Research &
Pharma Solutions

Whole Genome Sequencing on Our Illumina Platforms

Bioinformatic Note

Whole Genome Sequencing on Our Illumina Platforms

The genome represents the entire genetic information of an organism. The most comprehensive collection of an individual's genetic information is provided by analyzing the whole genome using next-generation sequencing (NGS) technology. By comparing the individual's genetic information to a reference genome, single nucleotide variants (SNVs), small insertions and deletions (indels), copy number variations (CNVs), and structural variants (SVs) can be studied.

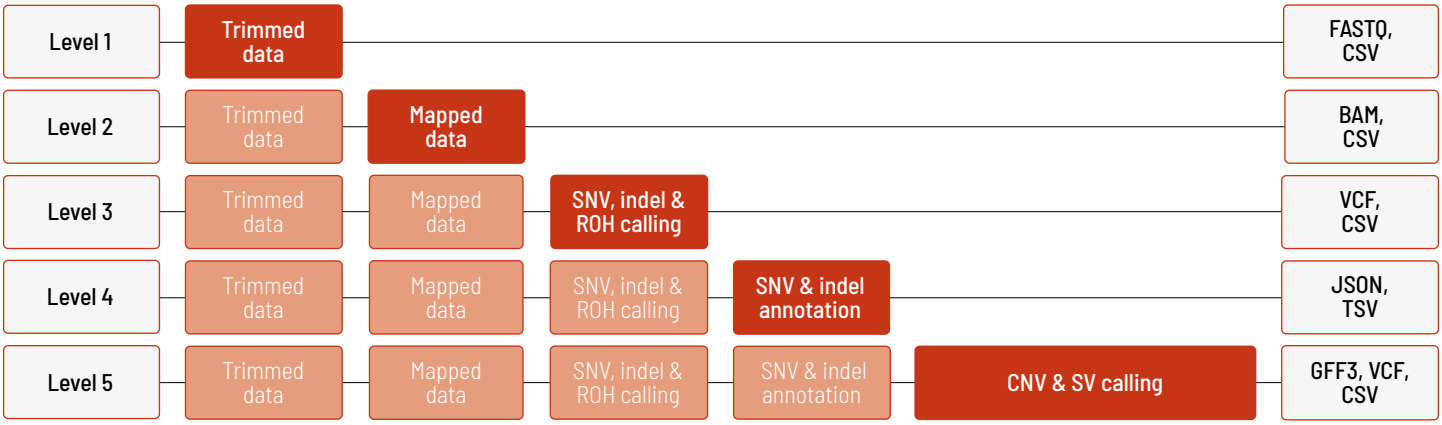
Whole genome sequencing (WGS) studies contribute to:

- cancer studies, personalized medicine approaches, and translational research
- discovery of biomarkers and understanding of pharmacogenetics
- disease research
- plant and animal breeding programs
- examination of microorganisms

The analyses of our WGS Large Classic and WGS Flex are performed with the Illumina DRAGEN Bio-IT Platform. For human samples, we can analyze the sequencing data based on the human references hg19 and GRCh38.

With increasing bioinformatic level, more data is delivered. All higher levels include the data from the lower levels, e.g., in Level 2, trimmed data and mapped data are provided. In addition to the data, and independent of the analysis level, a project report is generated.

Different levels of bioinformatic data analysis are available:



Level 1

If you wish to analyze your data yourself, we recommend the Levels 1 or 2. The default level for raw data is Level 1, where trimmed reads in FASTQ format are delivered. In this level, the sequencing data are demultiplexed and trimmed. This level is provided for every project, regardless of additionally purchased bioinformatic analyses.

Quality control of the samples is performed, resulting in a metrics file in CSV format. This file contains different sections for each metric type, such as read mean quality, positional base mean quality, positional base content, read GC content and quality, sequence positions, or positional quality. Each of these sections includes separate rows for length, position, or other relevant category variables.

The generated project report provides information for every sample about the laboratory protocol, including data about quality control of the starting material, library preparation, sequencing parameters, and the Q30 value of the sequencing. For the trimmed data, the number of sequenced fragments and bases is reported, and the sequence length, quality of the reads, and the GC content are illustrated in bar plots for all samples.

Additionally, a multiQC report in HTML format is generated. It covers the results from the DRAGEN FastQC module and – if performed – further DRAGEN analyses. In contrast to the generated project report, this multiQC report facilitates the interactive exploration of the analysis results: samples can be highlighted, renamed, or hidden. Additionally, figures can be customized, edited, and saved.

Level 2

If you wish to receive Level 2, the trimmed reads are aligned to the reference genome and duplicates are marked. In addition to the trimmed reads in FASTQ format, you will receive the mapped reads as BAM files.

