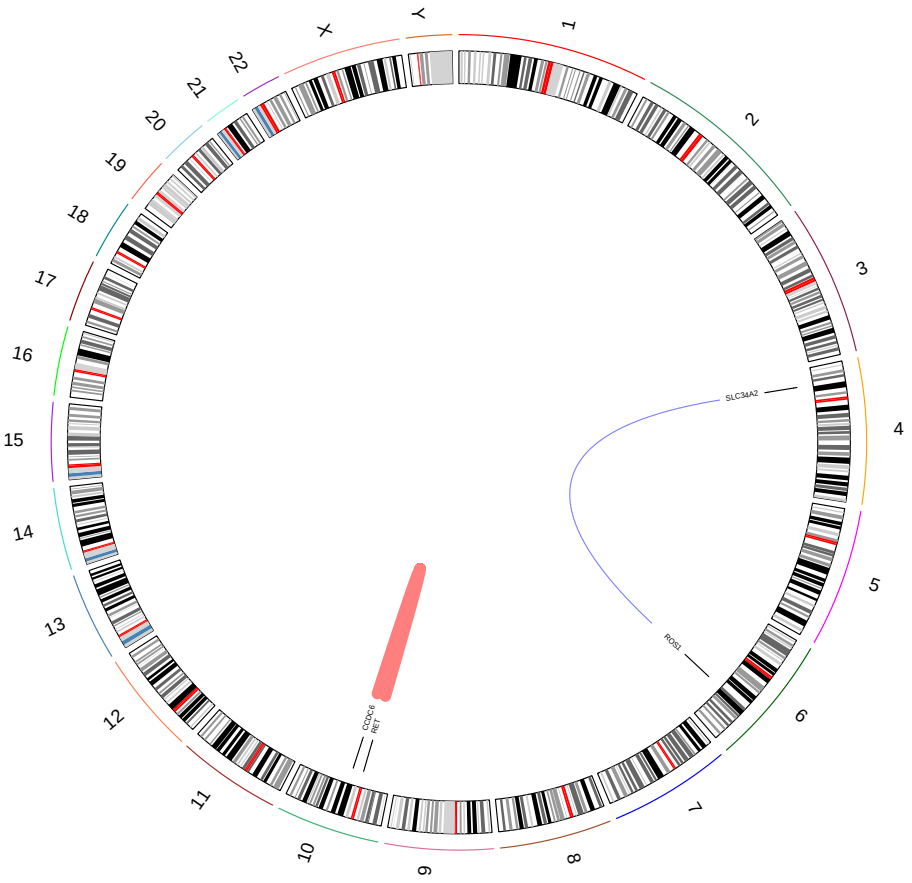

Figure 3: Circos plot displaying fusion events in relation to chromosome location for sample sample_2. Fusion events observed on the same chromosome are drawn in red whereas fusion events that are on different chromosomes are drawn in blue. Gene annotations are drawn at the tip of the arcs.

Table 10: Fusion events detected in sample sample_2. Associated disease and source of annotation are mentioned in Disease and Source column, respectively.

Fusion genes	Fusion location	Supporting fusion reads	Supporting paired reads	Disease	Source
RET–CCDC6	chr10:43114504-chr10:59878853	20	26	adenocarcinoma	Mitelman,OMIM,GenBank

Fusion genes	Fusion location	Supporting fusion reads	Supporting paired reads	Disease	Source
ROS1–SLC34A2	chr6:117337163-chr4:25665007	6	23	non small cell lung cancer, lung cancer, gastric adenocarcinoma, lung adenocarcinoma	Cosmic

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 11: Sequence quality metrics per sample

Sample	Total Reads	LQ Reads	Single Reads	HQ Reads
sample_1	43,787,136	394,056 (0.9%)	344,260 (0.8%)	43,048,820 (98.3%)
sample_2	42,247,494	365,427 (0.9%)	320,663 (0.8%)	41,561,404 (98.4%)

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).

