CeGaT

Research &
Pharma Solutions | HiFi WGS

Bioinformatic Note

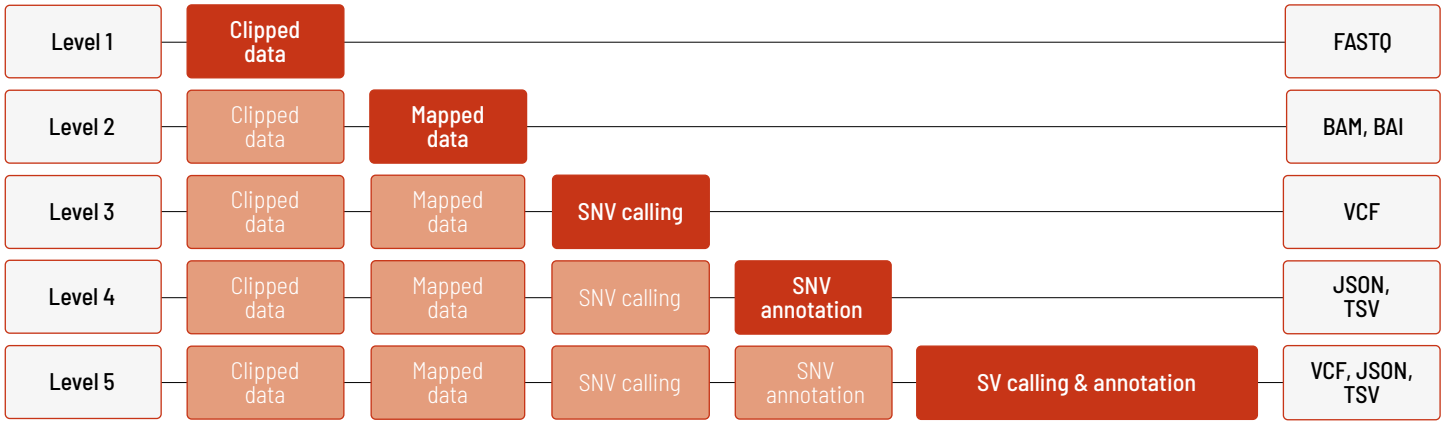
HiFi WGS

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 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

Our HiFi WGS product is analyzed with the PacBio Platform. Different levels of bioinformatic data analysis are available:



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

Level 1

If you wish to analyze your data yourself, we recommend Levels 1 or 2. In Level 1, the sequencing reads are demultiplexed. HiFi reads (CCS reads with a predicted accuracy ≥ Q20) are clipped and provided in FASTQ format, and the quality of the FASTQ files is analyzed. If required, clipped sequencing data in BAM format can additionally be delivered. This level is provided for every project, regardless of additional purchased bioinformatic analyses.

The additionally generated project report provides information about the laboratory protocol for every sample, including data about quality control of the starting material, library preparation, and sequencing parameters. For the demultiplexed data, the number of reads and bases is reported. The read length and the GC content per read are additionally illustrated in bar plots.

Level 2

If you wish to receive Level 2, the clipped reads are aligned to the reference genome. In addition to the reads in FASTQ format, you will receive the mapped reads as BAM files. Additionally, an index file in BAI format is provided.

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 median insert size, and the average coverage.

Level 3

In Level 3, small variants are called. The resulting VCF file contains the calls of the small variants. 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 1. For additional information on the VCF format, we refer the reader to the VCF specification.

In addition to the VCF file, a genomic VCF (gVCF) file is provided. It has the same underlying format specifications as the VCF file but includes additional records. These additional records distinguish regions with sequence coverage that seem to match the reference genome from regions without sequence coverage, resulting in an unknown genotype. Thus, the gVCF file contains variant sites and non-variant sites. The non-variant sites are represented by "<*>" in the alternate base column. To minimize the gVCF's output size, adjacent records with equal genotype qualities are merged into a single record. An excerpt of the small variant gVCF file is shown in table 2.

Table 1 | Excerpt of the small variants 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	10622	.	T	G	33.2	PASS	.	GT:GQ:DP:AD:VAF:PL	1/1:24:37:0,37:1:33,24,0
chr1	10623	.	T	G	34.6	PASS	.	GT:GQ:DP:AD:VAF:PL	1/1:26:37:0,37:1:34,26,0

Table 2 | Excerpt of the small variants gVCF file. The abbreviations in the INFO and FORMAT columns are explained in the respective gVCF file.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1
chr1	10830	.	G	<*>	0	.	END=10854	GT:GQ:MIN_DP:PL	0/0:50:36:0,111,1109
chr1	10855	.	G	G,<*>	3.3	PASS	.	GT:GQ:DP:AD:VAF:PL	0/1:3:37:29,8,0:0.216216, 0:0,0,23,990,990,990

To receive a complete picture of genetic variations, one needs to distinguish which variant comes from the same copy of a chromosome and which from a different copy. This process is called phasing – it resolves the genotypes into haplotypes. For the phasing process, a phased BAM file is generated. For the variant calling, a VCF file is provided. The PS ("phase set") tag is a unique identifier for a phase set in the VCF file. A phase set is a set of variants that were phased relative to each other. Usually, there are multiple phase sets in the file. Variants that belong to the same phase set do not need to be consecutive in the file. An excerpt of the phased VCF file is shown in table 3. In this excerpt, the lines of variants that belong to the same phase set are colored equally.