Together with the mapped reads, you will receive a mapping metrics file in CSV format that includes mapping and aligning metrics. The metrics are available over all input data as well as on a per-read-group level. Examples of the mapping and aligning metrics are the number of total input reads, the number of duplicate marked reads, the number of unique reads, reads with mate sequenced, QC-failed reads, and mapped reads. Table 1 shows an excerpt of the mapping metrics file.

Another file in CSV format reports the coverage metrics. It provides metrics over the target region, such as the aligned bases in the genome, average alignment coverage over the genome, or the uniformity of the coverage. Like the mapping metrics file, the coverage metrics file provides the metrics and respective values per row.

For the coverage distribution, three files are provided: a BEDGRAPH file, a PDF file, and a CSV file. The BEDGRAPH file contains the coverage of the genome in bins of bases. An excerpt is provided in table 2. Similar to a BED file, it contains the chromosome, start, and end position. In a fourth column, additional information is stored. In this case, it is the coverage in the respective region. This file is the basis for the coverage plots that are provided in the PDF file. The CSV file contains the coverage distribution of the genome in bins of coverage ranges. Table 3 shows an excerpt of the coverage distribution CSV file. For a given coverage range, the respective percentage of bases that have this coverage is indicated. In both tables, the headers are added for explanatory reasons and are not part of the delivered files.

Table 1 | Excerpt of the mapping metrics file in CSV format (header line added for the sake of clarity).

All input data/per read group level	Metrics	Value	Percentage
Mapping/Aligning summary	Total input reads	596571702	100
Mapping/Aligning summary	Number of duplicate marked reads	49986195	8.38
Mapping/Aligning summary	Number of duplicate marked and mate reads removed	NA	
Mapping/Aligning summary	Number of unique reads (excl. duplicate marked reads)	546585507	91.62

Table 2 | Excerpt of the coverage distribution BEDGRAPH file (header line added for the sake of clarity).

Chromosome	Start	End	Coverage
chr1	0	16384	30.51
chr1	16384	32768	42.98
chr1	32768	49152	26.91
chr1	49152	65536	27.52

Table 3 | Excerpt of the coverage distribution CSV file (header line added for the sake of clarity).

PCT of bases in wgs with coverage	
Range	Percentage
[100x:500x)	0.03
[50x:100x)	0.20
[20x:50x)	80.86
[15x:20x)	8.26
[10x:15x)	3.64
[3x:10x)	1.70
[1x:3x)	0.46
[0x:1x)	4.84

GC biases can arise from library prep, capture kits, sequencing system differences, and mapping. We can perform a GC bias correction. In a Gnu Zipped file (GZ), the GC-corrected target counts are stored. The file contains columns with the contig identifier, start position, end position, target interval name, (GC-corrected) count of alignments in this interval, and (GC-corrected) count of improperly paired alignments in this interval. An excerpt of this file can be found in table 4.

A ploidy estimator calculates the sequencing depth of the coverage for each autosome and allosome to subsequently estimate the sex karyotype of the sample. The result is provided in a ploidy estimation metrics file in CSV format. In this file, the median coverage of the autosomal chromosomes, the X chromosome, and the Y chromosome is provided, as well as the ratio of each chromosome and the autosomal median. Based on these ratios, a ploidy estimation is provided eventually.

For Level 2, the project report also includes a table with statistics of the mapped reads, including the number of mapped reads, the proportion of sequenced reads, the proportion of PCR duplicates, the median insert size, and the average coverage.

Table 4 | Excerpt of the GC-corrected target counts file.

contig	start	stop	name	sample1	improper_pairs
chr1	818022	819840	target-wgs-chr1-818022:819840	94.91390439	0
chr1	819840	821337	target-wgs-chr1-819840:821337	89.7734526	16
chr1	821337	822485	target-wgs-chr1-821337:822485	102.0582408	0

Level 3

In Level 3, single nucleotide variants (SNVs) and small insertions and deletions (indels) are called. For human samples, regions of homozygosity (ROH) can additionally be called.

Calling of the SNVs and indels is performed with default germline parameters. The resulting VCF file contains the quality-filtered calls of the SNVs and small indels. It includes information for every small variant about its location (chromosome and position), the identifier, the reference base(s) and the alternate base(s), the quality and filter status, and additional information in several columns. In the header of the VCF file, additional information about the FILTER, INFO, and FORMAT columns is provided, and the abbreviations are explained. An excerpt of the small variant VCF file is shown in table 5. For additional information on the VCF format, we refer the reader to the VCF specification. In addition to the VCF file, a respective index file in TBI format is delivered. This index file facilitates the automated processing of large variant files.