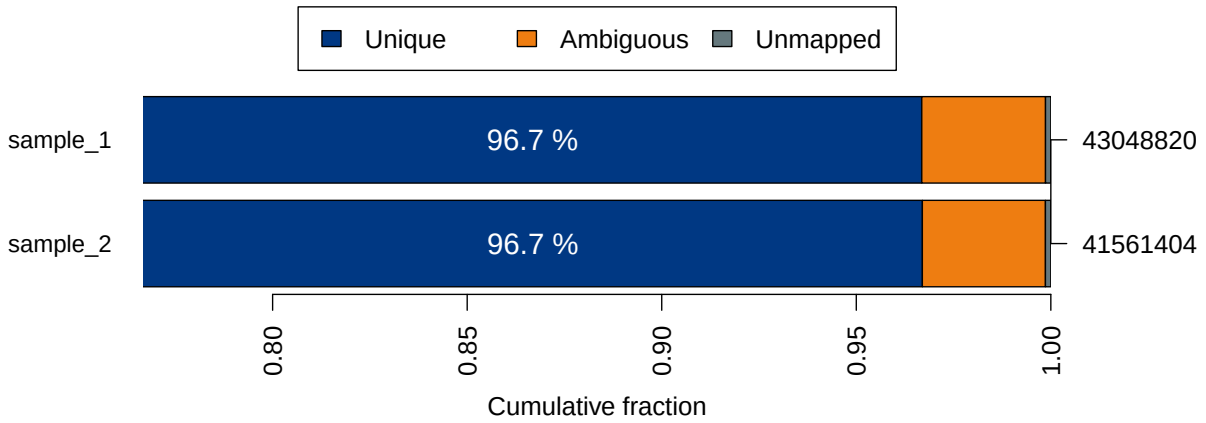

Figure 4: Summary of alignment results. For each sample, the fraction of uniquely mapped, non-uniquely mapped (ambiguous) and unmapped reads relative to the total number of reads per sample (right y-axis) is shown.

Table 12: 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	sample_1	42,989,734 (99.86%)	41,620,836 (96.82%)	33,935,022 (81.53%)	24,560,244 (72.37%)
2	sample_2	41,503,699 (99.86%)	40,189,165 (96.83%)	32,354,472 (80.51%)	23,837,862 (73.68%)

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)