In addition to the three VCF files, a visual report in HTML format is provided. This visual report includes visualizations of several overall statistics from the first VCF file: The variant types, depth, quality score, genotype quality, variant allele frequency for all genotypes, biallelic base changes, biallelic Ti/Tv ratio, and the biallelic indel size distribution. A detailed explanation of the different plots is given here.

Table 3 | Excerpt of the phased VCF file. The abbreviations in the INFO and FORMAT columns are explained in the respective VCF file. For explanatory purposes, the lines holding variants of the same phase set are colored equally.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample1
chr1	10855	.	C	G	3.3	PASS	.	GT:GQ:DP:AD:VAF:PL:PS	1|0:3:37:29,8:0.216216:0,0,23:10855
chr1	10884	.	C	G	23.5	PASS	.	GT:GQ:DP:AD:VAF:PL	1/1:23:37:0,37:1:23,32,0
chr1	11565	.	G	T	48.4	PASS	.	GT:GQ:DP:AD:VAF:PL	1/1:28:43:0,43:1:48,27,0
chr1	11863	.	C	A	18.4	PASS	.	GT:GQ:DP:AD:VAF:PL:PS	1|0:17:44:23,21:0.477273:18,0,21:10855
chr1	11922	.	T	A	14.5	PASS	.	GT:GQ:DP:AD:VAF:PL:PS	1|0:14:43:26,17:0.395349:14,0,27:10855
chr1	106052	.	G	GA	28.3	PASS	.	GT:GQ:DP:AD:VAF:PL	1/1:22:39:3,36:0.923077:28,23,0
chr1	106157	.	A	G	26.6	PASS	.	GT:GQ:DP:AD:VAF:PL:PS	0|1:5:38:3,35:0.921053:24,0,2:106157
chr1	106164	.	T	C	26.3	PASS	.	GT:GQ:DP:AD:VAF:PL:PS	0|1:6:38:4,34:0.894737:24,0,4:106157

Level 4

In Level 4, the small variants are annotated. Please note that the annotated variant list is a filtered variant list: it contains the variants that passed the quality filter. 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.

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 (Cline and Karchin, 2011). As SNVs, indels can have an impact on certain diseases. With the annotated small variants files, many research questions regarding functional impacts on diseases, disease susceptibility, and drug sensitivity might be answered.

Level 5

In Level 5, structural variations, such as large insertions, deletions, inversions, duplications, or translocations, are called and annotated.

Before the structural variations are called, signatures of structural variations are identified. These signatures reduce the aligned reads to those relevant for calling the structural variations. These signatures are stored in a separate SV signature (svsig) file, which is subsequently used to call the structural variations.

The structural variations are provided in a VCF file. The VCF file looks like the small variant VCF file displayed in table 1. 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.

As for the small variants, the annotations for the structural variations are available as compressed JSON files and as TSV files. As these two files were already described for the small variant annotations, we will not go into detail here again. However, the annotation files for structural variations do not include information about the observed variant in the global population.

References

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

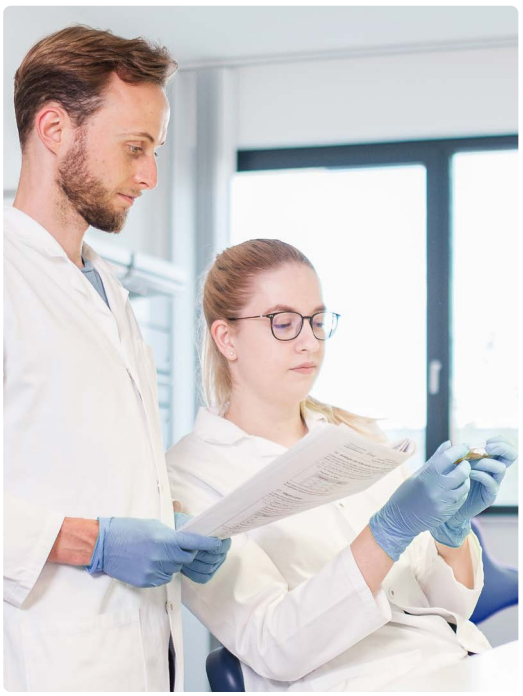
About Us

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.

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



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

2024/11