Table 5 | Excerpt of the SNVs and indels VCF file. The abbreviations in the INFO and FORMAT columns are explained in the respective VCF file.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1
chr1	10469	.	C	G	4.50	PASS	AC=1;AF=0.500;AN=2;DP=7; FS=0.000;MQ=44.24;MQRank-Sum=0.198;QD=4.02;ReadPosRank-Sum=0.198;SOR=1.802; FractionInformativeReads=1.000	GT:AD:AF:DP:F1R2:F2R1: GQ:PL:GP:PRI:SB:MB	0/1:5,2:0.2857:7:4,1:1,1:4: 37,0,13:4.4992e+00,2.000 4e+00,1.8470e+01:0.00,34. 77,37.77:2,3,0,2:3,2,1,1
chr1	10616	.	CCGCCGTT GCAAAGGC GCGCCG	C	4.11	PASS	AC=2;AF=1.000;AN=2;DP=3; FS=0.000;MQ=25.83;MQRank-Sum=0.000;QD=12.78;ReadPosRank-Sum=0.000;SOR=2.833; FractionInformativeReads=1.000	GT:AD:AF:DP:F1R2:F2R1: GQ:PL:GP:PRI:SB:MB	1/1:0,3:1.0000:3:0,2:0,1:3: 33,11,0:4.1116e+00,1.0465e+ 01,2.8221e+00:0.00,29. 00,32.00:0,0,0,3:0,0,1,2

An SNV and indels metrics file in CSV format includes information about the variant calling statistics. Amongst others, this metrics file contains information about the number of processed reads, the number of insertions, deletions, and SNVs before and after the quality filtering. Every entry consists of a section description, the sample, the metric, its value as count or ratio, and, where applicable, the percentage.

The regions of homozygosity (ROH) are called based on continuous runs of homozygous SNVs. They are provided as BED files. Each row corresponds to one region of homozygosity. The score is a function of the number of homozygous and heterozygous variants. The numbers of homozygous and heterozygous variants in the region are indicated in the last two columns, respectively.

Along with the ROH BED file, a metrics file for the ROH is provided in CSV format. It lists the number of large ROH and percentage of SNPS in these large ROH. A large ROH has more than 3 Mb.

Table 6 | Excerpt of the ROH BED file ((header line added for the sake of clarity).

Chromosome	Start	End	Score	#Homozygous	#Heterozygous
chr1	2340425	2409620	1.35	54	0
chr1	13625241	13807941	1.32	53	0
chr1	15045775	15145873	1.47	59	0
chr1	15961547	16030324	1.30	52	0

Level 4

In Level 4, the SNVs and indels are annotated. The annotations are available as compressed JSON files and as TSV files. The JSON files are usually very large and not optimized for human readability. However, it is useful for automated processing steps. We provide an additional annotation file in a tabular format that contains selected information from the JSON annotation file in a tabular format.

The annotations in tabular format include, amongst others,

- information about the chromosomal position and the observed variant,
- functional consequences of the variant in the context of a transcript,
- position and sequence changes in the context of the most affected transcripts,
- and information about the observed variant in the global population.

Together with the annotation list, you will receive a file that contains further information on the annotation of variant lists. Therefore, we will not go into detail about the provided TSV files. A database-versions file in TXT format provides information about the names, versions, and short descriptions of the used databases.

SNVs are commonly used markers in case-control association studies. Furthermore, some SNVs can have functional impact leading to disease susceptibilities and drug sensitivities. These functional impacts can concern the transcriptional machinery of a cell, alternative or aberrant splice isoforms when located at a splice site, or the translational machinery leading to protein folding, localization, stability, binding, or catalysis interference[1]. As SNVs, indels can have an impact on certain diseases. With the annotated SNVs and indels files, many research questions regarding functional impacts on diseases, disease susceptibility, and drug sensitivity might be answered.

Level 5

In Level 5, copy number variations (CNVs) can be called by identifying deviations from the expected coverage. For the CNVs, three files are provided: A VCF file, a GFF3 file, and a metrics file in CSV format.

The VCF file looks like the small variant VCF file displayed in table 5. Due to its similarity and the additional explanations in the header of the VCF file, we do not show an excerpt of the file here again.

The information about the CNVs is also stored in a GFF3 file. Like the VCF file, the GFF3 file includes information about the chromosome, the position of the feature, and additional information. An excerpt of the GFF3 file is shown in table 7. The header line is added for explanatory reasons and is not provided in the delivered file.