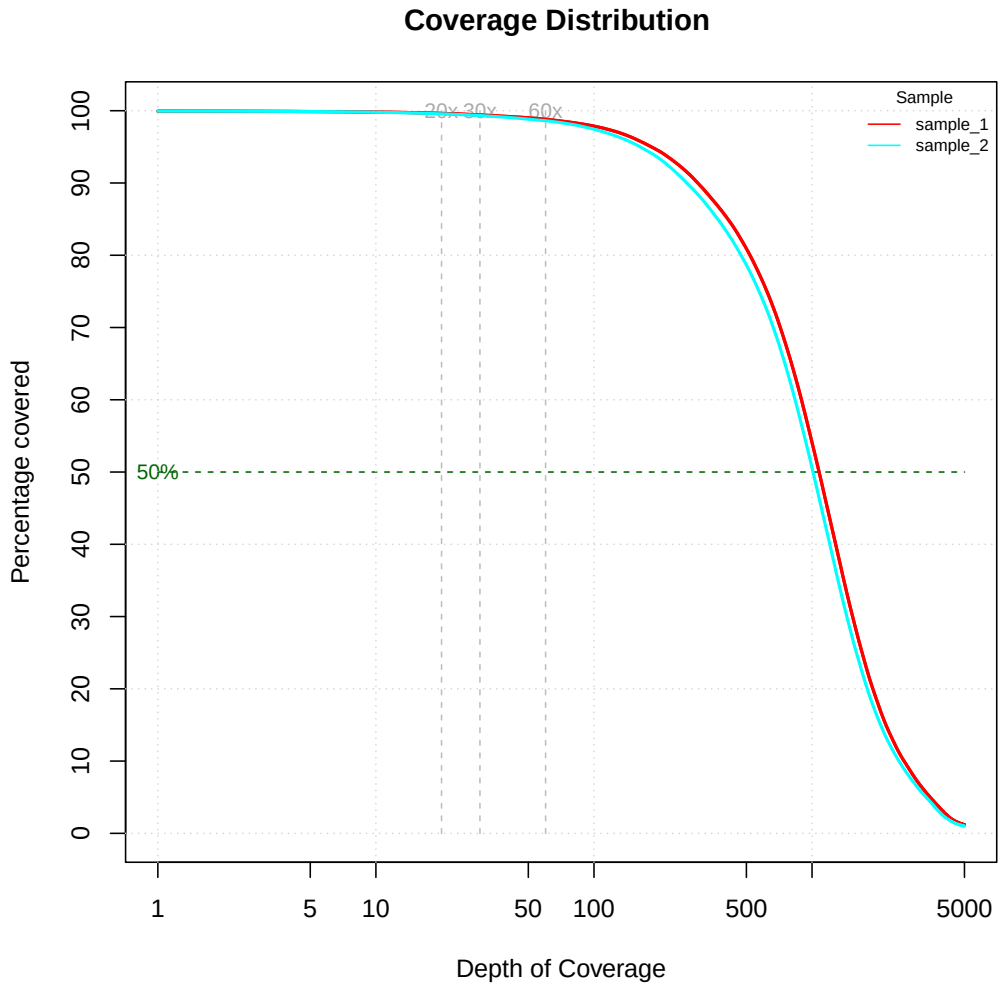

Figure 5: Coverage plot (including duplicated fragments).

Table 13: 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
sample_1	3.94 GB	1334.66	99.9	99.0	97.9	90.0	80.9
sample_2	3.71 GB	1257.91	99.9	98.8	97.4	88.4	78.7