Like the other metrics files, the metrics file for CNVs is in CSV format and contains information about the CNV calling statistics, including, e.g., the number of called amplifications, deletions, or segments. Every entry consists of a section description, the metric, its value as count or ratio, and, where applicable, the percentage.

The PDF file that includes the coverage distribution also contains plots of the CNVs. For every chromosome, the position in kilobases is indicated on the x-axis, and the number of copies on the y-axis.

Additionally, structural variations (SVs) such as translocations, inversions, and large and medium-sized indels are called. For the SVs, a VCF file that lists all structural variants and a metrics file in CSV format are provided. As both file types were already described, we refer you to the descriptions above.

Similar to the SNVs and indels, the CNV and SVs are annotated. The annotation results are available as JSON and TSV files. As these two files were already described for the SNV and indel annotation, we will not go into detail here again. However, the annotation files for CNVs and SVs do not include information about the observed variant in the global population.

Table 7 | Excerpt of the CNVs' GFF3 file (header line added for the sake of clarity).

Sequence ID	Source	Feature type	Feature start	Feature end	Score	Strand	Phase	Attributes
chr1	DRAGEN	CNV	818023	1604140	56	.	.	Alt=REF;LinearCopyRatio=1.04317;CopyNumber=2;Genotype=./.; Qual=56;Filter=PASS;Start=818022;Stop=1604140;Length=786118; BinCount=607;ImproperPairsCount=0,0;color=#00FF00;
chr1	DRAGEN	CNV	1604141	1606616	20	.	.	Alt=DEL;LinearCopyRatio=0.512123;CopyNumber=1;Genotype=0/1; Qual=20;Filter=cnvLength;Start=1604140;Stop=1606616;Length=2476; BinCount=2;ImproperPairsCount=0,39;color=#DDDDDD;
chr1	DRAGEN	CNV	1606617	2227209	55	.	.	Alt=REF;LinearCopyRatio=1.03512;CopyNumber=2;Genotype=./.; Qual=55;Filter=PASS;Start=1606616;Stop=2227209;Length=620593; BinCount=454;ImproperPairsCount=39,1;color=#00FF00;

Resources

For further information, the DRAGEN Bio-IT Platform manual can be browsed:
https://support-docs.illumina.com/SW/dragen_v42/Content/SW/FrontPages/DRAGEN.htm

References

1. Cline, Melissa S; Karchin, Rachel (2011): Using bioinformatics to predict the functional impact of SNVs. In Bioinformatics 27 (4), pp. 441 – 448.

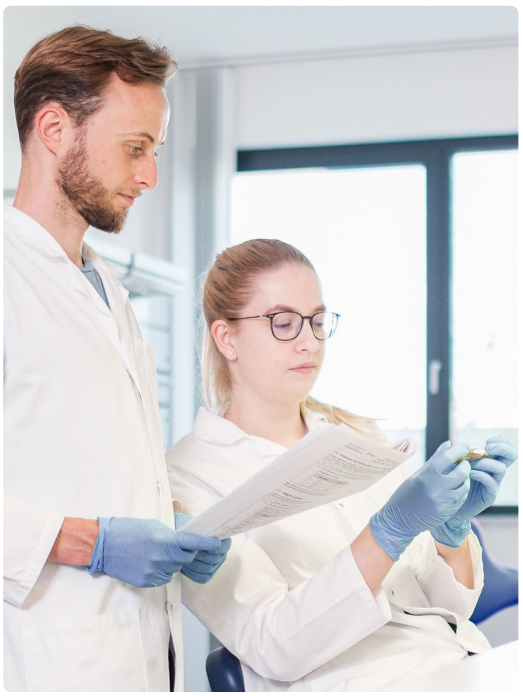

Sequencing complete
Sequencing complete
illumina

About Us

CeGaT was founded in 2009 in Tübingen, Germany. Our scientists are specialized in next-generation sequencing (NGS) for genetic diagnostics, and we also provide a variety of sequencing services for research purposes and pharma solutions. Our sequencing service portfolio is complemented by analyses suited for microbiome, immunology, and translational oncology studies.

Our dedicated project management team of scientists and bio-informaticians works closely with you to develop the best strategy to realize your project. Depending on its scope, we select the most suitable library preparation and conditions on our sequencing platforms.

We would be pleased to provide you with our excellent service.
Contact us today to start planning your next project.



CeGaT GmbH
Research & Pharma Solutions
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-333
Fax: +49 7071 56544-56
Email: rps@cegat.com

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS
CLIA CERTIFIED ID: 99D2130225

DAkkS Deutsche Akkreditierungsstelle D-PL-13206-01-00
Accredited by DAkkS according to DIN EN ISO/IEC 17025:2018

For more details please visit
www.cegat.com/rps

2024/08