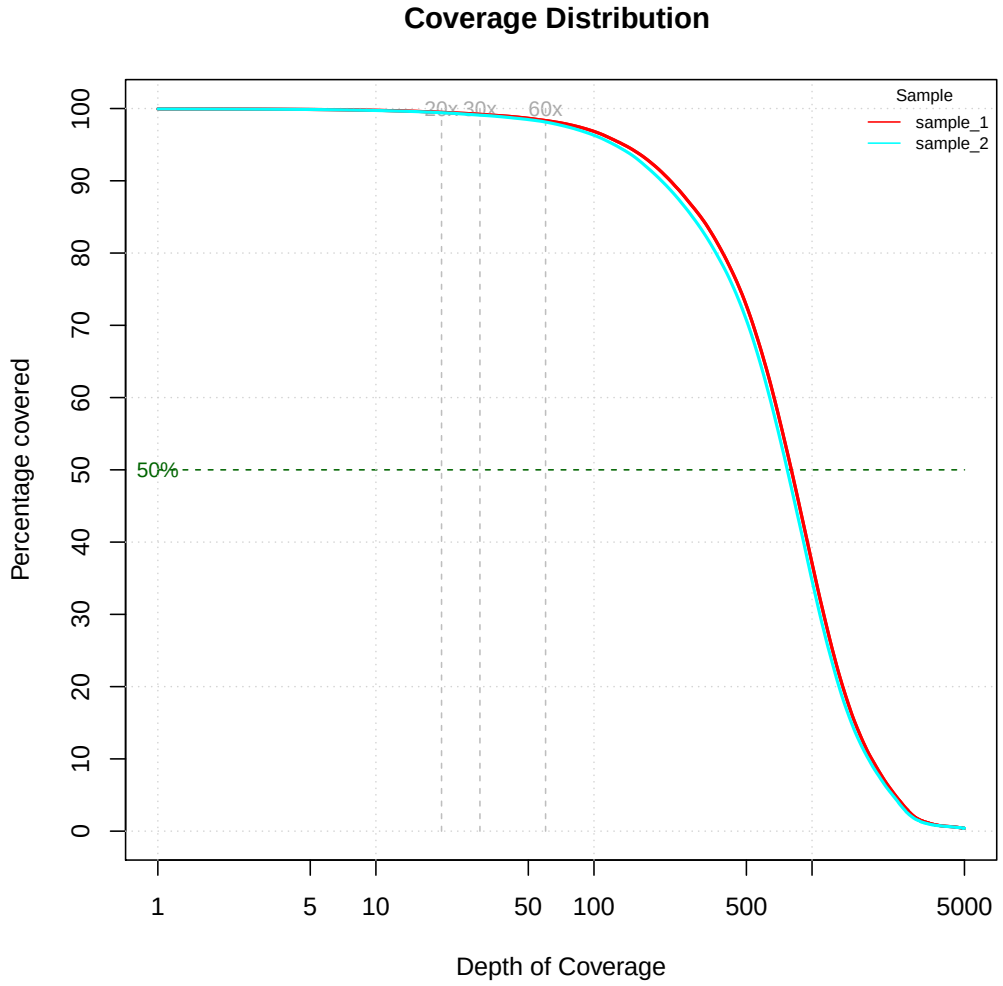

Figure 6: Coverage plot (excluding duplicated fragments).

Table 14: 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
sample_1	2.84 GB	964.17	99.9	98.6	96.8	85.5	72.7
sample_2	2.73 GB	925.92	99.9	98.5	96.3	84.0	70.7

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.

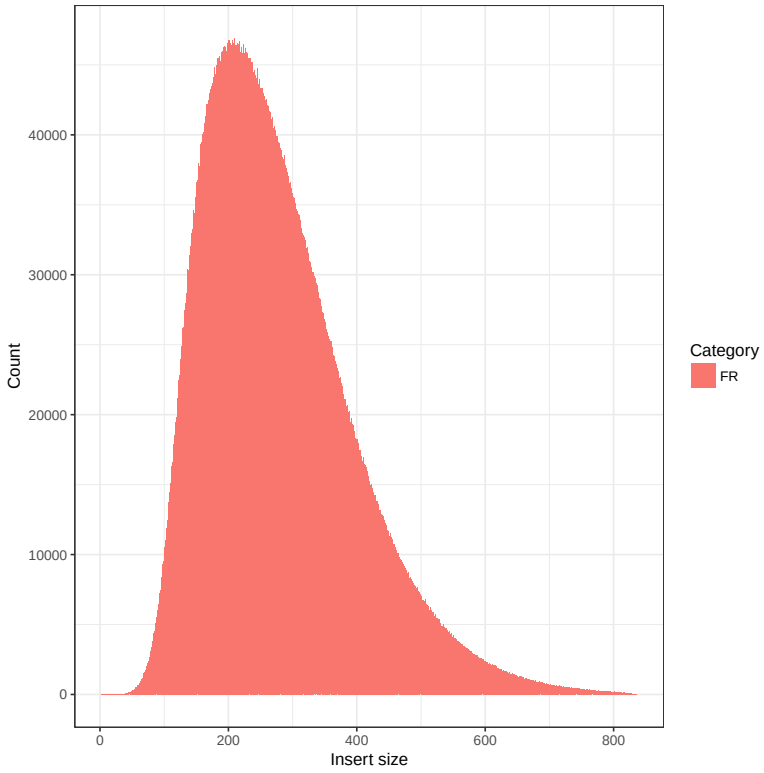

Figure 7: sample_1 .

Table 15: 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
sample_1	FR	278	117	12,264,599
sample_2	FR	289	122	11,906,871

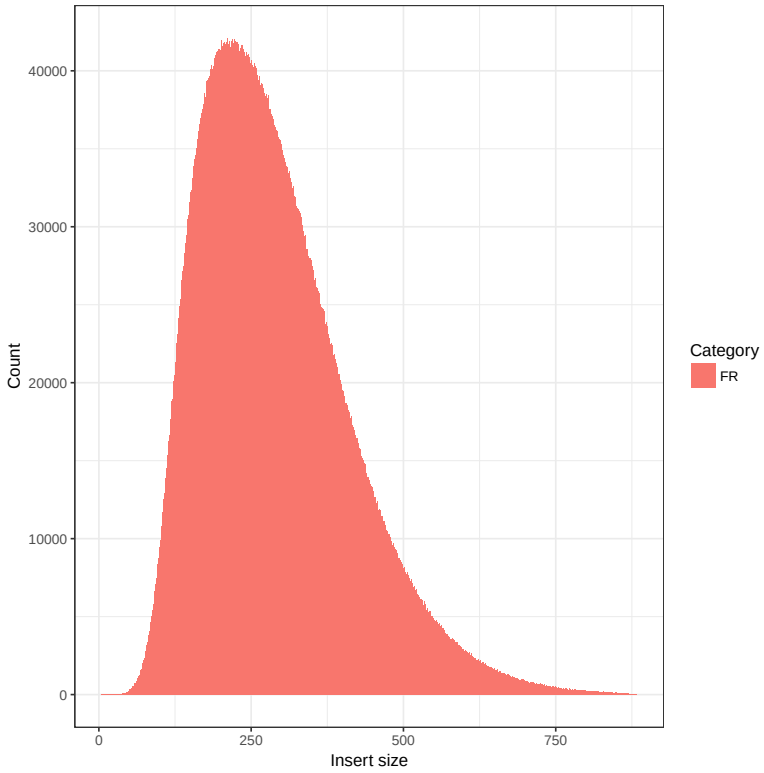

Figure 8: sample_2 .

3 Deliverables

Table 16: 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.hg19.HQ.alignment.bam	BAM	IGV, Tablet
SAMPLE.hg19.HQ.alignment.bam.bai	BAI	None
SAMPLE.hg19.alignment.bam	BAM	IGV, Tablet
SAMPLE.hg19.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.hg19.alignment.bam was used for Fusion Gene discovery (see chapter 1.5)
SAMPLE.hg19.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 17: 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.

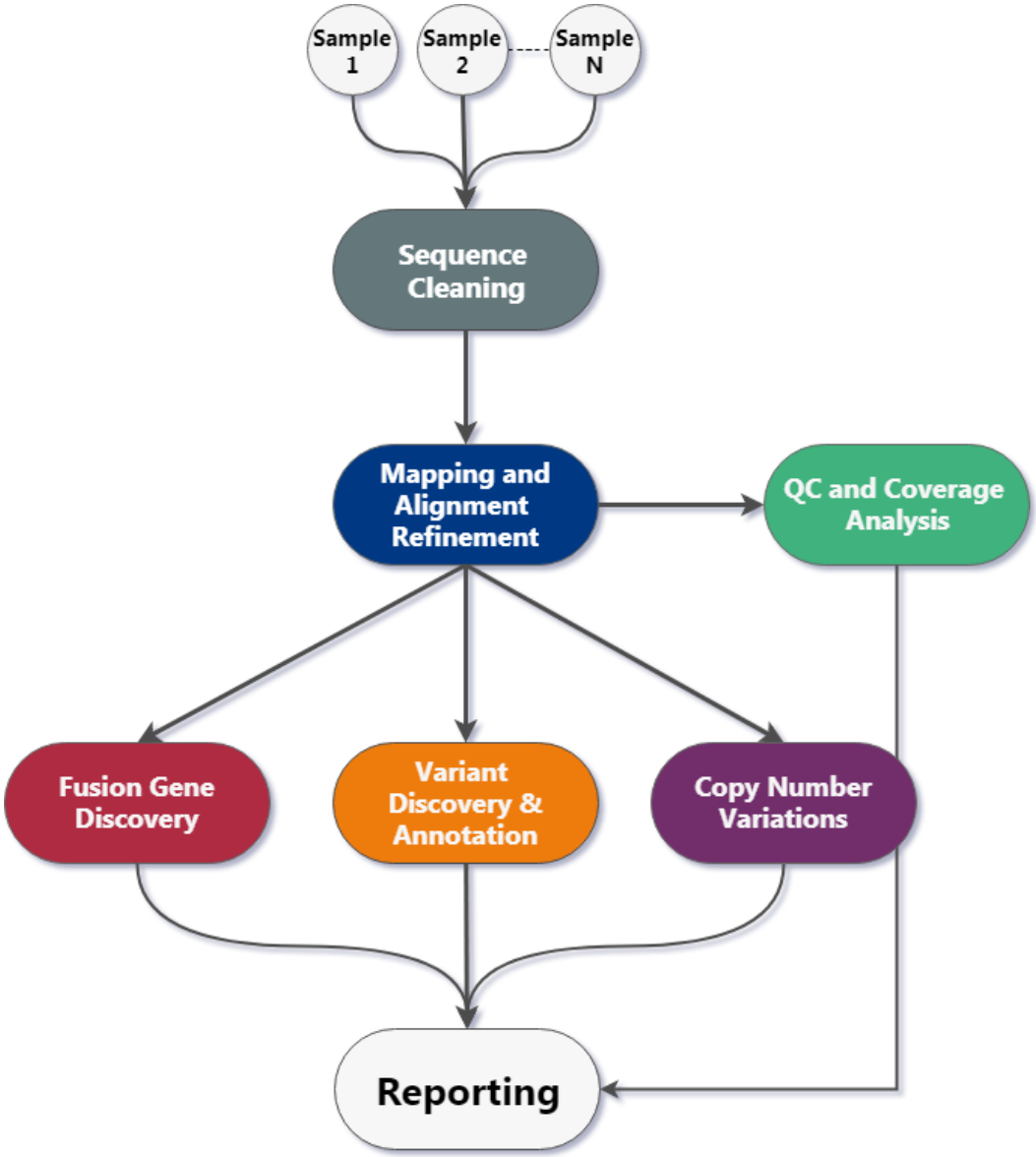

Figure 9: ONCOPANEL ALL-IN-ONE v2.6 Workflow

B Sequence Data Used

Table 18: Analysed samples (SE = single end, PE = paired end).

Sample	Read Type	File Name
sample_1	PE	EF-DEMO_sample_1_lib344027_6507_1_1.fastq.gz.gz
		EF-DEMO_sample_1_lib344027_6507_1_2.fastq.gz.gz
sample_2	PE	EF-DEMO_sample_2_lib344032_6507_1_1.fastq.gz.gz
		EF-DEMO_sample_2_lib344032_6507_1_2.fastq.gz.gz

C Reference Database

Table 19: 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 20: Information about additional reference data used.

Type	Version	Source
Annotation	22	GENCODE
dbSNP[6]	142	NCBI
ClinVar[3]	28.01.19	NCBI
COSMIC[2]	71	Sanger Institute
gnomAD[7]	2.1.1	Broad Institute
ChimerDB[11]	2.0	ERCSB

Table 21: 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 22: 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]	2.1.2	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 23: 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.

Table 24: 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' products, services and applications reach the best quality and safety levels. They are carried out under strict QM and QA systems and comply with the following standards:

ISO 17025	Accredited analytical excellence	GLP	The gold standard to conduct non-clinical safety studies
ISO 13485	Oligonucleotides according to medical devices standard	GCP	Pharmacogenomic services for clinical studies
cGMP	Products and testing according to pharma and biotech requirements		

Eurofins Genomics Europe Sequencing GmbH • Jakob-Stadler-Platz 7 • 78467 Constance • Germany